MODELING VISUAL RHETORIC AND SEMANTICS IN MULTIMEDIA

by

Christopher Lee Thomas

Bachelor of Science, University of Pittsburgh, 2013

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2020

UNIVERSITY OF PITTSBURGH

KENNETH P. DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Christopher Lee Thomas

It was defended on

July 14, 2020

and approved by

Adriana Kovashka, Department of Computer Science, University of Pittsburgh

Diane Litman, Department of Computer Science, University of Pittsburgh

Rebecca Hwa, Department of Computer Science, University of Pittsburgh

Abhinav Gupta, Robotics Institute, Carnegie Mellon University

MODELING VISUAL RHETORIC AND SEMANTICS IN MULTIMEDIA

Christopher Lee Thomas, PhD

University of Pittsburgh, 2020

Recent advances in machine learning have enabled computer vision algorithms to model complicated visual phenomena with accuracies unthinkable a mere decade ago. Their high-performance on a plethora of vision-related tasks has enabled computer vision researchers to begin to move beyond traditional visual recognition problems to tasks requiring higher-level image understanding. However, most computer vision research still focuses on describing *what* images, text, or other media literally portrays. In contrast, in this dissertation we focus on learning *how* and *why* such content is portrayed. Rather than viewing media for its content, we recast the problem as understanding *visual communication* and *visual rhetoric*. For example, the same content may be portrayed in different ways in order to present the story the author wishes to convey. We thus seek to model not only the content of the media, but its authorial intent and latent messaging. Understanding *how* and *why* visual content is portrayed a certain way requires understanding higher level abstract semantic concepts which are themselves *latent* within visual media. By latent, we mean the concept is not readily visually accessible within a single image (e.g. right vs left political bias), in contrast to explicit visual semantic concepts such as objects.

Specifically, we study the problems of modeling photographic style (how professional photographers portray their subjects), understanding visual persuasion in image advertisements, modeling political bias in multimedia (image and text) news articles, and learning cross-modal semantic representations. While most past research in vision and natural language processing studies the case where visual content and paired text are highly aligned (as in the case of image captions), we target the case where each modality conveys complementary information to tell a larger story. We particularly focus on the problem of learning cross-modal representations from multimedia exhibiting weak alignment between the image and text modalities. A variety of techniques are presented which improve modeling of multimedia rhetoric in real-world data and enable more robust artificially intelligent systems.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

The poet Louise Bogan said, "The initial mystery that attends any journey is: How did the traveler reach his starting point in the first place" [30]. Out of a perhaps whimsical notion that future readers may wish to know about my personal journey to completing my PhD, as well as my own personal need to retrospect on what has been achieved, I first provide a (short) summary of how I got to this point. As a child, I had an innate fascination with technology, always tearing things apart (and often not being able to reassemble them) in order to understand their inner workings. While many were satisfied by a cursory "high-level" explanation, I was not. I needed to understand *precisely* how things operated, particularly when it came to technology. While I sit quite dumbstruck at having just completed a PhD in Computer Science, in hindsight it perhaps is not so surprising at all that this is the path my life has taken.

From an early age, I toyed extensively with computers, wantonly ignoring warnings not to change operating system settings in order to learn their effects. I had an insatiable urge to learn and go where the adults around me at the time did not or feared to. This curiosity led me consistently to push the limits of my own knowledge. I taught myself to program before entering high school during a summer vacation at the beach. I much preferred learning new commands to the hot sun and water of the ocean. I flourished during high school despite innumerable obstacles, the details of which I will spare the reader. The computing courses at my high school at the time were extremely limited and I felt quite disappointed and somewhat surprised how far ahead I was of the other students. After graduation, I chose Pitt primarily because it was a great school and was also close to home. I wandered somewhat desultorily through my first year of college without the slightest idea of what my major would be. I began taking computer science courses primarily out of curiosity, never thinking that it would ultimately become my major. Four years later, I completed my degree in Computer Science.

Along the way, a number of faculty members strongly influenced my post-graduate directions. Professor Diane Litman announced in class one day the availability of an undergraduate research position. I was intrigued (at the time knowing nothing about research). To my surprise, I found it quite enjoyable and, more importantly, found it stimulated my curiosity and enabled me to push the limits of my knowledge. I also began working with Professor Janyce Wiebe at the time (during my undergraduate years). Finding that I truly enjoyed research, I decided that graduate school would be the next step and was accepted to Pitt's doctoral program. During one of my courses, I expressed that I had no clue what my research area would be. Jie Guo, a friend, suggested that I talk to Professor Steven Levitan about working in his lab. I followed up and was introduced to computer vision which, prior to this introduction, I had known essentially nothing about. However, I quickly found I was enthralled by the field and truly enjoyed what I was doing. After Professor Adriana Kovashka joined the department, I began working with her as my full-time advisor and the rest is history.

I wish to first acknowledge the *tremendous* efforts of my advisor in assisting me during my academic career. She provided exceptionally detailed line-by-line feedback on many of my early manuscripts, helping me to become a better writer and researcher. Our countless discussions greatly deepened my understanding and honed my critical thinking skills. She tolerated my slightly nettlesome tendencies and ultimately helped me become more focused. Perhaps most importantly, she taught me that doing hard work and great scholarship doesn't need to come at the expense of one's humanity. She encouraged me to take much needed breaks after deadlines and remained constantly accessible and grounded, despite her stature within the computer vision community. In sum, the work in this dissertation, as well as my growth as a researcher but also as a person, would not have been completed without my advisor's extensive mentorship. For this, I am grateful beyond words.

Pitt has been a truly supportive place and there are countless individuals that intersected with my journey that I wish to acknowledge. Professors Diane Litman, Rebecca Hwa, and Taieb Znati all provided extraordinary support along my journey and assisted me in many ways. I also wish to acknowledge Professor Donald Chiarulli who, along with Professor Steven Levitan, helped introduce me to computer vision. I wish to thank Yahoo Research! for a research internship and Yale Song for his mentorship, collaboration, and the unforgettable opportunity he provided for me in New York City during my residency there. I also wish to especially thank my dissertation committee for their invaluable feedback. Their comments helped organize and vastly improve this dissertation's draft. Similarly, I'd like to acknowledge Professor Stefan Lee of Oregon State University, who discussed the theme of this dissertation with me at length and helped me coherently organize my research. I also wish to acknowledge the National Science Foundation for providing the financial support which enabled me to complete the research in this dissertation.

I'd especially like to thank all my former and current labmates and fellow graduate students who were supportive during my academic career. In particular, I'd like to thank Nils Murrugarra-Llerena who has been a good friend and co-worker. Similarly, I'd like to thank my labmates Keren Ye, Mingda Zhang, and Narges Nazari for their encouragement, assistance, and countless discussions that we shared. I'd also like to thank Yuyu Zhou for her hilarious but also insightful comments and recommendations throughout graduate school. Similarly, I'd like to thank Jie Guo for her friendship which ultimately helped me discover computer vision. All will be sincerely missed!

Finally, I would like to thank my entire family for their support and assistance. The road to completing this dissertation has been bumpy, but they remained steadfast in their support. Most importantly, my mother's tireless and unacknowledged hours of assisting me during high school, as well as the support and love she provided throughout college and graduate school despite overwhelming obligations of her own, helped me achieve and become what and who I am today. I also wish to thank my father for his support and providing me with a safe home to live in during the latter part of my graduate career. Similarly, my grandmother Evelyn's endless prayers, love, thoughts, advice, and ironing / starching of my clothes in addition to all her other contributions meant so much to me. My grandmother grew up in a world where girls didn't often go to high school. Much later in her life, my grandmother insisted she obtain her GED (which she did obtain, with distinction), because she knew (and taught me) the value of education. She was so proud of me finishing this degree. I therefore dedicate this dissertation my parents and my grandmother. I'd also like to thank my uncle Terry, cousin Mandy, and late grandmother Marcella for their encouragement along the way, as well as Sarge, the most loving dog ever.

1.0 INTRODUCTION

Recently, the availability of large scale computer vision databases such as Imagenet [303], Visual Genome [183], and MSCOCO [211] have enabled remarkable progress on a range of computer vision tasks, from the more fundamental such as object recognition [292, 210, 143, 291, 297, 105] and image segmentation [49, 409, 105, 205, 301] to more challenging tasks requiring multimodal learning, such as visual question answering [8, 230, 382, 366, 138, 378, 12] or visual dialogue [236, 311]. There is essentially "unanimous agreement" that much of this improvement is due to the fact that such large data resources enable the training of increasingly complex deep neural networks [330]. These more powerful models have enabled researchers to begin to model ever more abstract phenomena, such as predicting image virality [6, 67, 115], understanding the effect of physical forces or processes in images [21, 249, 89], and predicting the motivations of actions in images [285].

Though specific tasks widely vary, the overwhelming majority of computer vision research focuses on using the literal visual content of an image, either by classifying or describing such content or leveraging the content for some downstream task. For example, image captioning approaches focus on producing text which literally describe the content of an image (e.g. "a dog sitting on a bed"), action recognition methods classify the actions occurring in images or video (e.g. jumping, sitting), and object detectors localize and classify specific objects appearing in an image. All of these methods directly map certain visual patterns to certain classes or co-occurring words (in the case of image captioning). There has been some recent effort to apply vision methods to tasks requiring higher-level reasoning. The visual commonsense reasoning task, for example, requires vision methods to answer questions about images which require commonsense knowledge and inference (e.g. "Q: Why is the person pointing? A: To tell the waiter where to put the food."). However, these problems still focus on understanding the literal visual content of images, even though they aim to require external reasoning about such content. These tasks continue to view visual media through the lens of understanding the content of images *for its literal value* and then leveraging that content, either in a descriptive manner (as in the case of image classification or captioning) or in a task-driven manner (as in visual question answering).

In contrast, in this dissertation we view images through the lens of *visual communication*, that is, as communicative tools used by humans to tell visual stories and to convey messages. We seek to understand and model not only *what* that message is, but *how* and *why* images convey it. For examples, while many different images could be used to illustrate a news event about a protest, the choice of image to illustrate a news article may depend on the overall political leaning of the author. One author may attempt to portray the protest as violent or dangerous, while another may wish to portray it as peaceful. While both images capture and illustrate the same event, *how* and *why* the event is portrayed in varying ways differ among the two authors. It is this authorial intent and latent messaging, which forms a "subtext" around the image, which we seek to model in this dissertation. We term such subtle messaging the *visual rhetoric* of the image; that is, the subtle ways in which visual content is used to tell a story which ultimately lies outside the image itself.

The fundamental challenge we address in this dissertation is modeling high-level, often *visually incoherent* or *latent* semantic concepts (e.g. political bias) within highly visually diverse, noisy (ground-truth labels may be incorrect), and/or small datasets. By *visual incoherence* or *latent visual concepts*, we mean that while the concept of interest has a common-sense semantic meaning occurring throughout the dataset, it is visually manifested in different ways such that isolating particular visual features present in multiple images is elusive for existing algorithms. For example, training a deep neural network to distinguish between different classes of objects is straightforward. After seeing many examples of each class via backpropagation, the network may learn that dogs have pointy noses and birds have beaks and use those patterns to discriminate between the two classes. Such patterns are fairly straightforward to extract from the data, since each object has a regular, consistent visual manifestation across multiple images.

But what if the underlying phenomenon to be modeled is a more abstract or semantic visual concept? For example, imagine a dataset of photographers where each photograph is labeled with the photographer who took it. In this case, each "class" contains much more visually diverse data than a single animal, i.e. many different objects, patterns, places, and visual settings. The "class" is meaningful to humans, but visually incoherent from the perspective of a machine classifiers. A human might observe, for example, “I notice this photographer tends to portray the poor in a more positive light.” From this, a human may be able to extrapolate political messages which then seem to occur consistently across the photographer’s body of work. Extracting such a signal requires viewing the image more hollistically, not for its content, but as a form of visual communication. This requires understanding how the specific content shown in an image interlaces with some broader, overarching concept, such as the author’s views of the poor or child labor. What a particular photographer’s “style” is only makes sense in retrospect, after viewing numerous examples of his or her work. As we will see across several such problems, this type of visual signal is *much* harder for a machine to extract from data on its own. Without careful design to force a model to rely on semantically meaningful concepts, a network trained on this problem is unlikely to capture the underlying concepts of interest and will instead likely fall into the trap of using low-level cues (such as color) to make its decisions.

This dissertation’s primary focus of modeling visual communication and semantics as described above can broadly be characterized as discriminative learning, wherein models learn to map an image to some discrete representation: class labels, words, semantic embeddings, etc. However, a collateral theme of our research is the use of *generative* techniques to generate new visual data exhibiting the semantic concepts we have modeled. The importance of generative models for our purposes is their ability to automatically learn a natural feature representation for any given dataset, with or without labels on it. For example, in our work on modeling political bias, we show an experimental result that externally introduced information about facial semantics helps with predicting the political bias of a face. We then use a generative model to visualize how political bias changes how politicians’ faces are portrayed. Thus, a secondary, recurring theme of this dissertation is the use of generative models to aid humans in understanding aspects of our datasets or the semantics we are modeling (Chapters 3, 4, 5).

	Visually incoherent semantics	**Limited data**	**Noisy data**	**Multimodal data**	**Figurative / symbolic visual rhetoric**
Modeling...					
Photographic style Chapter 3	X	X	X		
Visual persuasion Chapter 4	X	X	X		
Visual political bias Chapter 5	X		X	X	X
Abstract semantics in multimedia Chapters 6 and 7	X		X	X	X

Table 1: We highlight the common challenges our methods address.

1.1 COMMON CHALLENGES, COMMON METHODS

At a high level, this dissertation explores the problem of modeling, both discriminatively and generatively, visually incoherent phenomena within noisy and visually diverse datasets. In particular, we propose five methods for modeling four distinct types of visual semantics. While the concepts we study differ, the overarching challenges posed by each setting are similar and the methods we develop to address them thus build upon shared themes. We summarize these commonalities in Table 1. We observe that all problems we study involve modeling visually incoherent semantics which are challenging for networks to capture. All of our tasks also feature noisy data, which stems from the fact that the data was scraped from the web en masse without human supervision. This noise introduces labeling errors, whereby images labeled with a particular class may actually be of a different class (or even no class at all in the case of irrelevant retrievals). For example: our photographic style work features noisy data in the form of images incorrectly assigned to a particular photographer, in addition to image modifications such as unnatural borders, edges, or other distortions; our methods for modeling visual persuasion and visual political bias must contend with data whose ground truth labels are potentially incorrect; and our work on learning multimodal

	Structured or guided training	**Explicitly inject external semantics**	**Control for noisy data**	**Multi-stage learning**	**Generative models for analysis**
Modeling...					
Photographic style Chapter 3	X		X		X
Visual persuasion Chapter 4	X	X	X	X	X
Visual political bias Chapter 5	X	X		X	X
Abstract semantics in multimedia Chapters 6 and 7	X	X	X	X	

Table 2: Commonalities between the different methods we develop in this dissertation.

semantic representations is faced with noisy image-text pairings (e.g. image and text don't actually go together) and incorrectly scraped webpage text. Additionally, limited data is a challenge for our work on photographic style and visual persuasion. This stems from the fact that there are only a relative limited number of photos taken by a particular photographer and from the size of the pre-existing ads dataset we use in our work on visual persuasion. As discussed before, the concepts we study manifest themselves when data is viewed across the entire dataset (i.e. collectively), but may be extremely hard to localize from individual images. Thus, learning with limited data poses a significant obstacle to extracting meaningful semantic signals and thus, special care must be taken in these cases. Our work on political bias and learning semantic cross-modal representations both feature unique challenges and opportunities in that they feature *multimedia* documents, i.e. i.e. images paired with text. This is particularly challenging because we seek to model visually incoherent concepts, which do not explicitly align with the text paired with the image. Similarly, these methods must also cope with figurative and artistically modified images, as well as understand visual symbolism.

The shared challenges posed by these different problem settings result in similarities across the methods used to solve them. As we show in Table 2, the methods we develop for each of these settings share a number of common features and use cases. We see that all of our methods involve explicitly controlling what our classifiers learn through structured or guided training. This can be accomplished by designing custom loss functions to capture only the phenomenon of interest, imposing limitations inherent within the architecture of the model, or only finetuning higher-layers to prevent lower-layers from fitting to irrelevant phenomenon. Several of our methods involve explicitly providing our methods with external semantics believed to be relevant (by humans), either in the form of semantic attributes (for visual persuasion or political bias), or through the use of a general-purpose semantic representation (for multimedia semantics). Some methods control for noisy data in some way, using both simple and elaborate techniques. Additionally, some of our methods employ multi-stage learning techniques, typically in the form of a multi-stage training procedure in which models are first trained on different tasks and then combined in some way. Finally, a number of our methods use generative models to aid analysis by visualizing what the models have learned. We believe the insights used in each of these settings form a toolbox of techniques which can be applied to other domains where standard methods fail to adequately model visually incoherent semantic concepts. For example, when encountering a new problem which requires modeling non-literal semantics, one should first consider imposing structure on the training process.

Despite being developed for and applied on particular tasks, the methods we propose are generalizable to other tasks within this dissertation. We show the consonance of our methods cross-dataset in Table 3. We observe first that our cross-modal methods (Chapters 5, 6, 7) are broadly applicable to settings where paired text is available with images. The image ads dataset we use in Chapter 4 does contain a variety of types of paired text for a number of ads and thus these methods are potentially applicable in that setting. Similarly, our techniques of using explicit object (Chapter 3) or attribute (Chapter 4) representations are applicable to all the datasets we study. We remark however, that constraining learning to particular semantic representations (such as objects or attributes learned on other datasets), does preclude models from learning dataset-specific features which ultimately may be better

		Method				
		Transferring high-level object representations (Chapter 3)	Leveraging external semantic attributes (Chapter 4)	Textual semantics guide visual training (Chapter 5)	Enforcing cross-modal semantic proximity (Chapter 6)	Identifying semantically informative samples (Chapter 7)
Dataset	Photographic style (Chapter 3)	✓	⊛			
	Image advertisements (Chapter 4)	⊛	✓	⊛	⊛	⊛
	Communicative multimedia (Chapter 5)	⊛	✓	✓	✓	✓

Table 3: We show the generalizability of the methods we propose across different datasets. Methods which were actually performed on the dataset are shown with a ✓, while methods which are applicable (but which we do not actually apply) are shown with a ⊛.

suited to the particular problem and data. Thus, our goal throughout is to learn semantics for the specific problem from the actual data itself to the maximum extent practicable, in light of challenges such as dataset size, visual diversity of the concept to be learned, and noise.

In the remainder of this chapter, we more thoroughly describe the broader themes of our work, the shared challenges, and the individual methods we propose. The remainder of this chapter is as follows. In section 1.2 we introduce the problem of modeling latent concepts in visual data. In section 1.3, we more thoroughly discuss the problems encountered when working with visually diverse, noisy, and/or small datasets and outline several strategies we use to solve them. In section 1.4, we examine the unique challenges (as well as opportunities) faced when modeling semantics in communicative multimedia, as opposed to traditional vision and language used in vision. Section 1.5 introduces the scientific hypotheses which we seek to address in this dissertation. Finally, in section 1.6, we more explicitly delineate the contributions of our work.

1.2 LATENT VISUAL CONCEPTS

Throughout this dissertation, we consider images' content not as an end in itself (i.e. for its literal content), but rather as a communicative medium used by authors to convey visual stories. It is these stories which we seek to understand and model, not the content of images itself. Of course, understanding how an image fits into the "bigger picture" also requires recognizing the contents of the image. However, computationally modeling visual communication requires taking a step beyond the literal content of the image, and understanding broader semantic concepts and how a particular image's contents interplays with or reinforces them. For example, the Statue of Liberty can be used as a symbol representing justice, freedom, or immigration. However, the visual manifestation of these abstract concepts is itself missing from the literal object of the statue in the image. Only by looking at the broader dataset may one learn correspondences between the object and these various semantic topics. These semantic concepts are thus *latent* within the data, despite having various visual expressions. Similarly, understanding and modeling the overall themes defining a professional photographer's work requires reasoning over multiple images taken by the photographer and connecting and learning how specific objects and their portrayals cohere into a larger semantic narrative or theme. We call these broader semantic themes *latent visual concepts* because although they manifest themselves in individual images, they can only be understood in the context of other images viewed collectively. Thus, though a latent visual concept may be exhibited by a single image, it is simultaneously absent without the context of the broader dataset to give it meaning. Only by considering Lewis Hine's body of photographic work as a whole, for example, does one understand why and how a particular image portrays working children and what sets his photographs apart from others (Mr. Hine was a vocal advocate against child labor).

We illustrate the four different "visual concepts" explored in this dissertation in Figure 1. In Figure 1(a), we illustrate the problem of photographer classification, where the system must learn a model of a photographer's photographic "style"; (b) models the concept of *visual persuasion* by illustrating the different ways faces are portrayed in different types of ads; (c) shows a latent multimedia visual concept (political bias) with groundings in both the

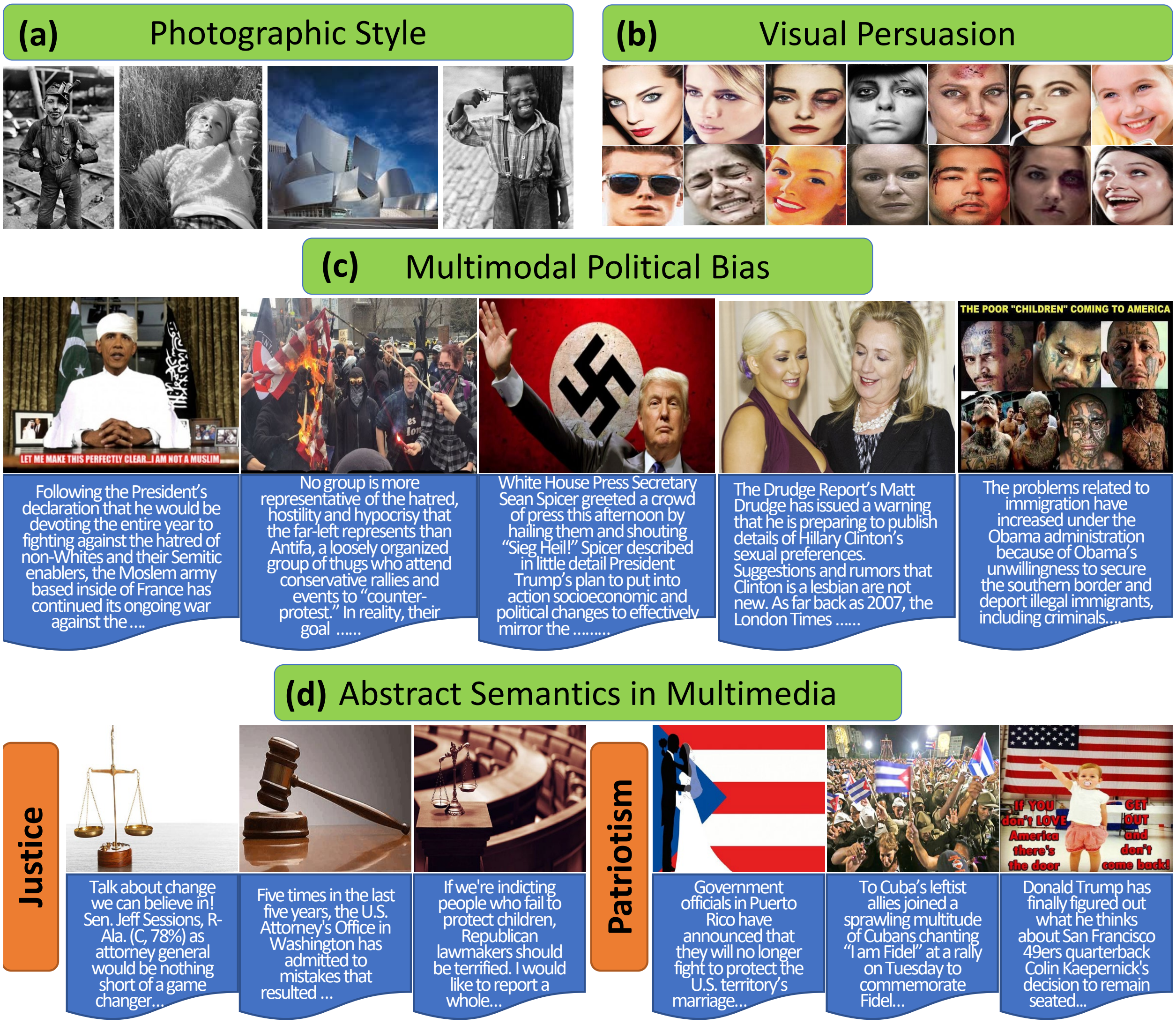

(a) Photographic Style
(b) Visual Persuasion
(c) Multimodal Political Bias
LET ME MAKE THIS PERFECTLY CLEAR...I AM NOT A MUSLIM
THE POOR "CHILDREN" COMING TO AMERICA
Following the President's declaration that he would be devoting the entire year to fighting against the hatred of non-Whites and their Semitic enablers, the Moslem army based inside of France has continued its ongoing war against the ....
No group is more representative of the hatred, hostility and hypocrisy that the far-left represents than Antifa, a loosely organized group of thugs who attend conservative rallies and events to "counter-protest." In reality, their goal ......
White House Press Secretary Sean Spicer greeted a crowd of press this afternoon by hailing them and shouting "Sieg Heil!" Spicer described in little detail President Trump's plan to put into action socioeconomic and political changes to effectively mirror the ....
The Drudge Report's Matt Drudge has issued a warning that he is preparing to publish details of Hillary Clinton's sexual preferences. Suggestions and rumors that Clinton is a lesbian are not new. As far back as 2007, the London Times ......
The problems related to immigration have increased under the Obama administration because of Obama's unwillingness to secure the southern border and deport illegal immigrants, including criminals....
(d) Abstract Semantics in Multimedia
Justice
Talk about change we can believe in! Sen. Jeff Sessions, R-Ala. (C, 78%) as attorney general would be nothing short of a game changer...
Five times in the last five years, the U.S. Attorney's Office in Washington has admitted to mistakes that resulted ...
If we're indicting people who fail to protect children, Republican lawmakers should be terrified. I would like to report a whole...
Patriotism
Government officials in Puerto Rico have announced that they will no longer fight to protect the U.S. territory's marriage...
To Cuba's leftist allies joined a sprawling multitude of Cubans chanting "I am Fidel" at a rally on Tuesday to commemorate Fidel...
IF YOU don't LOVE America there's the door
GET OUT and don't come back!
Donald Trump has finally figured out what he thinks about San Francisco 49ers quarterback Colin Kaepernick's decision to remain seated...

Figure 1: We showcase examples of the five latent visual concepts we model in this dissertation. **(a)** shows examples from our photographer style dataset, **(b)** shows examples of visual persuasion in image advertisement faces for various categories of ads, **(c)** shows an example of a latent concept (political bias) expressed in multimedia (images and text), and **(d)** shows abstract semantic concepts ("justice" and "patriotism") and their multimodal manifestations in communicative media.

image and text domains; and (d) depicts abstract multimodal semantic concepts ("justice", "patriotism") manifested in both visual and text space, exhibiting how images are used as communicative tools to illustrate abstract concepts in the text.

While the problems we study may seem different at first, the unifying theme reoccurring across problem settings is that the semantic visual concepts we wish to model are challenging for existing models to grasp without some form of human intervention. In all of the domains we study, models trained without some custom supervision, either in the form of forcing a classifier to rely on semantic knowledge transferred from other sources or designing custom loss functions to capture concepts of interest, results in a failure of the model to capture the concepts of interest. The CNN may simply fit itself to lower-level phenomena, such as textures, colors, or particular objects which reoccur in some classes, without actually modeling anything semantically meaningful. While the multimodal case of images and text poses some unique challenges because the semantics are abstract and manifested in both the image *and* text spaces and there is a lack of direct alignment between the text and the image (i.e. the text doesn't literally describe the contents of the image), ultimately we find that general intuitions and techniques used for preventing models from latching onto irrelevant image phenomena can be leveraged to facilitate multimodal learning of abstract semantic concepts.

Generative modeling. Our discussion thus far has mainly assumed a discriminative context. However, a secondary objective is *generating* visual data which exemplifies the visual concepts we have modeled. In other words, we are not simply interested in training a generative model on our different datasets and achieving synthetic images of good visual quality. Instead, we want our generations to emblematize the semantics that we are modeling in that particular domain. For example, if we are generating synthetic faces appearing in ads, it is not sufficient that the generator learns to generate a crisp face; we insist that the face must also bear the visual concepts that we have learned are inherent in that particular type of ad. This is a particularly challenging constraint to formulate because of the "hidden" nature of the concepts we consider. While conditional generative models [151, 61, 386, 327, 53] are now common, when the visual concepts in question defy vanilla discriminative classifiers (as previously discussed), the usual techniques of simply conditioning such classifiers on the class label does not work. As we will later discuss, state-of-the-art conditional GANs trained on such problems thus face two major obstacles: (1) the failure of auxiliary discriminative components to capture the visual concepts of interest (for the same reason that training discriminative models fail to capture the concepts without our techniques and (2) achieving a good fit of the underlying distribution of the data when the visual patterns occurring within a class (e.g. right-leaning political bias) are too varied. In Chapters 3 and 4 we present increasingly complex strategies to account for these challenges.

Multimodal modeling. Our problems of modeling political bias and learning representations of abstract multimodal semantics continues the theme of modeling latent visual concepts, but also extends it in interesting ways. For example, political ads with a pro-immigration agenda may portray immigrants as happy, contributing members of society, while anti-immigration ads may portray immigrants as hardened criminals. This problem is more challenging than that of modeling the difference in faces found in beauty ads vs. soda ads, for example, due to the fact that we seek to understand political bias broadly, rather than just how it is exhibited on specific objects. Moreover, unlike our prior contributions, we study the concept of bias in *both the image and texts domains.* We thus build a model which synergistically accounts for both textual and visual bias, learns correspondences between them for predicting political bias, and also exhibits these learned correspondences in generated data. We also show how a model of *intra*modal bias (on text alone) can serve as a supervisory signal to help models trained on multimedia (images and text) actually capture the phenomenon of interest. Our final two methods build upon our work in modeling multimodal semantics, but are also somewhat distinct from the rest of our work in this dissertation in that rather than study a particular latent visual concept, they instead propose *general solutions* for modeling abstract semantics in multimedia, in the absence of any specific semantic task. Instead, these methods seek to learn a robust semantic representation which captures such higher-level semantics automatically, without the need for human tuning of the method. We thus desire the learned semantic representations to preserve such broader abstract semantics *automatically* in a data-driven way. These representations can then be used for downstream tasks requiring inference on the latent concept (e.g. to train a political bias classifier or any other such task), without the need for task-dependent approaches for each dataset and task one seeks to model. To do so, these methods work by incorporating complementary intermodal semantics from the paired image-text domains into the models' learned semantic representations or by exploiting the pairwise nature of multimodal data to automatically identify and then emphasize image-text pairs which contain abstract semantics. We believe these two final methods serve as strong general purpose starting points for modeling latent visual concepts in visual multimedia.

1.3 DIVERSE, NOISY, AND LIMITED DATA

Obtaining enough data to train modern machine learning models is a perennial problem in computer vision. Fundamentally, this need for ever-more data stems from an increase in the number of parameters to be learned in ever-more complicated architectures. Contemporary deep learning architectures have hundreds of layers, yielding models with millions of parameters [127, 184, 321]. The recent InceptionResnetV2 network, for example, has nearly 56M parameters which must be optimized [335]. Applying such architectures to a sufficiently complex problem yields an extremely non-convex error surface. Finding an acceptable saddle point on such a surface thus requires a large amount of data [170]. Similarly, training data which is too noisy to compute such smooth gradients inevitably results in the model falling into unacceptable local minima or divergence.

Visually diverse datasets contain images and objects whose visual appearance significantly differs across the dataset. This poses challenges for learning as models struggle to find consistent visual patterns across images of the same class, e.g. right/left leaning images or images taken by a particular photographer. For the dataset of faces we consider in Chapter 4, this visual diversity amounts to variations in pose, gender, color, facial attributes, image backgrounds, etc. Our political bias dataset (Chapter 5) is also extremely diverse. It contains, for example, the same politicians in highly varied scenes, spans many different political topics, and features countless images of the same politically relevant events taken from different perspectives or showcasing different aspects of the event. In general, visually communicative media, which is intended for human viewers, features much richer visual diversity than is traditionally found in curated object detection and image captioning datasets featuring repetitive objects such as MSCOCO [211] or Imagenet [184]. In contrast, multimedia used in visual communication features an open vocabulary of topics, scenes, and objects (i.e. essentially any topic or object may appear). Moreover, some images and objects are often modified through image editing, or the image itself could be a graphic illustration. The conventional wisdom is that diverse datasets are beneficial when there is sufficient data to adequately train models on because they enable CNNs to become significantly more robust to noisy test cases. In fact, many techniques [77, 368, 93] have been proposed to add diversity to datasets to improve the performance of discriminative models. However, these approaches assume models are learning concrete, visually consistent object representations (i.e. the class is an object, but the scene type or pose of the object is changing). In contrast, we seek to model *latent* semantic concepts (such as political bias). In this case, having a highly diverse "open world" dataset spanning many different subjects with vastly different visual appearances makes extracting any meaningful signal of the visual rhetoric we seek to model extremely difficult.

Compounding matters further, when datasets are small *and* diverse, achieving acceptable results can be nearly impossible without providing some form of external structure or guidance on the learning process, as models simply memorize low-level patterns within the training data, which are not actually discriminative. The primary reason for limited data often comes from the nature of the problem itself. For example, there are only so many photos taken by a professional photographer. For our work on ads, finding a large number of non-duplicate images with a particular label (i.e. soda ad, car ad) is challenging, thus the dataset we use for this work is relatively small. In contrast, there are essentially limitless numbers of potentially politically biased images and paired text available on news websites. One solution to dealing with limited data is to simplify the model by reducing the number of parameters by reducing depth in order to achieve a more general, but less powerful model. However, in the case of generative models this simplification ultimately results in models which either output images at unsatisfyingly small resolutions (which are often too small to demonstrate the semantic concepts we wish to capture) [228]. In the case of discriminative models, reducing model complexity results in models lacking sufficient representative power to model the phenomena of interest.

One common solution to the problem of limited data is to leverage existing pre-trained models, trained on external large datasets. Pre-trained models have found use in a wide variety of applications and have proven to be a formidable baseline for many tasks [315]. Other techniques, such as domain adaptation and transfer learning aim to allow such pre-trained models to generalize to other domains with less data. Domain adaptation techniques enable models trained on one type of data to be re-applied on a separate type of data from a different distribution [274]. Transfer learning enable models to generalize across different tasks, dispensing with the need to relearn lower-layer texture filters, for example, and thereby reduce data requirements [220, 264]. The standard practice is to first train a powerful CNN model on a dataset for which labeled data is plentiful (e.g. Imagenet), even though the problem of interest may be completely unrelated to Imagenet or even image classification. Nevertheless, this training phase allows the model to learn powerful, general purpose semantic features which can be applied readily to a number of tasks, either by training a secondary classifier on top of them or by finetuning only some layers of the model, which requires much less data [148]. This allows classifiers to conduct inference over features with semantics "baked-in" from pre-training, without requiring the model to learn them again from a visually diverse, limited, and possibly unlabeled dataset. Many of the methods we devise in this dissertation make use of some form of transfer learning. However, unlike many transfer learning methods, we seek to transfer representations to model higher-level abstract semantic concepts where it is not immediately obvious the transferred semantic representation is appropriate. For example, it is not obvious that transferring object semantics is appropriate for photographer classification or that visual persuasion can be captured through facial attributes.

Several of the datasets we utilize in this work are also *noisy*, in that the ground truth labels of the dataset may be incorrect. For example, in the datasets we collect for our work on image ads and political bias, we download a large number of images and (for political bias) associated lengthy text from the internet. Though we endeavor to combat this, the images retrieved from the page may actually be irrelevant ads, completely unrelated to the article text. Moreover, the text parser may fail to correctly parse the article text and instead include irrelevant text from ads or other links to unrelated content. The presence of such "confounding" ground truth data poses a significant challenge to training models on this data without explicitly controlling for the fact that much of the data may not actually contain any learnable signal. Many such "webly" or weakly supervised methods have been proposed to combat this problem. Some works have shown that label noise has a deleterious effect on recognition performance [381], while other works have actually shown the opposite [294, 349]. Interestingly, [182] show that highly fine-grained recognition models can be learned using noisy data. We more thoroughly discuss related work related to learning from noisy data in Chapter 2.3.3. We control for various types of noisy data throughout this dissertation. However, the problem of label noise and irrelevant data is particularly pronounced in the case of modeling abstract, latent semantics. In order to tease out the effects of noise in our test data, we evaluate several of our methods on external, cleaner datasets as well as on human annotated data. In Chapter 7, we present several methods for emphasizing semantically informative samples within the dataset, while simultaneously de-emphasizing outliers. This approach automatically downweights samples which are significantly different from the overall dataset's distribution and has the effect of mitigating the effects of dataset noise, while simultaneously boosting the model's attention to nuanced samples containing the semantics of interest.

All of the problems studied in this dissertation feature highly diverse datasets and two of them feature limited data (photographic style and visual persuasion) which requires particularly careful consideration. Importantly, though our collected dataset on political bias is the largest dataset we collect, it is by far the noisiest, consisting entirely of data scraped from the web with automatically assigned labels (with no human supervision). Thus, the challenges posed by each of the distinct problems we study interlock and are complementary. Examples of the high visual diversity exhibited by some of the datasets we analyze can be seen in Figure 1(a)-(d). The classes within our photographer dataset (i.e. all photographs taken by a single photographer) span photographers' entire careers over many different projects and time periods and thus exhibit extreme diversity. Our visual persuasion dataset is highly diverse (and also somewhat noisy), so we restrict ourselves to working on faces, which we found exhibited the most appearance variation of any object category in the dataset. The number of faces per category is also extremely limited, with several categories having less than 500 faces in total. Our work on political bias relies on a dataset we collected which is not only extremely diverse in visual space, but which is also noisy in both the image and text domains. Finally, the two general methods we propose for learning abstract semantic representations in multimedia are applied on our political bias dataset, but also on several other large-scale, webly-supervised datasets such as Conceptual Captions [316] and GoodNews [27], which themselves feature a high-degree of visual diversity and some noise given they are automatically harvested from the Internet, with no human supervision. We note that we prefer learning on larger, noisier datasets than smaller, clean, and human annotated ones. This is because a larger dataset allows us to learn models on the data directly rather than relying on transferred semantic representations, which ultimately allows us to better capture the semantics within the dataset better. Additionally, we seek to develop techniques which allow learning abstract semantic concepts without the necessity of expensive human annotations. Further, large-scale datasets contain more diverse expressions of the concept we seek to model (e.g. political bias exhibited in many different ways) allowing us to draw insights about the concept which may be impossible in a small dataset.

1.4 COMMUNICATIVE MULTIMEDIA

It is difficult to believe that a mere few decades ago the Internet was dominated by primarily text-only webpages, perhaps sparingly interspersed with clipart and a few pixelated images. Today's Internet is dominated by pages imbued with embedded images, embedded video, sounds, GIFs, and other media which accompanies text as a way of illustrating it. For example, webpages containing recipes feature images demonstrating each step of the cooking process, or even provide an video which shows the entire preparation process of the dish from start to finish. Visual media adds entertainment interest, as well as informational value to the text which it accompanies, by explaining concepts and conveying visual stories to the reader which supplement the messages contained in the text. Similarly, images paired with news articles allow the reader to actually observe the subjects or incidents being discussed in the text. Likewise, the ability to discuss matters of interest and share visual media rapidly on social media have fundamentally changed the nature of human communication. Platforms like Twitter, Facebook, and Reddit have become the dominant vehicle through which the collective human experience is mediated, shared, and discussed. The ability for the average user to instantly disseminate content to the world at large without the costly requirements of a printing press or the need to convince an editorial board to disseminate it has facilitated the democratization of human communication. Visual media has thus become an enormously powerful tool for human communication and persuasion. It has been found, for example, that tweets with images have been found to get 89% more likes and 150% more retweets than purely text-based posts [401]. The public clearly crave such content as it is entertaining, more emotionally impactful, and ultimately more persuasive. In this dissertation, we term content composed of multiple *modalities* of data *multimedia*. For example, a news article may contain images, text, embedded video, and other graphics, but still forms a cohesive unit as a *multimedia document*. As a cohesive unit, the individual media components of a multimedia document make sense when viewed collectively, rather than independently, since they each contribute to the overall message conveyed by the document. While this is a natural way of viewing such multimedia for humans, we will see that existing machine learning approaches process the modalities quite differently.

Working with multimedia poses unique challenges for machine learning research because multimedia documents are composed, by definition, of multiple modalities of data. To reason about visual multimodal data, one must not only learn the visual concepts portrayed in images (e.g. people, dogs, or other objects) and their semantic relationships, but also learn how those concepts relate to specific *words* or tokens in the text. To do so the model must learn that a specific set of visual patterns tends to correspond with the co-occurrence of a particular word in the text, in order to establish a connection between the visual concept with its symbolic representation in text. Compounding matters further, many different words can apply to the same visual object, as a result of synonymy. For instance, the President of the United States can be referred to as “The President”, “Donald Trump”, the “Commander in Chief”, “POTUS”, etc. and the model must learn to associate these tokens with their associated visual manifestation (i.e. the person Donald Trump in a photo). While this may be straightforward for objects which have literal and recurrent visual groundings in image space

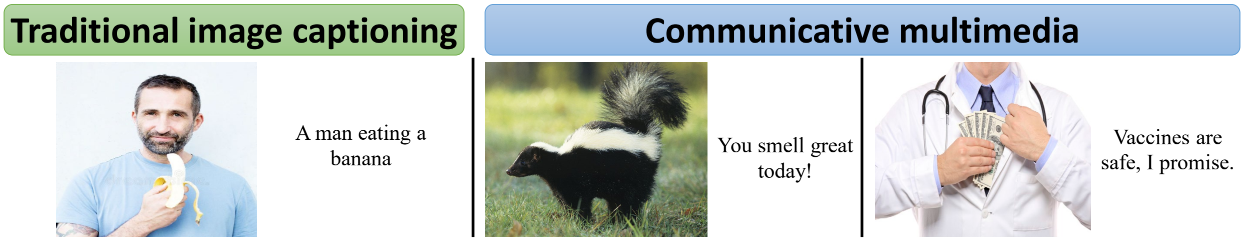


Figure 2: We illustrate the differences between our work in modeling multimedia semantics compared to existing visual semantic embedding methods which target traditional image-text datasets. On the left, we show examples of traditional image-caption pairs. In this case, the text literally describes the content of the image. In contrast, in real-world communicative multimedia, authors combine images and text to convey persuasive messages, without necessarily making the message explicit in text. Images and text thus carry complementary information, which contribute to the overall multimodal message.

(which objects such as people, structures, or places often do), the problem becomes much more challenging for abstract concepts encountered in real-world communicative multimedia such as "conservative" / "liberal" or "freedom".

Existing methods for learning representations of visual and text semantics (known as visual semantic embeddings) assume that the relationship between images and text is essentially literal and straightforward. Moreover, most popular image-text datasets [211, 183, 2] on which visual semantic embeddings are learned and evaluated, provide text in the form of image *captions*, which literally describe the content of the image. For example, an image caption might be "A man is eating a banana," and would be paired with an image showing the same content. As a consequence, existing methods for learning cross-modal representations rely on the fact that each modality provides essentially the same *redundant* information in order to learn representations of each modality that are close within the space. For example, continuing the above example of a man eating a banana, it is straightforward to see how the representation of the image and text could be close within the space, as they both express the same message. However, *most visual multimedia encountered in daily life* *is not like this.* Real-world media, such as news articles, social media posts, memes, etc. leverage images and text as a form of multimodal messaging. In these cases, the image often provides visual rhetoric undergirding arguments made in the text, but not necessarily literally mirroring them. As an example, the image described before could be paired with text describing grocery store shortages of fruits and vegetables as a result of coronavirus. In this case, the image content is more *illustrative* than redundant. More realistic cases have even *more* complex relationships between the modalities, where the semantics of each modality depend on understanding *both* modalities together. That is to say, that each modality's semantic meaning can only be understood when viewed holistically with the other paired modality(ies). Consider the examples shown above in Figure 2. In the example with the skunk, the text's actual meaning is that the person *does not* smell good, given that the text is paired with a skunk. However, this semantic message is lost in the absence of the image's context. Similarly, in the example on the right, the image undercuts the statement made in the text by implying that the medical profession is only suggesting vaccines are safe for selfish financial gains. In this case, the overall message conveyed by the multimedia about vaccines is lost unless the pair are considered collectively. This is because each modality provides a *separate* contribution to the overall message, rather than merely redundant information.

In this dissertation, we focus on understanding and modeling real-world multimedia used for communication and persuasion. As the above examples illustrate, real-world multimedia violates many assumptions commonly used by most visual semantic embedding methods. Specifically, 1) that the image and text contain redundant information, and 2) that the semantics of each modality can be understood and modeled independently. These assumptions guide the design of existing visual semantic embedding methods, and while they make sense for traditional image captioning scenarios, they do not for communicative multimedia, which creates obstacles to modeling semantics in real-world multimedia. This is because if one attempts to learn embeddings of the text and image comprising a multimedia document by constraining the two's embeddings to be close in the learned space (as standard methods do), the method may be forced to discard useful nuances and subtleties of the text or image from the representation in order to ensure the image and text are close within the space. The assumption that image and text carry redundant information thus results in discarding potentially useful information from the semantic representation. In actuality, each modality contributes *complementary* information which should be preserved in each modality's representation. Thus, leveraging standard approaches suitable for captioning datasets on real-world multimedia documents runs the risk of losing useful subtleties and nuance from the representation space.

Three of the methods that we propose in this dissertation specifically target communicative multimedia, that is, multimedia used for human communication. Specifically, in Chapter 5, we study the problem of modeling multimodal political bias. That is, we seek to computationally model how news articles from biased sources manifest political bias in both the visual and textual domains. We explicitly leverage our observation that complementary information is carried in the text domain, in order to guide our model towards learning visual concepts that are politically biased. Because our dataset is so highly visually diverse and the concept of political bias so semantically complex, models tend to fixate on lower-level cues like logos which "give-away" the politics of the image, rather than learn more intuitive notions of bias. However, by leveraging the complementarity of image and text as a form of semantic guidance (i.e. guiding the visual model to learn purely visual features which complement text features), the model can successfully be targeted to capture semantic concepts of interest which are useful for *purely visual* models of bias.

Our final two methods in Chapter 6 and Chapter 7 also leverage the complementarity of image and text in communicative multimedia. In Chapter 6, we propose a general purpose method for learning visual semantic embeddings in non-traditional multimedia domains (i.e. where images and text are not literally aligned as in the case of captions). Our approach exploits the pairwise nature of multimedia data by leveraging the text domain to find neighboring multimedia documents *in text space* which are semantic neighbors. These neighbors multimedia documents contain images which are *semantic*, but not necessarily *visual* neighbors of the image in the original multimedia document. We extend the common ranking loss formulation to explicitly account for the fact that *visually dissimilar* images may still be *semantic* neighbors. We show that our approach significantly outperforms standard visual semantic embedding approaches and learns much more semantically coherent representations of semantics on real-world data in our experiments. Finally, in Chapter 7, we propose another general purpose approach for learning nuanced semantic representations in real-world data. Our approach automatically weights multimedia samples according to their predicted semantic utility in preserving abstract semantic concepts within the data which we seek to preserve and model within the space. For example, we show how visual samples of abstract concepts like "justice" or "patriotism" have more visual dissimilarity in their semantic neighbors. We enhance these images' impact in learning the shared space thereby better preserving their semantics, while simultaneously decreasing the space's reliance on straightforward / literal image-text pairs. We propose several methods for measuring the semantic utility of samples and show that they all outperform numerous state-of-the-art baselines on multiple datasets. Additionally, the semantic spaces learned by our methods better preserve subtleties and abstract concepts in their representations relative to existing methods.

Collectively, our approaches for modeling semantics in communicative multimedia exploit unique aspects of how multimedia is used in visual communication, in contrast to typical image-text datasets. By explicitly providing for these differences, our methods are able to outperform existing work which assumes more literal image-text alignments. Our approaches all leverage the fact that for purposes of communication and persuasion, images and text actually convey complementary information. We leverage this complementarity explicitly to guide the training of our cross-modal models, which enable our methods to learn more powerful semantic representations of vision and language. Overall, our approaches advance the state-of-the-art in computer vision and natural language processing by enabling more robust, cross-modal representations of semantics in real-world multimedia.

1.5 HYPOTHESES

Apart from our methodological and technical contributions, this dissertation more broadly seeks to answer the following hypotheses:

- **H1:** We hypothesize that communicative visual media differs from standard computer vision datasets in the use of visual rhetoric and visual argumentation. We hypothesize computationally understanding such media is difficult for existing vision methods, due to the lack of visual consistency across the semantic classes we study.

- **H2:** We hypothesize that the use of guided training, injection of semantics from external sources, multi-stage learning, and controlling for dataset visual diversity noise (where applicable) can empower models to capture high-level semantic concepts in visual media which would otherwise be insuperable.
- **H3:** We hypothesize that communicative multimedia differs from standard computer vision image-text datasets in an important additional way. Standard multimodal vision datasets largely contain captions which literally describe image content. In contrast, we hypothesize that images and text in real-world media convey complementary, rather than redundant information. We hypothesize that modeling the complementarity of image and text is important for understanding such media.

1.6 CONTRIBUTIONS

In recent years, computer vision has made enormous strides on a number of challenging tasks, from autonomous driving to visual question answering. We are witnessing, and part of, an artificial intelligence revolution. While traditional tasks such as object recognition and classification remain an ongoing subject of research, many new abstract, subjective, and higher-level research problems have become popular. Despite this progress, most existing vision research constrains itself to viewing images for their literal content, that is, *what* the images contain. Human beings, however, understand most images they encounter in daily life within the context of the larger social and societal milieu. We believe focusing only on images' literal content limits the applicability and utility of vision methods on a wide number of important social and societal problems. In this dissertation, we view images as tools of human communication and focus on learning *why* and *how* visual content is portrayed in the manner it is. Doing so requires going beyond a surface-level understanding of the contents of visual media, but also modeling latent messaging and abstract semantic concepts. This dissertation presents a number of techniques used to model a variety of abstract semantics within diverse, limited, and noisy datasets of real-world visual media.

We believe that computationally addressing increasingly higher-level tasks using machine learning techniques is important for several reasons. First, from a scientific perspective, recognizing emergent latent semantics inherent within visually incoherent, real-world data represents a step towards building more AI-complete machine learning systems. The ability to generalize semantic concepts from image collections, even when those visually represented concepts are not consistently rendered with the same edges, patches, colors, objects, or other representation that is amenable for convolutional neural networks to discover within the data, represents a step towards building common-sense vision systems which move beyond purely pixel-driven learning. To accurately model such semantics, systems must not merely focus on what is in an image, but must also infer *why* such concepts are present and *how* the concepts are presented. Doing this requires conceptual reasoning over not just single images, but the ability to reason over collections of images. Understanding *how* and *why* certain visual content is portrayed or relates to broader themes within the data is an important step in building more human-like vision methods capable of high-level, commonsense reasoning.

We summarize the collective contributions of this dissertation below:

- We study the problem of modeling latent visual concepts in domains whose visual manifestations are highly varied or incoherent in image space. In some cases, our domains are also noisy or limited in data. We present methods for modeling five types of latent visual concepts: photographic style, visual persuasion, multimodal political bias, and abstract multimodal semantics.
- We experimentally demonstrate the benefit of imposing constraints into the learning process for all of our various problems. We show that without these learning guideposts, models often fail to learn representations of the latent semantics of interest, and instead resort to relying on simple cues such as learning color differences or high-frequency noise within the dataset (e.g. edges, recurring patches, etc.). Discovering which guideposts to use involves exploiting unique aspects of the problem or data, such as leveraging the complementarity of images and text in multimedia.
- Though generative models are notoriously challenging to train even on traditional, visually coherent problems, we show how generative models can be trained to generate

synthetic data containing latent visual concepts. We achieve this by training our generative models on constrained data settings or by forcing our generators to use semantics features known to capture the concepts or phenomenon of interest. The synthetic data produced by our models allows us to visualize exactly what aspects of the problem the model's features are capturing.

- We present two novel methods for modeling abstract semantics in multimedia consisting of images and text. Our first method learns a feature space which captures groups of related textual words and nearby visual content. Our embedding space allows us to encapsulate within a unified joint space, complementary information which may only exist within one domain. Non-visual textual concepts relevant to paired visual concepts are embedded close in this space. We show our unified embedding has several advantages over our existing methods, such as better preservation of cross-modal neighbors and improved semantic consistency within the learned space. Our second method for learning semantic representations in communicative multimedia learns to emphasize semantically informative image-text pairs. To do so, we propose three novel metrics for measuring a multimedia sample's semantic utility and subtlety, which leverage the innate image-text complementarity within real-world multimedia. Our method results in an improved training procedure which naturally encourages abstract semantics to be captured by the learned space and provides inherent robustness to noise. We demonstrate our technique better captures abstract cross-modal semantic concepts, while outperforming multiple image-text matching methods.

Broader impacts. As the public continues to interact with online platforms, the demand for responsive platforms which tailor content to fit users' interests is growing. Methods which can automatically analyze the visual or textual content users share using the platform can form an important part of building a positive and engaging user experience by learning a profile of user interests, views, preferences, etc. There are also significant financial incentives to building systems which can infer such latent semantics from visual media. Users are more likely to remain on sites serving them content they are interested in than content they disagree with. For example, social media sites can deploy such methods to build a customized feed, where the images presented to users emblematize their bias, style, preferences, etc.

Similarly, visual ads could be better targeted by automatically understanding the visual rhetoric within the ad and then targeting ads towards users likely to engage or respond to such rhetoric. Finally, we believe that these systems fulfill a growing social need. Facebook, for example, has recently hired thousands of online moderators (at great financial expense) to combat the spread of disinformation and hate speech on the platform [343]. Social media platforms are replete with multimedia content being created and shared, to the tune of 350M new photos uploaded to Facebook per day, along with associated text posts. The abundance of such communicative multimedia poses unique challenges to existing vision algorithms due to the high-level reasoning required to understand it and its differences from the typical data encountered in vision datasets, but also provides fertile ground for research progress and opportunities for innovation. Algorithms which automatically detect fraudulent, altered, misleading, or biased multimedia being shared on online platforms could suppress the virality of such deceptive content or inform users to be aware that the content they are seeing does not represent a neutral or mainstream viewpoint. We believe that developing methods which can perceive visual media for its extrinsic semantics is an important part to building more human-like vision systems and our work is a step in that direction.

1.7 ORGANIZATION

The remainder of this dissertation is organized as follows. In Chapter 2 we present related work relevant to our completed and proposed work. In Chapter 3, we present our method and results for modeling photographic style. Chapter 4 explores the problem of modeling visual persuasion in image advertisements. In Chapter 5, we present our method for modeling visual political bias in multimedia. In Chapter 6, we discuss image-text complementarity in multimedia and propose a general purpose method for learning visual semantic embeddings by imposing novel constraints on the learned embedding space. Chapter 7 continues our work on modeling multimedia semantics and proposes an another general purpose technique for preserving abstract semantics in cross-modal embeddings by identifying and emphasizing semantically informative samples using weights. Finally, Chapter 8 discusses limitations of our methods, ideas for future work, and concludes this dissertation.

2.0 RELATED WORK

In this chapter, we present related work and differentiate our contributions in relation to the relevant literature. As we develop the relevant research context to each of these themes we compare and contrast our contributions to each of them.

The remainder of this chapter is as follows. Section 2.1 discusses visual recognition, i.e. modeling high-level semantics in visual media, as well as research related to our work on modeling visual style and visual persuasion. In Section 2.2, we examine visual concepts expressed in multi-modal data and strategies enabling machine learning techniques to generalize their models across domains. Section 2.3 discusses research related to the challenges we face across the problems we study in this dissertation, such as diverse, limited, and noisy data.

2.1 VISUAL RECOGNITION

Developing machines which can fully understand the semantics of the visual world has long been considered the "holy grail" of computer vision [183]. While complete understanding of images remains out of reach, an enormous amount of work has been undertaken in the direction of modeling many aspects of image semantics. The availability of large, annotated datasets [209, 183, 303] coupled with advances in image classification algorithms [321, 184, 127, 303], object detection [292, 368, 210, 291, 297], image segmentation [49, 409, 50, 219, 105, 205, 301], saliency [405, 193, 186, 40, 146, 159] and in other fundamental vision problems, has enabled researchers to bootstrap such approaches to model ever more complex phenomena.

While improving performance on backbone vision tasks remains an active area of research [114], the performance of currently available methods is sufficient enough that perceptive vision (i.e. what is in the image and where) research has begun to give way to those focusing on cognitive visual understanding. The Visual Genome project, for example, was designed to facilitate such research by providing a large scale dataset densely annotated with objects, their attributes and relations, region descriptions, and captions [183]. Many diverse higher-level applied vision tasks have recently been proposed and include predicting image virality [6, 67, 115], specificity [153], persuasion [406, 340, 145, 164], visual question answering [8, 338, 230, 382, 366, 423, 139, 112, 378, 379, 383, 225, 337, 389, 422, 317, 12], understanding visual humor [393, 45], and predicting physical processes in images [81, 249, 89]. The visual concepts we seek to model in this dissertation are best situated in this line of applied visual recognition research. However, we more broadly present a series of *techniques* for studying such phenomena, independent of the actual problems we study in this work. Several of the problems we study make use of prior visual recognition methods retrofitted to our domains. We first describe related research in semantic modeling and then discuss relevant applied research for modeling visual persuasion and artistic style.

2.1.1 Modeling mid-level visual semantics

Attribute-based representations have long been considered a mid-level semantic representation [217, 191, 48]. Attributes are visual concepts shared across multiple objects or scenes. Attributes are a particularly attractive image representation because they capture generalizable intuitive semantics [94] and as such have found broad use in applications ranging from object recognition [188, 87, 371, 3], action recognition [42, 419, 76, 106, 352, 370], image captioning [37, 9, 8, 393, 354, 224, 277, 298, 161, 394, 356, 384], image search [181, 162], and visual categorization [188, 54, 331]. Such a representation is particularly useful for the types of high-level phenomena we propose to model; a representation capturing "attractiveness" may be particularly useful for distinguishing images advertising makeup products, for example. By explicitly using a mid-level semantic representation, we enable our models to learn associations of visual concepts with categories. Because of the diversity within our datasets, finding an easily generalizable representation of semantics which our models can reason about is crucial.

More specifically, our work on generating faces in advertisements (Chapter 4) and our work of modeling bias in political ads (Chapter 5) rely on attribute representations. We find that the faces we encounter in advertisements are so diverse that our models have difficulty learning discriminative properties of those faces which are class-indicative. In order to model such semantic properties of faces within our dataset, we found it useful to condition our generative model on facial attributes we predict on faces from ads. The facial attributes we use include semantic visual concepts like “bald,” “rosy cheeks,” “smiling” or “attractive” [217, 188, 87]. We also include facial expressions and emotions as part of our facial attribute bank. Facial expression and emotion recognition is an established and popular topic [84, 166, 314, 215, 247]. We include seven canonical expressions as part of our face modeling pipeline of ads: happiness, sadness, surprise, fear, disgust, anger, and contempt. For modeling political bias, we believe semantic attributes, as well as facial attributes, may provide a useful signal. For example, modeling how immigrants are portrayed in politically diametrically opposed media sources could be used to identify attributes which are strongly correlated with one political position. Such attributes could then be integrated into a larger pipeline for generating synthetic biased images.

2.1.2 Visual rhetoric and communication

Three of the visual concepts that we model in this dissertation are related to works that study visual persuasion and visual argumentation [164, 145, 155, 178, 324]. Visual persuasion is the use of visual media to persuasde viewers to make decisions [164]. [164] propose to study the “persuasive intent” of images by predicting how media sources seek to portray politicians. The approach relies on facial attributes and gestures to assess the communicative intent of photos, predicting, for example, that an image shows a politician as competent, energetic, trustworthy, etc. [145] extends the work of [164] by incorporating a wider range of features, such as body pose and image setting. [165] train classifiers to predict the outcomes of elections based on the candidates’ faces, but none of these works create generative models. [149] propose a dataset of advertisements, and predict what message the ad conveys (e.g. “buy this car because it is spacious”) but they do not model or generate the visual appearance of the same object across ad topics. Also related is work in modeling style in fashion and architecture [72, 198] but none of these build generative models as we do. All of these works are based on careful and expensive human annotations, while we aim to discover facets of visual rhetoric in a weakly supervised or automatic way.

Our work in modeling visual rhetoric is much more subtle, challenging, and less "visual" than many existing methods. For example, our work on modeling photographic style is different from past work in that professional photographers painstakingly construct their photos to convey an emotion or tell a story about their subjects. The message told by such photographs is often less overt and subjective. Thus, modeling how subjects are portrayed in photographs is an important part of developing techniques which take seriously authorial intent and move beyond merely perceptive vision, significantly differentiating our work from past work. Similarly, our work on modeling bias requires understanding *how* and *why* different media sources portray different subjects. For example, anti-immigration sources may frequently show criminal immigrants in a disproportionate number compared to other sources. Such a correspondence is likely not accidental: the biased nature of the portrayal may be discovered, for example, by finding such disproportions in both the visual and textual domains from that source compared to other sources. Our work on bias therefore extends our techniques for modeling rhetoric and persuasion into a multimodal setting.

2.1.3 Visual style

In this section, we introduce the concept of visual style modeling. It has long been noted in the media studies community that the concept of style suffers from an overabundance of interpretations. One working definition, of particular relevance to the machine learning community, is given by David Bordwell, a distinguished film theorist, who defines style as the "patterned use of a medium's techniques" [38]. We find this definition pleasing because it is not restrictive of the type of medium and also suggests the recurrence of patterns which a machine may be able to model. We note that other vision works have modeled other visual styles beyond those discussed in this dissertation, including fashion style [5, 173], urban style [78], and product styles in e-commerce images [116]. We consider two visual styles in this work: photographic style, in which a professional photographer's particular artistic traits are reflected in his or her work, and persuasive or communicative style, which describes *how* a visual artist uses visual rhetoric within the image to persuade or communicate with viewers.

Photographic style modeling is also of relevance to our work on modeling bias. For example, photographic styles may belie the authorial intent of the photographer, which is of interest when assessing bias in photographs. We contrast our work on modeling visual style with work in modeling artistic style, which is commonly defined as the colors, textures, brush strokes, or dominating geometric patterns comprising artistic works, such as paintings or drawings [98, 97].

Modeling artistic and photographic style. The task of automatically determining the author of a particular work of art has always been of interest to art historians whose job it is to identify and authenticate newly discovered works of art. The problem has been studied by vision researchers, who attempted to identify Vincent van Gogh forgeries, and to identify distinguishing features of painters [287, 88, 163, 59]. While the early application of art analysis was for detecting forgeries, more recent research has studied how to categorize paintings by school (e.g., "Impressionism" vs "Secession") [305, 171, 163, 313, 15, 20, 29]. [29] experimented with a simple dataset of 7 painters with very different styles and achieved good results with low-level features due to the dataset's simplicity. [305] explored a variety of features and metric learning approaches for computing the similarity between paintings and styles. Features based on visual appearance and image transformations have found some success in distinguishing more conspicuous painter and style differences in [29, 313, 171], all of which explored low level-image features on simple datasets. Recent research has suggested that when coupled with object detection features, the inclusion of low-level features can yield state-of-the-art performance [20]. [15] used the Classeme [344] descriptor as their semantic feature representation. While it is not obvious that the object detections captured by Classemes would distinguish painting styles, Classemes outperformed all of the low-level features. This indicates that the objects appearing in a painting are also a useful predictor of style.

This dissertation considers photographic authorship identification, but the change of domain from painting to photography poses novel challenges that demand a different solution than that which was applied for painter identification. The distinguishing features of painter styles (paint type, smooth or hard brush, etc.) are inapplicable to the photography domain. Because the photographer lacks the imaginative canvas of the painter, variations in photographic style are much more subtle. Complicating matters further, many of the photographers in the dataset we collect are from roughly the same time period, some even working for the same government agencies with the same stated job purpose. Thus, photographs taken by the subjects tend to be very similar in appearance and content, making distinguishing them particularly challenging, even for humans.

There has been related work in computer vision that studies aesthetics in photography [234, 251, 68]. Some work also studies style in architecture [72, 197], vehicles [198], or yearbook photographs [102]. However, all of these differ from our goal of *identifying authorship* in photography. Most related to our work on predicting photographic authorship is the study of visual style in photographs, conducted by [168]. Karayev et al. [168] conducted a broad study on both paintings and photographs. The 20 style classes and 25 art genres considered in their study are coarse (HDR, Noir, Minimal, Long Exposure, etc.) and much easier to distinguish than the photographs in our dataset, many of which are of the same types of content and have very similar visual appearance.

While [168] studied style in the context of photographs and paintings, we explore the novel problem of *photographer identification*. We find it unusual that this problem remained unexplored for so long, given that photographs are more abundant than paintings, and there has been work in computer vision to analyze paintings. Given the lower level of authorial control that the photographer possesses compared to the painter, we believe that the photographer classification task is more challenging, in that it often requires attention to subtler cues than brush stroke, for example. Besides our experimental analysis of this new problem, we also contribute the first large dataset of well-known photographers and their work. We also propose a method for generating a new photograph in the style of an author. Similarly, in our work on modeling visual persuasion and political bias we generate new photographs containing persuasive or communicative styles which we model in our work. We note that this problem is distinct from artistic style transfer (discussed below) [17, 39, 16] which adjusts the tone or color of a photograph.

Transferring artistic style. In this dissertation, we focus on modeling meaningful semantic concepts, such as photographic or persuasive style. We then use generative to generate synthetic data in order visualize what our models have learned. In our work on visual persuasion and political bias, we show how an existing image can be modified to bear the persuasive (or biased) styles we model. Our work is thus a type of *semantic*, rather than artistic, style transfer. Artistic style transfer methods attempt to render the content of one image in the artistic style of another. For example, we might modify a portrait to have the same artistic style as "Starry Night" by Van Gogh. Importantly, these methods do not seek to change the *semantics* of the image, but instead focus on changing low-level details, such as textures or colors. Early methods primarily rely on low-level (and often handcrafted) patch-based texture features [80, 131, 187, 318]. More recently, impressive results have been achieved using features extracted from pre-trained convolutional neural networks (CNNs) [97, 160, 83, 227, 367]. Gatys et al. [97] showed how style transfer can be formulated as iterative optimization that seeks an image which produces the same CNN activation statistics of both the "content" and "style" images. Follow-up works [160, 346] improve efficiency by performing style transfer in a single feed-forward pass, but these can only transfer towards those styles present during training [47]. Recent methods [47, 144, 204] combine the speed of feed-forward networks with the flexibility of optimization-based approaches, enabling fast style transfer on arbitrary styles. Unlike our work, all of these approaches focus on transferring low-level textures, while keeping the semantics of the produced image the same. In our setting, we seek on transferring high-level semantics, which change the meaning of the image itself.

2.1.4 Generative models of semantics

We described above how artistic style transfer methods seek to transfer only low-level visual details, without changing the semantics of images. We now discuss related work on *generative* models, which are capable of transferring or generating images with certain semantics. Generative models are models which seek to learn the underlying data distribution they are trained on such that they can generate synthetic data which appears to come from that distribution [240]. Many such models learn from target data distributions without the need for any labeled data, and thus share many features found in unsupervised learning. In this dissertation, however, while we too are interested in producing qualitatively good generations that appear to come from the visual distribution of images on which they are trained (i.e. are realistic), our *primary* focus is producing generations which capture the underlying *semantic distribution* of the data. For example, in our work on modeling political bias, while we wish to generate realistic looking images, we desire that the images we generate manifest the underlying semantics of politically biased images which we have modeled (i.e. we desire the images we generate to exhibit the specified type of political bias).

In computer vision, many generative architectures have been proposed, including Restricted Boltzmann Machines [256, 334], Deep Belief Networks [194, 215, 246], PixelCNNs [263, 306, 347], Plug & Play networks [254, 268, 293, 392], and real NVP [70, 62]. However, the two most prominent generative models by far are variational autoencoders (VAEs) [137, 137, 288, 359, 176] and Generative Adversarial Networks (GANs) [240, 385, 26, 214, 77, 120, 110, 293, 151, 233, 387, 411, 402, 53, 420, 32, 14]. Each of the applied problems we study in this dissertation (photographic style, visual persuasion, political bias) involve some element of generative modeling and image generation, particularly using VAEs or GANs. In this section, we discuss related research relating to autoencoders and generative adversarial networks, variations of which are used in several of our methods.

Fundamentally, a variational autoencoder [176] is probabilistic graphical model [137] comprised of two distinct components: the encoder network transforms the input into a low-dimensional latent representation, while the decoder network is tasked with reconstructing the input from the latent representation. Regularization is imposed on the latent distribution, typically by imposing a constraint to enforce that the distribution approximates a unit Gaussian [190]. The two components are trained end-to-end via a reconstruction loss which penalizes deviations from the output of the decoder to the input. Autoencoders [133, 359, 288, 229, 189, 386] are an older type of generative model compared to generative adversarial networks, but perform reasonably well when trained with recent perceptual loss functions [137]. Numerous variations of VAEs have been proposed. Similar to our task of generating faces in ads and modeling political bias, [386] condition their model on facial attributes in a conditional variational autoencoder (CVAE) framework. We rely on a custom conditional variational autoencoder conditioned on facial attributes and expressions in Chapter 4 and Chapter 5. [75] uses the structural similarity score to improve visual quality of the reconstructions. Several works [190, 229, 240] for example, have worked to combine the sharpness of generations typical of generative adversarial networks with the flexibility of the VAE framework.

Generative adversarial networks (GANs) have enjoyed enormous popularity after their introduction in 2014 by Goodfellow et al. [110]. GANs are generative models which learn to transform vectors of random numbers into images [110]. GANs consist of two components, the generator is trained to produce outputs which are indistinguishable from the real distribution (in the vision case, this equates to generating realistic images), while the discriminator is trained to detect if a given image is real or fake. In order to become better at detecting fakes, the discriminator continuously learns aspects of the real and synthetic distributions. The generator, in order to fool the discriminator, attempts to cause the synthetic image distribution to match the real distribution. In practice, this equates to the generator producing increasingly realistic fakes. Many approaches [254, 386, 135, 373, 270] have also introduced semantic conditioning into GAN models, where GANs are tasked with not only producing realistic images, but also producing images bearing certain semantics. However, essentially all of these study generating "surface-level" semantic phenomena which has clear visual groundings, such as gender [369], attractiveness [69], or age [11]. In this dissertation, we leverage a custom GAN based on [26] in our work on modeling visual persuasion in order to generate synthetic objects which capture their appearance in different categories of ads.

2.2 MULTIMODAL LEARNING

The same semantic concept can be expressed in many different types of data. For example, the concept of dog can be expressed in a photograph, spoken and recorded in audio form, represented in natural language, drawn in a sketch, or illustrated in cartoons or clipart. Our overarching goal in this dissertation is to develop methods for modeling abstract visual phenomena in diverse datasets. In addition to visual diversity within data of the same type (e.g. photographs), however, there may be other *modalities* or *domains* of data which complement one another. For example, in our work on modeling bias, images are accompanied by their textual descriptions and the text of the article from which they came. In such a scenario, with interacting text and image data, the entire problem can be described as being *multimodal*, in that it features data from more than one modality. We believe that by exploiting the complementarity of such parallel modalities, we can better model visual bias than models which rely on single-modal data alone.

Our task of modeling multimodal bias with groundings in image and text relates to a number of works which model inter-modal correspondences between natural language and images. We believe that much, but possibly not all, of the bias expressed in text has examples of corresponding visual groundings. Racist language, for example, may correspond to certain portrayals of minorities. One of the features of this problem we propose to study is to learn multimodal alignments of bias in image and text. In so doing, we intend to build models which account for examples of bias which may have been missed without the complementarity of the other domain.

Our work in Chapter 6 and Chapter 7 build on the idea of modeling complementarity in multimedia in a more general way. Rather than target a specific task, such as political bias, we instead learn *general* purpose representations of abstract semantics that can be used for any such task. We then evaluate these representations in the context of cross-modal retrieval. We thus include related work to learning multimodal alignments between images and text and visual semantic embeddings. Unlike standard multimodal representation learning approaches which assume that each modality carries redundant semantics as in the case of image captions, we target the much more challenging domain of communicative multimedia, where each modality is complementary to the other and plays a part in conveying the overall semantic message of the image-text pair. We discuss in detail the relationship of work in modeling abstract semantics in multimedia to traditional methods in metric learning and for learning visual semantic embeddings in this section. We first begin with a discussion of task-specific methods, such as for modeling multimodal political bias before moving into task-agnostic semantic representation learning methods.

2.2.1 Integrating text and vision

Our task of modeling visual bias in images is somewhat different from the other problems we study because of the presence of the complementary text domain. Recently, there has been a great amount of research into multimodal data fusion of images and text [226]. Image captioning methods [177, 8, 298, 394, 355, 226] learn to output textual descriptions of images by learning from human-written captions. Visual question answering (VQA) techniques [12, 138, 112, 230, 8] seek to allow machines to answer natural language questions posed about visual concepts in a given image. VQA models typically work by learning an embedding of the question, combining it with a neural image representation, and then allowing a recurrent model to generate the textual answer [378]. Other visual-textual grounding problems have been proposed, such as visual dialogue [63, 46, 311, 237], in which the system discusses visual concepts within an image with a human user, and text to image synthesis [385, 402, 411, 44], which generate images from given textual descriptions. All of these techniques require cross-modal grounding of visual concepts from images with natural language.

Other works [232, 286, 396, 141, 208, 179] more explicitly study grounding natural language concepts with visual media. [141] ground natural language objects with images in the context of retrieval. [396] learn the meaning of natural language phrases by watching associated short video clips. Recent work [222, 140, 223, 123] propose to integrate text and vision using end-to-end trained multimodal transformer-based architectures. [179, 286] model coreference between images and text to determine what visual objects are being mentioned in associated text. Our work of modeling cross-modal bias between images and paired text is similar to these works but differs in that many of the visual concepts we seek to model are much more abstract. These works presuppose strong correspondences between objects appearing in images and associated text. In our case, the concept is *implicit* within the text and visual domains. Moreover, representations of bias are likely to be more diverse than variations in appearance of the same object, due to the many ways that bias can be expressed. Collectively, these two aspects of our problem make modeling bias significantly more challenging than typical visual-text grounding tasks. We more thoroughly discuss works which attempt to model bias below.

2.2.2 Learning visual semantic embeddings

A fundamental problem in cross-modal inference is the creation of a shared semantic manifold on which multiple modalities may be represented. The goal is to learn a space where content about related semantics (e.g. images of "border wall" and text about "border wall") projects close by, regardless of which modality it comes from. We note that such embeddings are generally *task-agnostic*, that is, they seek to learn a representation preserving cross-modal semantics, in the absence of any particular applied task (e.g. bias detection). However, such *visual-semantic embeddings* (VSE) have received tremendous interest due to their broad down-stream applications such as retrieval [43, 328], captioning [169, 384], tagging [92], and visual question answering [383]. Most VSE approaches learn a joint visual-text space where some distance metric between embedded samples reflects their semantic relationship [377]. Following the early deep VSE models [92, 241] research has focused on improving the learning objectives [353, 364, 113, 360, 361], e.g. to preserve order [353] rather than distance, to preserve structure within modalities [364], to ground embeddings via generation [113], or to provide modality invariance [360, 361]. Others leveraged properties of text to improve the visual representation, e.g. through cross-modal attention techniques [195, 253] which consider all possible alignments between detected regions and words. [147] extract visual concepts from images and organize them semantically using the paired text (to determine their correct semantic order).

Unlike the above approaches which rely on additional tasks, losses, and may require extra annotated data, our approaches exploit the structure of each unimodal space (image and text) by leveraging the semantic complemetarity found in communicative multimedia. We propose two approaches for learning task-agnostic visual semantic embeddings, one relying on a complementarity-based loss which imposes constraints to preserve intra and inter-modal semantics, and another relying on a sample weighting strategy which leverages complementarity between the image and text modalities to assess whether samples are semantically informative. Both our methods use traditional, well-understood two-stream visual semantic embedding models trained via ranking losses, such as [92, 85, 328].

We first discuss work related to our novel loss contributions for learning visual semantic embeddings (Chapter 6). Most image-text embedding methods rely on a two-stream architecture, with one stream handling visual content (e.g. captured by a CNN) and the other stream handling textual content (e.g. through an RNN). Both streams are trained with paired data, e.g. an image and its captions, and a variety of loss functions are used to encourage both streams to produce similar embeddings for paired data. One common loss used to train such retrieval models is triplet loss, which originates in the (single-modality) metric learning literature, e.g. for learning face representations [309]. In cross-modal retrieval, the triplet loss has been used broadly [252, 418, 245, 271, 390, 85]. Alternative choices include angular loss [363], N-pairs loss [326], hierarchical loss [99], and clustering loss [259]. Triplet loss [309, 134] takes into account the *relative* similarity of positives and negatives, such that positive pairs are closer to each other than positives are to negatives.[408] generalize triplet loss by fusing it with classification loss. [260] propose a lifted structure loss which integrates all positive and negative pairs within a minibatch, such that all pair combinations are updated jointly rather than independently. [364] propose a structural loss, which pulls multiple pieces of text paired with the same image together, but requires more than one ground truth caption per image (which most datasets lack). In contrast, our approach pulls semantically similar images *and* text together and only requires a single caption per image.

While single-modality losses like triplet, angular and N-pairs have been used across and within modalities, they are not sufficient for cross-modal retrieval. First, these losses do not ensure that the general semantics of the text are preserved; thus, the cross-modal matching task might distort them too much. This phenomenon resembles forgetting [207, 111] but in the cross-modal retrieval domain. Second, these losses do not exploit the complementary relationship between images and text found in communicative multimedia. In particular, two images might depict substantially *different visual* content but nonetheless be *semantically related.* For example, one image of a wedding might show a couple dancing, and another show a large number of guests eating at several tables; these images are visually diverse but still semantically related. However, there is no component in standard metric learning losses that enforces this semantic coherence at the image level. This is less of a problem in the case of traditional image captioning datasets featuring literal image-text descriptive relationships. In contrast, in real-world communicative multimedia, the complementarity of image and text is much more pronounced. Note that we do not propose new *models* for image-text alignment, but instead propose cross-modal embedding *constraints* or weighting metrics which can be used to train any such model. For example, we compare to Song et al. [328]'s recent polysemous visual semantic embedding (PVSE) model, which uses global and local features to compute self-attention residuals. Our loss and weighting based approaches improve upon [328]'s performance. Our work is also related to cross-modal knowledge distillation [92, 325, 119, 103], which transfers supervision across modalities. None of these approaches exploit cross-modal complementarity, e.g. the semantic signal that text neighborhoods carry for the image space, to constrain a learned metric space as we do. Finally, [406, 185] detect different types of image-text relationships (e.g. parallel, complementary) but do not retrieve across modalities.

We propose a second approach (Chapter 7) for learning semantically robust embeddings in communicative multimedia which relies on weighting samples judged to be abstract (i.e. exhibiting latent visual concepts) and therefore important for learning. Our work again exploits image-text complementarity in order to estimate the emphasis the model should pay to a given sample. Our work is thus related to work on sample mining and weighting-based methods. For example, it has long been known that triplet loss can be challenging to train [129] due to the difficulty of choosing informative dissimilar samples. Many have exploited hard negative mining [85, 130, 309, 320, 397, 142], while others have tackled issues stemming from negative sample choice [362, 272, 60, 326, 395], e.g. by pushing multiple negatives away [326]. For example, [326] push multiple negatives away at a time, lessening the need to pick a single hard negative, while [395] correct the distribution shift on the chosen triplets relative to the dataset. Other approaches [43, 360, 413, 65] rely on the use of classification labels or metadata, e.g. to ensure negatives in the triplet belong to different classes than the positive. Unlike these, our approach works in self-supervised settings without the requirement of additional labels. Rather than hard selection of negatives, others have used soft weights over samples. In [243], positive samples which violate the margin but are still correctly retrieved are weighted less, while others incur a larger penalty. [212] use sample weights to address hubness (a phenomenon where a small number of embeddings remain undesirably close to many others), such that samples which are hubs receive more attention. Our weights are designed to improve the semantic properties of the learned space by emphasizing samples where the relation between image and text is *abstract*, not necessarily “hard” samples. This is an important distinction, because some “hard” samples may actually be noisy; we found using hard negative mining prevented methods from training successfully on several of our challenging datasets. Our method outperforms [243] and [212]. We show that our method significantly better preserves challenging, abstract and latent semantic concepts such as “justice” or “freedom” in real-world multimedia in a task-agnostic, data-driven manner.

2.2.3 Multimodal bias

There has been significant prior work [1, 64, 91, 90, 96] in modeling bias in text from the natural language processing community. [96] trains word embeddings on 100 years of text data and tracks how the embedding changes over time, showing correlations with changes in attitudes towards minorities and other groups. [90] studies how stereotypes are expressed in language, finding for example, that terrorism and political conflicts are often mentioned in respect to Muslims. [345] trains models to detect racism in text. Other relevant work [350] explore bias in image descriptions, finding for example, that the race of a person isn’t mentioned unless the person is black or Asian. [412] models gender bias in semantic roles (e.g. cooking is associated with females) and leverages text data to impose corpus-level constraints to correct for bias in classifiers. [71] study stereotypes in media sources, finding for example, that Latinos are disproportionately represented as illegal aliens, while Muslims are portrayed as terrorists 81% of the time they appeared on television [90, 71]. [302] demonstrate gender-based bias in a natural language corpus. [41] discover phrases that lead humans to incorrectly guess social features of the author. [300] study bias in news sources, demonstrating that the selection of topics and issues discussed originates from a male-centric perspective. [37] seek to correct bias in image captioning, stemming from learned gender priors. For example, they aim to correct captioning models which incorrectly assume the gender of actors in the image based on the task being performed (e.g. snowboarding=male) without looking at the actual person. Our work is related to, but distinct from, these works in that we seek to identify bias in media photographs by leveraging cross-modal information from biased text. We show that the understanding of bias we obtain by multimodal fusion of the image and text domains allows us to model types of bias single domain models would otherwise miss. Our work is also different from these in that we *generate* photos containing particular types of bias, rather than just identify or describe them. [290] and [22] use carefully designed dictionary, lexical, grammatical and content features to detect biased language, using supervision over short phrases. We leverage [290]'s technique to discover biased word usage in our dataset. Others [281, 56, 57, 58, 376, 358] have studied predicting political affiliation from text, mainly in the context of social media. In contrast, it is not clear what "lexicon" of biased content to use for images.

We investigate the bias in how events, topics, and people are portrayed in the media. This type of bias is directly related to bias in human relationships, i.e. human perceptions that people of a particular group (demographic, political, etc.) have certain qualities or beliefs. This bias over human qualities is evident in data that can be used to train machine learning algorithms, and has thus been tackled in a few prior works [37, 412, 304, 24, 31]. For example, [37] ensure that the same classifier is equally likely to fire on images of men and women when the relevant property (e.g. "snowboarding") is present. In contrast, rather than *debiasing* models, we aim to *model* and predict the *type* of political bias.

Other works [310, 262, 255, 82] have analyzed the bias inherent in human annotated data introduced by crowdsourcing annotations. [267] show that sexist workers are less likely to find image search results biased. [74] show that different ethinic groups tend to label the same images differently. In contrast to these works, we show *how* media sources already believed to be politically biased then exhibit that bias, both in terms of the visual content they choose to accompany article text and in terms of the text of the article itself. We also explicitly ask our workers to provide their rationale for their predictions and then *leverage* the stereotypical and biased notions used by the workers to model bias in visual media.

2.3 FACILITATING LEARNING IN CHALLENGING DATA SETTINGS

All of the problems we study in this dissertation involve modeling abstract semantic concepts in communicative media. These datasets pose unique challenges for learning compared to traditional computer vision datasets. For example, most computer vision datasets for supervised learning feature images containing recurring types of visual content, such as objects [211], faces [118], or scenes [415]. Though how the content is portrayed may vastly vary from image-to-image, the content itself remains the same (i.e. images are known to contain or not contain a particular type of content) and form a type of closed visual world. In contrast, in the datasets we consider, all that is known is that the content contains a certain latent semantic concept (e.g. a particular photographer took the photo, the image came from a right-wing website, etc.). In this case, the images come from an open visual world, where the image content could show any imaginable scene type, any imaginable object, contain illustrations, etc. This results in an enormous amount of visual *diversity* to the data, which can make it difficult for models to extract meaningful patterns corresponding with the semantics we seek to model. Compounding matters further, some of our datasets are also *limited* in size. These datasets exhibit high visual diversity, but also relatively few images containing the semantic concepts we wish to model. This renders our models particularly vulnerable to overfitting to low-level details in the data which models memorize (but which do not generalize beyond the small dataset). Finally, all of the datasets we target in this dissertation feature data harvested from the web, with minimal to no human supervision. While we know that overall there is a semantic signal within the semantic labels (e.g. type of political bias) assigned to our data, in many cases, the label itself may be incorrect or the image paired with the label may be irrelevant and incorrectly scraped from the webpage. Thus, several of our datasets contain a high amount of both label and image *noise*. We discuss related work to each of these challenges below, as well as their impact on learning generative models.

2.3.1 Visual diversity

There has been a substantial amount of research exploring the impact of visual diversity in computer vision datasets [19, 322, 18, 278]. Many approaches [329, 372, 121] seek to *increase* visual diversity within training data. By showing the same object in multiple poses, scenes, and visual settings, models become more capable of recognizing the object in varied settings [19]. [278] study the relationship of visual diversity to *semantic* diversity and show that visual variability of the data has a significant impact on models' performance for image classification. However, all of these works study traditional computer vision problems, where visual diversity might be expected to help recognize the same object or scene type. In contrast, in our setting we study *latent* semantic concepts, where images lack consistent visual patterns, objects, and scenes. Visual diversity poses a substantial challenge in our case, because it makes extracting a meaningful signal more difficult. For example, models may be more likely to capture meaningful visual patterns of political bias if the dataset consisted entirely of political protest images on both sides of the political spectrum, since the content between images is the same. In contrast, our dataset contains images of all types, including illustrations, making extracting a repetitive signal much more difficult. [109] propose an approach for learning binary classifiers when the negative class is too visually diverse. In contrast, our methods seeks to distinguish between multiple positive classes. Most relatedly, [218] propose "open long-tailed recognition" for learning object representations which capture known objects as well as unseen objects in an open-world, visually diverse setting. In contrast, our datasets lack any object annotations and we primarily focus on particular abstract tasks, such as photographic style, visual persuasion and political bias. Additionally, our general-purpose representations capture abstract semantics in *multimedia* (i.e. images and text), while [218] focus on capturing specific objects (both seen and unseen) in images only.

2.3.2 Limited data

Most recent computer vision methods rely on deep learning methods trained on large-scale datasets in order to learn discriminative patches and semantics [403]. While obtaining large amounts of data is possible for many traditional computer vision problems, such as object recognition, by leveraging pre-existing datasets such as Imagenet [303] and Visual Genome [183] or by crawling image search engines directly [182], it is not possible to obtain large amounts of visual data containing the concept one seeks to model in all cases. For example, if one seeks to model the photographic style of a particular photographer, there are likely to be only a relatively small number of photographs by the photographer (hundreds to at most thousands), much less than the millions required to train modern deep CNNs [330].

Training models to quickly grasp semantic concepts from limited data is known as few-shot learning [333, 289]. Most few-shot learning methods themselves can be characterized as a form of transfer learning, where models first learn semantics from large-scale, often labeled datasets, and then leverage the pre-learned semantic representations on the target task by fine-tuning [117, 180]. All of our methods in this dissertation make use of some form of transfer learning, by either relying on semantic categories such as facial attributes [25, 217], or by initializing models with features obtained by pre-training on Imagenet [125]. Other approaches to few-shot learning include adding auxiliary tasks for which there is abundant labeled data [332] and leveraging semi-supervised learning where models learn using both labeled and unlabeled data [400]. While we too make use of knowledge transferred from other datasets, our work conceptually differs from most work in few-shot learning because the types of semantic concepts we seek to model are much more abstract. Most few-shot learning approaches focus on recognizing objects [167] or scenes [154] with little to no target data. In contrast, we seek to model highly abstract and latent concepts like photographic style or visual persuasion.

2.3.3 Noisy data

Noisy or outlier data is a perennial problem in computer vision when relying on data automatically harvested from the web. Even when crowdsourcing is used to obtain human annotations on such webly-harvested data, they too often contain some noise and require special handling [33] or acquiring multiple annotations per image, which increases expense. All of the datasets studied in this dissertation contain some level of noise due to automatic harvesting. One way uncertain or noisy data can be mitigated is through the use of noise-robust losses [235, 36, 351, 275, 410]. For example, [410] propose generalized cross entropy which down-weighs the gradient on highly incorrect samples. Others [158, 296, 196] predict weights for samples based on the estimated reliability of the data, but require some clean data. [122, 203] leverage self-learning approaches where models first train on noisy labels, then predict*pseudo-labels* on the data which are also used for training. However, self-learning can suffer from error-amplification if the model incorrectly learns from the initial noisy labels. Moreover, in our settings the problem of noisy labels is more pronounced due to the fact that the semantics we seek to model are *latent* within the data and lack a consistent visual appearance across the class. This makes it more difficult for models to learn the concept, particularly in the presence of label noise, due to the visual incoherency within the class. We indirectly handle noise in the case of our photographic style and visual persuasion projects by restricting the types of features the model can learn (for photographs) or by restricting the data by training only on faces detected within the images (for ads). We explicitly handle noise for our task of modeling political bias by applying automated techniques to clean the data.

In our work in modeling multimodal political bias (Chapter 5), we propose a two-stage approach where the text is used to guide the visual model towards semantics of interest. Then, in a second stage, we remove the requirement of text and learn to make purely visual predictions. Our method thus leverages the text domain as a form of guidance to contend with high noise and visual diversity. Simiarly, our work on learning general abstract semantics in multimedia also address noise. Because of the challenge of cross-modal image-text matching, approaches for contending with label noise have also been adopted in the retrieval setting. [244] learn image-text embeddings on noisy web data by exploiting metadata (tags) while [421] conditions a generative model on noisy texts. In contrast, both of our methods for learning abstract semantics require no annotations or metadata beyond image-text co-occurrence. Our first approach (Chapter 6) relies on the image-text complementarity found in communicative multimedia and makes use of semantically neighboring images, which is inherently robust to noise. Our second method (Chapter 7) also relies on complementarity and *explicitly* handles noise by enhancing semantically informative samples, while down-weighting samples suspected to be outliers. Most similar to ours, [7] estimate density by computing the correlation between samples from different modalities. We too aim to detect outliers, but we model density in both the image and text spaces independently through modality-specific variational Gaussian mixtures [28]. This has the benefit of taking global statistics into account, e.g. a sample from a small tight cluster of outliers would be weighted low by our approach, but high by [7]. We show our approach outperforms [7].

Weakly supervised learning. Recently, weakly supervised approaches have been proposed for classic topics such as object detection [265, 55, 414, 375, 391], action localization [365, 299], etc. Researchers have also developed techniques for learning from potentially noisy web data, e.g. [51]. Also related to our work is work in unsupervised discovery of patterns and topic modeling. For example, [323, 416] use an iterative clustering-detection pipeline to discover patterns that occur frequently but are discriminative. [199, 206] and [319] leverage deep networks to mine discriminative patterns. [152] and [72] discover patterns informative for the architectural style of a city or the evolving design of cars over the decades. Both of these rely on finding clusters of image patches that are compact in terms of the top-level weak label (e.g. "Paris" or "1950s car"), i.e. clusters that primarily contain samples from a given label, and ignore clusters with near-uniform label distribution.

Our work is related to weakly supervised discovery methods in the sense that other than often noisy labels, our method does not receive information about what makes an image contain the latent visual concepts we seek to model. In contrast to these weakly supervised discovery works though, the problems we study exhibit *much* larger within-class variance (e.g. with the classes being photographer's identities, types of ads, or whether an image is politically biased). Unlike objects and styles, the differences between our classes live in semantic space as much (if not moreso) than they do in visual space, thus these methods do not guarantee success. Nevertheless, we borrow intuitions from these methods and help our methods by focusing them on the higher-level semantics of the problem, such as by injecting external semantics or leveraging guided training.

Curriculum learning. Several of our methods use multi-stage training as a strategy to facilitate learning (Table 2). Thus, also relevant to our work are self-paced and curriculum learning approaches [157, 282, 398, 409, 158]. These attempt to simplify learning by finding "easy" examples to learn with first or by leveraging multi-stage training procedures. Several of our methods employ a type of curriculum learning. For example, we first train a multimodal classifier to predict bias, using the assumption that the relation between text and bias is more direct. We then leverage this model as a feature extractor by adding an image-only politics classifier on top of it. Thus, our method focuses the model on relevant visual concepts using text. Related work by [164] and [108] both learn semantic concepts on a separate, auxiliary training task, which aid the classifier in performing inference on the target task. Because prior work [266, 128] has shown that using a larger-batch size improves classification performance on noisy data by smoothing the gradient, we compare against a baseline curriculum-learning approach designed to alleviate the problem of noisy minibatches in our work on photographic style and predicting political bias. To do so, we freeze the lower-layers of the model after training and then perform a second stage of training of just the classifier using all features in the train set for optimization, which we show slightly improves performance on both of these problems.

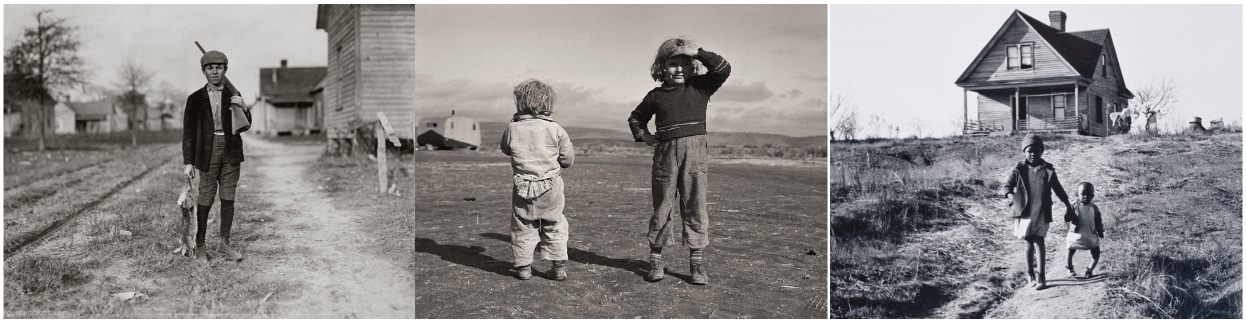
2.3.4 Impact on generative models

Though we primarily seek to model latent semantic concepts discriminatively in this dissertation, a secondary focus of our work is generating synthetic data which visualizes the semantics we have learned. However, the above challenges pose major obstacles to training generative models on our datasets. Despite enormous progress having been made in generative modeling in recent years, many problems persist [13]. Even when advanced stabilization techniques are used, GANs are notoriously challenging to train, particularly when the underlying data distribution is highly noisy [120]. In such cases, the generative component of the network, which must learn a complex distribution, is overpowered by the discriminative component, which much merely tell apart the generator's output from the real distribution. This destabilization of the equilibrium between both the generative and discriminative components of the framework results in training failures which are frustratingly complex if not impossible to overcome. This difficulty is particularly relevant in the domains that we study in this dissertation, in which the *manifold in visual space* of a higher-level semantic concept is highly cragged, making fitting the distribution all the more challenging [120, 172]. A simpler problem, though no less fatal one, is that generative models require a large amount of data to achieve a good fit of the target data distribution [120]. While discriminative models are also vulnerable to over-fitting, they do not need to learn the contours of the underlying distribution. Oftentimes, when limited data is available, discriminative models can obtain surprising performance by training a linear classifier using features extracted from a network trained on another domain [315]. This is not possible in the generative case, where the underlying target distribution must be learned. Without adequate data, such models simply memorize the target dataset and reproduce the samples without any diversity or generalization [66, 213]. [312] present a highly interesting approach for learning a generative model which can be learned from a single natural image, but can change the semantics of the image by reshuffling objects and structures within the image. In contrast, we seek to learn the semantic distribution of the visual concept *across* the dataset, not just within a single image. For example, we wish to learn that particular photographers tend to shoot specific objects at specific sizes and locations. In this dissertation, we utilize generative models in multiple settings, in order to visualize semantics of our problems (Chapters 4, 5), either by constraining our models on subsets of data or by imposing conditioning on the models to force them to learn specified semantic concepts from our data.

3.0 MODELING PHOTOGRAPHIC STYLE

Summary. *In this chapter, we introduce the novel problem of identifying the photographer behind a photograph.*[1] *We thus seek to model the concept of photographic style. To explore the feasibility of current computer vision techniques to address this problem, we created a new dataset of over 180,000 images taken by 41 well-known photographers. Using this dataset, we examined the effectiveness of a variety of features (low and high-level, including CNN features) at identifying the photographer. We also trained a new deep convolutional neural network for this task. Our results show that high-level features greatly outperform low-level features. We provide qualitative results using these learned models that give insight into our method's ability to distinguish between photographers, and allow us to draw interesting conclusions about what specific photographers shoot. We also demonstrate two applications of our method.*

3.1 INTRODUCTION



(a) (b) (c)

Figure 3: Three sample photographs from our dataset taken by Hine, Lange, and Wolcott, respectively. Our top-performing feature is able to correctly determine the author of all three photographs, despite the very similar content and appearance of the photos.

[1] The work presented in this chapter was published in our CVPR 2016 paper, "Seeing Behind the Camera: Predicting the Authorship of a Photograph" [339].

The ability to accurately extract stylistic and authorship information from artwork computationally enables a wide array of useful applications in the age of massive online image databases. For example, a user who wants to retrieve more work from a given photographer, but does not know his/her name, can speed up the process by querying with a sample photo and using "Search by artist" functionality that first recognizes the artist. Modern search engines can currently only "recognize" the artist of a photograph if there is a page that includes both this particular photograph and the artist's name, but many photographs are not available on annotated websites. Automatic photographer identification can be used to detect unlawful appropriation of others' photographic work, e.g. in online portfolios, and could be applied in resolution of intellectual property disputes. It can also be employed to analyze relations between photographers and discover "schools of thought" among them. The latter can be used in attributing historical photographs with missing author information. Finally, understanding a photographer's style might enable the creation of novel photographs in the spirit of a known author.

While researchers have made progress towards matching the human ability to categorize paintings by style and authorship [305, 20, 15], no attempts have been made to recognize the authorship of *photographs*. This is surprising because the average person is exposed to many more photographs daily than to paintings. We believe one possible reason for this is because the stylistic cues (such as brush stroke) available for identifying a particular painter are greatly reduced in the photographic domain due to the lessened authorial control in that medium (we do not consider photomontaged or edited images in this study). This makes the problem of modeling the visual concept of photographic authorship significantly more challenging than that of identifying the author of a painting. Fig. 3 shows photographs taken by Lewis Hine, Dorothea Lange, and Marion Wolcott, three iconic American photographers. Both Lange and Wolcott worked for the Farm Security Administration (FSA) documenting the hardship of the Great Depression, while Hine worked to address a number of labor rights issues. All three images depict child poverty and there are no obvious differences in style, yet our method is able to correctly predict the author of each.

This chapter makes several important contributions: 1) we propose the problem of modeling the concept of photographic style, which no existing work has explored; 2) due to the

Adams	245	Brumfield	1138	Capa	2389	Bresson	4693	Cunningham	406	Curtis	1069	Delano	14484
Duryea	152	Erwitt	5173	Fenton	262	Gall	656	Genthe	4140	Glinn	4529	Gottscho	4009
Grabill	189	Griffiths	2000	Halsman	1310	Hartmann	2784	Highsmith	28475	Hine	5116	Horydczak	14317
Hurley	126	Jackson	881	Johnston	6962	Kandell	311	Korab	764	Lange	3913	List	2278
McCurry	6705	Meiselas	3051	Mydans	2461	O'Sullivan	573	Parr	20635	Prokudin-Gorsky	2605	Rodger	1204
Rothstein	12517	Seymour	1543	Stock	3416	Sweet	909	Van Vechten	1385	Wolcott	12173		

Table 4: Listing of all photographers and the number of photos by each in our dataset.

lack of a relevant dataset for this problem, we create a large and diverse dataset which tags each image with its photographer (and possibly other metadata); 3) we investigate a large number of pre-existing and novel visual features and their performance in a comparative experiment in addition to human baselines obtained from a small study; 4) we provide numerous qualitative examples and visualizations to illustrate: the features tested, successes and failures of the method, and interesting inferences that can be drawn from the learned models; 5) we apply our method to discover schools of thought between the authors in our dataset; and 6) we show preliminary results on generating novel images that *look like* a given photographer's work.

3.2 APPROACH

3.2.1 Dataset

A significant contribution of our work is our photographer dataset.[2] It consists of 41 well known photographers and contains 181,948 images of varying resolutions. We searched Google for "famous photographers" and used the list while also choosing authors with large, curated collections available online. Table 4 contains a listing of each photographer and their associated number of images in our dataset. The timescale of the photos spans from the early days of photography to the present day. As such, some photos have been developed from film and some are digital. Many of the images were harvested using a web spider with permission from the Library of Congress's photo archives and the National Library of Australia's digital

[2] It can be obtained from `http://www.cs.pitt.edu/~chris/photographer`.

collection's website. The rest were harvested from the Magnum Photography online catalog, or from independent photographers' online collections. Each photo in the dataset is annotated with the ID of the author, the URL from which it was obtained, and possibly other meta-data, including: the title of the photo, a summary of the photo, and the subject of the photo (if known). The title, summary, and subject of the photograph were provided by either the curators of the collection or by the photographer. Unlike other datasets obtained through web image search which may contain some incorrectly labeled images, our dataset has been painstakingly assembled, authenticated, and described by the works' curators. This rigorous process ensures that the dataset and its associated annotations are of the highest quality.

3.2.2 Features tested

Modeling the visual concept of photographic style is a complex problem and relies on multiple factors. Thus, we explore a broad space of features (both low and high-level). The term "low-level" means that each dimension of the feature vector has no inherent "meaning." High-level features have articulatable semantic meaning (i.e. the presence of an object in the image). We also train a deep convolutional neural network from scratch in order to learn custom features specific to this problem domain.

Low-level features.

- **L*a*b* Color Histogram:** To capture color differences among the photographers, we use a 30-dimensional binning of the L*a*b* color space. Color has been shown useful for dating historical photographs [269].
- **GIST:** GIST [261] features have been shown to perform well at scene classification and have been tested by many of the prior studies in style and artist identification [168, 305]. All images are resized to 256 by 256 pixels prior to having their GIST features extracted.
- **SURF:** Speeded-up Robust Features (SURF) [23] is a classic local feature used to find patterns in images and has been used as a baseline for artist and style identification [20, 29, 15]. We use k-means clustering to obtain a vocabulary of 500 visual words and apply a standard bag-of-words approach using normalized histograms.

High-level features.

- **Object Bank:** The Object Bank [200] descriptor captures the location of numerous object detector responses. We believe that the spatial relationships between objects may carry some semantic meaning useful for our task.
- **Deep Convolutional Networks:**
  - **CaffeNet:** This pre-trained CNN [156] is a clone of the winner of the ILSVRC2012 challenge [184]. The network was trained on approximately 1.3M images to classify images into 1000 different object categories.
  - **Hybrid-CNN:** This network has recently achieved state-of-the-art performance on scene recognition benchmarks [415]. It was trained to recognize 1183 scene and object categories on roughly 3.6M images.
  - **PhotographerNET:** We trained a CNN with the same architecture as the previous networks to identify the author of photographs from our dataset. The network was trained for 500,000 iterations on 4 Nvidia K80 GPUs on our training set and validated on a set disjoint from our training and test sets.

  To disambiguate layer names, we prefix them with a C, H, or P depending on whether the feature came from CaffeNet, Hybrid-CNN, or PhotographerNET, respectively. For all networks, we extract features from the Pool5, FC6, FC7 and FC8 layers, and show the result of using those features during SVM training in Table 5. The score in the TOP column for PhotographerNET is produced by classifying each test image as the author who corresponds to the dimension with the maximum response value in PhotographerNET's output (FC8).

Low			**High**													
				CaffeNet				**Hybrid-CNN**				**PhotographerNET**				
Color	GIST	SURF-BOW	Object Bank	Pool5	FC6	FC7	FC8	Pool5	FC6	FC7	FC8	Pool5	FC6	FC7	FC8	TOP
0.31	0.33	0.37	0.59	*0.73*	0.7	0.69	0.6	***0.74***	0.73	0.71	0.61	0.25	0.25	*0.63*	0.47	0.14

Table 5: Our experimental results for photographic authorship prediction. The F-measure of each feature is reported. The best feature overall is in **bold**, and the best one per CNN in *italics*. Note that high-level features greatly outperform low-level ones. Chance performance is 0.024.

3.3 EXPERIMENTAL EVALUATION

To test the effectiveness of the aforementioned features at modeling the concept of photographic style, using our new photographer dataset, we randomly divided our dataset into a training set (90%) and test set (10%). Because a validation set is useful when training a CNN to determine when learning has peaked, we created a validation set by randomly sampling 10% of the images from the training set and excluding them from the training set for our CNN only. The training of our PhotographerNET was terminated when performance started dropping on the validation set.

For every feature in Table 5 (except TOP which assigns the max output in FC8 as the photographer label) we train a one-vs-all multiclass SVM using the framework provided by [86]. All SVMs use linear kernels.

Table 5 presents the results of our experiments. We report the F-measure for each of the features tested. We observe that the deep features significantly outperform all low-level standard vision features, concordant with the findings of [168, 20, 305]. Additionally, we observe that Hybrid-CNN features outperform CaffeNet by a small margin on all features tested. This suggests that while objects are clearly useful for photographer identification given the impressive performance of CaffeNet, the added scene information of Hybrid-CNN provides useful cues beyond those available in the purely object-oriented model. We observe that Pool5 is the best feature within both CaffeNet and Hybrid-CNN. Since Pool5 roughly corresponds to parts of objects [399, 374, 146], we can conclude that seeing the *parts* of objects, not the *full* objects, is most discriminative for identifying photographers. This is intuitive because an artistic photograph contains many objects, so some of them may not be fully visible.

The Object Bank feature achieves nearly the same performance as C-FC8 and H-FC8, the network layers with explicit semantic meaning. All three of these features encapsulate object information, though Object Bank detects significantly fewer classes (177) than Hybrid-CNN (978) or CaffeNet (1000). Despite detecting fewer categories, Object Bank encodes more fine-grained spatial information about where the objects detected were located in the image, compared to H-FC8 and C-FC8. This finer-grained information could be giving it a slight advantage over these CNN object detectors, despite its fewer categories.

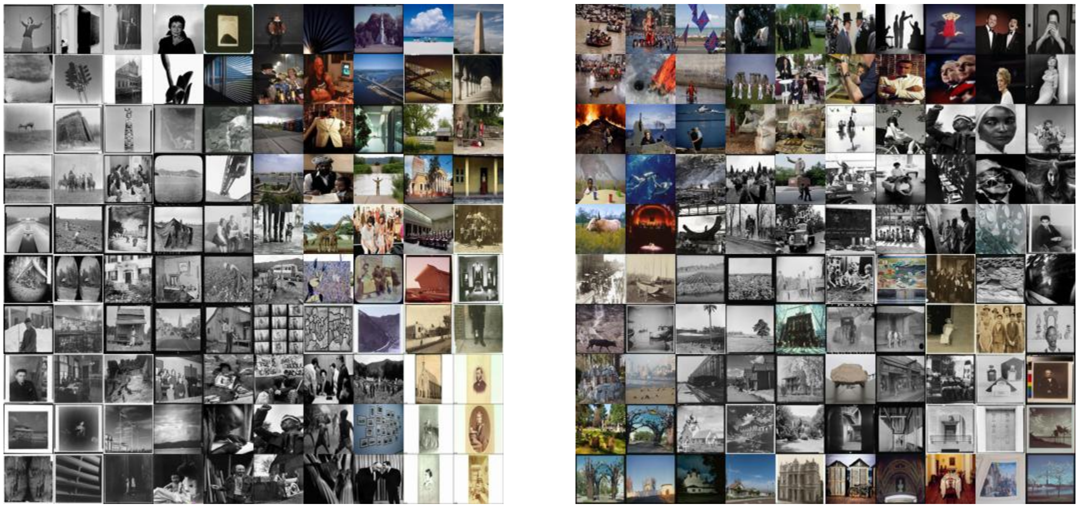
One surprising result from our experiment is that PhotographerNET does not surpass either CaffeNet or Hybrid-CNN, which were trained for object and scene detection on different datasets. We also tried fine-tuning the last three layers of CaffeNet and Hybrid-CNN with our photographer data, but we did not obtain an increase in performance. PhotographerNET's top-performing feature (FC7) outperforms the deepest (FC8) layers in both CaffeNet and Hybrid-CNN, which correspond to object and scene classification, respectively. However, P-FC7 performs worse than their shallower layers, especially H-Pool5. Layers of the network shallower than P-FC7, such as P-FC6 and P-Pool5, demonstrate a sharp decrease in performance (a trend opposite to what we see for CaffeNet and Hybrid-CNN), suggesting that PhotographerNET has learned different and less predictive intermediate feature extractors for these layers than CaffeNet or Hybrid-CNN. Attributing a photograph to the author with highest P-FC8 response (TOP) is even weaker because unlike the P-FC8 method, it does not make use of an SVM.

This result provides a key insight that is of particular relevance to this dissertation's focus of modeling abstract visual concepts. It may be that the task PhotographerNET is trying to learn is *too* high-level and challenging. Because PhotographerNET is learning a task even more high-level than object classification and we observe that the full-object-representation is not very useful for this task, one can conclude that for photographer identification, there is a mismatch between the high-level nature of the task, and the level of representation that is useful.

In Fig. 4, we provide a visualization that might explain the relative performance of our top-performing PhotographerNET feature (P-FC7) and the best feature overall (H-Pool5).

We compute the t-distributed stochastic neighborhood embeddings [348] for P-FC7 and H-Pool5. We use the embeddings to project each feature into 2-D space. We then plot the embedded features by representing them with their corresponding photographs.

We observe that H-Pool5 divides the image space in semantically meaningful ways. For example, we see that photos containing people are grouped mainly at the top right, while buildings and outdoor scenes are at the bottom. We notice H-Pool5's groupings are agnostic to color or border differences. Rather, nearby photos are closer in *semantic* meaning. In con-



(a) P-FC7 t-SNE embeddings.

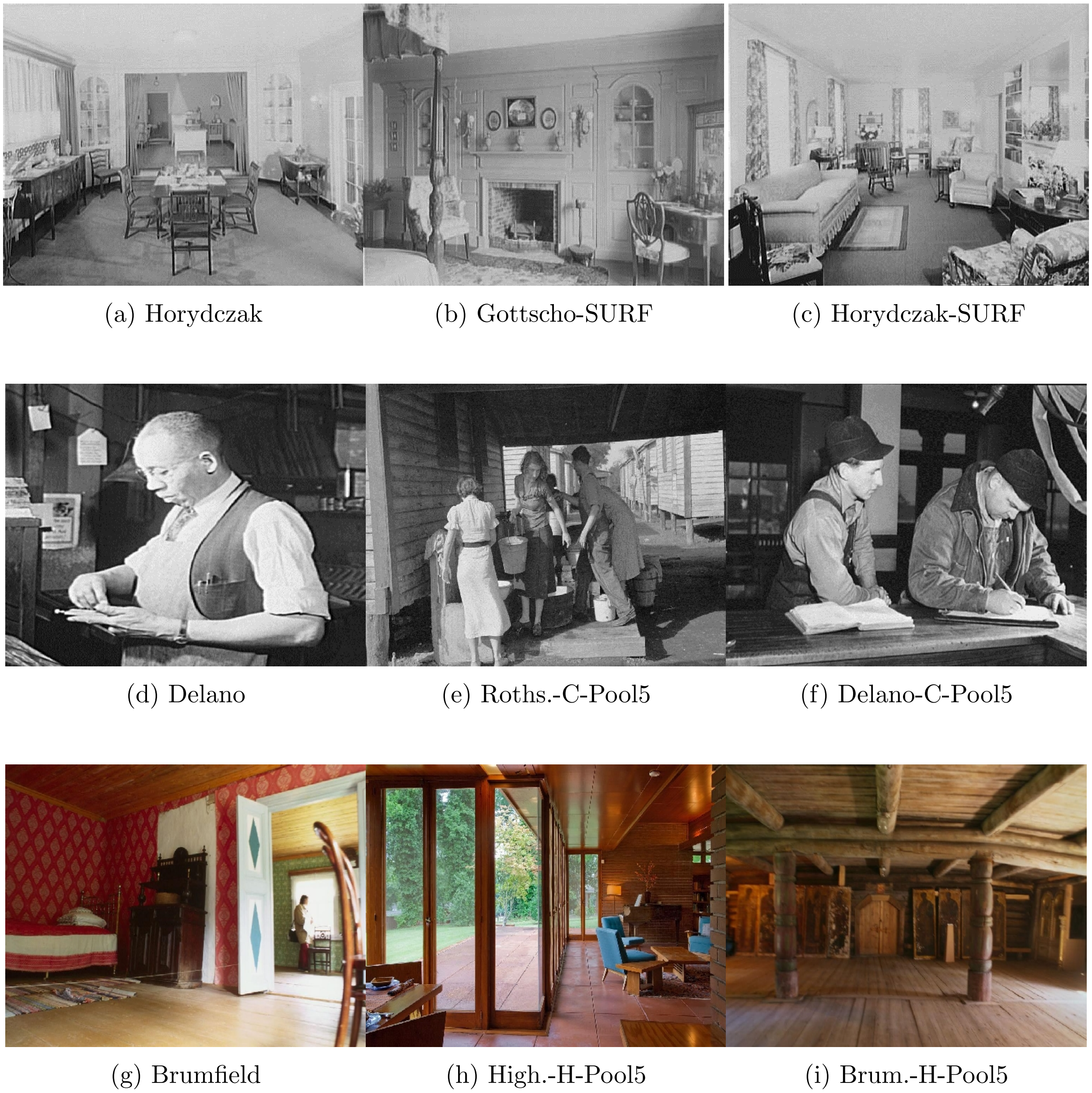
(b) H-Pool5 t-SNE embeddings.

Figure 4: t-SNE embeddings for two deep features. We observe that PhotographerNET relies more heavily on lower-level cues (like color) than higher-level semantic details.

trast, PhotographerNET's P-FC7 divides the image space along the diagonal into black and white vs. color regions. It is hard to identify semantic groups based on the image's content. However, we can see that images that "look alike" by having similar borders or similar colors are closer to each other in the projection. This indicates that PhotographerNET learned to use lower-level features to perform photographer classification, whereas Hybrid-CNN learned higher-level semantic features for object/scene recognition. One possible explanation for this is that because the photos within each class (photographer) of our dataset are so visually diverse, the network is unable to learn semantic features for objects which do not occur frequently enough. In contrast, networks trained explicitly for object recognition only see images of that object in each class, enabling them to more easily learn object representations. Interestingly, these semantic features learned on a different problem outperform the features learned on our photographer identification problem.

To establish a human baseline for the task of photographer identification, we performed two small pilot experiments. We created a website where participants could view 50 randomly chosen images training images for each photographer. The participants were asked to review these and were allowed to take notes. Next, they were asked to classify 30 photos chosen at random from a special balanced test set. Participants were allowed to keep open the page containing the images for each photographer during the test phase of the experiment. In our first experiment, one participant studied and classified images for all 41 photographers and obtained an F1-score of 0.47. In a second study, a different participant performed the same task but was only asked to study and classify the ten photographers with the most data, and obtained an F1-score of 0.67. Our top-performing feature's performance in Table 5 (on all 41 photographers) surpasses both human F1-scores even on the smaller task of ten photographers, demonstrating the difficulty of the photographer identification problem on our challenging dataset.

Finally, to demonstrate the difficulty of the photographer classification problem and to explore the types of errors different features tend to make, we present several examples of misclassifications in Fig. 5. Test images are shown on the left. Using the SVM weights to weigh image descriptors, we find the training image (1) from the incorrectly predicted class (shown in the middle) and (2) from the correct class (shown on the right), with minimum distance to the test image. The first row (Fig. 5a-5c) depicts confusion using SURF features. All three rooms have visually similar decor and furniture, offering some explanation to Fig. 5a's misclassification as a Gottscho image. The second row (Fig. 5d-5f) shows a misclassification by CaffeNet. Even though all three scenes contain people at work, CaffeNet lacks the ability to differentiate between the scene types (indoor vs. outdoor and place of business vs. house). In contrast, Hybrid-CNN was explicitly trained to differentiate these types of scenes. The final row shows the type of misclassification made by our top-performing feature, H-Pool5. Hybrid-CNN has confused the indoor scene in Fig. 5g as a Highsmith. However, we can see that Highsmith took a similar indoor scene containing similar home furnishings (Fig. 5h). These examples illustrate a few of the many confounding factors which make photographer identification challenging.



(a) Horydczak

(b) Gottscho-SURF

(c) Horydczak-SURF

(d) Delano

(e) Roths.-C-Pool5

(f) Delano-C-Pool5

(g) Brumfield

(h) High.-H-Pool5

(i) Brum.-H-Pool5

Figure 5: Confused images. The first column shows the test image, the second shows the closest image in the predicted class, and the third shows the closest image from the correct class. Can you tell which one doesn't belong?

3.3.1 Additional analysis

The experimental results presented in the previous section indicate that classifiers can exploit semantic information in photographs to differentiate between photographers at a much higher fidelity than low-level features. At this point, the question becomes not *if* computer vision techniques can perform photographer classification relatively reliably but *how* they are doing it. What did the classifiers learn? In this section, we present qualitative results which attempt to answer this question and enable us to draw interesting insights about the photographers and their subjects.

Photographers and objects. Our first set of qualitative experiments explores the relationship of each photographer to the objects which they photograph and which differentiate them. Each dimension of the 1000-dimensional C-FC8 vector produced by CaffeNet represents a probability that its associated ImageNet synset is the class portrayed by the image. While C-FC8 does not achieve the highest F-measure, it has a clear semantic mapping to ImageNet synsets and thus can be more easily used to reason about what the classifiers have learned. Because the C-FC8 vector is high-dimensional, we "collapse" the vector for purposes of human consideration. To do this, we map each ImageNet synset to its associated WordNet synset and then move up the WordNet hierarchy until the first of a number of manually chosen synsets[3] are encountered, which becomes the dimension's new label. This reduces C-FC8 to 54 coarse categories by averaging all dimensions with the same coarse label. In Fig. 6, we show the average response values for these 54 coarse object categories for each photographer. Green indicates positive values and red indicates negative values. Darker shades of each color are more extreme.

We apply the same technique to collapse the learned SVM weights. During training, each one-vs-all linear SVM learns a weight for each of the 1000 C-FC8 feature dimensions. Large positive or negative values indicate a feature that is highly predictive. Unlike the previous technique which simply shows the average object distribution per photographer, using the learned weights allows us to see what categories specifically *distinguish* a photographer from

[3] These synsets were manually chosen to form a natural human-like grouping of the 1000 object categories. Because the manually chosen synsets are on multiple levels of the WordNet hierarchy, synsets are assigned to their deepest parent.

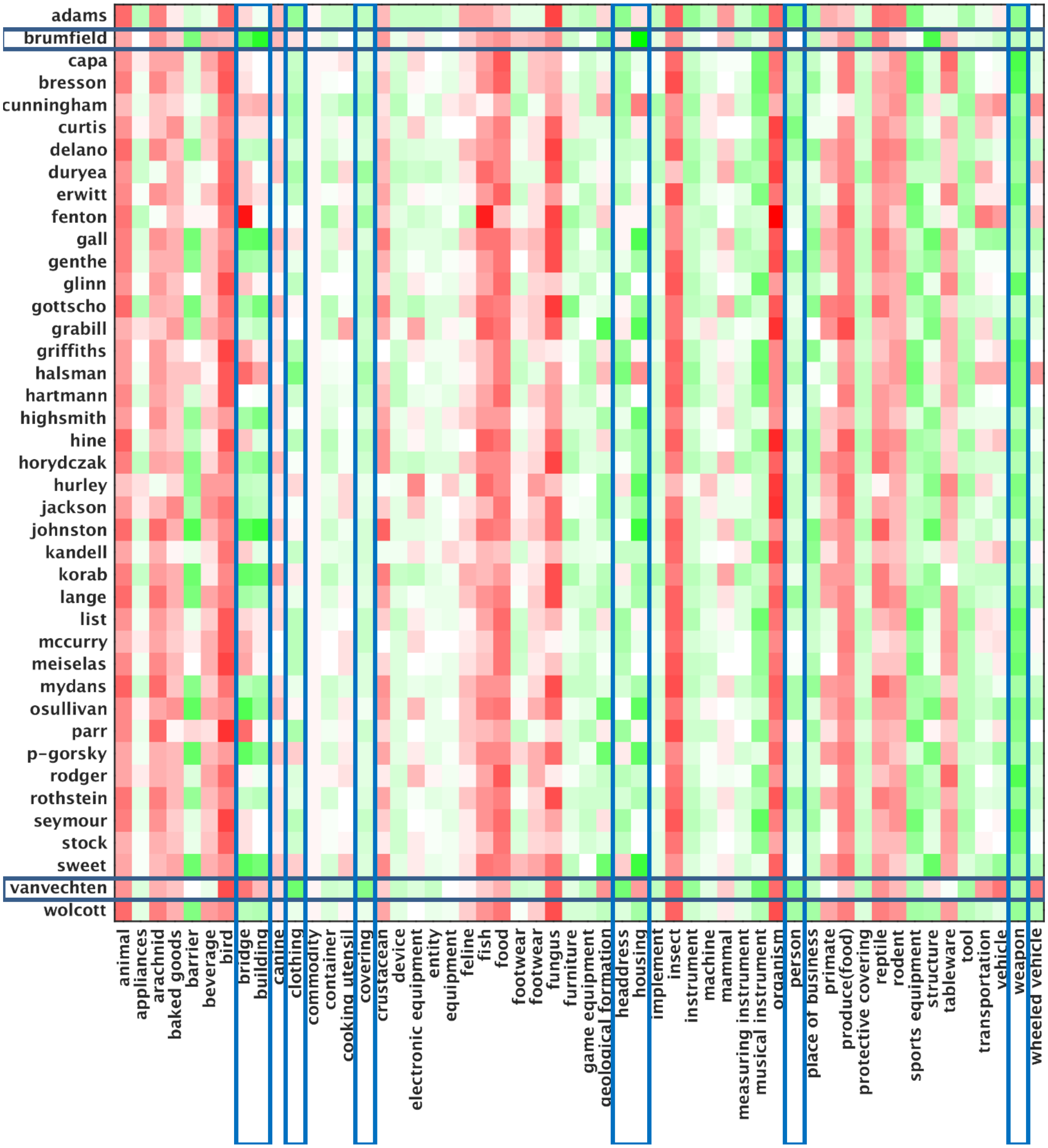


Figure 6: Average C-FC8 collapsed by WordNet.

others. We show the result in Fig. 7.

Finally, while information about the 54 *types* of objects photographed by each author is useful, finer-grained detail is also available. We list the top 10 individual categories with

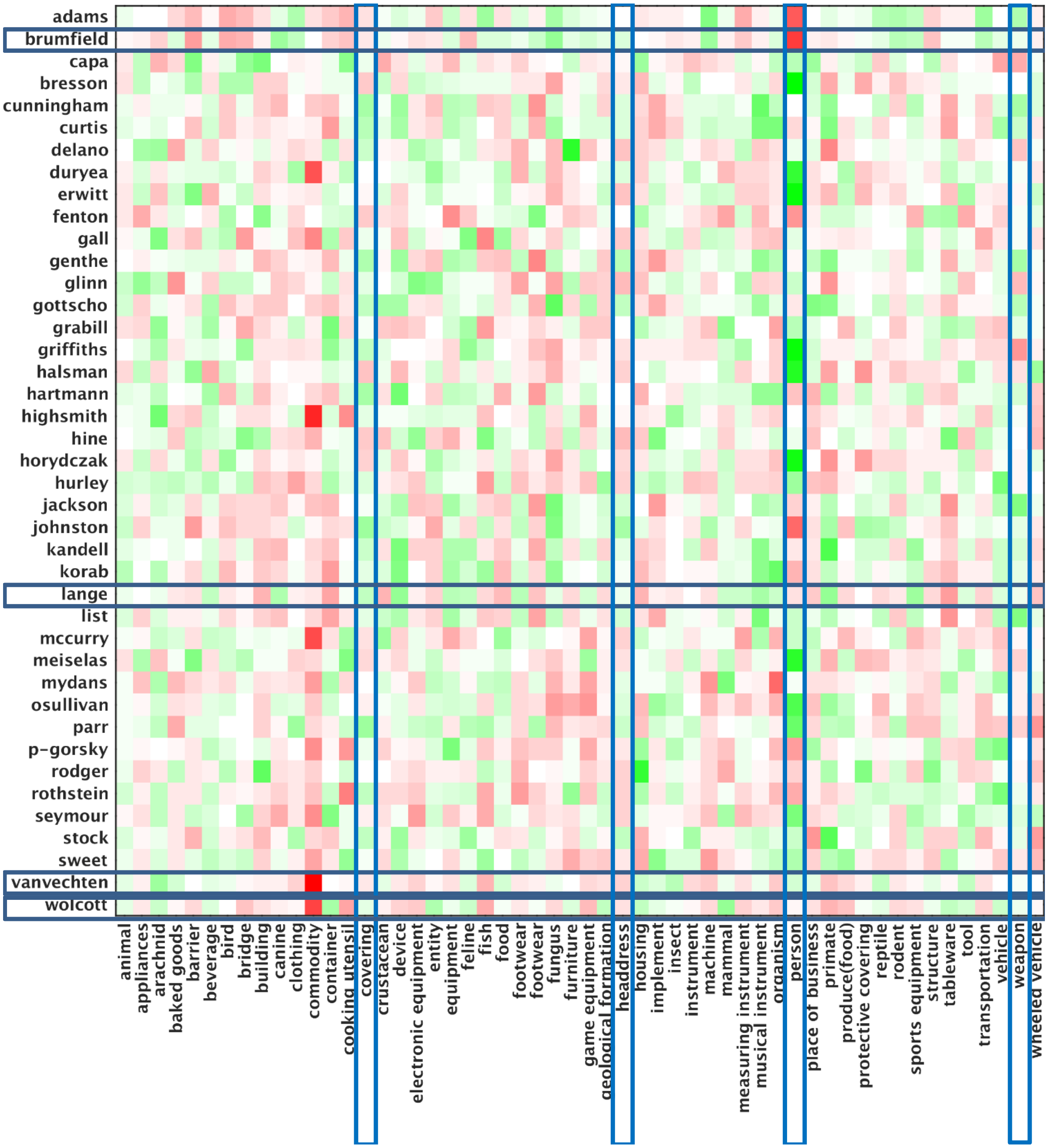


Figure 7: C-FC8 SVM weights collapsed by WordNet.

Adams	hospital room	hospital	office	mil. uniform	bow tie	lab coat	music studio	art studio	barbershop	art gallery
Brumfield	dome	mosque	bell cote	castle	picket fence	stupa	tile roof	vault	pedestal	obelisk
Delano	hospital	construction site	railroad track	slum	stretcher	barbershop	mil. uniform	train station	television	crutch
Hine	mil. uniform	pickelhaube	prison	museum	slum	barbershop	milk can	rifle	accordion	crutch
Kandell	flute	marimba	stretcher	assault rifle	oboe	rifle	panpipe	cornet	mil. uniform	sax
Lange	shed	railroad track	construction site	slum	yard	cemetery	hospital	schoolhouse	train railway	train station
Van Vechten	bow tie	suit	sweatshirt	harmonica	neck brace	mil. uniform	cloak	trench coat	oboe	gasmask

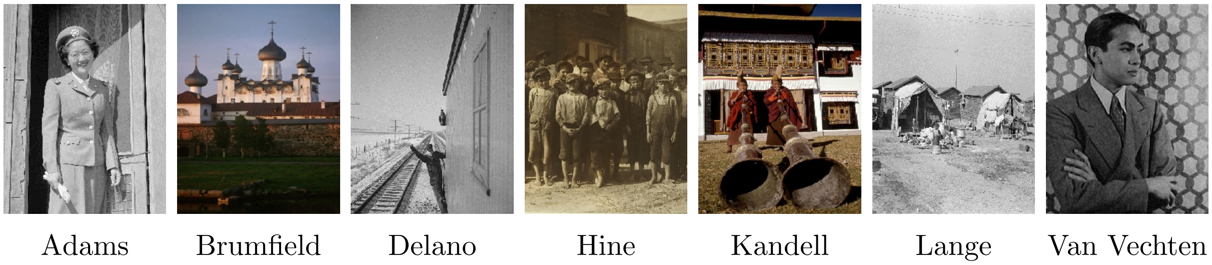


Table 6: Top ten objects and scenes for select photographers, and sample images.

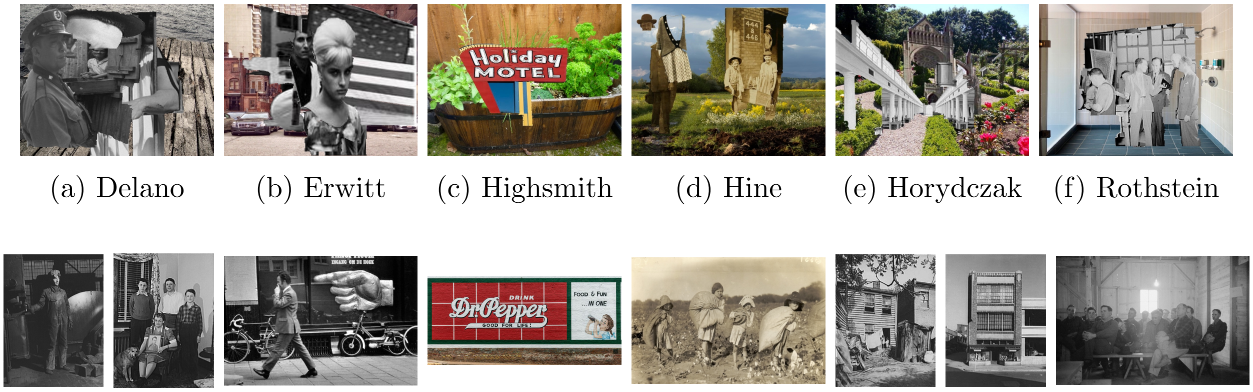
highest H-FC8 weights (which captures both objects and scenes). To do this, we extract and average the H-FC8 vector for all images in the dataset for each photographer. We list the top 10 most represented categories for a select group of photographers in Table 6, and include example photographs by each photographer.

We make the following observations about the photographers' style from Figs. 6 and 7 and Table 6. From Fig. 6, we conclude that Brumfield shoots significantly fewer people than most photographers. Instead, Brumfield shoots many "buildings" and "housing." Peering deeper, Brumfield's top ten categories in Table 6 reveal that he frequently shot architecture (such as mosques and stupas). In fact, Brumfield is an architectural photographer, particularly of Russian architecture. In contrast, Van Vechten has high response values for categories such as "clothing", "covering", "headdress" and "person". Van Vechten's photographs are almost exclusively portraits of people, so we observe a positive SVM weight for "person" in Fig. 7.

Comparing Figs. 6 and 7, we see that there is not a clear correlation between object frequency and the object's SVM weight. For instance, the "weapon" category is frequently represented given Fig. 6, yet is only predictive of a few photographers (Fig. 7). The "person" category in Fig. 7 has high magnitude weights for many photographers, indicating its utility as a class predictor. Note that the set of objects distinctive for a photographer does not fully depend on the photographer's environment. For example, Lange and Wolcott both worked for the FSA, yet there are notable differences between their SVM weights in Fig. 7.

Schools of thought. Taking the idea of photographic style one step further, we wanted to see if meaningful genres or "schools of thought" of photographic style could be inferred from our results. We know that twelve of the photographers in our dataset were members of the Magnum Photos cooperative. We cluster the H-Pool5 features for all 41 photographers into a dendrogram, using agglomerative clustering, and discover that nine of those twelve cluster together tightly, with only one non-Magnum photographer in their cluster. We find that three of the four founders of Magnum form their own even tighter cluster. Further, five photographers in our dataset that were employed by the FSA are grouped in our dendrogram, and two portrait photographers (Van Vechten and Curtis) appear in their own cluster. These results indicate that our techniques are not only useful for describing individual photographers but can also be used to situate photographers in broader "schools of thought."

Generating new photographs. Our experimental results demonstrated that object and scene information is useful for distinguishing between photographers. Based on these results, we wanted to see whether we could take our photographer models yet another step further by generating new photographs imitating photographers' styles. Our goal was to create "pastiches" assembled by cropping objects out of each photographer's data and pasting them in new scenes obtained from Flickr. We first learned a probability distribution over the 205-



(a) Delano (b) Erwitt (c) Highsmith (d) Hine (e) Horydczak (f) Rothstein

Figure 8: Generated images for six photographers (top row) and real photographs by these authors (bottom row). Although results are preliminary, we observe interesting similarities between the synthetic and real work.

scene types detected by Hybrid-CNN for each photographer. We then learned a distribution of objects and their most likely spatial location for each photographer, conditioned on the scene type. To do this, we trained a Fast-RCNN [104] object detector on 25 object categories which frequently occurred across all photographers in our dataset using data we obtained from ImageNet. We then sampled from our joint probability distributions to choose which scene to use and which objects should appear in it and where. We randomly selected a detection (in that photographer's data) for each object probabilistically selected to appear, then cropped out the detection and segmented the cropped region using [205]. We inserted the segment into the pastiche according to that photographer's spatial model for that object.

We show six pastiches generated using this approach in Fig. 8. The top row shows generated images for six photographers, and the bottom shows real images from the corresponding photographer that resemble the generated ones. For example, Delano takes portraits of individuals in uniforms and of "common people," Erwitt photographs people in street scenes without their knowledge or participation, and Rothstein photographs people congregating. Highsmith captures large banner ads and Americana, Hine children working in poor conditions, and Horydczak buildings and architecture. While these are preliminary results, we see similarities between the synthetic and authentic photos.

3.4 DISCUSSION

In this chapter, we explored the problem of modeling photographic style and generating synthetic photos with those styles. Our experiments reveal that high-level features perform significantly better overall than low-level features or humans. While our trained CNN, PhotographerNET, performs reasonably well, early proto-object and scene-detection features perform significantly better. This result demonstrates that the concept of photographic style is more *semantically* exhibited in terms of the subjects portrayed rather than visually manifested through colors or textures The inclusion of scene information provides moderate gains over the purely object-driven approach explored by [168, 305]. Our results confirm **H1** that models focus on low-leval details which preclude learning absent guidance (**H2**).

Specifically, we show that training a CNN on the task of classifying which photographer took a photographer causes the model to learn features which do not capture semantics. However, by guiding our model's training by transferring semantics from a pre-trained method our method was able to perform significantly better. Our primary contribution here is a technique for qualitative analysis by determining which objects respond strongly to each photographer in the feature values and learned classifier weights. Using our techniques, we were able to draw interesting conclusions about the photographers we studied as well as broader "schools of thought" between those photographers. We also presented a probabilistic generative approach that creates new photographs in the spirit of a given author.

The broader implications of our work in this chapter for this dissertation is a demonstration that even when higher-level semantics are unable to be extracted from a dataset using a naïve training strategy, by forcing classifiers to rely on semantics which we have an *a priori* belief are useful, we may still be able to model the phenomena (**H2**). Likewise, even though the visual concept we wished to model was too high-level to train a generative network on, we were still able to generate synthetic data by enforcing a simple probabilistic model which we believed was justified from our analysis. Thus, in many cases where visual phenomena is too challenging to model automatically, by enforcing external structure on the generative process one can still achieve semantically sensible image synthesis.

4.0 MODELING VISUAL PERSUASION

Summary. *In this chapter, we examine the visual variability of objects across different ad categories, i.e. what causes an advertisement to be visually persuasive.*[1] *This work applies and extends the techniques we develop for modeling implicit visual concepts in Chapter 3. We focus on modeling and generating* faces *which appear to come from different types of ads. For example, if faces in beauty ads tend to be women wearing lipstick, a generative model should portray this distinct visual appearance. Training generative models which capture such category-specific differences is challenging because of the highly diverse appearance of faces in ads and the relatively limited amount of available training data. To address these problems, we propose a conditional variational autoencoder which makes use of predicted semantic attributes and facial expressions as a supervisory signal when training. We show how our model can be used to produce visually distinct faces which appear to be from a fixed ad topic category. Our human studies and quantitative and qualitative experiments confirm that our method greatly outperforms a variety of baselines, including two variations of a state-of-the-art generative adversarial network, for transforming faces to be more ad-category appropriate. Finally, we show preliminary generation results for other types of objects, conditioned on an ad topic.*

4.1 INTRODUCTION

The task of modeling persuasive visual media is of particular relevance to our proposed work of modeling bias in media photos. Advertisements are persuasive tools that affect people's habits and decisions. They often advertise products and establishments, such as cosmetics and beauty, clothing, alcohol, automobiles, or restaurants. However, they can also be public service announcements that aim to educate the public about important social

[1] The work presented in this chapter was published in our BMVC 2018 paper, "Persuasive Faces: Generating Faces in Advertisements" [340].

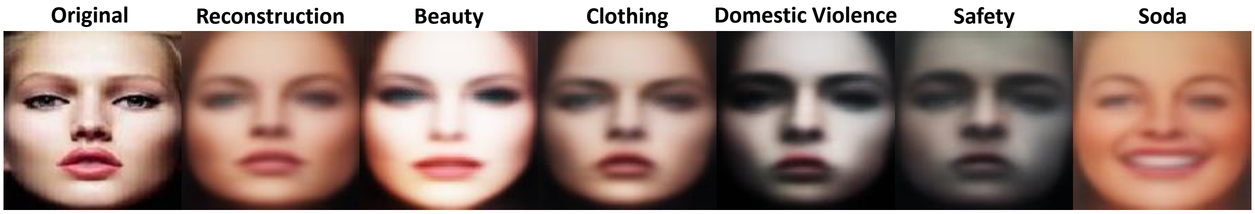


Figure 9: We transform faces so they appear more persuasive and appropriate for particular ad categories. We show an original face on the left, followed by our method's reconstruction without any transformation. We then show the face transformed according to five types of ads. Notice how the beauty face contains heavy make-up, the domestic violence face is sad and possibly bruised, the safety face is somewhat masculine, and the soda face is happy.

issues, such as domestic violence or environmental protection. Many topics advertised by ads contain distinctive objects, e.g. the most common object in car ads might be cars, bottles for alcohol ads, and faces for cosmetic ads. There is more to ads than what objects they contain, however. It is *how* objects are portrayed that makes an ad persuasive. For example, faces frequently appear in both beauty and domestic violence ads but their portrayal is vastly different. Thus, our work on modeling visual persuasion in ads is similar to our work on modeling photographic style (Chapter 3): photographers choose to portray the same scenes in vastly different ways for interpretable reasons.

What is it that makes a face become a beauty ad or a domestic violence prevention ad? This is what we set out to discover in this chapter. As the visual phenomenon we seek to model is high-level, diverse, and only limited data is available, we begin by utilizing the same analytical framework we used for modeling photographic style and that we presented as a general technique for modeling other such problems. Thus, we first analyze the distribution of objects in common ad topics (beauty, soda, domestic violence, safety, etc.) Based on the object distributions, we select to model the appearance of faces, since faces are the most frequent object across all ad categories and have the most distinctive appearance per category. We then learn a generative model capable of transforming faces into each ad topic. We note, however, that faces have significantly more regularity in structure and appearance that entire photographs. Thus, it is feasible to train a generative model on our ad faces, while it was not on our photographer dataset in Chapter 3. Because ads are rarer than general images, we must work with a sparser dataset than modern generative approaches usually assume. Thus, we propose a method for *transferring knowledge* from faces in other datasets, in order to mimic the variability of faces in the ads domain. We validate our approach qualitatively, by morphing the same face according to different ad categories, and quantitatively, using human judgments and classifier accuracy.

Our method works as follows. We first train facial expression and facial attribute classifiers using existing datasets. We detect faces in ads and predict their attributes and expressions. Next, we train a conditional variational autoencoder (CVAE) on our dataset of ad faces. The model learns to reconstruct an ad face from a vector comprised of a learned latent representation, facial attributes, and facial expressions. At test time, we embed all ad faces into vector space using our encoder and then compute how faces differ in that space across ad topics. Using these per-topic learned differences, we transform embeddings of other ad faces into each ad topic. Finally, we use our decoder on the transformed embeddings to generate distinct faces across ad topics. We show examples of our transformations in Fig. 9. Note that prior work has modeled the conceptual rhetoric that ads use to convey a message [149, 390], but no work models the visual variance in the portrayal of the same object across different ad categories, nor attempts to generate such objects.

This method proposed in this chapter makes the following three contributions:

- We propose the problem of studying what makes an object visually persuasive and generating objects which convey appropriate visual rhetoric for a given ad topic.
- We analyze object frequency and appearance in ads, and discover objects with class-dependent appearances, which we then generate with promising quality.
- We develop a novel generative approach for modifying the appearance of faces into different ad categories, by elevating visual variance to a semantic level without the need for new semantic labels (**H2**). Rather than directly modeling how faces in different ad categories differ on the pixel level, we model how they differ in terms of predicted attributes and facial expressions, then use these distinctions to create faces appropriate for a given ad category. Our method outperforms relevant baselines at this task.

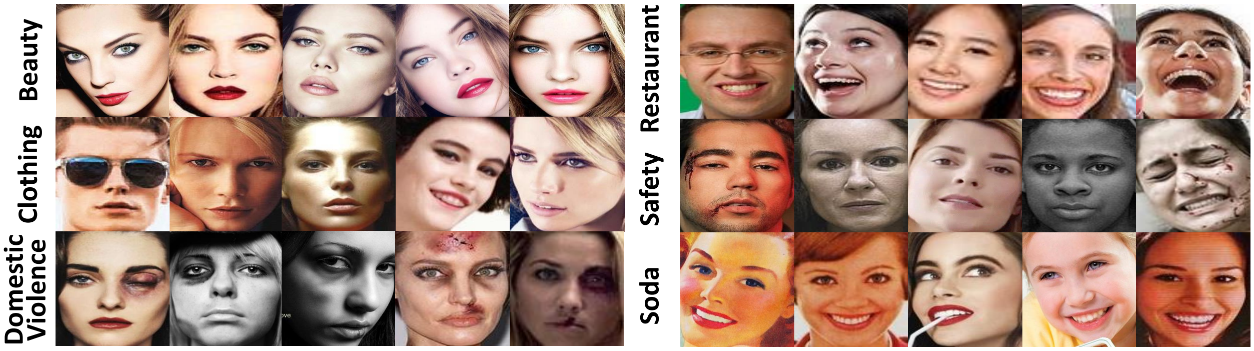


Figure 10: We show examples of real faces from different categories of ads. We notice significant differences, many of which can be captured through facial attributes and expressions.

4.2 APPROACH

In this section, we begin by describing how we extract faces from ads. We then describe how we predict attributes and facial expressions on the detected faces. Next, we present our autoencoder architecture and then describe how we use it to transform faces across ad categories.

4.2.1 Dataset

We focus on the Ads Dataset of [149]. It contains ads belonging to 38 topic categories: beauty, soda, restaurants, etc. (called product ads) and domestic violence, safety, etc. (called public service announcements, or PSAs). We chose to study the ten most frequent product topics in the dataset, as well as all PSA topics, resulting in a set of 17 ad topics.

4.2.2 Face detection on ads

Our first step is to extract faces from ads. The remaining steps of our model work on this dataset of ad faces, rather than operating on whole ads images. This allows our model to concentrate on modeling and modifying facial appearance, without having to reconstruct the entire ad. We train Faster-RCNN [297] on the Wider Face dataset [388]. We remove face detections whose confidence is less than 0.85 or whose width or height is less than 60 pixels. We show examples of detected faces in different ad categories in Fig. 10. In total, we detected 20,532 faces. We observe, for example, that beauty ads often have brighter skin tones and feature women wearing makeup. Domestic violence faces are often darker and not smiling. Many soda faces appear vintage and smiling. Clothing ads are similar to beauty, but don't feature as bright of skin or makeup. Finally, safety ads feature more men and are not as dark as domestic violence ads. Importantly, many of the differences we observe are captured by facial attributes and expressions datasets.

4.2.3 Predicting facial attributes and expressions

We want our method to model the most relevant characteristics of faces in each ad topic category. As we observed in Fig. 10, the differences between faces in different ad categories can naturally be described in terms of facial attributes and expressions. Because our dataset is small and diverse, our model may not have enough signal to reliably learn to model facial attributes and expressions without explicitly being directed to do so. In other words, it may devote its modeling power to matching the precise vintage or cartoon appearance of ad faces (i.e. low-level details) without learning a high-level model of recognizable semantic differences. Thus, rather than formulating our task as modeling the unconstrained distribution of pixels from the faces in each ad group, we manually inject high-level knowledge to facilitate manipulation of specific semantic attributes and expressions across ad topics.

We use the CelebA dataset [217] of 40 facial attributes and the AffectNet dataset [247] of eight facial expressions plus valence and arousal scores. We train Inception-v3 [336] on each dataset. We train each classifier using a cross-entropy loss for classification. For the network trained on expressions, we add an additional classifier for the regression task of predicting the valence and arousal of the facial expression and also use a mean-squared error loss.

Formally, let $\mathbf{I}_t$ represent the dataset of ad faces extracted from each ad topic t (e.g. beauty faces, domestic violence faces, etc.). We use our trained attributes and expressions classifiers to predict these properties on our entire ad faces dataset. This results in an automatically labeled ads face dataset $\mathbf{I}_t = \{\mathbf{x}_t^i, \mathbf{y}_t^i\}_{i=1}^{N_t}$, where $\mathbf{x}_t^i$ represents face i from ad topic t, $\mathbf{y}_t^i$ represents the image's associated 50-dimensional vector (composed of 40 facial attributes and eight facial expressions with their accompanying valence and arousal scores), and N_t represents the total number of faces per topic. We binarize our facial attribute predictions and represent our facial expressions in a one-hot fashion. The valence and arousal scores are real numbers from $[-1, 1]$.

4.2.4 Conditional variational autoencoder

Given an image $\mathbf{x}_t^i$ and conditional vector $\widehat{\mathbf{y}_t^i}$, which may differ from the image's ground truth signature, we seek a model θ parameterizing the following transformation function:

$$f_\theta\left(\mathbf{x}_t^i, \widehat{\mathbf{y}_t^i}\right) = \widehat{\mathbf{x}_t^i} \tag{1}$$

where $\widehat{\mathbf{x}_t^i}$ is a face retaining the overall appearance of $\mathbf{x}_t^i$, but now bearing the attributes and expressions encoded in $\widehat{\mathbf{y}_t^i}$. If $\mathbf{y}_t^i = \widehat{\mathbf{y}_t^i}$, we seek an unmodified reconstruction of $\mathbf{x}_t^i$. To modify the original appearance, we would like the reconstructed face to bear the provided set of attributes. If we denote our attribute and expression classifiers from Sec. 4.2.3 jointly as C, we wish to enforce the following constraint:

$$C\left(f_\theta\left(\mathbf{x}_t^i, \widehat{\mathbf{y}_t^i}\right)\right) = \widehat{\mathbf{y}_t^i} \tag{2}$$

Thus, any modifications done by our model should result in our classifiers producing the same conditional vector that was provided to the transformation model.

We also seek the capability of transforming ad topic-wise facial appearance *beyond* what is captured by our conditional vector. For example, if one topic features a predominant ethnicity, we would like our model to be capable of transforming a face into that ethnicity, even though it is not presented in our conditional vector. We thus seek a model capable of learning latent facial appearance information from our dataset. Autoencoders, which project an image into a low-dimensional space and then learn to reconstruct it from the sparse representation, are a natural choice. However, because we wish to interpolate faces across ad topics, enforcing that the learned space is smooth is important. We thus propose a custom conditional variational autoencoder, which enforces a Gaussian prior on the latent space [327].

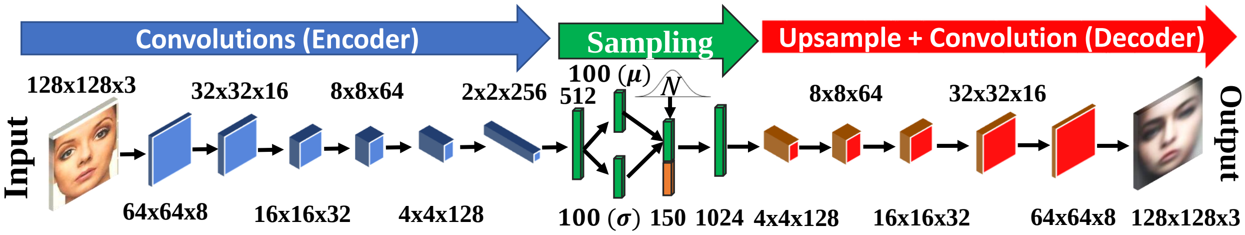


Figure 11: We show our model transforming a beauty ad face into a domestic violence face. The conditional vector (orange bar) is appended to the sampled latent vector (see Eqs. 3,4).

We present our model's architecture in Fig. 11. It contains two distinct components, an encoder and decoder, which are trained end-to-end to reconstruct ad faces.

Encoder. Our encoder g_ϕ encodes any image $\mathbf{x}$ into the latent space $\mathbf{z}$ as follows:

$$\mathbf{z} = g_\phi(\mathbf{x}, \epsilon), \epsilon \sim \mathcal{N} \tag{3}$$

where ϵ represents a vector sampled at random from $\mathcal{N}$, a standard normal distribution. Specifically, g_ϕ encodes an image by predicting μ and σ for each dimension of the latent space. The latent embedding for an image is produced by combining ϵ with the predicted latent distribution parameters as follows: $\mathbf{z} = \mu + e^{\frac{\sigma}{2}}\epsilon$. This mechanism of predicting the latent variable (coupled with the smoothness constraint discussed later) represents an image as a sample drawn from a Gaussian image space. Thus, the same image's latent embedding will differ each forward pass of the encoder due to random sampling of ϵ. This exposes our decoder network to a degree of local variation because the decoder learns that a larger space of embeddings map to the same face. This encourages smoothness in the latent space, which is important for the interpolation on latent vectors performed later.

Decoder. We concatenate each image's latent vector with its associated conditional vector (attributes and expressions) to produce the final representation given to our decoder p_ψ:

$$\mathbf{q}_t^i = \left[\widehat{\mathbf{y}}_t^i, \mathbf{z}_t^i\right] = \left[\mathbf{y}_t^i, g_\phi\left(\mathbf{x}_t^i, \epsilon\right)\right] \tag{4}$$

During training, $\widehat{\mathbf{y}_t^i} = \mathbf{y}_t^i$. Our decoder network learns to reconstruct the original image from the embedding:

$$\widehat{\mathbf{x}_t^i} = p_\psi\left(\mathbf{q}_t^i\right) = p_\psi\left(\left[\mathbf{y}_t^i, g_\phi\left(\mathbf{x}_t^i, \epsilon\right)\right]\right) \tag{5}$$

Learning. We train our model end-to-end to reconstruct the image provided to the encoder. However, because L2 reconstruction losses have been shown to produce blurry predictions [238], we instead use a perceptual loss similar to [137]. Rather than compute the distance between the reconstruction and original image in pixel space, we compute the distance in *feature space* of a pretrained VGG classification network following [407]. In our experiments, using a perceptual loss substantially improved the quality of reconstructions. Formally, let $\Phi\left(\mathbf{x}_t^i\right)$ and $\Phi\left(\widehat{\mathbf{x}_t^i}\right)$ represent the activations of layer `relu2_2` of a pretrained VGG-19 [321] network on the original and reconstructed images. The reconstruction loss $\mathcal{L}_r$ is given by:

$$\mathcal{L}_r = \left\|\Phi\left(\mathbf{x}_t^i\right) - \Phi\left(\widehat{\mathbf{x}_t^i}\right)\right\|_2^2 \tag{6}$$

We provide our decoder with the predicted facial attributes and expressions $\mathbf{y}_t^i$ so that we know these aspects of faces will be represented and thus modifiable across ad categories. However, the decoder might ignore less conspicuous attributes, so we force it to use the conditional information. The model should produce samples that cause our classification networks to output the same vectors provided to the decoder. If C_a and C_e represent attribute and facial expression classifiers, our conditional classification loss $\mathcal{L}_c$ is given by:

$$\mathcal{L}_c = l_{bce}\left(C_a\left(\mathbf{x}_t^i\right), C_a\left(\widehat{\mathbf{x}_t^i}\right)\right) + l_{nll}\left(C_{e_{exp}}\left(\mathbf{x}_t^i\right), C_{e_{exp}}\left(\widehat{\mathbf{x}_t^i}\right)\right) + l_2\left(C_{e_{va}}\left(\mathbf{x}_t^i\right), C_{e_{va}}\left(\widehat{\mathbf{x}_t^i}\right)\right) \tag{7}$$

where $C_{e_{exp}}$ and $C_{e_{va}}$ represent the facial expression and valence and arousal predictions from C_e respectively, l_{bce} represents the binary cross entropy loss, l_{nll} represents the negative log-likelihood loss after softmax is applied to the inputs (for multiclass classification), and l_2 represents the l_2 loss (for regression). In practice, we found our classification constraint improved reconstructions and made them more responsive to changes in the conditional vector.

To encourage smoothness in the latent space, we use a standard KL divergence term which measures the relative entropy between a spherical Gaussian distribution and the latent distribution [327]. The KL term $\mathcal{L}_{KL}$ can be analytically integrated [176] into a closed form equation as follows:

$$\mathcal{L}_{KL} = \frac{1}{2}\sum e^{\sigma} + \mu^2 - 1 - \sigma \tag{8}$$

We found the KL constraint critical to producing smooth faces. Our final loss is:

$$\mathcal{L} = \alpha\mathcal{L}_r + \beta\mathcal{L}_c + \gamma\mathcal{L}_{KL} \tag{9}$$

where α, β, and γ are hyperparameters weighting the contribution of each loss component.

4.2.5 Cross-category facial transformation

We described how to reconstruct a face, using an encoder, decoder, and fixed attributes and expressions. We now define what we input to our decoder, to translate a face to an ad class.

Notice that our model never accesses the ad topic category each face comes from. This is because the faces within topic categories are too varied for the model to make use of topic information. However, in order to transform faces so they appear to come from different topics, we first must learn how faces differ in each topic. We compute a vector for each ad topic, which, when added to an image's embedding, makes the reconstruction appear more appropriate for that topic. Specifically, we compute the *topic transformation vector* $\mathbf{v}_t$ for each topic t as follows, where the horizontal bar indicates computing the mean per dimension:

$$\mathbf{v}_t = \sum_{i}^{N_t} \overline{\mathbf{q}_t^i} - \sum_{t' \neq t} \sum_{i}^{N_{t'}} \overline{\mathbf{q}_{t'}^i} \tag{10}$$

In order to make the transformations more visible, we increase the magnitude of the vector by multiplying the conditional portion of $\mathbf{v}_t$ by 10 and the latent portion by 2.5. We found this visibly improved the distinctiveness across topic categories. To translate a face $\mathbf{x}$ into ad category t', we modify the embedding of $\mathbf{x}$ using $\mathbf{v}_{t'}$ and then reconstruct it as follows:

$$\widehat{\mathbf{x}_{t \to t'}^i} = p_\psi\left(\mathbf{q}_t^i + \mathbf{v}_{t'}\right) \tag{11}$$

4.2.6 Implementation details

We train our encoder and decoder end-to-end, but we do not train the VGG-19 network. We train the two classification Inception networks offline, before training our autoencoder. We train using the Adam optimizer [175] with learning rate 5.0e-4. We use minibatch size of 32 and train for 200 epochs. To ensure robustness to the highly varied ads faces dataset, we perform aggressive data augmentation. We randomly horizontally flip the training data and also randomly zoom into or out of the images. We then crop the zoomed images to 128x128. This allows our models to be less sensitive to facial alignment. We empirically found using 100 dimensions for $\mathbf{z}$ to work well. We set $\alpha = 1$ and β and γ to 0.0001; larger values caused poor reconstructions. We use Xavier initialization [107] and leaky ReLU activation [126] for inner layers with negative slope 0.01. We find using batch normalization [150] with eps 1e-4 helps stabilize training. We implement all components of our model in PyTorch [273].

4.3 EXPERIMENTAL EVALUATION

We conduct our experiments on the image advertisement dataset of [149]. We initially sought to study general object appearance across ad topics, but our analysis below revealed faces were by far the most distinct object per topic. We thus focus primarily on modeling faces.

4.3.1 Objects in ads

We ran a 50 layer residual RetinaNet [210] trained on the COCO dataset [211] on all ads in the 17 ad topics defined in Sec. 4.2.1. We first studied the *distributions* of objects across ad topics. We found many object-topic correlations, e.g. cars are most frequent in car ads, bottles occur frequently in alcohol and soda ads, animals are often found in animal rights and environment ads, etc. Overall, we found that people tended to occur 13 times more frequently than the second most common object (car). We next studied how objects' *appearance* differed across ad topic categories. We extracted SIFT [221] features for each object and computed BoW histograms with $k = 100$. We then analyzed the "visual distinctiveness" of objects, by measuring how each object's appearance changed within and across ad topics. We found that cars are highly visually distinct in car ads. This makes sense because cars in car ads are the *focus* of the ad, not just a background object. We also found dogs were distinct in animal rights ads, cell phones in electronics ads, cake and bowl in chocolate ads, and bottle in soda ads. Faces were the single object category which occurred frequently enough across topics to train a model on, thus we primarily focus on modeling faces in this work.

4.3.2 Qualitative results

We compare our method against two baselines inspired by attribute autoencoders [137, 189, 386], one of which has access to attributes and one which does not, as well as two variations of a state-of-the-art adversarial network for transforming images and attributes [53]:

- **Conditional+Latent (Ours)** - Our full model, described in Sec. 4.2.
- **Conditional** - Our model trained with latent and conditional information (attributes, expressions, and valence/arousal), however only the 50 conditional dimensions are changed when translating a face across topics, while the latent dimensions stay fixed.
- **Latent** - Our model *without* the conditioning on attributes and expressions.
- **StarGAN [53] (Conditional)** - We train StarGAN to modify faces to a given 50-dimensional conditional vector (facial attributes, expressions, valence/arousal).
- **StarGAN [53] (Topics)** - We train StarGAN to modify faces into a given topic. At training time, we train the model on the ground truth ad topic categories the faces are from. The model thus explicitly learns how facial appearance changes across topics.

In Fig. 12, we observe that our method **Conditional+Latent** produces the most noticeable and dramatic changes in visual appearance. We observe changes in gender, skin tone, facial expression, and facial shape. Alcohol ads feature smiling men, beauty ads tend to have light skin with lipstick, and clothing ads are similar, but with less skin brightness and less smiling. Faces in domestic violence ads are often frowning and darker, while those in safety

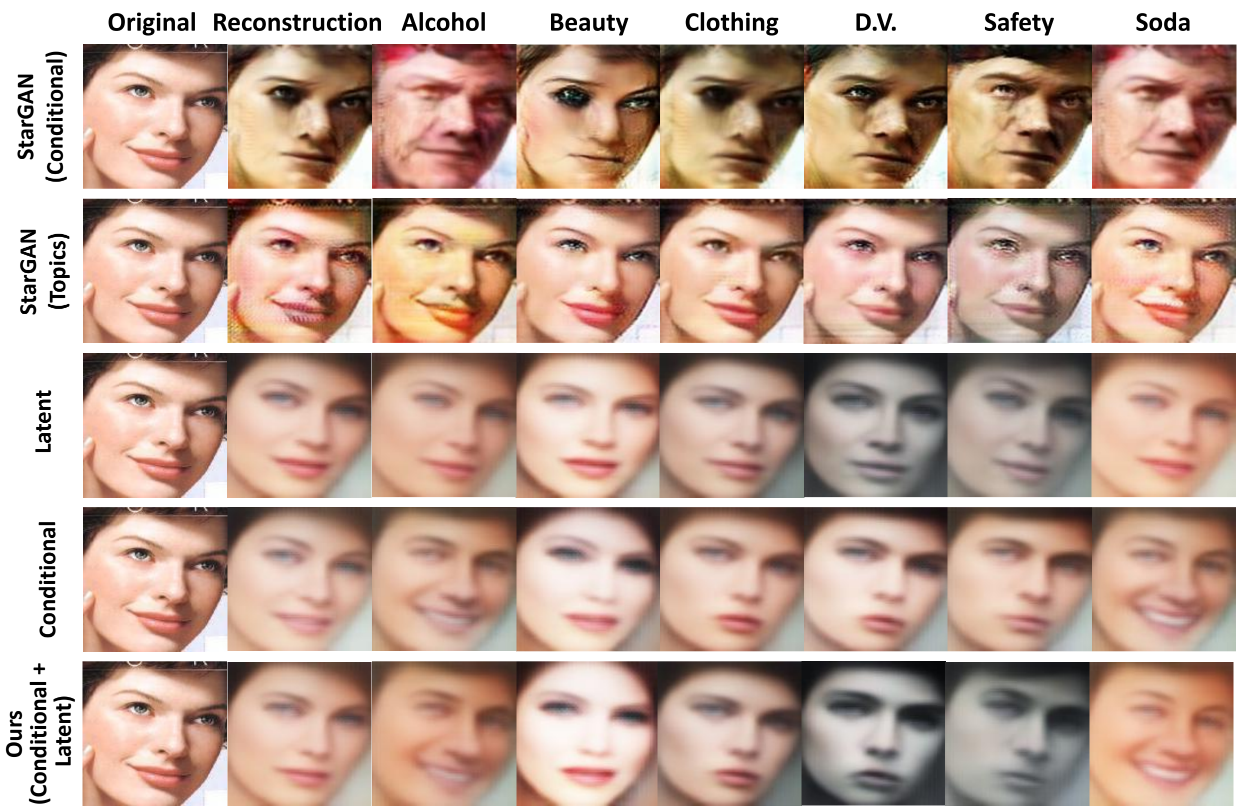


Figure 12: We show the result of transforming the same face using five different methods. Our method (bottom row) most faithfully transfers the topic-specific facial appearance as judged by our human study.

ads tend to appear more masculine. Finally, soda ads have a vintage appearance with a large smile. For **Conditional**, we find that many aspects of the face change appropriately. However, the model is unable to transform other features not captured by the conditional vector: for example, making the face appear darker for domestic violence ads. For **Latent**, we find that while facial appearance overall changes, facial expressions and many facial attributes remain fixed, leaving a smiling face in inappropriate categories such as domestic violence.

We observe that both versions of StarGAN maintain the original image's appearance, but do not change the image much per topic. We notice that **StarGAN (Conditional)** tends to produce smoother skin and highlighted eyes for "beauty," but its other categories are harder to discern. **StarGAN (Topics)** adds low-level details into the generated images in order to achieve a lower topic prediction loss rather than changing the facial appearance of the image.

4.3.3 Quantitative evaluation of generated faces

In addition to our qualitative results, we perform two quantitative experiments to assess how well our method transforms faces into each ad topic. For our first experiment, we perform a human study to assess how well humans perceive each method to do in terms of data transformation. Eight non-author participants participated in our study. We first show them examples of real faces from five ad categories: beauty, clothing, domestic violence, safety, and soda. To ensure our judges pay attention to the visual distinctions, we ask them to classify 10 rows of real faces into the correct ad topic. We then show participants the same image translated by five randomly sorted methods into the five topics, and ask them to select the method which best portrays the distinct visual appearance of faces across the five topics.

For our second experiment, we transform the same faces into all 17 ad categories. Next, we finetune a pretrained AlexNet [184] on the transformed faces to predict which topic the face is supposed to portray. Finally, we evaluate our model on *real* faces. Thus, methods which reliably transform faces in ways which capture the distinct traits of each topic will achieve higher classification accuracy. This metric assesses how well discriminative features are translated into generated ads but does not assess the visual quality of the results or the task we are ultimately interested in, namely producing visually distinct faces across topics.

We present quantitative results in Table 7. **Ours** performs best at the objective we set out to accomplish, and does competitively on the objective but less informative classifier accuracy task. In our human study, human judges found that our method best generates topic-specific faces nearly ***4 times as often*** as the next best method, **Conditional**. Interestingly, humans rarely prefer the **Latent** model, demonstrating the importance of including attributes and facial expressions. For the classification task, the classifier trained on **StarGAN (Topics)**'s data performs best, followed by our method. This makes sense because

Method	Human Judgments: Best At Topic Transformation	Classifier Topic Prediction Accuracy
StarGAN (Conditional)	0.100	0.069
StarGAN (Topics)	0.113	**0.100**
Latent	0.038	0.086
Conditional	0.144	0.080
Conditional + Latent (Ours)	**0.606**	0.092

Table 7: We present two quantitative results. The first shows in what fraction of examples humans chose each method as the best, for generating visually distinct and appropriate faces in each topic. The rightmost column shows the accuracy of a classifier when trained on each row's synthetic training data and tested on real images from 17 categories.

StarGAN (Topics) sees images labeled with topics at train time and learns what features are useful for topic classification. However, we see that the method only changes low-level details (e.g. color) without changing any semantics. Our method changes face semantics, never sees topic information at train time, yet performs on par with **StarGAN (Topics)** on this task, confirming our method does transfer topic-specific appearance. We observe that the accuracy for all models is similar and fairly low. This is most likely because many faces are impossible to classify since they are generic, non-persuasive background faces.

4.3.4 Generating other objects

We wanted to see whether we could generate other objects besides faces as they appear in different ad categories. We conditioned BEGAN [26] on ad topics and trained on bottles from alcohol, beauty, and soda ads. We used an image size of 64x64 due to the limited amount of training data per class. We observe that the model does learn meaningful topic-wise differences in object appearance. For example, alcohol bottles look like liquor bottles with a long stem, beauty bottles are wider with a short stem (perfume), soda bottles have a

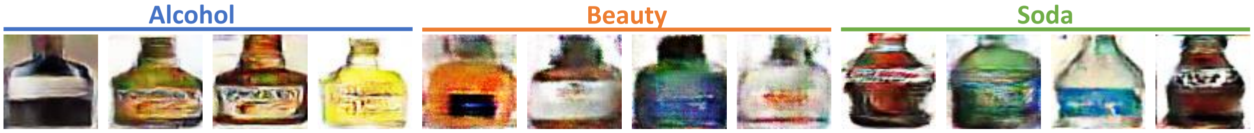


Figure 13: Alcohol, beauty, and soda bottles generated using our implementation of a conditional BEGAN [26] trained on bottles. We observe interesting differences across ad topics.

soda bottle shape and label. These results show that intra-topic object appearance can be modeled, but future work is needed to address problems such as mode-collapse.

4.4 DISCUSSION

In this chapter, we studied modeling cross-category object appearance in ads and how ads use these objects for persuasion. Based on our object analysis, we focused on faces and explored how faces could be generated across different types of ads. We proposed a conditional variational autoencoder for this task, which we augment by providing high-level facial attributes and expressions; experiments showed this auxiliary supervision was critical to achieving good results. Our experiments confirm that our method greatly outperforms a variety of baselines. We also show early results on how topic-specific objects beyond faces may be generated. Our results confirm our hypothesis that both object and facial appearance substantially differs in persuasive media relative to standard vision datasets (**H1**), by reflecting the notions and desires of target audiences. We further demonstrate that restricting training to more consistent objects like faces and using explicit semantic representations can facilitate learning abstract concepts such as visual persuasion which would otherwise be unlearnable in noisy, limited, and diverse settings (**H2**). In Chapter 5, we propose to build upon on work here to model *implicit* persuasion and bias within media sources, a signal which is less explicit than that found within ads.

5.0 MODELING POLITICAL BIAS IN MULTIMEDIA

Summary. *In this chapter, we extend our work on modeling latent visual concepts by leveraging multimodal information in the form of lengthy news articles paired with images. We model multimodal visual and textual political bias in contemporary media sources at scale using webly supervised data.*[1] *We collect a dataset of over one million unique images and associated news articles from left- and right-leaning news sources, and develop a method to predict the image's political leaning. This problem is particularly challenging because of the enormous intra-class visual and semantic diversity of our data. We propose two stages of training to tackle this problem. In the first stage, the model is forced to learn relevant visual concepts that, when joined with document embeddings computed from articles paired with the images, enable the model to predict bias. In the second stage, we remove the requirement of the text domain and train a visual classifier from the features of the former model. We show this two-stage approach that relies on an auxiliary task leveraging text, facilitates learning and outperforms several strong baselines. We present extensive quantitative and qualitative results analyzing our dataset. Our results reveal disparities in how different sides of the political spectrum portray individuals, groups, and topics. This problem is well-situated within the theme developed in this dissertation of modeling abstract visual phenomena from noisy, diverse datasets and extends our completed work by incorporating multimodal information. Subsequent chapters focusing on modeling abstract semantics in multimedia build upon the dataset gathered in this chapter.*

5.1 INTRODUCTION

We have previously studied techniques for modeling latent visual phenomena: by using images alone (Chapter 3) and by using images and explicitly engineered visual semantics

[1] The work presented in this chapter was published in our NeurIPS 2019 paper, "Predicting the Politics of an Image Using Webly Supervised Data" [341] and a journal version is in submission to IJCV.

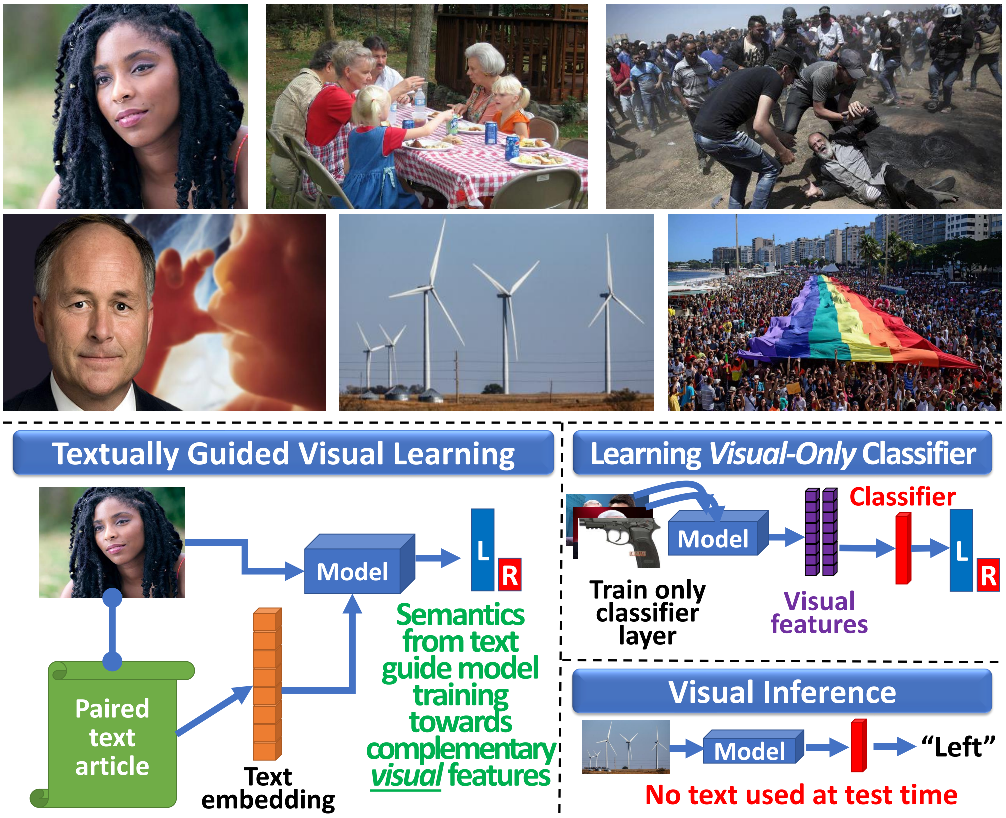

Textually Guided Visual Learning
Model
L
R
Paired text article
Text embedding
Semantics from text guide model training towards complementary *visual* features
Learning *Visual-Only* Classifier
Model
Classifier
L
R
Train only classifier layer
Visual features
Visual Inference
Model
"Left"
No text used at test time

Figure 14: **Top:** Can you guess whether each image appears in a far-left or far-right media source? Use your bias: What are the left and right stereotypically associated with? See the footnote[2] for answers. **Bottom:** Our method relies on text paired with images to guide the model towards learning relevant *visual* semantics. We then freeze our model and learn a purely visual classifier using features extracted from our pre-trained model. At test time, our method makes purely visual classifications, without requiring any text for inference.

(Chapter 4). In this chapter, we extend our work of modeling latent visual concepts by proposing a method applicable to real-world *multimedia* (i.e. content consisting of both images and text), rather than images alone. In particular, we consider the problem of modeling multimodal visual political bias.

One of the goals of the media is to inform, but in practice, the media also shapes opinions

[2] Top: Left, Right, Left; Bottom: Right, Left, Left

[124, 284, 10, 101, 307, 250]. The same issue can be presented from multiple perspectives, both in terms of the text written in an article, and the visual content chosen to illustrate the article. For example, when speaking of immigration, left-leaning sources might showcase the struggles of well-meaning immigrants, while right-leaning sources might portray the misdeeds of law-breaking immigrants. The type of topics portrayed is a strong cue for the left or right bias of the source media —for example, tradition is primarily seen as a value on the right, while diversity is seen as a value on the left [79].

In this chapter, we present a method for recognizing the political bias of an image, which we define as whether the image came from a left- or right-leaning media source. This requires understanding: 1) what visual concepts to look for in images, and 2) how these visual concepts are portrayed across the spectrum. Note that this is a very challenging task because many of the concepts that we aim to learn show serious visual variability within the left and right. For example, the concept of "immigration" can be illustrated with a photo of a border wall, children crying behind bars while detained, immigration agents, protests and demonstrations about the issue, politicians giving speeches, etc. Human viewers account for such within-class variance by generalizing what they see into broader semantic concepts or themes using prior knowledge, deduction, and reasoning.

On the other hand, modern CNN architectures learn by discovering recurring textures or edges representing objects in the images through backpropagation. However, the same objects might appear and be discussed *across* the political spectrum, meaning that the simple presence or absence of objects is not a good indicator of the politics of an image. Thus, model training may fall into poor local minima due to the lack of a recurring discriminative signal. Further, it is not merely the presence or absence of objects that matters, but rather *how* they are portrayed, often in subtle ways.

In order to capture the visual concepts necessary to predict the politics of an image, we propose a method which uses an auxiliary channel at training time, namely the article text that the image is paired with. Our method contains two stages. In the first one, we learn a document embedding model on the articles, then train a model to predict the bias of the image, given the image and the paired document embedding. To be successful on this task, the model learns to recognize visual cues which complement the textual embedding and suggest the politics of the image-text pair. At test time, we want to recognize bias from images alone, without any article text. Thus, in the second training stage of the model, we use the first stage model as a feature extractor and train a linear bias classifier on top. The article text serves as a type of privileged information to help guide learning.

Since recognizing the right semantic and visual concepts amidst intra-class variance requires large amounts of data, we train our approach on webly supervised data: the only labels are in the form of the political leaning of the source that the image came from. However, for testing purposes, we collect human annotations of bias (political leaning) and test on images where annotators agreed on the label. We experimentally show that our method outperforms numerous baselines on both a large held-out webly supervised test set, and the set of human-annotated images.

We present many qualitative results, studying different types of bias inherent within our dataset, including both visual and text bias. Our results show different political groups present different subjects (incl. politicians, political groups, individuals, etc.) in significantly disparate ways. We also present generative results in which we explicitly model, and then generate, faces exhibiting the disparities our method captures.

Ethical ramifications. We believe that recognizing the political bias of a photograph is an important step towards building computer vision systems that are aware of matters of social importance. Such awareness is necessary if we hope to use computer vision systems to automatically tag or describe images (e.g. for the visually impaired) or to summarize large collections of potentially biased visual content. Social media companies or search engines may deploy such techniques to automatically identify the political bent of images or even entire news sites being spread or linked to. Progress has already been made in this space in domains other than images. For example, Facebook automatically determines users' political leanings from site activity and pages liked [239]. Other works have studied predicting political affiliation from text [58, 376, 358] or even MRI scans [308]. However, *visual* bias understanding has been greatly underexplored. While some work examines *visual persuasion* [164, 149] or how political figures are portrayed in the media [280], none analyze predicting the political leaning of general images as we do.

The goal of our work is not to enable or further discrimination or reinforce stereotypes about individuals or groups. Rather, our work seeks to use machine learning techniques to *reveal* disparities in visual media which already exist. By raising awareness, we hope individual consumers of media are better able to approach material they are presented with (with a more skeptical eye) and question whether the portrayal of a subject they are seeing is politically skewed. Our work can also be used to combat, rather than reinforce, bias. One of many possibilities is a "balanced" image search engine, where our method is used to predict the political bias for each image returned. Studies [258] show that search engine algorithms may perpetuate bias. The bias score accompanying each image could be directly presented to the user. Another possible option would be to explicitly present users with images from both sides of the political spectrum, allowing the user to get a broader view of the subject. By returning images from across the political spectrum and/or explicitly revealing the inherent bias of images, we can help users be more informed consumers of visual media.

To summarize, our contributions in this chapter are as follows:

- We collect and make available[3] a very large dataset of biased images with paired text, and a large amount of diverse crowdsourced annotations regarding political bias.
- We propose a weakly supervised method for predicting the political leaning of an image by using noisy auxiliary textual data at training time.
- We perform detailed experimental analysis of our method on both webly supervised and human annotated data, and demonstrate the factors humans use to predict bias in images.

5.2 APPROACH

5.2.1 Dataset

Because no dataset exists for this problem, we assemble and release[4] large dataset of images and text about contemporary politically charged topics. We got a list of "biased"

[3] Our dataset, code, and additional materials are available online for download here: `http://www.cs.pitt.edu/∼chris/politics`

[4] `http://www.cs.pitt.edu/∼chris/politics`

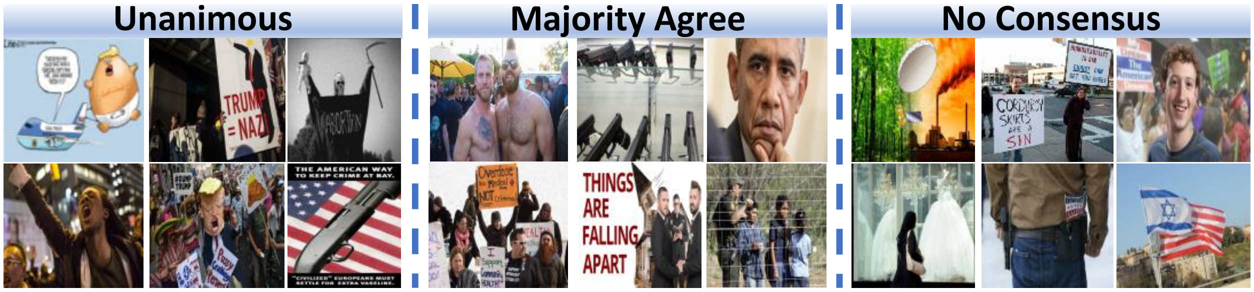

Unanimous
Majority Agree
No Consensus
TRUMP = NAZI
THE AMERICAN WAY TO KEEP CRIME AT BAY.
"CIVILIZED" EUROPEANS MUST SETTLE FOR EXTRA VASELINE.
THINGS ARE FALLING APART
CORDUROY SKIRTS ARE A SIN

Figure 15: We asked workers to predict the political leaning of images. We show examples here where all annotators agree, the majority agree, and where there was no consensus.

sources from `mediabiasfactcheck.com` which places news media on a spectrum from extreme left to extreme right. We used a list of "hot topics" e.g. immigration, LGBT rights, welfare, terrorism, the environment, etc from [276]. We crawled the media sources that were labeled left/right or extreme left/right for images using each of these topics as queries. After identifying images associated with each keyword and the pages they were on, we used [283]'s method to extract articles. We obtained 1,861,336 images total and 1,559,004 articles total. We manually removed some boilerplate text (headers, copyrights, etc.) which leaked into some articles. However, because of the large diversity of HTML formats across the media sources, boilerplate text could not be completely removed in all cases.

5.2.2 Data deduplication

Because sources cover the same events, some images are published multiple times. To prevent models from "cheating" by memorization, all experiments are performed on a "deduplicated" subset of our data. We extract features from a Resnet [127] model for all images. Because computing distances between all pairs is intractable, we use [231] for approximate kNN search ($k = 200$). We set a threshold on neighbors' distances to find duplicates and near-duplicates. We determine the threshold empirically by examining hundreds of kNN matches to ensure all near-duplicates are detected. From each set of duplicates, we select one image (and its associated article) to remain in our "deduplicated" dataset while excluding all others. If the same image appeared in both left and right media sources, we keep it on the side where it was more common, e.g. one left source and three right sources would result in preserving one of the image-text pairs from the right sources. After removing duplicates, we are left with 1,079,588 unique images and paired text on which the remainder of this chapter is based.

5.2.3 Crowdsourcing annotations

We treat the problem of predicting bias as a weakly supervised task. For training, we assume all image-text pairs have the political leaning of the source they come from. In Sec. 5.3.3 we show that this assumption is reasonable by leveraging human labels, though it is certainly not correct for all images / text, e.g. a left-leaning source may publish a right-leaning image to critique it, or a photo in a biased source may contain no bias at all (e.g. an image of a cat). In order to better explore the viability of the weak labels, and understand human conceptions of bias, we ran a large-scale crowdsourcing study on Amazon Mechanical Turk (MTurk). We asked workers to guess the political leaning of images by indicating whether the image favored the left, right, or was unclear. In total, we showed 3,237 images to at least three workers each. We show examples of different levels of agreement in Fig. 15. In total, 993 were labeled with a clear L/R label by at least a majority. The remaining images were labeled as some combination of "Unclear" labels with "Left"/"Right" labels, e.g. "UUL" or "ULR".

We also asked our annotators what image features were used to make their guess. The features workers could choose (and the count of each agreed upon) was: closeup-90 (closeup of specific person's face), known person-409 (portrays public figure in political way), multiple people-237 (group or class of people portrayed in political way), no people-81 (scenes or objects associated with parties, e.g. windmill/left, gun/right), symbols-104 (e.g. swastika, pride flag), non-photographic-130 (cartoons, charts, etc.), logos-77 (logo of e.g. CNN, FOX, etc.), and text in image-267 (e.g. text on protest signs, captions, etc.). We also asked workers to provide a free-form text explanation of their politics prediction for a small number of images. We extracted semantic concepts from these explanations and later use them to train one of our baseline methods (Sec. 5.3.1). Humans often mentioned using the positive/negative portrayal of public figures and the gender, race and ethnicity of photo subjects. We provide a demonstration of differences in portrayal across L/R in Sec. 5.3.4. Absent these cues, workers used stereotypical notions of what issues the left/right discuss or their values. For example, for images of protests or college women, annotators might guess "left".

We next showed workers the image's article and asked a series of questions about the image-text pair, such as the political leaning of the *pair* (as opposed to image only), the topic (e.g. terrorism, LGBT) the pair is related to, and which part of the article text is best aligned with the image. We computed agreement scores and found that 2.45 out 3 annotators agreed on the bias label of an image on average (including the "unclear" label), while 1.71 out of 3 agreed on topic, on average.

To ensure quality, we used validation images with obvious bias to disqualify careless workers. We restricted our task to US workers who passed a qualification test verifying familiarity with recent news and persons in the news, who had $\geq$ 98% approval rate, and who had completed $\geq$1,000 HITs. In total, we collected 14,327 sets of annotations (each containing image bias label, image-text pair bias label, topic, etc.) at a cost of $4,771. We include a number of experimental results on this human annotated set of images in Sec. 5.3.3.

5.2.4 Relationship between weakly supervised and human annotations

In order to ensure that our weakly supervised labels are actually capturing a meaningful signal which approximates human understandings of political bias, we perform the following test of weak-to-human label correlation. We evaluated the impact of text on humans' bias predictions. To do so, we compared how humans *changed* their predictions (made originally using the image only) after they saw the text paired with the image.

We found that when workers picked a L/R label, the label was strongly correlated with the weakly supervised label. Moreover, after seeing the text, humans became even more correct with respect to the noisy labels, switching many "unclear" predictions to the "correct" label (i.e. the noisy label). Specifically, in Table 8, we show the number of images labeled L/R before/after showing the worker the text paired with the image. Rows represent the

		Human Label After Seeing Image + Text							
		Weakly Supervised = Left				**Weakly Supervised = Right**			
		Left	**Right**	**Unclear**	**SUM**	**Left**	**Right**	**Unclear**	**SUM**
Human Label On Image Only	**Left**	**67**	**2**	**13**	**82**	28	**20**	6	54
	Right	**17**	22	10	49	**9**	**25**	**2**	**36**
	Unclear	**95**	8	207	310	37	**47**	121	205
	SUM	**179**	32	230	441	74	**92**	129	295

Table 8: Counts of how many users labeled an image Left / Right / Unclear. Rows show the label of the human on the image alone, while columns show the label after seeing the text. We further divide the table into two larger columns, which represent images with a weak label of Left / Right. Our results show the text helps annotators, and provides evidence that our weakly supervised labels are meaningful. See text for further discussion. We shade rows and columns corresponding to the "correct" label (with respect to the weakly supervised label) both before and after seeing the text.

image-only label of humans, and the columns represent the label after seeing both the image and the paired text. Any off-diagonal number represents a change in labeling between seeing the image only and seeing the image and text. When the weakly supervised label is Left, for example, we can see that of the 82 people who initially voted Left, 15 changed their vote (incorrectly, i.e. diverging from the source-derived bias) after seeing the text. Of the 49 who voted Right initially, 22 kept their initial vote, 17 changed their vote to Left, while 10 changed it to Unclear. Finally, for the 310 who initially voted an image was Unclear, 95 changed their vote to Left after seeing the text, 8 changed it to Right, and a majority kept it Unclear. When the weak label is Left, we see that while 82 initially voted left, after seeing the text 179 voted left. When the weak label is Right, 36 agreed with the weak label before, and 92 after seeing the text. In other words, we can conclude that after seeing the (disambiguating) text, annotators do in fact align more with the weak label of the image, which indicates that the weakly supervised label captures a meaningful notion of bias.

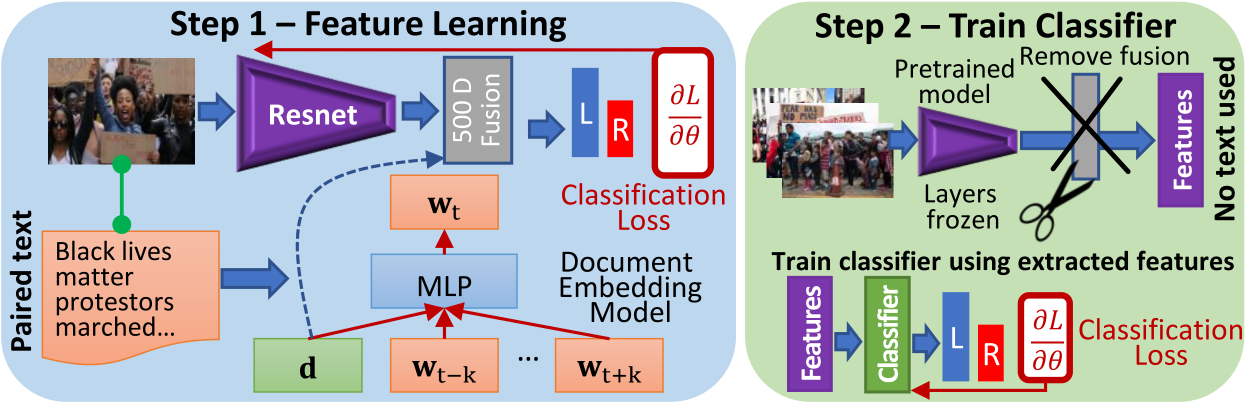


Figure 16: We propose a two-stage approach. In stage 1, we learn visual features jointly with paired text for bias classification. In stage 2, we remove the text dependency by training a classifier on top of our prior model using purely visual features. We show that this approach significantly outperforms directly training a model to predict bias. See Sec. 5.2.5 for details.

Overall, this analysis indicates that: 1) our noisy labels are a good approximation of the true bias of the images (and thus can be used for training a method); and 2) the paired text is useful for predicting bias (a result also later borne out by our experiments).

We hypothesize that the complementary textual domain provides a useful cue to guide the training of our visual bias classifier. The text of the articles includes words that clearly correlate with political bias, e.g. “unite”, “medicaid”, “donations”, “homosexuality”, “Putin”, “Antifa” and “brutality” strongly correlate with left bias according to our model, while “defend”, “retired”, “NRA”, “minister” and “cooperation” strongly correlate with right bias. By factoring out these semantic concepts into the auxiliary text domain, we enable our model to learn complementary visual cues. We use information flowing from the visual pipeline, and fuse it with the document embedding as an auxiliary source of information. Because we are primarily interested in *visual* political bias, we next remove our model's reliance on textual features, but keep all convolutional layers fixed. We train a linear bias classifier on top of the first model, using it as a feature extractor. Thus, at *test time,* our model predicts the bias of an image *without using any text.* We illustrate our method in Fig. 16.

5.2.5 Method details

We wish to capture the implicit semantics of an image by leveraging the association between images and text. More specifically, let

$$\mathcal{D} = \{\mathbf{x}_i, \mathbf{a}_i, \mathbf{y}_i\}_{i=1}^{N} \tag{12}$$

denote our dataset $\mathcal{D}$, where $\mathbf{x}_i$ represents image i, $\mathbf{a}_i$, represents the textual article associated with the i^{th} image, and $\mathbf{y}_i$ represents the political leaning of the image. In the first stage of our method, we seek the following function:

$$f_\theta\left(\mathbf{x}_i, \Omega\left(\mathbf{a}_i\right)\right) = \mathbf{y}_i \tag{13}$$

where $\Omega\left(.\right)$ represents transforming the article text into a latent feature space. We train Doc2Vec [192] offline on our train set of articles to parameterize Ω. Specifically, Ω is trained to maximize the average log probability

$$\frac{1}{T}\sum_{t=1}^{T} \log p\left(\mathbf{w_t}|\mathbf{d}, \mathbf{w_{t-k}}, \ldots, \mathbf{w_{t+k}}\right) \tag{14}$$

where T is the number of words in article $\mathbf{a}$ (we omit the index i to simplify notation), p represents the probability of the indicated word, $\mathbf{w_t}$ is the learned embedding for word t of article $\mathbf{a}$, $\mathbf{d}$ is the learned document embedding of $\mathbf{a}$ (200D), and k is the window around the word to look when training the model. We use hierarchical softmax [248] to compute p. We train Doc2Vec on our corpus of news articles, and observe more intuitive embeddings than from a pretrained model.

We show examples of the learned Doc2Vec space in Table 9. In the top row of the table, we show several query words which we embed using our model. We then compute the distance from each query word to all other learned words in our dataset's vocabulary and rank the words in order of increasing distance. Thus, retrieved words near the top are more closely related to the query word in the learned space than words below. We observe meaningful relationships within the space which are model can potentially exploit. For example, we see for the topic "Stoneman" (a school shooting), the model has learned that "Parkland" (another school shooting), "NRA" (the National Rifle Association which

Charlottesville	**Clinton**	**dreamers**	**fascist**	**FBI**	**FOX**	**Obama**	**Stoneman**	**supremacist**	**terrorism**	**Trump**
parkland	o'reilly	daca	fascism	cia	nbc	trump	parkland	supremacists	extremism	obama
antifa	maher	immigrants	racists	comey	cbs	bush	nra	supremacy	islamophobia	bush
ferguson	obama	undocumented	racist	doj	abc	reagan	gunman	nationalist	extremists	duterte
dallas	bush	aliens	nationalist	irs	breitbart	erdogan	shooter	house	racism	erdogan
rally	huckabee	immigration	extremist	investigation	cable	bashar	morning	privilege	extremist	sterling
nfl	merkel	deportation	supremacist	mueller	fake	clinton	separating	dana	fascism	reagan
islamophobia	trump	illegally	democrat	intelligence	buzzfeed	duterte	shootings	fascist	fbi	corbyn
berkeley	blasio	deferred	supremacy	flynn	cnn	macron	cbs	extremist	bigotry	macron
spencer	davis	shutdown	bigotry	wikileaks	hannity	carter	sheriff	conspiracy	immigration	clinton
shootings	treasury	amnesty	supremacists	dhs	outlet	vice	ripped	racist	shootings	bashar
tweeted	benghazi	bipartisan	islamophobia	epa	msnbc	obamacare	outrage	evangelical	russia	cameron

Table 9: Word relationships learned by our trained document embedding. The top row are query words and the words below are the nearest words in the learned space.

protested gun measures following the shooting), "gunman", "shooter" and related words all relate to the broader topic of school shootings and violence in general. By providing this semantic supervision to our model, we wish to discover relevant *visual* cues which relate to the broader subject matter, and which are predictive of the politics of the image.

After training, we compute Ω for a given article $\mathbf{a}$ by finding the embedding $\mathbf{d}$ that maximizes Eq. 14. Ω thus projects each article into a space where the resulting vector captures the overall latent context and topic of the article. We provide $\Omega(\mathbf{a})$ to our model's fusion layer for each train image. The fusion layer is a linear layer which receives concatenated image and text features and learns to project them into a multimodal image-text embedding space which is finally used by the classifier.

The formulation of $f_\theta(.)$ described above requires that the *ground-truth* text be available at test time and also does not ensure that our model is learning *visual* bias (i.e. the classifier may be relying primarily on text features and ignoring the visual channel completely). To address this problem, in the second stage of our method, we finetune f_θ to directly predict the politics of an *image only*, without the text, as follows:

$$f'_{\theta,\theta'}(\mathbf{x}_i) = \mathbf{y}_i \tag{15}$$

Specifically, we freeze the trained convolutional parameters of f_θ and add a final linear classifier layer to the network, whose parameters are denoted θ'. Because f_θ's convolutional layers have already been trained jointly with text features, they have already learned to extract visual features which complemented the text domain; we now learn to use those features *alone* for bias prediction, as shown in Fig. 16.

5.2.6 Additional method for faces

We wish to explore whether the same people were shown in disparate ways across the political spectrum. We thus began by detecting faces in our dataset using DLIB's [174] CNN-based face detector. Observationally, we found there is strong visual variability in the faces that left/right-leaning sources choose for popular figures, such as Donald Trump, Barack Obama and Hillary Clinton. We later provide quantitative and qualitative demonstrations of this in Sec. 5.3.4.

We also seek to capture the semantics behind these differences in facial portrayals. To do so, we leverage existing datasets containing labeled facial attributes and expressions. We train two residual networks on the datasets of [217] and [247], and use them to predict facial attributes and expressions for every image in our dataset. After detecting faces in our dataset, we *recognize* faces of known political figures because we expect popular political figures to recur throughout the dataset and be indicative of bias. In order to decide which figures to recognize, we leverage the text paired with images. We ran [136]'s named entity recognizer on our text articles and narrowed the list of detected "Person" entities to the 96 most frequent politicians (and other celebrities) to form a vocabulary of "known" faces. We downloaded images for each face and used [309] to perform face recognition on our detected faces.

Formally, let f_e, f_a, f_r be our facial attribute, expression, and recognition networks, respectively. For each image in our dataset, $\mathbf{x}^i$, we obtain automatically predicted attributes, expression labels, and one-hot identity labels as follows:

$$\left\{\mathbf{x}^i, f_e\left(\mathbf{x}^i\right), f_a\left(\mathbf{x}^i\right), f_r\left(\mathbf{x}^i\right)\right\}_{i=1}^{N}. \tag{16}$$

We later use these predicted facial attribute and expression features for analysis and as input for our baseline networks for predicting the bias of faces.

In addition to "Person" entities detected by our named entity recognizer in the text, we also examine the nationalities, religious, and political groups (NORP) entities detected by our recognizer. As we did with person entities, we use the detected vocabulary of NORPs and download data for each of the top 200 from Google Image Search. We train a separate residual network [127] to perform image classification on this train set. We then use this model to provide predictions of each concept on each image in our dataset: $P(n_j|\mathbf{x}_i)$ denotes the probability that image $\mathbf{x}_i$ exhibits NORP n_j. We then use the probability of each predicted concept, as a feature vector $\mathbf{x}_i = [P(n_1|\mathbf{x}_i), P(n_2|\mathbf{x}_i), \ldots, P(n_{|N|}|\mathbf{x}_i)]$, where N is our vocabulary of NORPs. We later use these predictions for analyzing how different sides of the political spectrum portray different NORPs.

Note that we do not use demographic information to predict bias. Instead, we use it to show that some demographic factors (e.g. certain ethnic groups) are portrayed in notably different ways on the left and right in our crawled dataset. While we firmly believe this result is problematic, our goal is to point out the problem so it may eventually be addressed.

5.2.7 Implementation details

All methods use the Resnet-50 [127] architecture and are initialized with a pretrained Imagenet model. We train all models using Adam [175], with learning rate of 1.0e-4 and minibatch size of 64 images. We use cross-entropy loss and apply class-weight balancing to correct for slight data imbalance between L/R. We use an image size of 224x224 and random horizontal flipping as data augmentation. We use Xavier initialization [107] for non-pretrained layers. We use PyTorch [273] to train all image models. For our text embedding, we use [295], with $\mathbf{d} \in \mathcal{R}^{200\times 1}$ and train using distributed memory [192] for 20 epochs with window size $k = 20$, ignoring words which appear less than 20 times.

5.3 EXPERIMENTAL EVALUATION

In this section, we present experimental results on a number of tasks. We introduce the baselines we compare against for politics prediction, in Section 5.3.1. In Section 5.3.2, we present our results for predicting left/right bias on full images for a variety of methods, and perform a detailed analysis of factors the model uses for prediction. We test on a large held-out test set from our dataset, whose left/right labels come from the leaning of the news source containing the image. We also perform ablations of our method using weakly supervised labels to test the soundness of our method and experimental design for politics prediction. After testing on weakly supervised labels, in Section 5.3.3 we test our methods using the per-image labels provided by humans. We show results on test images for which a majority of human annotators agreed on the bias. We also further discuss the relationship between our weakly supervised and human labels and further analyze how humans reason about visual bias.

Because we find humans strongly relied on identifying known public figures and how people were portrayed in guessing the politics of an image, we then perform an analysis of faces alone, without any context from the image, in Section 5.3.4. We first trained models to predict the bias of faces. We show results for both well-known politicians and for faces in general. We then analyze the differences in facial portrayals across the left/right for a variety of facial features. We present results showing that faces are portrayed significantly different for popular figures and ethnic groups on opposite ends of the political spectrum. We also show the most important semantic facial features for predicting the politics of an image.

In Section 5.3.5, we perform a similar analysis of the text paired with images, to discover how the text itself manifests political bias. We note political figures and some ethnic groups that appear disproportionately on one side vs. the other. Similarly, we leverage existing techniques for discovering biased word usage in language to analyze our dataset, discovering biased words used by each side of the political spectrum.

In Section 5.3.6, we present several results exploring the relationship between image and text. We show the most "visually consistent" words in our dataset (i.e. the words where the paired image content is more consistent across images which the word was paired with). We also show results for directly predicting the words that appeared in the article given an image.

Finally, in Section 5.3.7, we examine the topic annotations (e.g. abortion, gun rights, etc.) within our dataset. We also show visual consistency across topics, with some visually grounded topics (e.g. gun rights) being more consistent in visual space than abstract topics, illustrating the challenging semantic nature of the problem of modeling visual political bias.

5.3.1 Methods compared

For quantitative results, we show the accuracy of each method on predicting left/right bias. Note that we apply some baselines to either full images or faces. For example, "facial semantics" only applies to faces. Similarly, OCR is not applicable to faces detections. We compare against the following baselines:

- Resnet [127] - A standard 50-layer classification Resnet, trained for left/right classification.
- Curric - Our approach is a two-stage curriculum method, which first learns features coupled with text features and then learns to predict bias without the text. We wanted to see whether a Resnet trained in the same way would gain any benefit. We thus first train a Resnet on our task. We then freeze the lower layers and train *only* the classifier on all train features in the second stage. Optimizing over all train features at once vs. minibatches can mitigate noisy gradients from our diverse and noisy data [266, 128].
- Joo [164] - Adaptation of Joo et al.'s method for our task. We use [164]'s dataset to train predictors for 15 attributes and nine "intents" (qualities the photo subject is estimated to have, e.g. trustworthiness, competence). We then use the predictions for these attributes and intents on images from our dataset as additional features to a Resnet to predict a left/right leaning.
- HumConc - We use the manually extracted vocabulary of bias-related concepts (e.g. "confederate", "African-American") from the human-provided explanations (Sec. 5.2.3) and download data for each from Google Image Search. We train a separate Resnet to

predict concepts, and use it on each image in our dataset: $p(c_j|\mathbf{x}_i)$ denotes the probability that image $\mathbf{x}_i$ exhibits concept c_j. We then use the confidence of each detected concept, as a feature vector to predict bias.

- OCR - We use [242] to recognize free-form scene text in images. Because images contain words not found in the default lexicon (e.g. "Manafort"), we create our own lexicon from the 100k most common words in our articles. We use [95] for spelling correction. We represent each recognized word as its learned word embedding, denoted $\mathbf{w}_i'$, weighed by the confidence of the recognition $p\left(\mathbf{w}_i'\right)$ as provided by the recognition model. The feature is thus given by $\frac{1}{n}\sum_{i=1}^{n} p\left(\mathbf{w}_i'\right)\mathbf{w}_i'$.
- Gomez [108] - Similar to our method, Gomez leverages text to guide the learning process, without requiring text at test time. Gomez first trains a Resnet to predict the text embedding of the article paired with the image, from the image alone. Note that in our case, we do not predict the text embedding, but rather use it as a source of auxiliary information. In the second stage, a classifier is trained to predict the left / right label from features predicted by the model.
- Zhang [404] - We compute nearest neighbors for each image in visual space and formulate the inference problem as a graph. We compute attention using the features of neighboring images extracted from the last layer of Resnet. We leverage neighbors' features to assist in inferring the political label. The intuition behind using this approach as a baseline is that images in our dataset are ambiguous, hence neighbors may make the learning task easier.
- FacialSem - We predicted facial attributes, expressions, and identities for every face in our dataset (see Sec. 5.2.6). We create a feature vector by appending the predicted identity of the portrayed person to the facial attributes and expressions, resulting in the following vector which is fused with Resnet image features: $\mathbf{x}_i = [f_a(a_1|\mathbf{x}_i), \ldots, f_a(a_m|\mathbf{x}_i), f_e(e_1|\mathbf{x}_i), \ldots, f_e(e_n|\mathbf{x}_i), f_r(p_1|\mathbf{x}_i), \ldots, f_r(p_o|\mathbf{x}_i)]$, where $f_a(a_j)$ and $f_e(e_j)$ denote the confidence of attribute/expression a_j or e_j being present in image $\mathbf{x}_i$. Further, $f_r(p_k|\mathbf{x}_i)$ is a 1 or 0 depending on whether person identity marker p_k is predicted in $\mathbf{x}_i$.

For reference, we also show three "upper-bound" methods which use the ground truth text paired with the images *at test time*. Because these methods rely on text at test time, we thus consider them upper-bounds to the task of visual-only prediction.

- Text uses the document embeddings computed from the text paired with the image, without using the image at test time.
- Words is a two stage method. We first train a model to predict words from an image and its paired document embedding. We trained a Resnet to predict, for the 1000 most visually consistent words (see Sec.5.3.6, which words appeared in the first two sentences of the image's paired article. To make the modeling task easier, we also conditioned the model on the Doc2Vec vector of the image. Specifically, the model is trained to predict whether each word is/is not present in the image's article given the image and text embedding. We then use train a second model using *just the predicted words* to predict the left/right label for the image.
- Im+Text uses the text paired with the images (to compute a document embedding), in addition to the image. It is the same as the first stage of our approach (see Fig. 16, left), without the addition of the image classifier layer in step 2.

All methods use the same residual network architecture. For methods relying on additional features, we use the fusion architecture in Fig. 16.

5.3.2 Evaluating on weakly supervised labels

In this section, we present our experimental results for predicting the political leaning of visual media. We first present results on full images using our weakly supervised labels. We also show ablations of our method to test assumptions about our weakly supervised labels and experimental design. Later, in Section 5.3.3, we will evaluate on human labels and perform an analysis of the relationship between human labels and our weakly supervised labels. In Section 5.3.4, we show results on face crops only.

In Table 10, we show the results of evaluating our methods on 75,148 held-out images with weakly supervised labels. Our method performs best overall. The top two performing methods rely on semantics discovered in the text domain (Ours and OCR). OCR is unique in that it is able to explicitly use text information at test time, by discovering text within the image and then using word embeddings. Ours improves over OCR by 2.6% (relative 3.8%,

Method	**Resnet**	**Curric**	**Joo**	**HumConc**	**OCR**	**Gomez**	**Zhang**	**Ours**	**Text**	**Words**	**Im+Text**
Accuracy	0.678	*0.687*	0.670	0.675	0.686	0.547	0.566	**0.712**	0.825	0.626	0.803

Table 10: Accuracy on weakly supervised labels with the best visual-only prediction method in bold, and second-best in italics. These results are computed on full images, hence face-specific methods are excluded.

reduction in error of 8%). The improvement of Ours over Resnet is 3.4% (relative 5%, error reduction of 11%). This amounts to classifying an additional ~2,555 images correctly. We also observe that Curric performs nearly 1% better than Resnet. One reason for this is because of the high visual diversity of our dataset. A classifier that is being trained while the lower layers of the model continue to change (because the model keeps shifting features because it is unable to settle on consistent patters) must continually readjust itself to the changing features. However, by freezing the lower layers of the model, we allow the classifier to optimize for a stationary set of visual features over the entire dataset. The classifier is thus able to obtain a better and more stable classification, resulting in a slight gain in performance.

We observe that Joo, which leverages features learned on an external dataset, performs worse than Resnet (i.e. not using those features). This is likely because [164]'s data mainly features closeups of politicians, while ours contains a much broader image range, thus the predicted features are not useful in our setting. Further, relying on the concepts humans identified actually slightly *hurt* performance compared to Resnet. This may be because of a disconnect between humans' preconceived notions about left/right and those required by the dataset.

In addition, we note that Gomez performs much worse than our method, even though both try to exploit information in the paired text domain as a source of privileged information. We believe one reason for this is because of the many-to-many relationship of images with topics (e.g. image of the White House can be paired with text about Trump's children, border control, LGBT rights, etc.). Thus, it is much harder to predict the document embedding paired with the image since the text could be about many different issues. It is likely that the text embedding prediction model of Gomez is unable to accurately predict an embedding, thus the model's features are not discriminative of politics. We observe that Zhang, which relies on nearest neighbors computed in image space, also performs poorly. One reason for this is because our problem lies highly in semantic, rather than visual space. Thus, visual nearest neighbors are not necessarily indicative of the politics of a particular image. Furthermore, because of the high visual variability of our dataset, providing additional (potentially irrelevant) features for the classifier to consider at train time distracts it from recognizing the weak signal from the query image features.

Quantitative ablations. In order to test the soundness of our method and our experimental design, we performed several ablations. We first tested the importance of the second stage of our method (right side of Fig. 16). To do so, we used Im+Text, the result of the first stage of our method, and instead of performing stage 2, we removed the dependency on text by zeroing out all text embedding weights in the fusion layer. We evaluated on our weakly supervised test set and obtained 0.677, a result significantly worse than our full method, underscoring the importance of stage 2. We next tested how the performance of our method varied given the length of the article text. We thus trained our method with the first k sentences of the article and obtained these results: $k = 1 \rightarrow 0.672$, $k = 2 \rightarrow 0.669$, $k = 5 \rightarrow 0.668$, $k = 10 \rightarrow 0.669$. All choices of k tested performed worse than using the full article (0.712). We finally examined how reliant our method was on images from a particular media source being in our train set (i.e. to test if the model was learning non-generalizable, source-specific features). We experimented with leaving out all training data harvested from a few popular sources. The result was (before excluding → after excluding): Breitbart (0.607→0.566), CNN (0.873→0.866), CommonDreams (0.647→0.636), DailyCaller (0.703→0.667), DemocraticUnderground (0.713→ 0.700), NewsMax (0.685→0.628), and TheBlaze (0.746→0.742). We observed only a slight decrease for all sources we tested, suggesting our method is not dependent on seeing the source at train time.

Feature/Method	**Resnet**	**Curric**	**Joo**	**HumConc**	**OCR**	**Ours**	**Text**	**Im+Text**	**# Ims**
Closeup	0.567	0.589	0.544	*0.622*	0.578	**0.656**	0.667	0.578	90
Known Person	*0.567*	0.558	0.550	**0.570**	0.560	0.521	0.558	0.575	409
Multiple People	0.722	*0.738*	0.671	0.688	0.730	**0.768**	0.709	0.705	237
No People	0.556	0.531	**0.605**	0.494	0.580	*0.593*	0.642	0.667	81
Symbols	0.558	0.587	*0.596*	0.548	0.577	**0.606**	0.625	0.587	104
Non-Photographic	0.577	**0.585**	0.569	*0.584*	0.577	**0.585**	0.631	0.654	130
Logos	0.545	*0.636*	0.584	0.597	**0.662**	0.623	0.546	0.584	77
Text in Image	0.629	**0.652**	0.625	0.596	*0.637*	0.607	0.648	0.659	267
Average	0.590	0.610	0.593	0.587	*0.613*	**0.620**	0.628	0.626	

Table 11: Accuracy on human consensus labels with the best visual-only prediction method in bold, and second-best in italics. The results are computed on full images grouped into eight categories by our human annotators.

5.3.3 Evaluating on human labels

We next tested our methods on test images which at least a majority of MTurkers labeled as having the same bias, i.e. those that humans agreed had a particular label. We describe this dataset in Sec. 5.2.3. Because workers also labeled images with what features of the image they used to make their prediction, we break down each method's performance by feature. We show the result in Table 11. Note that in this table, we only include the competitive methods from Table 10 for brevity; we include methods which achieve at least 65% accuracy.

Ours performs best on average across all categories and performs best (or ties) on four out of eight categories. Categories where Ours is outperformed are reasonable: OCR performs best or second-best when text can be relied on in the image, i.e. “logos” and “text in image”. We note that while the overall result for OCR approaches Ours, Ours works better on a broader set of images than OCR and is thus a more general method for predicting *visual* bias. Ours is also outperformed by HumConc when humans relied on a known face (politician, celebrity, etc.). This may be because HumConc relies on external training data

(Sec. 5.3.1) which feature many known individuals, e.g. “rappers” and “founding fathers”. Perhaps counterintuitively, Joo outperforms our method when the prediction depends on scene context (“no people”), but note that some of the attributes that Joo uses do capture the scene/background (e.g. indoor, background, national flag, etc.). Further, unlike Ours, this method uses an external human-labeled dataset to learn features, including the scene attributes. Curric improves upon Resnet (whose features it uses in its second stage of training) in nearly every category and performs best or ties in two categories. This result suggests that label noise and high visual diversity within minibatches may prevent the classifier from converging on the best local minima for a given set of features. By fixing the model’s layers and optimizing the classifier layer across the entire train set at once, the classifier converges on a better solution. This technique can be applied to any method to potentially improve their performance as well.

In terms of the “upper bound” methods, we note that Im+Text performs significantly worse on human labels vs. weakly-supervised labels. This is likely because some words in text point to a specific bias (e.g. abortion vs. pro-choice), which the model may be over-relying on to predict the bias of the image. In contrast, the relationship between image features and bias is often more ambiguous. Further, because of the noisy data collection, Im+Text may have learned to exploit dataset-specific features (e.g. author names, header text, etc.) for prediction, which does not actually translate into humans’ commonsense understanding of political bias. This also explains why Im+Text does not improve upon Text alone on average (but it does for five of eight categories).

We next test whether our assumption that all images harvested from a right- or left-leaning source exhibit that type of bias is reasonable. Several results computed from our ground-truth human study suggest that our web labels are a reasonable approximation of bias. First, we observe that the relative performance of the methods across Table 10 and 11 is roughly maintained; Ours is best, followed by OCR and Curric essentially tied. The results are also sound, e.g. when humans used text, OCR tends to do better, which indicates the model’s concept of bias correlates with humans’. Earlier, in Table 8, we showed that human labels agree with our weak labels more, when text information is presented to disambiguate the image’s bias.

Human label consensus's effect on performance. In an additional experiment, we explored the difference between the performance of our method on images on which the *majority* of humans agreed vs. those on which humans *unanimously* agreed. We found that our method worked better when humans unanimously labeled the images vs. simple majority (gain of 4.8%). This suggests that as humans become more certain of bias, our model (trained on noisy data) also performs better.

Detecting images with ambiguous bias. Our result in Table 8, showed that humans become more correct with respect to our weakly supervised labels after seeing the text. Next, we learn to predict whether a human will change their political label of the image *after seeing the text*. Thus, we model whether the politics of the image is *predictable without the text* (i.e. unambiguous). We use the F_1 score rather than accuracy due to class imbalance. We find that our model is fairly accurate at predicting whether images are unambiguous (0.731), but less accurate at detecting ambiguous images (0.308), i.e. images that humans change their label on. In other words, given an image and the text paired with the image, our model is able to accurately able to predict when the human label of the image *will not* change after humans see the text. These are likely simpler cases where the image has more apparent political bias. However, the model performs worse at predicting when the human label of the image *will* change, after seeing the text. For example, if the model suspects the image has a right bias, but suspects the text has a left bias, it is difficult for the model to decide whether humans will rely more on the image (unambiguous label) or the text (ambiguous label) to make their final decision for the image. These cases are likely more semantically complex (and therefore more difficult to model) and require one to reconsider what the image is intended to portray in the context of the text.

Politically discriminative words. Given the strong performance of models relying on text and the fact that workers became more confident after seeing the text, we wanted to discover words in the text which are indicative that the image-text pair leans left or right. We used the classifier from the WORDS model. In Table 12, we show words for both the left and the right that had the highest predictive weight (i.e. the word's appearance caused the classifier to be more likely to predict that category). We note several interesting results: "bob" (likely from Robert Mueller's name) and "unite" (from the Unite The Right protest)

Left							
bob	television	unite	views	speakers	irs	medicaid	putin
homosexuality	outlets	gary	enforce	donations	doj	opposition	broadcast
speaks	antifa	adhere	westminster	lobby	achievements	networks	pelosi
reactions	labour	venezuela	supporter	meeting	memoir	warrant	outlet
brutality	misleading	hall	sharon	prominent	illegal	angela	referring
raped	absurd	berkeley	spoke	donald	qaeda	karl	rejected
brad	quit	roe	intelligence	candidate	evan	hosted	comedian

Right							
teresa	defend	hopeful	survivor	indicted	immigration	colleagues	retired
theresa	refuse	political	roger	caucus	nba	bipartisan	williams
rand	nra	withdraw	trump	minister	racist	ratings	longtime
sexually	fox	joins	cruz	deputy	unilaterally	sentenced	denial
dana	pleaded	declaring	exposing	victories	planned	ballot	hannity
russians	juan	guests	hashtag	cooperation	establishment	chancellor	network
sarah	recording	blaming	deportation	roy	supporting	don	erdogan

Table 12: Words for both the left and the right that had the highest predictive weight (i.e. the word's appearance caused the classifier to be more likely to predict that category).

are among the strongest predictors of "left", while "teresa" (likely from Teresa May) and "immigration" strongly indicate right. Many of the words used by the model suggest topics frequently mentioned by their respective sides, but which are not mentioned by the other side, possibly because they are politically damaging / advantageous to one side. For example, "irs", "putin", "doj", and "antifa" are predictive of left, while "nra", "trump", "fox" predict right. In sum, this result allows us to see disparities in the issues covered in left vs. right articles, as well as the different words used by the articles which are politically discriminative.

5.3.4 Evaluating on faces

Many workers noted how politicians were portrayed in making their decision (Sec. 5.2.3). We thus wished to analyze how well our methods could do at predicting the politics of faces

Method	**Resnet**	**Joo**	**FacialSem**	**Ours**	**Im+Text**
Accuracy	0.588	0.579	**0.607**	0.590	0.723

Table 13: For a subset of methods, we show the accuracy of predicting the politics of *just faces* detected in our images and evaluate on weakly supervised labels. We show the best visual-only prediction method in bold.

alone, in the absence of any context from the image. We thus trained models to predict the political bias of the faces we detected in our images (see Sec. 5.2.6). We assume all detected faces in an image have the same political bias as the image itself (e.g. a right leaning image with 10 detected faces results in 10 individual samples all with the "right" weakly supervised label).

Predicting the bias of all faces. We present our results in Table 13 for a subset of the previously evaluated baselines as well as the face-specific method FacialSem which relies on predicted facial attributes, expressions, and identity. Note that the OCR model is inapplicable to cropped faces because those do not contain text. We observe that FacialSem substantially improves over other baselines and achieves the strongest performance (0.607). We also observe that Ours performs on par with (slightly better than) Resnet and Joo.

One possible reason for the lack of performance gain of our main method on faces is the lack of context from which the model can learn. In the full image setting, our method has a complete view of the image and the text (in stage 1) and is thus able to learn how the concepts in the image complement the text. However, in the face setting, our model has no visual context and is unable to learn relevant visual features to complement the shared text embedding. We note that even though the model sees no context outside of the cropped face, the FacialSem model is able to predict the political leaning of the face with 60.7% accuracy, which suggests that faces are portrayed in a biased manner which the models are capturing.

Predicting the bias of well-known vs. lesser-known faces. We next show the accuracy

Face Type	**Obama / Trump**	**One of 96 Politicians**	**Any NORP**	**All Faces**
Accuracy	0.830	0.820	0.670	0.590

Table 14: Accuracy of predicting the politics of different types of faces (from most well-known to least well-known).

no beard	blurry	young	social democrats	neutral	arousal	happy	narrow eyes
republican	valence	mouth slightly open	smiling	anger	gray hair	oval face	chubby
jordanian	haitian	balkan	qatari	catalan	pointy nose	sandinista	democrat
congolese	arab	latin	serbian	judeo-christian	semitic	iranian	leftists
marine	russian	latin american	canadian	saudi arabian	african american	anarchists	rwandan
wearing necktie	high cheekbones	eastern european	nigerian	veterans	evangelical	daesh	syrian
israeli	european	armenian	feminists	north african	ukrainian	soviet	protestant

Table 15: Most politically discriminative semantic features on faces in decreasing order (left to right, top to bottom).

of predicting the bias of a media source, based on of different types of faces, in Table 14. We wanted to test whether well-known public figures are portrayed in a substantially different way compared to lesser-known or unknown figures. We show our model's accuracy at predicting the politics of Obama / Trump faces (most well known), then on a much larger set of 96 politicians we detected in text as described in Sec. 5.2.6 (less well-known), then faces that were classified as being one of 196 nationalities, religions or groups (NORPs) (unknown person other than a known category or nationality), and finally all faces (most obscure / unknown). We observe our model is remarkably accurate for known political figures and that performance decreases as the face in question becomes less well known. Our results strongly suggest that public figures, and to a lesser extent, nationalities, religious, and political groups, are portrayed in politically biased ways. This is sensible because content creators may attempt to disparage or elevate political figures and groups of people who are politically opposed or aligned to their position.

Facial features most discriminative of politics. We next wanted to explore the relative importance of various semantic features for predicting the politics of faces. We trained a decision forest [100] to predict the bias indicated by a face, from concatenated predicted facial attributes, expressions, and NORP confidence scores. We then calculated the Gini importance [34] of each feature in the decision forest. The Gini importance, also known as the mean decrease in impurity, is a measure of the total decrease in subtree class impurity as a result of splitting on the feature (relative to the probability of reaching the node on which the split occurs). Thus, features which are more discriminative of classes (in our case left / right) have a higher mean decrease in impurity. The statistic is computed for all nodes and features in each tree and then the average value for each feature is computed across all decision trees in the forest. We show the most important (i.e. politically discriminative) features according to the model below in order of decreasing Gini importance (left to right, top to bottom) in Table 15.

We observe that two of the first three most important predicted features (the top row of Table 15) suggest that the model may be attempting to capture the age and gender of the faces: "no beard" is the most important feature (which may correlate to gender) and "young" is the third most important. The fourth is "social democrats" and the ninth is "republican" (both predicted NORP features), which suggests that the NORP model has learned some discriminative concept of what democrats' and republicans' faces look like. We also note that the "arousal" of the face (how intense the expression is) and whether the face is portrayed as "happy" is also strongly indicative of political bias. We believe these results are sensible in that media publishers may portray individuals or groups on their side of the political spectrum as happier. Similarly, we note that there exist age and gender disparities between the two major political parties, with the Democratic party (associated with the left) featuring younger and more female members than the Republican party [73]. We believe our models may be thus leveraging age, gender, ethnicity, etc. as cues for predicting the political alignment of faces, much as humans indicated they did in our crowdsourcing study.

Modeling facial differences across politics. So far we have seen strong evidence that faces are presented in substantially different ways across the political spectrum, particularly political figures. We next seek to actually *visualize* the differences in how well-known individ-

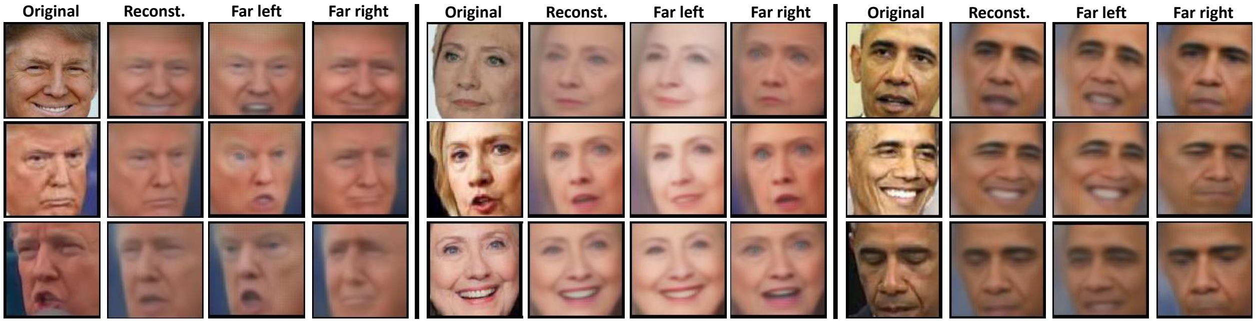


Figure 17: We modified photos to be more left/right-leaning, using a generative model trained on our noisy data. We show the model's "reconstruction" of each face next to the original sample, followed by the sample transformed to the far left and right.

uals are portrayed within our dataset. To this end, we trained a generative model to modify a given Trump/Clinton/Obama face, and make it appear as if it came from a left/right leaning source. We use a variation of the autoencoder-based model from [340], which learns a distribution of facial attributes and latent features on ads, not political images. We train the model using the features from the original method on faces of Trump/Clinton/Obama detected in our dataset. To modify an image, we condition the generator on the image's embedding and modify the distribution of attributes/expressions for the image to match that person's average portrayal on the left/right, following [340]'s technique. We show the results in Fig. 17. Observe that Trump and Clinton appear angry on the far-left/right (respectively) end of the spectrum. In contrast, all three appear happy/benevolent in sources supporting their own party. We also observe Clinton appears younger in far-left sources. In far-right sources, Obama appears confused or embarrassed. These results further underscore that our weakly supervised labels are accurate enough to extract a meaningful signal.

Discovering biased features for public figures. Though our results in Table 15 have indicated which semantic facial features are discriminative of political bias and Figure 17 visualized how those differences are expressed for several known politicians, we still have not shown quantitatively *which* semantic facial attributes differ for which politicians. We next show in Table 16 *which* features are different for *which* politicians, using a subset of all

Facial Attributes							
	Barack Obama	Chuck Schumer	Donald Trump	Hillary Clinton	Mitch Mcconnell	Nancy Pelosi	Paul Ryan
5 o'Clock Shadow	🟦	🟦	🟦				
Attractive	🟦			🟦			
Bags Under Eyes	🟦		🟦	🟦			🟦
Bald	🟦			🟦			
Big Lips	🟦		🟦			🟦	
Big Nose	🟦		🟦	🟦	🟦		🟦
Chubby	🟦			🟦	🟦		
Double Chin	🟦			🟦	🟦		
Gray Hair	🟦			🟦			
Heavy Makeup				🟦			
High Cheekbones	🟦		🟦	🟦	🟦		
Mouth Slightly Open	🟦	🟦	🟦				
Mustache	🟦		🟦				
Narrow Eyes	🟦		🟦	🟦		🟦	🟦
No Beard	🟦		🟦				
Pointy Nose			🟦	🟦			
Receding Hairline				🟦	🟦		
Rosy Cheeks				🟦		🟦	
Smiling	🟦			🟦			
Young	🟦		🟦	🟦	🟦	🟦	
Facial Expressions							
Anger	🟦		🟦	🟦			
Contempt	🟦				🟦		
Disgust				🟦		🟦	
Fear		🟦	🟦	🟦	🟦	🟦	
Happy	🟦		🟦	🟦		🟦	
Neutral			🟦		🟦		🟦
Sad	🟦	🟦	🟦	🟦	🟦		
Surprise			🟦	🟦	🟦		
Arousal	🟦		🟦	🟦		🟦	
Valence	🟦		🟦	🟦		🟦	

Table 16: Facial attributes and expressions which significantly differed (shown in blue) across the left/right per politician.

features.

We predicted facial attributes and expressions for the most frequent politicians which appeared in our dataset. We then performed a per-feature T-test to discover which attributes and expressions are portrayed differently across the left and the right for each politician. We highlight cells in blue whose feature differences in portrayal are significant ($p \leq 0.05$)

Facial Attributes and Expressions							
	African American	**Arab**	**Asian**	**Muslim**	**Mexican**	**Hispanic**	**White**
Median p-val.	**0.000**	**0.003**	**0.018**	0.071	0.099	0.185	0.370

Table 17: Frequently detected NORP's facial semantic attributes and their differences in portrayal across the left/right.

across the political spectrum. Empty cells indicate the difference across left/right was not significant.

We observe that Obama, Trump, and Clinton have the most facial differences. We observe a number of significant differences which were also reflected in our generation results (Figure 17). We see Hillary Clinton differs in "attractive", "bags under eyes", "chubby", "double chin", "heavy makeup", and numerous other attributes which suggest she is being portrayed as older and less attractive on one side vs. the other. We observe similar attribute patterns for Obama and Trump, with Obama and Trump likewise being portrayed differently in terms of their age ("young") and attractiveness (e.g. "5 o'clock Shadow", "bags under eyes"). For facial expressions, we see Obama, Trump, and Clinton all differ in the "anger" "happy", and "sad" facial attributes, as well as their facial expressions' arousal and valence scores. As was shown from our generation results, negative expressions (e.g. "anger", "disgust", etc.) are used to portray figures from the opposite side of the spectrum, while positive expressions (e.g. "happy") are used for political figures on the same side. Interestingly, we also note significant differences in both the arousal and valence scores for several politicians. Arousal is a measure of the intensity of a given facial expression and measures whether a given face is exciting / agitating vs. calm / soothing, while valence is a measure of the "pleasantness" of the face [247]. Thus, our results suggest that not only are the expressions themselves different, but the degree to which those expressions are shown is also different (i.e. through their arousal) as well as the overall pleasantness of the face.

Discovering biased features for NORPs. We next expand the analysis that we performed in Table 16, to faces detected to be NORPs by our classifier described in Section 5.2.6. We predict facial attribute and expression values on each face and discover features which significantly differ in their portrayal across left/right. We found many NORPs had features which significantly differed. We thus show a condensed version of the table for a subset of NORPs which were most commonly detected in our dataset. In Table 17, we show the median p-values of the features (we found that the average p-value was too strongly influenced by several features with large p-values). We highlight significant differences ($p \leq 0.05$) in bold. We see that "African American", "Arab" and "Asian" all have median p-values which are significant, indicating at least half of the features significantly differ across the left and the right. We observe that the most significant differences occur with "African American". We note that the least significant p-value observed occurs for the "White" category, which implies this category's portrayal is most uniform across left/right. This result shows that groups of people, primarily a number of minority groups, are portrayed in significantly different ways in left vs. right media sources. While this result is to be expected based on what we know about media bias, it does quantitatively demonstrate a problem that needs to be tackled by media content creators, search engine designers and machine learning researchers, and society more broadly.

5.3.5 Bias in text

In this section, we extend our analysis of political bias to the text paired with each image, without considering the image. We observe in Table 10 that the Text model is highly accurate at predicting bias, suggesting that the text contains a highly discriminative signal. We thus wish to understand precisely how the text is biased, both in terms of disparities in the frequency in which certain subject matter is discussed, as well as the choice of words to discuss those subjects. We first consider what political figures are mentioned disproportionately on each side of the political spectrum. We then consider the use of language by each side known to be biased from prior research.

Public figures with disproportionate mentions in text. In Section 5.2.6, we described

Left	**Right**
Richard Spencer	Brett Kavanaugh
Milo Yiannopoulos	Justin Trudeau
Scott Pruitt	Jesus Christ
Michael Flynn	Nancy Pelosi
Alex Jones	George Soros
Karl Marx	Joe Biden
Richard Bertrand Spencer	Rush Limbaugh
Moon Jae In	Barack Obama
Colin Kaepernick	Pope Francis
Steve Bannon	Al Gore
Jared Kushner	Jeremy Corbyn
Betsy Devos	Bill Clinton
Adolf Hitler	Ronald Reagan
Michael Cohen	Chuck Schumer
Doug Jones	Ron Paul

Table 18: Top-15 names across the left/right which were mentioned most on one side, relative to the other side.

how we performed named entity recognition on our text dataset and discovered frequently mentioned names which we then used to train a face recognition model. We also wanted to discover what names were lopsided in their frequency of occurrence on each side of the spectrum. We counted the number of occurrences for each name on the left vs. the right. Because of data imbalance between the left and the right, we normalized the number of occurrences of a name on each side by the total number of names mentioned on that side. In Table 18, we show the names with the largest difference between sides. We observe extreme and polarizing figures are mentioned significantly more disproportionately, e.g. Richard Spencer,

Left										
report	people	thing	work	king	way	very	white	right	try	revolution
say	movement	fascist	lack	fight	struggle	act	content	world	system	start
need	march	see	happen	comment	rights	know	make	write	national	society
film	racist	different	country	live	support	war	win	take	regime	justice
post	human	social	pat	article	action	violence	call	nationalist	quality	point

Right										
tax	end	child	year	use	law	government	conservative	lie	state	liberal
migration	left	life	illegal	policy	form	increase	abortion	provide	cost	author
pass	school	free	rate	serve	claim	believe	man	business	day	terrorist
new	vote	aim	old	order	prove	heart	individual	formation	church	economic
come	result	marriage	term	former	religious	faith	service	far	case	fact

Table 19: Most disproportionately used known biased words from [290] by the left/right.

Alex Jones, Adolf Hitler are mentioned much more on the left vs. the right. In contrast Brett Kavanaugh, Rush Limbaugh, Bill Clinton, etc. are mentioned more on the right relative to the left. We believe these results are sensible. For example, biased sources on the left may attempt to smear the right with Richard Spencer (a neo-Nazi), Milo Yiannopoulos (an alt-right figure), and Alex Jones (a conspiracy theorist). Discussing these figures disproportionately on one side suggests that relatively obscure public figures are being overemphasized for potentially politically biased reasons. Similarly, the right more frequently mentions Brett Kavanaugh (a Supreme Court justice accused of sexual misconduct) and George Soros (a large donor to political causes on the left). In the case of Kavanaugh, the right sources were likely trying to rally support behind his nomination to the Supreme Court. Right sources have also frequently attacked Soros's funding of leftist political causes with conspiracy theories [357]. Overall, Table 18 gives us a broad view of the political figures being discussed in our dataset and reveals which figures either appeal to each party's base or which galvanize the readership against the other party.

Imbalanced biased word usage. [290] studies the problem of detecting bias in text. The authors consider edits to Wikipedia made to remove biased language and develop a lexicon of words which suggest a biased or non-neutral point of view (e.g. McMansion vs. large home, murder vs. kill, pro-life vs. anti-abortion, etc.). We counted the number of times words appearing in the biased word lexicon were used by both the left and the right in our dataset. We then show the "biased" words that are used most by one political side relative to the other in Table 19. The most skewed words across the left/right were "report" (for the left), most likely in connection to the Mueller Report, while "tax" is used most by the right. On the left, we observe words which indicate potentially biased characterizations, e.g. revolution, movement, struggle, fascist, racist, nationalist, etc. On the right, we observe biased language about a different set of issues, e.g. lie, migration, illegal, abortion, terrorist, etc. Collectively, our results presented in Table 18-19 reveal interesting dynamics across biased sources from the left / right. It appears that each side of the political spectrum has a set of "hot-button" issues which they use to either galvanize their audience for their cause or which they use to attack the opposite side.

5.3.6 Exploring the relationship between image and text

Our method for predicting political bias leverages the text paired with articles to guide the training of our purely visual model. We thus seek to better understand the relationship between our images and text. We first discover and illustrate words whose visual representation is most consistent throughout our dataset. We then further examine the Words method (described above in Section 5.3.1), which directly predicts words from an image and discover which words the model is able to predict best. Finally, we study whether we can directly model the complex relationship between images and text within our dataset by training a model to predict whether a given image-text pair is properly aligned.

Modeling word-level visual consistency. We have argued that one of the challenges of modeling political bias in images is that the relationship between images and semantic topics and text is highly complex. For example, an image of the White House could be paired with an article about immigration or one about the US-Afghanistan war. Thus, unlike traditional image captioning tasks where the text directly describes the literal content of the image, the visual grounding of the text in the image is non-literal and consequently more challenging for a model to grasp. However, because we are exploiting our text to guide training of our

fox	cnn	gop	host	republican	republicans	donald	candidate	senate
clinton	hillary	democrats	trump	presidential	secretary	barack	interview	attorney
democratic	president	conservative	obama	immigration	liberal	committee	speech	campaign
election	house	twitter	congress	immigrants	party	leader	vote	bill
executive	racist	meeting	abortion	prime	george	paul	asked	white
conference	debate	minister	press	administration	chief	calling	john	washington

Table 20: Most visually consistent words in our data, in decreasing order (left to right, top to bottom). See text for details.

models, we wished to discover the most "visually consistent" words in our dataset, that is, words whose visual expression in the image is most consistent across different images. Our goal is similar to [132], which seeks to model the visual concreteness of topics within multimodal data, but our approach for quantifying visual consistency (described below) is different given the unique nature of our dataset. To discover the most visually consistent words in our dataset, we first performed tokenization using Spacy [136]. Then, for each word, we created a list of images in which that word appeared in the first two sentences. Next, again for each word, we performed $k-$means clustering, with $k = 5$, which we determined worked well empirically. The intuition behind performing $k-$means in our case is that many words may appear visually inconsistent if one simply takes the average distance between all pairs of images for a given word, because their visual grounding could be multimodal. We compute the visual consistency v for word w as:

$$v_w = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} \|x_i^j - c_j\|^2}{\sum_{j=1}^{k} n_j} \tag{17}$$

where k is the number of clusters, n_j is the number of images in cluster j, x_i^j are image features which have been assigned to cluster j by k-means and c_j is the centroid of cluster j. Equation 17 essentially measures how tightly the visual features for a given word fit the 5-modal distribution induced by our clustering. We computed the above metric for the 10,000 most common words in our dataset.

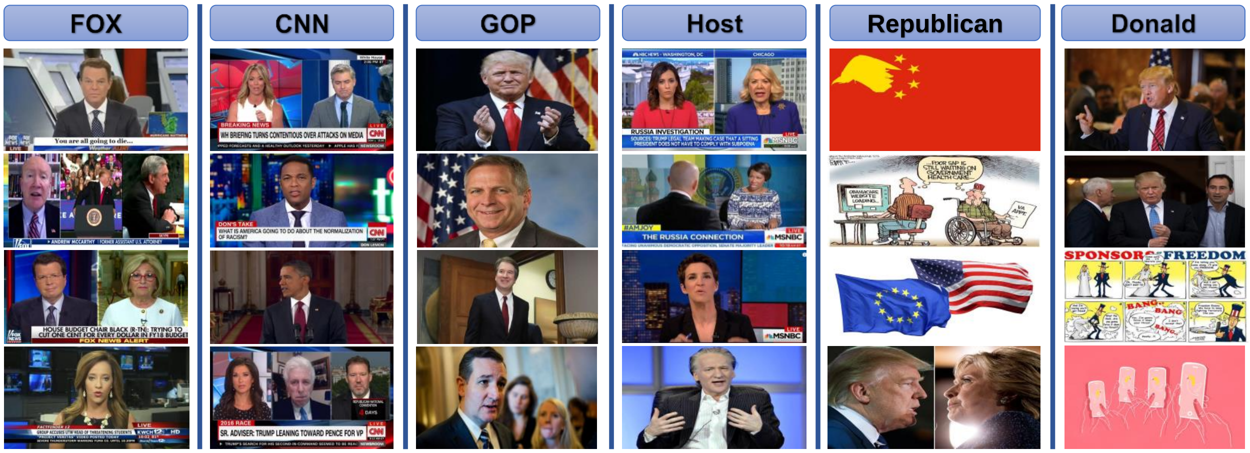


Figure 18: Most visually consistent words and image examples from the two tightest visual clusters computed for each word.

We show the most visually consistent words in our dataset in Table 20. We observe that news organizations, political groups, and candidates dominate: FOX, CNN, GOP, Republican, Donald, Clinton, Trump, Barack, etc. We observe that several political topics also emerge as visually consistent, e.g. immigration, immigrants, abortion. We next wanted to see what did our visually consistent images for each word actually look like. For six of our most visually consistent words, we sampled images from the top-2 tightest visual clusters computed for each word and present them in Figure 18. We observe that three of the top four (FOX, CNN, host) most consistent words primarily feature images of people on newscasts. The model has placed these images closest together in the learned space, most likely because the images feature similar visual content. We note that the most visually similar images are not necessarily semantically similar, as the news broadcasts are presenting a variety of unrelated topics. The tightest clusters for the word "GOP" feature portrait shots of political figures on the right (from top to bottom: Donald Trump, Mike Pompeo, Brett Kavanaugh, and Ted Cruz). For the last two words ("Republican" and "Donald"), we observe that the model has placed cartoons and illustrations closest together, in addition to clusters of political figures.

Word	$\mathbf{F_1}$ score	Word	$\mathbf{F_1}$ score	Word	$\mathbf{F_1}$ score	Word	$\mathbf{F_1}$ score	Word	$\mathbf{F_1}$ score	Word	$\mathbf{F_1}$ score
trump	0.590	abortion	0.534	immigration	0.497	president	0.434	hillary	0.430	gay	0.429
donald	0.413	clinton	0.378	immigrants	0.357	supreme	0.341	obama	0.317	republican	0.309
news	0.302	fox	0.302	party	0.299	republicans	0.295	racist	0.274	presidential	0.273
democratic	0.272	media	0.263	candidate	0.259	bill	0.257	white	0.252	illegal	0.252
election	0.248	conservative	0.246	justice	0.244	democrats	0.240	campaign	0.238	senate	0.232
tuesday	0.225	speech	0.223	deal	0.222	administration	0.218	house	0.217	debate	0.216
paul	0.211	vote	0.204	foreign	0.203	political	0.200	minister	0.199	washington	0.191
thursday	0.189	conference	0.188	voters	0.187	meeting	0.186	twitter	0.183	night	0.181
cnn	0.171	prime	0.169	congress	0.168	barack	0.162	host	0.157	committee	0.157

Table 21: Per-word F_1 scores of a model trained to predict whether each word is/is not present in the image's article given the image and text embedding. We consider a dictionary of the top-1000 most visually consistent words and show the performance of the model on the best-performing words below.

Predicting visually-consistent words from images. We have now discovered the words in our dataset that have the most consistent visual expression across images. We next wanted to see how well a model could exploit this word-level visual consistency. Note that we previously used the predictions from this model to train a word-based politics predictor in Table 10 (the Words model). We show the F_1 score of predicting words from an image on our test set in Table 21. We choose F_1 score because multiple words can be paired with each image. We observe that our model performs better at predicting visually consistent words on average vs. non-consistent words. We observe numerous words which appeared in Table 20 have relatively higher F_1 scores relative to other words, with all the highest scoring words appearing in the table as being visually consistent. For example, we see "president" : 0.434, "trump" : 0.590, "donald" : 0.413, "immigration" : 0.497, and "abortion" : 0.534. However, we observe that the visual consistency of images associated with a single word does not guarantee discriminativity. In other words, just because images associated with a word all share similar visual content, does not imply that all images with that type of visual content are exclusively associated with that particular word. For example, we observe relatively poor performance at predicting the word "CNN" and "FOX", even though these words

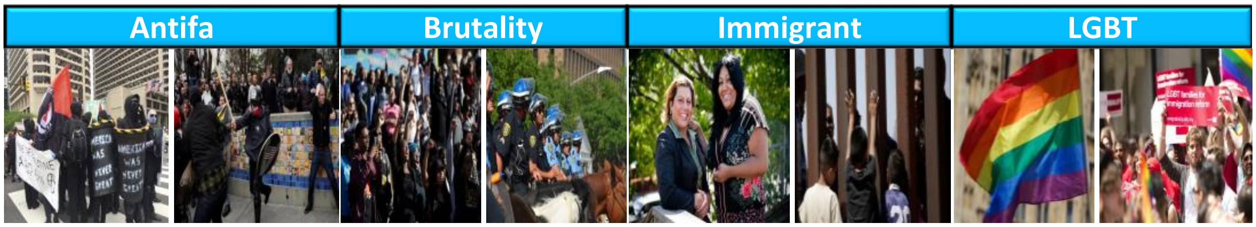


Figure 19: We train a model to predict words from images. The model learns relevant visual cues for each word, demonstrating the utility of exploiting text, even for purely visual classification.

have visually consistent images. This is likely because the model has trouble differentiating between many different news programs, given their similar visual content. That is, the model may recognize a newsanchor at a desk, but then become confused as to whether the image is from CNN, FOX, MSNBC, ABC, etc.

High-response images for visually consistent words. We next visualize what images our model responded strongest to for various words as a way of understanding what it learned. In Fig. 19, we show examples of images that were among the top-100 strongest predictions for that word. We observe, for example, that the model strongly predicts "antifa" for black-clad protesters and protesters holding banners. The model predicts "brutality" for images with African American protesters and for police scenes. The model predicts the word "immigrant" for images containing a border wall and Hispanic individuals, and "LGBT" for pride flags and rainbow like banners.

5.3.7 Visual variability across political topics

Each image in our dataset is also labeled with the political topic (e.g. abortion) that the media source was queried with when the image was scraped. We have seen initial results in Tables 12 and 19 revealing that different subjects are mentioned disproportionately across the left / right, suggesting that the topic of the image may be a useful cue for bias prediction. We now seek to further explore the topic annotations on our dataset. We first present

Topic	**F$_1$**	**Topic**	**F$_1$**
Abortion	0.688	ISIS	0.555
Animal Rights	0.540	LGBT	0.540
Black Lives Matter	0.426	Minimum Wage	0.504
Blue Lives Matter	0.053	Racism	0.526
Border Security	0.465	Religion	0.547
Climate Change	0.480	Terrorism	0.544
Fracking	0.455	Unemployment	0.511
Gun Control	0.627	Vaccines	0.596
Homelessness	0.527	War On Drugs	0.545
Immigration	0.578	Welfare	0.192
Average		0.534	

Table 22: F$_1$ score of predicting the political topics of an image-text pair on human annotations. Note that the same image-text pair can be labeled with multiple issues.

results on predicting the political topic of an image. We then discover topics which are most visually consistent in their portrayal across images. Finally, we present results illustrating the difficulty of classifying images as left/right, by showing images which are closest in visual space from each political side within each topic.

Predicting political topics from images. We trained a model to predict the weak political topic label for each image in our training set (assuming each image exemplifies the topic of the parent article), given the image and the document embedding of the text. To ensure that the weakly supervised topic labels were actually capturing the real political issue of the images (rather than dataset harvesting artifacts), we evaluated our model on our set of human annotated data. Each image can be labeled as being related to multiple topics, so we compute F$_1$ score rather than accuracy. We present the results in Table 22. We find that our model is able to predict most topics fairly accurately. For example, we observe that our model is most accurate at predicting images of "abortion" and "gun control". This makes sense because images about these topics share common scenes and objects: images about

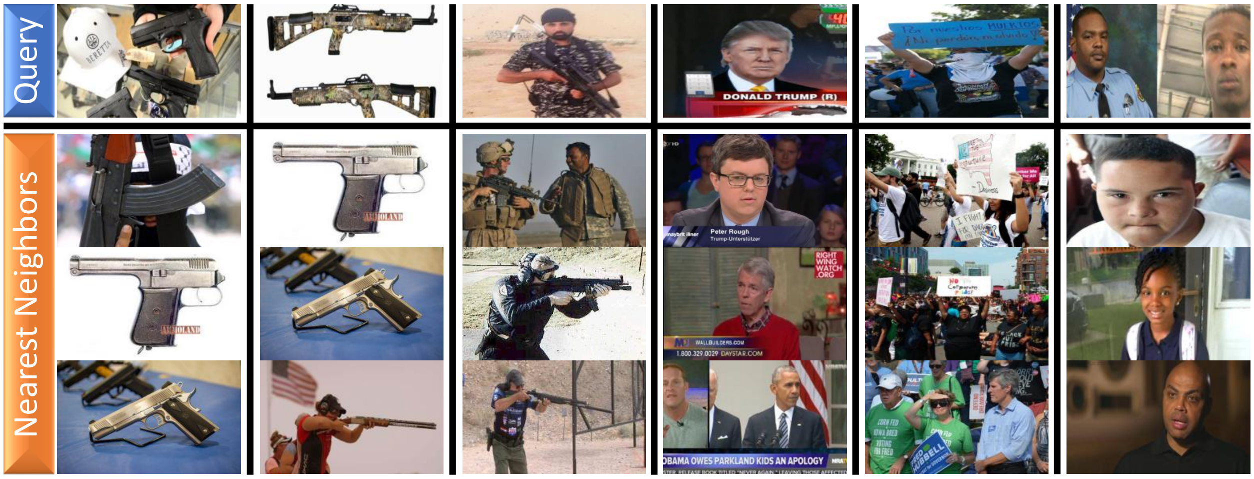


Figure 20: Images where neighbors in visual space are most consistent in terms of their topic. We show that some topics (e.g. gun control) have a consistent visual expression, while other topics are less visually cohesive.

abortion often feature protest scenes or images of babies, while gun control images often feature firearms. We find our model performs worst at predicting "blue lives matter". This is likely more an artifact of our annotations since few annotators picked "blue lives matter" and instead picked the more well-known "black lives matter."

Visually consistent topics. We next analyze which topics had the most consistent purely visual expression i.e. without considering the text. We computed the 20 nearest neighbors in visual space for several hundred randomly chosen images from our dataset. We then computed the entropy of the topic distribution of the retrieved neighboring images and sorted the results in order of increasing entropy. We show the result in Figure 20, with the first row showing the query image and the next three rows showing the top-3 closest images to the query in visual space. We see the left two columns all feature feature firearms. The retrieved neighbors in the first two columns are extremely consistent in their topic annotations and are almost all labeled "gun control". The third column also features military / law enforcement holding firearms, but are much more diffuse in terms of the neighbors' topics (e.g. ISIS, foreign policy, terrorism, etc.). The queries and their neighbors to the right are even more diffuse in terms of topics, e.g. the protest images (second to last column) all feature protests (or political rallies), but are about a number of disparate topics from welfare to immigration, even though they are close in visual space. Thus, predicting the political topic of an image is complex in that it requires not only recognizing the objects and scene type of an image (e.g. protest), but actually reasoning how the objects and individuals relate in more nuanced ways.

5.4 DISCUSSION

In this chapter, we introduced the problem of modeling a multimodal latent visual concept: political bias. We assembled a large dataset of biased images and paired articles and presented a weakly supervised approach for inferring the political bias of images. Our method leverages the image's paired text to guide the model's training process towards relevant visual semantics in a way which ultimately improves bias classification. Specifically, our method leverages a form of guided training, externally injected semantics, and multi-stage training to facilitate learning a high-level latent visual concept. We demonstrate the contribution of our method and dataset both quantitatively and qualitatively, including on a large crowdsourced dataset. We provide numerous qualitative examples illustrating the *types* of bias found in our dataset, including a generative result which transforms faces across the political spectrum. We performed a detailed experimental analysis demonstrating how bias in the media is expressed both visually and textually. Collectively, our results demonstrate that political bias is exhibited not just in text, but visually as well. We show that by exploiting unique aspects of the image-text relationship, the text domain can guide purely visual classifiers in order to improve visual inference (**H2**). This method works by exploiting the relationship between image and text in real-world media (**H3**). Our analysis reveals major differences in terms of visual portrayals of objects in communicative multimedia compared to conventional image datasets, further confirming our first hypothesis (**H1**).

Our work in this chapter has several important broader contributions to society. First, we believe studying and recognizing visual bias in images is an important step in building socially-informed machine learning systems. By recognizing how data is biased, researchers can actively work to combat biased portrayals learned by their models. Further, by automatically recognizing biased depictions, we can actively combat bias within the media. Our work can be used to expose and make users aware of discrimination and stereotypical portrayals of individuals or groups. For example, one possible solution is to automatically flag content for users so that they can become more informed that the perspective they are being presented with is non-neutral. Similarly, our work can be used to quantify not only that media sources (through the images they publish) are biased, but the types of bias that each media sources tend to purvey. Our work thus has implications for social media companies which may seek to prevent the spread of discriminatory content on their platforms. By revealing bias within content presented to users, we ultimately hope to help both users and publishers become more informed consumers of visual media.

The two subsequent chapters make further contributions in modeling abstract semantic concepts in real-world multimedia. However, unlike our work in this chapter (and the previous chapters), the methods presented therein are *task-agnostic*, in that they seek to capture high-level semantics in feature representations irrespective of any particular application (e.g. photographic style or political bias modeling). Thus, the following two chapters build upon the problem-specific intuitions developed in this chapter in order to advance *general-purpose* methods for modeling abstract semantics in real-world multimedia.

6.0 LEARNING SEMANTICALLY ROBUST EMBEDDINGS IN MULTIMEDIA THROUGH CROSS-MODAL COMPLEMENTARITY PRESERVING CONSTRAINTS

Summary. *In this chapter, we present a method for learning general-purpose representations of abstract semantics in communicative multimedia.*[1] *The abundance of multimedia (e.g. social media posts with text and images) have inspired interest in methods for learning joint representations where semantic concepts and their relationships from each modality are captured within a learned space. However, most prior methods have focused on the case where image and text convey redundant information; in contrast, real-world image-text pairs convey complementary information with little communicative overlap. Popular approaches to for learning shared semantic spaces rely on a variety of metric learning losses, which prescribe what the proximity of image and text should be, in the learned space. However, images in communicative multimedia such as news articles portray topics in a visually diverse fashion; thus, we need to take special care to ensure a meaningful image representation. We propose novel within-modality losses which ensure that not only are paired images and texts close, but the expected image-image and text-text relationships are also observed. Specifically, our method encourages semantic coherency in both the text and image subspaces, and improves the results of cross-modal retrieval in three challenging scenarios.*

6.1 INTRODUCTION

All of our prior work in this dissertation has focused on modeling particular types of abstract, latent visual concepts. Specifically, in Chapter 3, we studied the problem of modeling photographic style. Then, in Chapter 4, we modeled visual persuasion in image ads and learned facial differences across ad types. Next, in Chapter 5 we presented a method for

[1] The work presented in this chapter was published in our ECCV 2020 paper, "Preserving Semantic Neighborhoods for Robust Cross-modal Retrieval" [342].

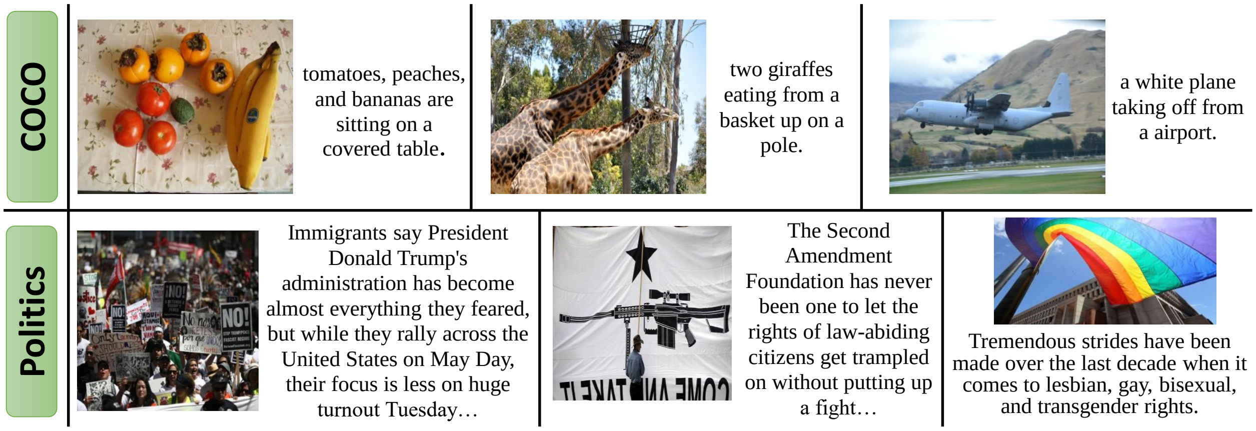

COCO
tomatoes, peaches, and bananas are sitting on a covered table.
two giraffes eating from a basket up on a pole.
a white plane taking off from a airport.
Politics
Immigrants say President Donald Trump's administration has become almost everything they feared, but while they rally across the United States on May Day, their focus is less on huge turnout Tuesday…
The Second Amendment Foundation has never been one to let the rights of law-abiding citizens get trampled on without putting up a fight…
Tremendous strides have been made over the last decade when it comes to lesbian, gay, bisexual, and transgender rights.

Figure 21: Image-text pairs from COCO [211] and Politics [341]. Traditional image captions (top) are descriptive of the image, while we focus on the more challenging problem of aligning images and text with a non-literal complementary relationship (bottom).

modeling a multimodal semantic concept, political bias, in multimedia news articles. Our work in this (and the following) chapter differs from all our previous work in that rather than study a particular abstract task, we propose methods for learning *task-agnostic* multimodal semantic representations. By task-agnostic we mean that rather than carefully tuning an approach for a particular problem, we instead propose methods which attempt to learn representations of images and text which capture nuanced semantic concepts in the absence of any target application (e.g. predicting political bias). We desire our methods to learn which semantics to preserve in a purely-data driven manner (i.e. automatically from the dataset without human intervention or guidance). To do so, the method in this chapter as well as in Chapter 7 exploit the idiosyncratic relationships between image and text uniquely found in communicative multimedia in different ways.

Vision-language tasks such as image captioning [394, 8, 226] and cross-modal generation and retrieval [293, 402, 411] have seen increased interest in recent years. At the core of methods in this space are techniques to bring together images and their corresponding pieces of text. However, most existing cross-modal retrieval methods only work on data where the two modalities (images and text) are well aligned, and provide fairly redundant information.

As shown in Fig. 21, captioning datasets such as COCO contain samples where the overlap between images and text is significant (both image and text mention or show the same objects). In contrast, real-world news articles contain image and text pairs that cover the same topic, but show complementary information (protest signs vs information about the specific event; guns vs discussion of rights; rainbow flag vs LGBT rights). While a human viewer can still guess which images go with which text, the alignment between image and text is abstract and symbolic. Thus, in captioning, cross-modal retrieval means finding the manifestation of a single concept in two modalities (e.g. learning embeddings such that the word "banana" and the pixels for "banana" project closeby in a learned space). In cross-modal retrieval on *news articles*, images are ambiguous *in isolation* (see Fig. 22), so we must first resolve any ambiguities in the image, and figure out "what it means".

We propose a metric learning approach where we use the semantic relationships between text segments, to guide the embedding learned for corresponding images. Our approach grounds the representation of the image in the meaning of corresponding text. In other words, to understand what an image shows, we have to look at what articles it appeared with. Unlike prior approaches, we wish to capture this information not only across modalities, but within the image modality itself, through new within-modality losses.

If texts y_i and y_j are semantically similar, we learn an embedding where we explicitly encourage their paired images x_i and x_j to be similar, using an additional unimodal loss. Note that in general x_i and x_j need not be similar in the original visual space. We show this in Fig. 22 where an image might be chosen to illustrate multiple related texts (shown in green), and each text in turn could be illustrated with multiple visually distant images (e.g. the four images on the right-hand side could appear with the border wall text). In addition, we encourage texts y_i and y_j, who were close in the unimodal space, to remain close.

Our novel loss formulation explicitly encourages *within-modality semantic coherence.* Fig. 23 shows the effect. On the left, we show the proximity of samples before cross-modal learning; specifically, while two texts are close in the document space, their paired articles may be far from the texts. In the middle, we show the effect of using a standard triplet loss, which pulls image-text pairs close, but does not necessarily preserve the similarity of related articles; they are now further than they used to be in the original space. In contrast, on the

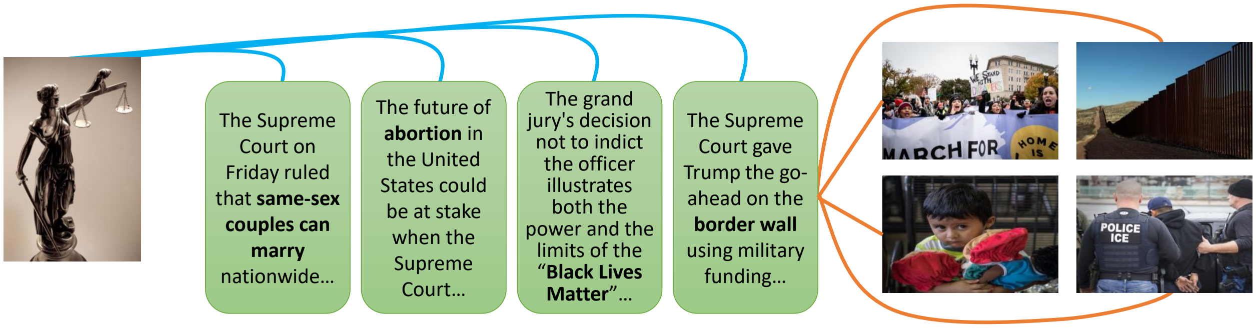

The Supreme Court on Friday ruled that **same-sex couples can marry** nationwide…
The future of **abortion** in the United States could be at stake when the Supreme Court…
The grand jury's decision not to indict the officer illustrates both the power and the limits of the **"Black Lives Matter"**…
The Supreme Court gave Trump the go-ahead on the **border wall** using military funding…
MARCH FOR
HOME IS
POLICE ICE

Figure 22: The image on the left symbolizes justice and may be paired with text about a variety of subjects (e.g. abortion, same sex marriage). Similarly, the text regarding immigration may be paired with visually dissimilar images. Our approach enforces that *semantically* similar content (e.g. images on right) is close in the learned space. To discover such content, we use semantic neighbors of the text and their paired images.

right, we show how our method brings paired images and text closer, while also preserving a semantically coherent region, i.e. the texts remained close.

In our approach, we use neighborhoods in the original text document space, to compute semantic proximity. We also experiment with an alternative approach where we compute neighborhoods using the visual space, then guide the corresponding texts to be close. This approach is a variant of ours, and is novel in the sense that it uses proximity in one unimodal space, to guide the other space/modality. While unimodal losses based on visual similarity are helpful over a standard cross-modal loss (e.g. triplet loss), our main approach is superior.

Next, we compare to a method [364] which utilizes the *set* of text annotations available for an image in COCO, to perform more robust captioning. We show that when these ground-truth annotations are available, using them to compute neighborhoods in the textual space is the most reliable. However, on many datasets, such sets of annotations (more than one for the same image) are not available. We show that our approach offers a comparable alternative.

Finally, we test our approach using PVSE [328], a state-of-the-art visual semantic embedding model. We show that our proposed loss further improves the performance of this

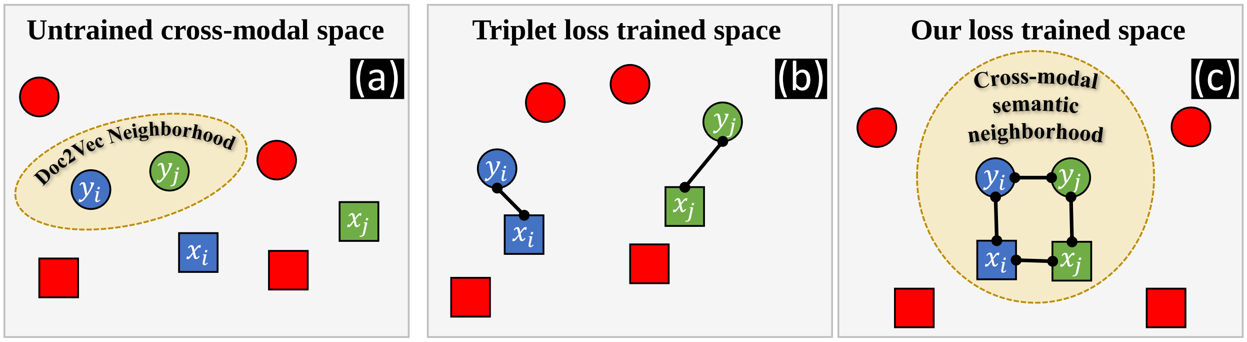


Figure 23: We show how our method enforces cross-modal semantic coherence. Circles represent text and squares images. In (a), we show the untrained cross-modal space. Note y_i and y_j are neighbors in Doc2Vec space and thus semantically similar. (b) shows the space after triplet loss training. y_i and x_i, and y_j and x_j, are now close as desired, but y_i and y_j have moved apart and x_i and x_j remain distant. (c) shows our loss's effect. Now, all semantic neighbors (both images and text) are pulled closer.

model.

To summarize, our contributions are as follows.

- We preserve relationships in the original semantic space. Because images do not clearly capture semantics, we explicitly use the semantic space to guide the image representation through a unimodal (within-modality) loss.
- We perform detailed experimental analysis of our proposed loss function, including ablations, on four recent large-scale image-text datasets. One [27] contains multimodal articles from New York Times, and another contains articles from far-left/right media [341]. We also conduct experiments on [316, 211]. Our approach significantly improves the state-of-the-art in most cases.
- We tackle a new cross-modal retrieval problem where the visual space is much less concrete. This scenario is quite practical, and has applications ranging from automatic caption generation for news images, to detection of fake multimodal articles (i.e. detecting whether an image supports the text).

6.2 METHOD

Consider two image-text pairs, $\{x_i, y_i\}$ and $\{x_j, y_j\}$. To ground the "meaning" of the images, we use proximity in a generic, pre-trained textual space between the texts y_i and y_j. If y_i and y_j are semantically close, we expect that they will also be relatively close in the learned space, and further, that x_i and x_j will be close also. We observed that, while intuitive, this expectation does not actually hold in the learned cross-modal space. No part of the common cross-modal losses enforces such "pull" within modalities, so the embeddings for semantically related images may be pulled in different directions. The problem becomes more severe when image and paired text do not exhibit literal alignment, as shown in Fig. 21, because images paired via text neighbors could be visually different, as is the case with news articles with more lengthy text, social media posts, or in artistic media such as posters with slogans. For example, an image of the U.S. border wall may be paired with text about immigration, while another image also about immigration shows an immigrant child at the border. Thus, even though the two images are fundamentally semantically related to immigration, their embeddings will only be brought close if their paired texts are close in the learned space. Our proposed method augments standard metric learning losses to preserve intra-modal semantic similarity. A graphical illustration of how our approach differs from standard metric learning losses is shown in Fig. 24. First, we provide the problem formulation, describe how several common existing loss functions tackle the problem, and discuss their limitations. Second, to address this issue, we propose two constraints which pull within-modality semantic neighbors close to each other within the manifold.

6.2.1 Problem formulation and existing approaches

We assume a dataset $\boldsymbol{\mathcal{D}} = \{\mathbf{I}, \mathbf{T}\}$ of n image-text pairs, where $\mathbf{I} = \{x_1, x_2, \ldots, x_n\}$ and $\mathbf{T} = \{y_1, y_2, \ldots, y_n\}$ denote the set of paired images and text, respectively. By pairs, we mean y_i is text related to or co-occurs with image x_i. Let f_I denote a convolutional neural network which projects images into the joint space and f_T a recurrent network which projects text. For brevity, in the remainder of this chapter, we use the notational shorthand $f_T(y) = y$ and

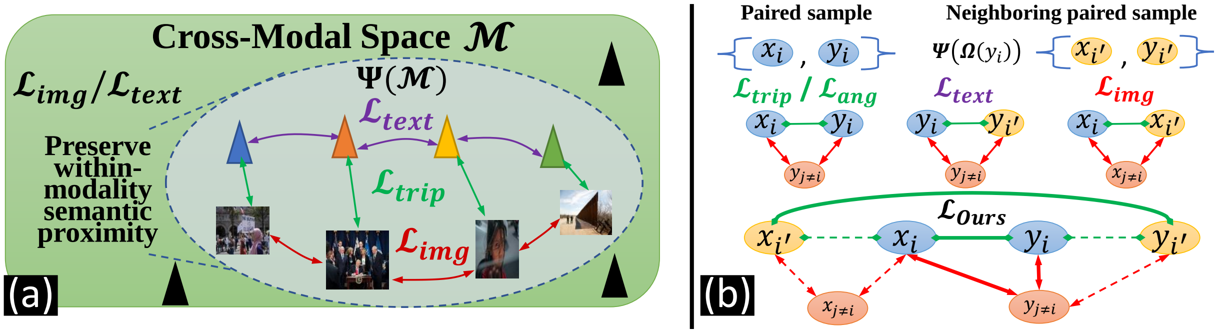


Figure 24: (a): $\mathcal{L}_{text}$ and $\mathcal{L}_{img}$ pull semantic neighbors of the same modality closer. Note the images are visually distinct, but semantically similar. (b): Pull connections are shown in green, and push in red. $\mathcal{L}_{trip}$ and $\mathcal{L}_{ang}$ operate cross-modally, but impose no within-modality constraints. $\mathcal{L}_{ours}$ exploits the paired nature of the data to enforce that the expected inter/intra-modal relationships are observed.

$f_I(x) = x$. The goal training both f_I and f_T is to learn a cross-modal manifold $\mathcal{M}$ where semantically similar samples are close. At inference time, we wish to retrieve a ground-truth paired text given an input image, or vice versa.

One common technique for training for this objective has been the triplet loss [309]. It posits that paired samples should be closer to one another than they are to non-paired samples in the metric space. Let $\mathcal{T} = \left(x_i^a, y_i^p, y_j^n\right)$ denote a triplet of samples consisting of an anchor (a), positive (p), and negative (n). Specifically, the image $x_i^a \in \mathbf{I}$ and text $y_i^p \in \mathbf{T}$ are paired samples, while the negative $y_j^n \in \mathbf{T}$ is chosen randomly such that $i \neq j$. The triplet loss $\mathcal{L}_{trip}$ is:

$$\mathcal{L}_{trip}(\mathcal{T}) = \left[\|x_i^a - y_i^p\|_2^2 - \|x_i^a - y_j^n\|_2^2 + m\right]_+ . \tag{18}$$

This loss is perhaps the most common one used in cross-modal retrieval tasks. However, consider the gradient of the triplet wrt. the anchor in Eq. 18. It can be easily seen that $\frac{\partial \mathcal{L}_{trip}}{\partial x_i^a} = 2\left(y_j^n - y_i^p\right)$. Thus, the gradient only depends on the other two components, but not their overall relationship. This is true for all three components of the loss, and allows for degenerate cases where the anchor and positive become further apart in order to increase the distance between positive and negative. Thus, angular loss $\mathcal{L}_{ang}$ [363] adds a third-order constraint which accounts for the angular relationship of all three points to the negative point:

$$\mathcal{L}_{ang}\left(\mathcal{T}\right) = \left[\|x_i^a - y_i^p\|_2^2 - 4\tan^2\alpha\|y_j^n - \mathcal{C}_i\|_2^2\right]_+ \,, \tag{19}$$

where $\mathcal{C}_i = \left(x_i^a + y_i^p\right)/2$ is the center of a circle around anchor and positive.

One challenging aspect of these losses is choosing a good negative term in the triplet. If the negative is chosen to be too far from the anchor, the loss becomes 0 and no learning occurs. In contrast, if negatives are chosen too close, the model may have difficulty converging to a reasonable solution as it continuously tries to move samples to avoid overlap with the negatives. How to best sample triplets to avoid these issues is an active area of research [77]. One recent technique, the N-pairs loss [326], proposes that instead of a single negative sample being used, all negatives within the minibatch of triplets used to train the model should be used. The N-pairs loss, denoted $\mathcal{L}_{ang}^{NP}$, thus pushes the anchor and positive embedding away from *multiple* negatives simultaneously:

$$\mathcal{L}_{ang}^{NP}\left(\mathcal{T}\right) = \sum_{\forall y^n \in \text{minibatch}} \mathcal{L}_{ang}\left(x_i^a, y_i^p, y_j^n\right) \,. \tag{20}$$

The symmetric constraint [417] can also be added to explicitly account for bidirectional retrieval, i.e. text-to-image, by swapping the role of images and text to form symmetric triplets $\mathcal{T}_{sym} = \left(y_i^a, x_i^p, x_i^n\right)$:

$$\mathcal{L}_{ang}^{NP+SYM}\left(\mathcal{T}, \mathcal{T}_{sym}\right) = \mathcal{L}_{ang}^{NP}\left(\mathcal{T}\right) + \mathcal{L}_{ang}^{NP}\left(\mathcal{T}_{sym}\right) \,. \tag{21}$$

Limitations. While these loss functions have been used for cross-modal retrieval, they do not take advantage of several unique aspects of the multi-modal setting. Only the thick/solid pull/push connections in the bottom of Fig. 24 (right) are part of a triplet/angular loss application. The thinner, dashed connections are intuitive, but only enforced in our novel formulation. We argue the lack of explicit *within-modality* constraints allows discontinuities within the space for semantically related content from the same modality.

6.2.2 Our proposed loss

The text domain provides a semantic fingerprint for the image-text pair, since vastly dissimilar visual content may still be semantically related (e.g. image of White house, image of protest), while similar visual content (e.g. crowd in church, crowd at mall) could be semantically unrelated. We thus use the text domain to constrain within-modality semantic locality for both images and text.

To measure ground-truth semantic similarity, we pretrain a Doc2Vec [192] model Ω on the train set of text. Specifically, let T denote the number of words in article y_i, $p(\cdot)$ be the probability of the given word, w_t represent the embedding learned for word t of article y_i, d be the document embedding of y_i, and k denote the look-around window. Ω learns word embeddings and document embeddings which maximize the average log probability: $\frac{1}{T}\sum_{t=1}^{T}\log p\left(w_t|d, w_{t-k}, \ldots, w_{t+k}\right)$. After training Ω, we use iterative backpropagation to compute the document embedding which maximizes the log probability for every article in the dataset: $\Omega(\mathbf{T}) = \left\{\Omega\left(y_1\right), \ldots, \Omega\left(y_n\right)\right\}$.

Because Doc2Vec has been shown to capture latent topics within text documents well [257], we seek to enforce that locality originally captured in $\Omega(\mathbf{T})$'s space also be preserved in the cross-modal space $\mathcal{M}$. Let

$$\Psi\left(\Omega(y_i)\right) = \langle x_{i'}, y_{i'}\rangle \tag{22}$$

denote a nearest neighbor function over $\Omega(\mathbf{T})$, where $\langle\cdot,\cdot\rangle$ is an image-text pair in the train set randomly sampled from the $k = 200$ nearest neighbors to y_i, and $i \neq i'$. $\Psi\left(\Omega(y_i)\right)$ thus returns an image-text pair semantically related to y_i.

We formulate two loss functions to enforce within-modality semantic locality in $\mathcal{M}$. The first, $\mathcal{L}_{text}$, enforces locality of the text's projections:

$$\begin{gathered}\mathcal{T}'_{text} = \left(y_i^a, y_{i'}^p, y_j^n\right)\ , \\ \mathcal{L}_{text}\left(\mathcal{T}'_{text}\right) = \mathcal{L}_{ang}\left(\mathcal{T}'_{text}\right)\ , \\ \mathcal{L}_{ang}\left(\mathcal{T}'_{text}\right) = \left[\|y_i^a - y_{i'}^p\|_2^2 - 4\tan^2\alpha\|y_j^n - \mathcal{C}_i\|_2^2\right]_+\ ,\end{gathered} \tag{23}$$

where y_j^n is the negative sample chosen randomly such that $i \neq j$ and $\mathcal{C}_i = (y_i^a + y_i^p)/2$. $\mathcal{L}_{text}$ is the most straightforward transfer of semantics from $\Omega(\mathbf{T})$'s space to the joint space, as it seeks to preserve nearest neighbors in Ω remain close in $\mathcal{M}$.

$\mathcal{L}_{text}$ also indirectly causes semantically related images to move closer in $\mathcal{M}$ as their associated text embeddings move closer. See Fig. 24 (right). This serves as a weak constraint on image semantic locality in $\mathcal{M}$, i.e. there is now a weak connection between x_i and x_i' through the now-connected y_i and y_i'. To directly ensure smoothness and semantic coherence between x_i and x_i', we propose a second constraint, $\mathcal{L}_{img}$:

$$\begin{aligned} \mathcal{T}'_{img} &= \left(x_i^a, x_{i'}^p, x_j^n\right) \ , \\ \mathcal{L}_{img}\left(\mathcal{T}'_{img}\right) &= \mathcal{L}_{ang}\left(\mathcal{T}'_{img}\right) \ , \\ \mathcal{L}_{ang}\left(\mathcal{T}'_{img}\right) &= \left[\|x_i^a - x_{i'}^p\|_2^2 - 4\tan^2\alpha\|x_j^n - \mathcal{C}_i\|_2^2\right]_+ \ , \end{aligned} \tag{24}$$

where x_j^n is the randomly chosen negative sample such that $i \neq j$ and $\mathcal{C}_i = (x_i^a + x_i^p)/2$. Note that x_i and $x_{i'}$ are often not going to be neighbors in the original visual space. We use N-pairs over all terms to maximize discriminativity, and symmetric loss to ensure robust bidirectional retrieval:

$$\begin{aligned} &\mathcal{L}_{ang}^{OURS}\left(\mathcal{T}, \mathcal{T}_{sym}, \mathcal{T}'_{text}, \mathcal{T}'_{img}\right) = \\ &\mathcal{L}_{ang}^{NP+SYM}\left(\mathcal{T}, \mathcal{T}_{sym}\right) + \alpha\mathcal{L}_{text}^{NP}\left(\mathcal{T}'_{text}\right) + \beta\mathcal{L}_{img}^{NP}\left(\mathcal{T}'_{img}\right) \ , \end{aligned} \tag{25}$$

where α and β are hyperparameters controlling the relative importance of the text/image semantic constraints.

Second variant. We also experiment with a variant of our method where the nearest neighbor function in Eq. 22 (computed in Doc2Vec space) is replaced with one that computes nearest neighbors in the space of visual (e.g. ResNet) features. Now $x_i, x_{i'}$ are neighbors in the original visual space before cross-modal training, and $y_i, y_{i'}$ are their paired articles (which may not be neighbors in the original Doc2Vec space). We denote this method as OURS (Img NNs) in Table 23, and show that while it helps over a simple triplet- or angular-based baseline, it is inferior to our main method variant described above.

Discussion. At a low level, our method combines three angular losses. However, note that our losses in Eq. 23 and Eq. 24 do not exist in the prior literature. While [364] leverages ground-truth neighbors (sets of neighbors provided together for the same image sample in a dataset), we are not aware of prior work that estimates neighbors. Importantly, we are not aware of prior work that uses the text space to construct a loss over the image space, as Eq. 24 does. We show that the choice of space in which semantic coherency is computed is important; doing this in the original textual space is superior than using the original image space. We show the contribution of both of these losses in our experiments.

6.2.3 Implementation details

All methods use a two-stream architecture, with the image stream using a Resnet-50 [127] architecture initialized with ImageNet features, and the text stream using Gated Recurrent Units [52] with hidden state size 512. We use an image size of 224x224 and random horizontal flipping and initialize all non-pretrained learnable weights via Xavier initialization [107]. All text models are initialized with word embeddings of size 200 learned on the dataset on which they are applied. We apply a linear transformation to each model's output features ($\mathbb{R}^{2048\times256}$ for image and $\mathbb{R}^{512\times256}$ for text) to get the final embedding. We perform L_2 normalization on embeddings produced by each model. We train all models using Adam [175] with minibatch size of 64, learning rate 1.0e-4, and weight decay 1e-5. We decay the learning rate by a factor of 0.1 after every 5 epochs of no decrease in validation loss. We use a train-val-test split of 80-10-10 for all datasets. For Doc2Vec, we use [295], with $d \in \mathbb{R}^{200}$ and train using distributed memory [192] for 20 epochs with window $k = 20$, ignoring words that appear less than 20 times. We use hierarchical softmax [248] to compute p. To efficiently compute approximate nearest neighbors for Ψ, we use [231]; our method adds negligible computational overhead as neighbors are computed prior to training. We determined the values for α and β empirically on a held-out val. set; we perform grid search in the range $\{0.1, 0.2, 0.3\}$ for each hyperparameter.

6.3 EXPERIMENTAL EVALUATION

In this section, we evaluate our method as well as four baselines on four recent large-scale public datasets. Our quantitative results consistently demonstrate the superiority of our proposed approach at bidirectional retrieval. We also show that indeed our method better preserves within-modality semantic locality by keeping neighboring images and text closer within the joint space. We show example images and text for which our method preserves semantic locality the most compared with the baseline. Finally, we present cross-view retrieval results showing the images/text retrieved by several methods for given words/images.

6.3.1 Datasets

We evaluate our methods on four large-scale image-text datasets. Two feature especially challenging and indirect relationships between image and text, compared to what is typically seen in standard image captioning datasets:

- **Politics** [341] (Chapter 5) consists of images paired with news articles. In Chapter 5, because we were primarily interested in purely visual predictions, we did not perform deduplication of the text domain (we did deduplicate within the image domain). However, in this and the subsequent chapter, we are interested in both visual and textual learning. We identified that in some cases, multiple images were paired with boilerplate text (website headliner, privacy policy text) due to failed data scraping. We thus also removed duplicates in text space using MinHash [35]. We were left with 246,131 unique image-text pairs. Because the articles are lengthy, we only use the first two sentences of each, which had the highest overlap with a set of sparse human annotations that indicate which piece of the text is most relevant to the image.
- **GoodNews** [27] consists of ∼466k images paired with their captions. All data was harvested from the New York Times. Captions often feature abstract or indirect text in order to relate the image to the article it appeared with. The method in [27] takes image and text as input, hence cannot serve as a baseline.

Both of the above datasets exhibit much longer text paired with images, compared to traditional image captions. For example, the average length of text paired with an image on COCO is 11 words, while the average in GoodNews is 18 words and the average in Politics is 59 words.

We also test on two large-scale standard image captioning datasets, where the relationship between image and text is typically more direct:

- **COCO** [211] is a large dataset containing numerous annotations, such as objects, segmentations, and captions. The dataset contains ∼120k images with captions. Unlike our other datasets, COCO contains more than one caption per image, with each image paired with four to seven captions.
- **Conceptual Captions** [316] is composed of ∼3.3M image-text pairs. The text comes from automatically cleaned alt-text descriptions paired with images harvested from the internet and has been found to represent a much wider variety of style and content compared to COCO [316].

6.3.2 Baselines

We compare our approach with N-Pairs Symmetric Angular Loss (Ang+NP+Sym, a combination of [363, 326, 417], trained with $\mathcal{L}_{ang}^{NP+SYM}$). For a subset of results, we also replace the angular loss with a weaker but more common loss, namely triplet, resulting in N-Pairs Symmetric Triplet Loss (Trip+NP+Sym). We show the result of choosing to enforce coherency within the image and text modalities by using images rather than text; this is the second variant of our method we described earlier. This method is denoted as Ours (Img NNs).

We also compare our approach against the deep structure preserving loss [364] (Struc), which enforces that captions paired with the same image are closer to each other than to non-paired captions. Because this approach requires multiple captions per image, we only show results for COCO.

Finally, we show how our approach can improve the performance of a state-of-the-art cross-modal retrieval model. PVSE [328] uses both images and text to compute a self-attention residual before producing embeddings. We include results for training this model

	Img-Text Non-Literal				**Img-Text Literal**			
	Politics [341]		**GoodNews** [27]		**ConcCap** [316]		**COCO** [211]	
Method	**I→T**	**T→I**	**I→T**	**T→I**	**I→T**	**T→I**	**I→T**	**T→I**
Ang+NP+Sym	0.6270	0.6216	0.8704	0.8728	0.7687	0.7695	**0.6976**	**0.6964**
Ours (Img NNs)	0.6370	0.6378	0.8840	0.8852	0.7636	0.7666	0.6819	0.6876
Ours	**0.6467**	**0.6492**	**0.8849**	**0.8865**	**0.7760**	**0.7835**	0.6900	0.6885
PVSE	0.6246	0.6199	0.8724	0.8709	0.7746	0.7809	0.6878	0.6892
PVSE+Ours	**0.6264**	**0.6314**	**0.8867**	**0.8864**	**0.7865**	**0.7924**	**0.6932**	**0.6925**
Trip+NP+Sym	0.4742	0.4801	0.7203	0.7216	**0.5413**	0.5332	**0.4957**	**0.4746**
Ours (Trip)	**0.4940**	**0.4877**	**0.7390**	**0.7378**	0.5386	**0.5394**	0.4790	0.4611

Table 23: We show retrieval results for image to text (**I→T**) and text to image (**T→I**) on all datasets. The best method per group is shown in bold.

both with and without our constraints.

6.3.3 Quantitative results

We first present results demonstrating our method's performance at cross-modality retrieval. To do so, we compute the embeddings of every image and text in our test set. We formulate a cross-modal retrieval task such that given a query image or text, the embedding of the paired image/text from the target modality must be closer to the query embedding than non-paired data (also of the target modality). We sample random (non-paired) samples (of the target modality) from the test set, along with the ground-truth paired sample. We then compute Recall@1 within each task: that is, whether the ground truth paired sample is closer to its cross-modal embedding than the non-paired embeddings. For our most challenging datasets (GoodNews and Politics), we use a 5-way task. For COCO and Conceptual Captions, we found this task to be too simple and that all methods easily achieved very high performance due to the literal image-text relationship. Because we wish to distinguish meaningful performance differences between methods, we used a 20-way task for Conceptual Captions and a 100-way task for COCO[2]. We report our result in Table 23.

The first and second group of results in Table 23 all use angular loss, while the third set use triplet loss. We observe that our method significantly outperforms all baselines tested for both directions of cross-modal retrieval for three of the four datasets. Our method achieves a 2% boost in accuracy (on average across both retrieval tasks) vs. the strongest baseline on GoodNews, and a 4% boost on Politics. We also observe that the recall is much worse for all tasks on the Politics dataset when compared to the GoodNews dataset, likely because the images and article text are much less well-aligned. Our method outperforms all baselines on ConcCap also, but not on COCO, since COCO is certainly the easiest of these datasets, with the most literal image-text alignment. Importantly, we also see that while the variant of our method using neighborhoods computed in image space (Ours Img NNs) does outperform Ang+NP+Sym, it is weaker than our main method variant (Ours). Overall, we conclude that our approach of enforcing within-modality semantic neighborhoods substantially improves cross-view retrieval performance, particularly when the relationship between image and text is complementary, rather than redundant.

We also observe that when adding our loss on top of the PVSE model [328], accuracy of retrieval improves. In other words, our loss is complementary to advancements accomplished by network model-based techniques such as attention.

In Table 24, we show a result comparing our method to Deep Structure Preserving Loss [364]. Since this method requires a *set* of annotations (captions) for an image, i.e. it requires ground-truth neighbor relations for texts, we can only apply it on COCO. In the first column, we show our method. In the second, we show [364] using ground-truth neighbors. Next, we show using [364] but with estimated neighbors, as in our method. We see that as expected, using estimated rather than GT text neighbors reduced performance (third vs. second cols). When estimated neighbors are used in [364]'s structural constraint, our method performs better. Interestingly, we observe using image neighbors in the structural constraint outperforms text neighbors. This may be because the structural constraint, which

[2] Task complexities were chosen before our method's results were computed based on the baseline's performance.

		Ours (Trip)	Struc (GT)	Struc (Text NN)	Struc (Img. NN$_\Omega$)
COCO	**I→T**	0.4790	**0.4817**	0.4635	0.4752
	T→I	0.4611	**0.4867**	0.4594	0.4604

Table 24: We show retrieval results for image to text (**I**→**T**) and text to image (**T**→**I**) on COCO using [364]'s loss vs. ours. GT requires multiple **G**round **T**ruth captions per image, while NN uses **N**earest **N**eighbors. The best method per row is shown in bold, while the best method which does not require a set of neighboring text is underlined.

requires the group of neighbors to be closer together than to others, is too strict for estimated text neighbors. That is, the constraint may require the text embeddings to lose useful discriminativity to be closer to neighboring text. In contrast, image neighbors are likely to be much more visually similar in COCO than in GoodNews or Politics (as they will contain the same objects). Note that in Table 23 image neighbors are computed in *visual* space, whereas here they are in *semantic* space (i.e. through neighboring text via Ω).

We next test how well each method preserves the *semantic neighborhood* given by Ω, i.e. Doc2Vec space. We begin by computing the embeddings in $\mathcal{M}$ (cross-modal space) for all test samples. For each such sample s_i (either image or text), we compute $\Psi_{\mathcal{M}}(s_i)$, that is, we retrieve the neighbors (of the same modality as s_i) in $\mathcal{M}$. We next retrieve the neighbors of s_i in Ω, $\Psi_{\Omega}(s_i)$, described in Sec. 6.2.2. For each sample, we compute $|\Psi_{\mathcal{M}}(s_i) \cap \Psi_{\Omega}(s_i)| \,/\, |\Psi_{\Omega}(s_i)|$, i.e. the percentage of the nearest neighbors of the sample in $\mathcal{M}$ which are also its neighbors in Ω. We consider the 200 nearest neighbors. That is, we measure how well each method preserves within-modality semantic locality by measuring the number of neighbors in Doc2Vec space which remain neighbors in the learned space. We report the result for competitive baselines in Table 25. We find that our constraints are, indeed, preserving within-modality semantic locality, as sample proximity in Ω is more preserved in $\mathcal{M}$ with our approach than without it, i.e. we better reconstruct the semantic neighborhood of Ω in $\mathcal{M}$. We believe this allows our model to ultimately perform better at

	GoodNews [27]		**Politics** [341]	
Method	**I**	**T**	**I**	**T**
Trip+NP+Sym	0.1183	0.1294	0.1135	0.1311
Ours (Trip)	**0.1327**	**0.1426**	**0.1319**	**0.1483**
Ang+NP+Sym	0.1032	0.1131	0.1199	0.1544
Ours (Ang)	**0.1270**	**0.1376**	**0.1386**	**0.1703**

Table 25: We test how well each method preserves the semantic neighborhood (see text) of Ω in $\mathcal{M}$. Higher values are better. Best method is shown in bold.

cross-modal retrieval.

We finally test the contribution of each component of our proposed loss. We test two variants of our method, where we remove either $\mathcal{L}_{text}$ or $\mathcal{L}_{img}$. We present our results in Table 26. In every case, *combining* our losses for our full method performs even better, suggesting that each loss plays a complementary role in enforcing semantic locality for its target modality.

6.3.4 Qualitative results

In this section, we present qualitative results illustrating how our constraints both improve semantic proximity and demonstrate superior retrieval results.

Semantic proximity. In Fig. 25, we perform an experiment to discover what samples our constraints affect the most. We randomly sampled 10k image-image and text-text pairs and computed their distance in $\mathcal{M}$ using features from our method vs. the baseline Ang+NP+Sym. Small ratios indicate the samples were closer in $\mathcal{M}$ using our method, relative to the baseline, while larger indicate the opposite. We show the samples with the top two smallest ratios for images and text. We observe that visually dissimilar, but semantically similar images have the smallest ratio (e.g. E.U. flag-Merkel, Judge's gavel-Supreme Court), which suggests our $\mathcal{L}_{img}$ constraint has moved the samples closer. For text, we

	GoodNews [27]		**Politics** [341]	
Method	**I→T**	**T→I**	**I→T**	**T→I**
OURS (ANG)	**0.8849**	**0.8865**	**0.6467**	**0.6492**
OURS (ANG)-$\mathcal{L}_{text}$	<u>0.8786</u>	0.8813	0.6387	<u>0.6467</u>
OURS (ANG)-$\mathcal{L}_{img}$	0.8782	<u>0.8817</u>	<u>0.6390</u>	0.6413

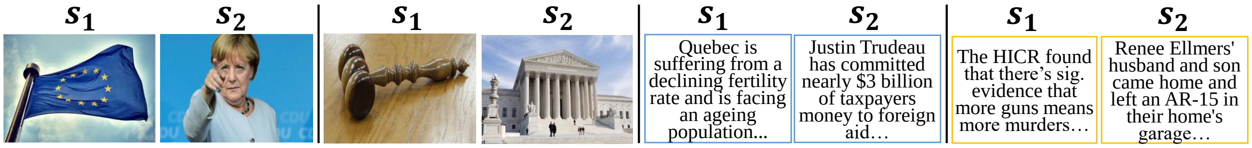
Table 26: We show an ablation of our method where we remove either component of our loss. The best method is shown in **bold** and the best ablation is underlined.

observe articles about the same issue are brought closer even though specifics differ.

Cross-modal retrieval results. In Fig. 26 we show the top-3 results for a set of queries, retrieved by our method vs. ANG+NP+SYM. We observe increased semantic homogeneity in the returned samples compared with the baseline. For example, images retrieved for "drugs" using our method consistently feature marijuana, while the baseline returns images of pills, smoke, and incorrect retrievals; "wall" results in consistent images of the border wall; "immigration" features arrests. For text retrieval, we find that our method consistently performs better at recognizing public figures and returning related articles.

6.4 DISCUSSION

In this chapter, we focus on learning task-agnostic semantic representations in real-world multimedia. We proposed a novel loss function which improves semantic coherence for cross-modal retrieval. Our approach leverages a latent space learned on text alone, in order to enforce proximity within *cross-modal* space. We constrain text and image embeddings to be close in joint space if they or their partners were close in the unimodal text space. Our technique explicitly controls for our observation that images and paired text in communicative multimedia may not be literally aligned and that concepts mentioned in text may



Quebec is suffering from a declining fertility rate and is facing an ageing population...

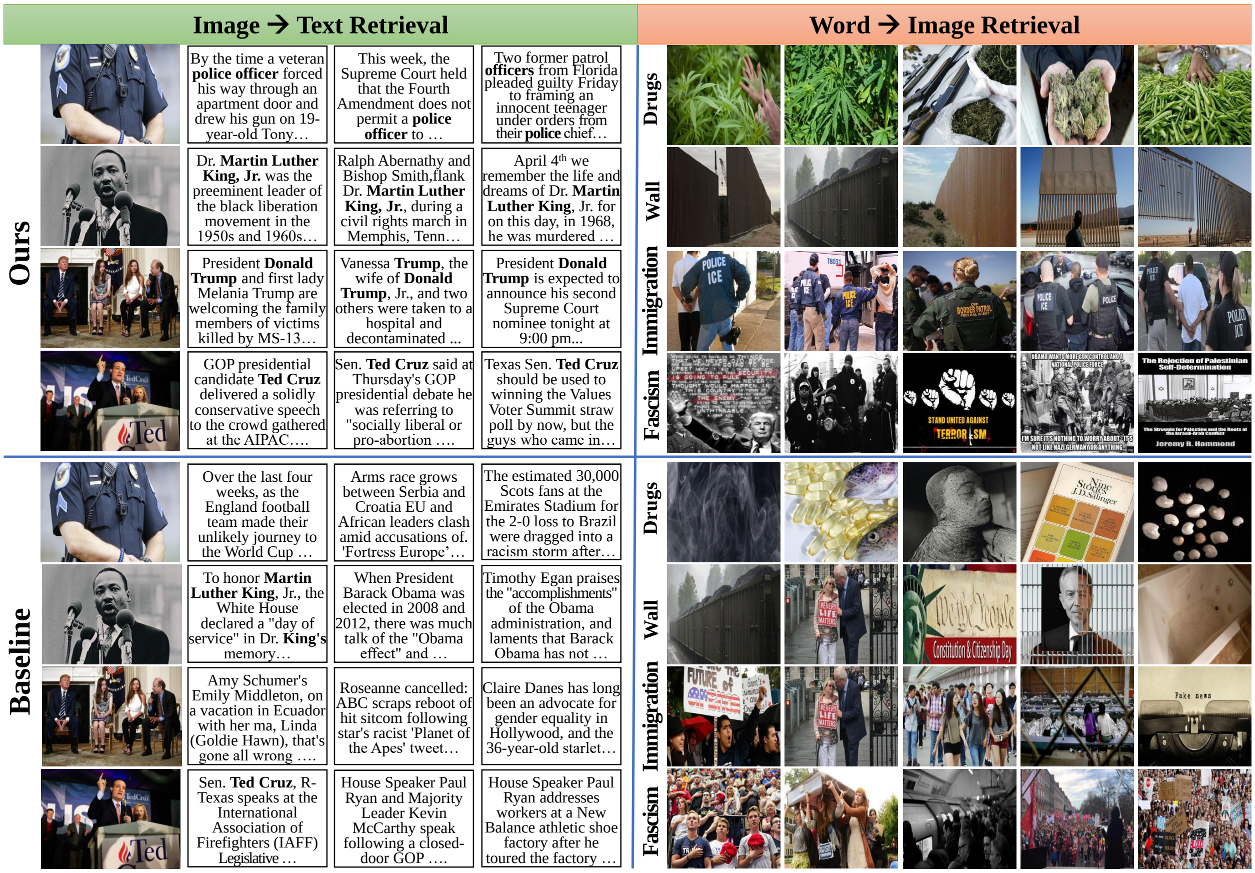
Justin Trudeau has committed nearly $3 billion of taxpayers money to foreign aid…

The HICR found that there’s sig. evidence that more guns means more murders…

Renee Ellmers' husband and son came home and left an AR-15 in their home's garage…

Figure 25: Uncurated results showing image/text samples that our method keeps closest in $\mathcal{M}$ compared to the baseline, i.e. pairs where $\frac{d_{ours}(s_1,s_2)}{d_{baseline}(s_1,s_2)}$ is smallest. Our method keeps semantically related images and text closer in the space, relative to the baseline. While the images are not visually similar, they are semantically similar (EU and Merkel; judge’s gavel and Supreme Court).

not have explicit visual groundings. Our novel loss constraints force models to account for such complementarity by learning representations which accomodate non-literal, symbolic, and polysemous usages, rather than just capturing straightforward literal visual and textural content. We experimentally demonstrated that our approach significantly improves upon several state-of-the-art loss functions on multiple challenging datasets, confirming our hypothesis regarding complementarity in communicative multimedia (**H3**). Our approach leverages the paired text domain as a sort of guidance for training and further proves our hypothesis regarding directed training (**H2**). We also presented qualitative results demonstrating increased semantic homogeneity of retrieval results. We observed many figurative visual usages and examples of visual argumentation, confirming communicative images substantially differ from traditional content (**H1**). Applications of our method include improving retrieval of non-literal text, visual question answering, and learning robust visual-semantic embeddings. The subsequent chapter also seeks to learn task-agnostic semantic representations in multimedia by exploiting image-text complementarity, but proposes an orthogonal approach which doesn’t require additional losses.



By the time a veteran **police officer** forced his way through an apartment door and drew his gun on 19-year-old Tony…

This week, the Supreme Court held that the Fourth Amendment does not permit a **police officer** to …

Two former patrol **officers** from Florida pleaded guilty Friday to framing an innocent teenager under orders from their **police** chief…

Dr. **Martin Luther King, Jr.** was the preeminent leader of the black liberation movement in the 1950s and 1960s…

Ralph Abernathy and Bishop Smith,flank Dr. **Martin Luther King, Jr.**, during a civil rights march in Memphis, Tenn…

April 4th we remember the life and dreams of Dr. **Martin Luther King**, Jr. for on this day, in 1968, he was murdered …

President **Donald Trump** and first lady Melania Trump are welcoming the family members of victims killed by MS-13…

Vanessa **Trump**, the wife of **Donald Trump**, Jr., and two others were taken to a hospital and decontaminated ...

President **Donald Trump** is expected to announce his second Supreme Court nominee tonight at 9:00 pm...

GOP presidential candidate **Ted Cruz** delivered a solidly conservative speech to the crowd gathered at the AIPAC….

Sen. **Ted Cruz** said at Thursday's GOP presidential debate he was referring to "socially liberal or pro-abortion ….

Texas Sen. **Ted Cruz** should be used to winning the Values Voter Summit straw poll by now, but the guys who came in…

Over the last four weeks, as the England football team made their unlikely journey to the World Cup …

Arms race grows between Serbia and Croatia EU and African leaders clash amid accusations of. 'Fortress Europe'…

The estimated 30,000 Scots fans at the Emirates Stadium for the 2-0 loss to Brazil were dragged into a racism storm after…

To honor **Martin Luther King**, Jr., the White House declared a "day of service" in Dr. **King's** memory…

When President Barack Obama was elected in 2008 and 2012, there was much talk of the "Obama effect" and …

Timothy Egan praises the "accomplishments" of the Obama administration, and laments that Barack Obama has not …

Amy Schumer's Emily Middleton, on a vacation in Ecuador with her ma, Linda (Goldie Hawn), that's gone all wrong ….

Roseanne cancelled: ABC scraps reboot of hit sitcom following star's racist 'Planet of the Apes' tweet…

Claire Danes has long been an advocate for gender equality in Hollywood, and the 36-year-old starlet…

Sen. **Ted Cruz**, R-Texas speaks at the International Association of Firefighters (IAFF) Legislative …

House Speaker Paul Ryan and Majority Leader Kevin McCarthy speak following a closed-door GOP ….

House Speaker Paul Ryan addresses workers at a New Balance athletic shoe factory after he toured the factory …

Figure 26: We show cross-modal retrieval results on Politics [341] using our method and the strongest baseline. We bold text aligning with the image. For text retrieval, ours returns more relevant (and semantically consistent) results. For image retrieval, our method exhibits more consistency (e.g. drug images are marijuana, immigration are arrests), while the baseline returns more inconsistent and irrelevant images.

7.0 MODELING ABSTRACT SEMANTICS IN MULTIMEDIA BY IDENTIFYING AND WEIGHTING SEMANTICALLY INFORMATIVE SAMPLES

Summary. *In this chapter, we continue our work in modeling task-agnostic semantics in communicative multimedia.*[1] *As we explain in Chapter 6, existing approaches for learning cross-modal semantic representations assume a straightforward relationship between images and text, with the text literally descriptive of the image. However, real-world multimedia often feature abstract semantics and symbolic relationships where image and text are complementary. In order to ensure the model learns a semantically robust space and nuanced relationships are accounted for, care must be taken to ensure that challenging, informative image-text pairs contribute to learning. In the previous chapter, we leveraged the mutually supplementary relationship between image and text to enforce novel loss constraints which we showed improved the semantic coherence in the learned space. In this chapter, we show how standard loss functions can be updated to better capture abstract semantics, without the addition of custom losses. We propose a novel approach which weighs image-text pairs in the loss and prioritizes informative and representative samples. Our method takes into account the diversity of the sample's semantic neighbors, the discrepancies in the broader neighborhood, and the relative density of the sample. All three measures take into account the unique mutually supporting nature of image and text in communicative multimedia. Experiments on three challenging datasets with image-text complementarity, as well as COCO, demonstrate significant performance gains compared to recent state-of-the-art models and weight-based approaches.*

[1] The work presented in this chapter is in submission to NeurIPS 2020.

7.1 INTRODUCTION

We live in a multimodal world; many people both see and hear their environment. Modern media use this multimodality to better convey stories: the same overall topic is expressed in text, images, video, audio, etc. However, the modalities (e.g. images and text) tell different *parts* of the story. For example, the text might describe the events in the Syrian War, while the image *illustrates one aspect,* the suffering of a refugee child. Given the enormity of multimodal data available on the web, intelligent systems must reason across modalities. The basic step is constructing a shared semantic feature space such that images and text corresponding to the same concept neighbor each other. However, to model complex multimodal media, cross-modal methods must also understand not just the *value of the relationship* (close/not) but also the *nature of the relationship* between the co-occurring image and text, as well as how the pair relates to the broader data pool.

Most cross-modal retrieval approaches assume the relationship between image and text is redundant and image and text align at a literal level: for example, the caption describes an airplane in the sky and the image shows airplane in the sky. This makes sense when the *purpose* of a caption is to provide the exact same content as the image, e.g. to serve a visually impaired user. In the more general case however, in real-world data, the type of relationship between image and text can vary.

Simple cases (airplane-airplane) provide strong training signal, but there are other samples, with a less direct (less overlapping, more complementary) relationship, that may be more informative. For example, an image might show tomato plants in a garden, while a caption reads "Plant the tomatoes at 5 inches from each other and water them." In other words, images may be used illustratively or even figuratively to underscore a point made in the text, or they may add nuance or subjective arguments to the text. Figs. 27, 28, and 29 show some examples. For instance, in Fig. 27 (left-top), a group of children with Israeli flags are paired with text about Netanyahu and "an unprecedented government decision". Without proper care, such examples will be drowned out by the easier cases (e.g. the laptop in Fig. 27).

To address this problem, we propose to dynamically weigh each image-text pair within

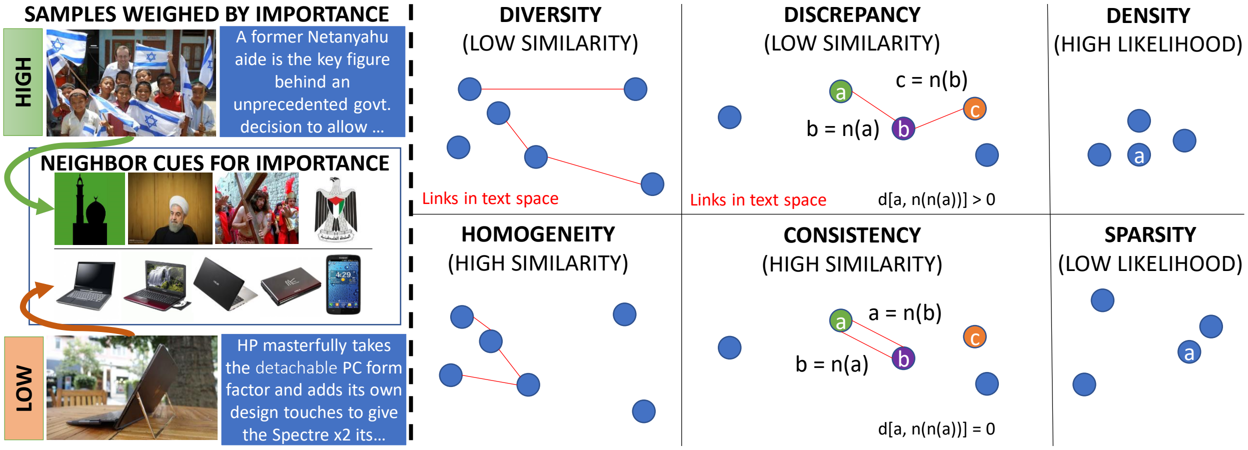


Figure 27: (Left of dashed line) Image-text pairs weighted by importance. One weighting uses the neighboring images, which are diverse for the "Israeli flag" image weighed high, and consistent (not diverse) for the laptop image weighed low. (Right) At the top, our three weighting cues; at the bottom, their inverses. Blue circles denote images in visual space. Red links connect images whose corresponding texts are close. Diversity vs homogeneity measures whether images whose corresponding texts are close (red links) are also close in image space (*no* for diversity, *yes* for homogeneity). Discrepancy vs consistency measures whether the neighbor of a sample's neighbor returns the original sample (*no*/*yes*, respectively); *n()* denotes first nearest neighbor in semantic space. Density vs sparsity measures whether sample a is likely or not given the dataset (*yes*/*no*).

a training batch. We model the relationship between images and text, as well as their surrounding samples, by measuring the extent to which (1) the image and text modalities come from a diverse neighborhood, and (2) the image and text pair is representative of the overall data space. Specifically, we first measure the diversity of the sample's semantic neighbors, which could suggest more abstract, complementary matches between image and text. Neighbors with low textual, but high visual diversity could imply the same semantic concept shares multiple visual expressions, i.e. these are samples that should be prioritized in cross-modal learning. Second, we consider the relationship of the sample to its broader neighborhood. We measure the sample's neighborhood discrepancy by computing the distance of the sample to the *neighbors of its neighbors*. This measures the symmetry of the sample-to-neighbor relationship, and whether the local neighborhood around a sample is compact or diffuse. If a sample is far from the neighbors of the sample's neighbors (i.e. has a diffuse neighbor space), this could indicate the sample is more likely to have multiple senses, much like samples with high diversity. Third, in order to gauge how representative or typical a sample is of the overall dataset and isolate the impact of spurious image-text connections, we compute a modified probability density function for each sample using a variational Gaussian mixture. Higher scores indicate that the sample's embedding lies in a denser region of the overall space, while lower scores suggest the sample may be an outlier and should be discounted. We illustrate our three weighting cues in Fig. 27.

Our main contribution in this chapter is a suite of three novel sample weighting methods for learning visual-semantic embeddings on challenging data consisting of complementary image-text pairs. Our approach can be easily integrated into standard ranking losses. We perform detailed experimental analysis on three datasets that exhibit complementarity, namely Conceptual Captions [316], GoodNews [27] and Politics [341]. We also evaluate on a more standard retrieval dataset, COCO [211], to put our method in the context of standard retrieval methods. We show strong results against four recent, state-of-the-art weighting-based approaches and a recent model-based approach. We show that the spaces learned by our methods better preserve abstract concepts and nuanced semantic notions relative to existing methods. More broadly, we present approaches for automatically identifying and quantifying the degree to which samples convey abstract semantics and latent visual concepts, in the absence of any particular task. We believe the method presented in this chapter provides a strong starting point which can facilitate learning in challenging real-world scenarios.

Our work in this chapter continues the theme from Chapter 6 of learning semantic representations in multimedia, in the absence of a particular application. In the prior chapter, we showed how the semantic relationships between concepts from a task-agnostic text space could be preserved on the learned multimodal space to better preserve semantic coherency in a noisy real-world multimedia dataset by imposing novel within-modal loss functions. While this approach ensures that the semantics of the text space may be preserved, it doesn't necessarily preserve the emergent semantics of the multimodal space, nor does it emphasize abstract and challenging samples. In this chapter, we present an approach to automatically discover and weight semantically informative multimedia samples according to their predicted semantically utility. We show that our method emphasizes samples conveying abstract concepts such as “justice” or “freedom”, which convey latent visual concepts. Thus, our method in this chapter presents a multimodally data-driven approach for quantifying the degree to which a sample carries abstract visual concepts, rather than enforcing external semantic structures learned from text as in Chapter 6.

7.2 METHOD

Let $\boldsymbol{\mathcal{D}} = \{\mathbf{I}, \mathbf{T}\}$ represent a dataset of n image-text pairs, where $\mathbf{I} = \{x_1, x_2, \ldots, x_n\}$ and $\mathbf{T} = \{y_1, y_2, \ldots, y_n\}$ represent the set of images and text, respectively, and y_i is text co-occurring with image x_i (the two are semantically related). We refer to (x_i, y_i) as positive pairs and either $(x_i, y_{j\neq i})$ or $(x_{j\neq i}, y_i)$ as negative pairs. In order to compare across modalities, we seek a common manifold $\mathcal{M}$. A convolutional network $f : \mathbf{I} \to \mathcal{M}$ is used to project images into the joint space, while a recurrent network $g : \mathbf{T} \to \mathcal{M}$ projects text. We use the notational shorthand $f\left(x\right) = x \in \mathbb{R}^{K\times H}$ and $g\left(y\right) = y \in \mathbb{R}^{K\times H}$, where K is the number of embeddings per sample and H is the dimension of the learned manifold. Most prior methods assume $K = 1$ but this may be too stringent when image and text have multiple meanings. Recently [328] propose a polysemous embedding model (PVSE), where every image and text are represented by K embeddings encouraged to be diverse; we adopt this formulation for all methods compared. When comparing two samples, we use the maximum cosine similarity across all K^2 pairs: $s(x_i, y_i) = \max\limits_{k\in K} \left\langle \frac{x_{i_k}}{\|x_{i_k}\|_2}, \frac{y_{i_k}}{\|y_{i_k}\|_2} \right\rangle : \mathbb{R}^{K\times H} \times \mathbb{R}^{K\times H} \to \mathbb{R}$. For notational simplicity, we omit the reference to the k embeddings in the remaining text.

7.2.1 Training objective

We assume the same pairwise ranking objective as other recent VSE methods [328, 7, 85, 212], but introduce a weighting constraint to emphasize semantically informative samples. We optimize a sample-weighted bidirectional n-pairs [326] triplet loss $\mathcal{L}_{\text{RANK}}$ given by:

$$\begin{aligned}\mathcal{L}_{\text{RANK}} = \frac{1}{2N^2}\Bigg(&\sum_{x_i \in I_B} \sum_{y_j \in T_B} \alpha_i \left[\, m - s\left(x_i, y_i\right) + s\left(x_i, y_{j\neq i}\right) \,\right]_+ + \\ &\sum_{y_i \in T_B} \sum_{x_j \in I_B} \alpha_i \left[\, m - s(x_i, y_i) + s(x_{j\neq i}, y_i) \,\right]_+ \Bigg)\end{aligned} \tag{26}$$

where m is a margin parameter, $[\cdot]_+ = \max(0, \cdot)$ and I_B / T_B are images and text, respectively, within a minibatch of samples. We introduce a per-positive sample weight α_i, given by our method. All methods and baselines use this loss to train, but vary in how α_i are computed.

Limitations of hard negative mining. Traditional VSE methods give all samples equal weight within a minibatch. To facilitate learning, most recent methods [328, 85, 364] also perform hard negative mining, where only the most challenging negative sample is used (e.g. $\max_j s(x_i, y_{j\neq i})$). While this makes sense in common captioning datasets, we found it prevented models from successfully training on our more challenging, complementary datasets. When using hard negatives, the problem becomes too hard since many negative image-text alignments are semantically plausible, even if technically incorrect. Moreover, relying only on hard negatives makes the model more vulnerable to noise, which is present within the webly-harvested datasets we consider. Mithun [243] proposes a soft (weighted) semi-hard negative mining approach to enable learning, and we outperform it on three datasets.

7.2.2 Measuring semantic diversity

In order to emphasize informative, complementary image-text samples, we first must detect them and determine how much weight to give each. We seek to quantify properties that such samples may posses. We first observe that semantic concepts with nonliteral portrayals are likely to be visually diverse. For example, a piece of text about justice could be paired with an image of the Supreme Court, an American flag, Themis the Goddess of

Justice, a judge's gavel, etc. In contrast, images paired with the caption "a bowl of apples on a table" would likely be much more visually similar. To measure the visual diversity of the semantic concept a sample illustrates, we first discover each image-text pair's *semantic* neighbors in text (Doc2Vec [192]) space $\Omega(\mathbf{T})$. We choose to compute neighbors in text space because the text domain provides the cleanest semantic representation of the image-text pair. In contrast, similar visual content (e.g. images of people or faces) could be semantically unrelated while visually dissimilar content may still be semantically related (justice example above). Let $\Psi\left(\Omega(y_i)\right) = \left\{\left\langle x'_{i_n}, y'_{i_n}\right\rangle\right\}_{n=1}^{N}$ represent the semantic nearest neighbor function over $\Omega(\mathbf{T})$, where $\left\{\left\langle x'_{i_n}, y'_{i_n}\right\rangle\right\}_{n=1}^{N}$ denotes the set of the N neighbors of $\langle x_i, y_i\rangle$ and $y_i \notin \Psi\left(\Omega(y_i)\right)$. Note the semantic neighbor images $\left\{x'_{i_n}\right\}^{N}$ are not necessarily visual neighbors of x_i.

We next measure the diversity of the semantic neighbors in both the image and text domains. Because our formulation is equivalent for both image/text neighbors, we let s_i represent a sample from either domain but require samples s_i and s_j come from the same domain. Let $\mathbf{s}'_i = \left[s'_{i_1}, s'_{i_2}, \ldots, s'_{i_N}\right]^{\intercal}$ denote the matrix of size $N \times H$ of the embeddings of the neighbors of of s_i found via Ψ, and $\mathbf{U} = \mathbf{s}'_i\mathbf{s}'^{\intercal}_i$ computes a cross-product between all semantic neighbors to obtain their pairwise similarities. We compute the *semantic diversity score* Υ_i^{DIV} for s_i as follows:

$$\Upsilon_i^{DIV} = \Gamma^{DIV} \times \frac{1}{N^2}\sum_{r=1}^{N}\sum_{c=1}^{N}\mathbf{U}_{(r,c)} \tag{27}$$

where r, c index over the rows and columns of $\mathbf{U} = \mathbf{s}'_i\mathbf{s}'^{\intercal}_i$ and $\Gamma^{DIV} \in \{1, -1\}$ is a switching parameter, which controls whether more weight is given to more similar or less similar samples. We finally enforce that all Υ_i^{DIV} in a minibatch form a proper attention vector $\boldsymbol{\alpha^{DIV}}$ as follows:

$$\boldsymbol{\alpha^{DIV}} = \left[\alpha_1^{DIV}, \alpha_2^{DIV}, \ldots, \alpha_B^{DIV}\right], \text{ where } \alpha_i^{DIV} = \lambda \times \frac{e^{\Upsilon_i^{DIV}}}{\sum_{j=1}^{B} e^{\Upsilon_j^{DIV}}} \tag{28}$$

and where λ is a scaling constant and B is the minibatch size. The weights can now be directly used in Eq.26 to weight samples by the semantic diversity measure. We compute $\boldsymbol{\alpha^{DIV}}$ for the image and text domains separately (i.e. $s_i = x_i$ or $s_i = y_i$), then combine the two vectors by addition: $\boldsymbol{\alpha^{DIV}} = \text{softmax}\left(\boldsymbol{\alpha_X^{DIV}} + \boldsymbol{\alpha_Y^{DIV}}\right)$ or by taking their absolute difference: $\boldsymbol{\alpha^{DIV}} = \text{softmax}\left(\left|\boldsymbol{\alpha_X^{DIV}} - \boldsymbol{\alpha_Y^{DIV}}\right|\right)$. We show results for both combination strategies in Tab. 28.

7.2.3 Measuring semantic neighborhood discrepancy

The previous measure quantifies the diversity within the semantic neighbors of a query sample s_i. It does not, however, consider the relationship of the neighborhood to the query sample itself. We next examine the following relationship: $s_i \in \Psi\left(\Omega(\Psi\left(\Omega(s_i)\right))\right)$, i.e. whether a sample is a neighbor of its neighbors. This criterion measures the relationship of s_i to its local neighborhood and quantifies whether the surrounding space is compact (high similarity) or diffuse (low similarity). Samples with diffuse neighborhood could suggest the image/text have multiple meanings or are used figuratively.

Formally, let $\Psi\left(\Omega(\Psi\left(\Omega(s_i)\right))\right) = \left\{s''_{i_n}\right\}_{n=1}^{N^2}$ represent the set of the semantic neighbors of s_i's semantic neighbors. Note that as in Sec. 7.2.2, we always compute Ψ in text space (for images, this amounts to using the ground truth text paired with the image). Let $\mathbf{s}''_i = \left[s''_{i_1}, s''_{i_2}, \ldots, s''_{i_{N^2}}\right]^\intercal$ denote the matrix of size $N^2 \times H$ of the embeddings of the neighbors of neighbors. Then, the *semantic discrepancy score* Υ_i^{DIS} and corresponding scaled score α_i^{DIS} of s_i is given as follows:

$$\alpha_i^{DIS} = \lambda \times \frac{e^{\Upsilon_i^{DIS}}}{\sum_{j=1}^{B} e^{\Upsilon_j^{DIS}}}\,, \text{ where } \Upsilon_i^{DIS} = \Gamma^{DIS} \times \frac{1}{N^2}\sum_{r=1}^{N^2}\mathbf{V}_{(r)}\,, \tag{29}$$

where $\mathbf{V} = \mathbf{s}''_i s_i'^\intercal$ is the matrix-vector product of the sample's neighborhood and the sample (size $N^2 \times 1$), r indices its entries, Γ^{DIS} is a switching parameter (see Sec. 7.2.2), and B is the minibatch size. The final attention vector is given by stacking sample weights: $\boldsymbol{\alpha^{DIS}} = \left[\alpha_1^{DIS}, \alpha_2^{DIS}, \ldots, \alpha_B^{DIS}\right]$. We compute $\boldsymbol{\alpha^{DIS}}$ in both image and text space then combine the two as in Sec. 7.2.2.

7.2.4 Measuring sample density

The previous two metrics rely on the semantic neighbors of a sample, but do not tell us the relationship of the sample to the dataset as a whole. Consider a small, tight cluster of outlier samples which lie far from other samples in the dataset (Fig. 27 right). The above metrics would capture that the sample resided in a compact region, but not that the region is atypical. Knowing how representative a sample is of the dataset is important for detecting and mitigating the impact of outliers. We train a Gaussian mixture model (GMM) on our learned embedding space and use it to quantify a sample's likelihood. Let ξ denote a GMM and $\mathbf{s}$ the set of embeddings of the train set. We train ξ using a variational approach [28], which allows the number of mixtures to be determined automatically. Given a sample s_i, the standard GMM is given by $\xi(s_i) = p(s_i|\boldsymbol{\theta}) = \sum_{m=1}^{M} \pi_m \mathcal{N}(s_i|\mu_m, \Sigma_m)$, where $\boldsymbol{\theta}$ is the set of model parameters, μ_m and Σ_m are the mean and covariance of the mth Gaussian mixture, and π_m are the mixing coefficients such that $\sum_m \pi_m = 1$. Let $z_i = [z_{i_1}, \ldots, z_{i_m}]$ be a latent binary indicator indicating s_i's membership in mixture m. ξ can then marginalized as: $\xi(s_i) = \sum_z p\left(s_i|z_i, \{\mu_m, \Sigma_m\}\right) p\left(z_i|\left\{\pi_m\right\}\right)$, where $p\left(s_i|z_i, \{\mu_m, \Sigma_m\}\right) = \prod_m \mathcal{N}\left(s_i|\mu_m, \Sigma_m\right)^{z_{im}}$ and $p\left(z_i|\left\{\pi_m\right\}\right) = \prod_m \left(\pi_m\right)^{z_{im}}$. ξ can be trained in a fully Bayesian manner by imposing a prior distribution over all parameters, using expectation maximization. We refer readers to [28] for details. Because embeddings change during training of f and g, every 5 epochs we train ξ on $\mathbf{s}$. We use a warm-start, initializing the model with the previously found solution to aid convergence.

After training ξ, we compute the probability density of each sample, $p(s_i)$. Let $\Upsilon_i^{DEN} = \xi\left(s_i\right)$ denote the *sample density score* and $\boldsymbol{\Upsilon}^{DEN} = \left[\Upsilon_1^{DEN}, \ldots, \Upsilon_n^{DEN}\right]$. We compute α_i^{DEN} as:

$$\begin{aligned} \boldsymbol{\rho} &= \log\left(\boldsymbol{\Upsilon}^{DEN} + \left|\min\left(\boldsymbol{\Upsilon}^{DEN}\right)\right| + 1\right) \\ \kappa_i &= \Gamma^{DEN} \times \operatorname{med}\left(\mu_{\boldsymbol{\rho}} - 2\sigma_{\boldsymbol{\rho}}, \rho_i, \mu_{\boldsymbol{\rho}} + 2\sigma_{\boldsymbol{\rho}}\right) \\ \alpha_i^{DEN} &= \lambda \times \frac{e^{\kappa_i}}{\sum_{j=1}^{B} e^{\kappa_j}} \end{aligned} \tag{30}$$

where $\boldsymbol{\kappa}$ and $\boldsymbol{\rho}$ are scaled and clipped copies of $\boldsymbol{\Upsilon}^{DEN}$, Γ^{DEN} is the switching parameter, B is the minibatch size, and λ is a scalar. The median function, med, clips each value within the range of $\pm 2\sigma$ of the mean of scaled densities $\boldsymbol{\rho}$. We found clipping necessary to stabilize the behavior of softmax. We compute densities in both image and text spaces and combine the two as in Sec. 7.2.2.

7.2.5 Implementation details

All methods use ResNet-50 [127] initialized with ImageNet features for images, and Gated Recurrent Units [52] for text, with hidden state size 512. All methods and baselines are built on top of PVSE [328]; we learn k per dataset. Images are scaled to 224x224 and augmented with random horizontal flipping. We use Xavier initialization [107] on all non-pretrained learnable weights. GRUs are init. with 200D word embeddings learned on the dataset on which they are applied. We perform L_2 normalization on embeddings produced by each model. We train all models using Adam [175] with minibatch size of 32, learning rate 1.0e-4 (decayed by a factor of 10 after every 5 epochs of no decrease in val loss), and weight decay 1e-5. We use a train-val-test split of 80-10-10 for all datasets. We use [295]'s implementation of Doc2Vec, $d \in \mathbb{R}^{200}$ and train on each dataset using distributed memory [192] for 20 epochs with window $k = 20$, ignoring words that appear less than 20 times. We use [231] to efficiently compute approximate nearest neighbors for Ψ and use $N = 200$ nearest neighbors. We use a Wishart [28] prior for our GMM and constrain each component to a diagonal covariance matrix. We probabilistically sample at most 1000 neighbors at a time from s_i'' in Eq. 29. We cache embeddings from the prior epoch for efficient computation of Eqs. 27 and 29. All method use $\lambda = 96$, which was determined empirically. We show the impact of $\Gamma = \pm 1$ for all methods in Tab. 28.

7.3 EXPERIMENTAL EVALUATION

We compare the three weighting strategies we propose, Ours-Diversity (Sec. 7.2.2), Ours-Discrepancy (Sec. 7.2.3) and Ours-Density (Sec. 7.2.4), to four very recent techniques:

- PVSE [328] which computes multiple embeddings to account for polysemy;

- HAL [212] which up-weighs samples likely to be the closest sample to multiple queries;
- Mithun [243] which weighs samples based on hardness (computed using ranks of matching images/text, larger values denoting worse match hence more challenging sample); and
- Amrani [7] weighs highly samples where both image/text in a sample belong to tight clusters (an intuition opposite to ours) and do not use semantic neighborhoods for images.

Datasets. We demonstrate our approach on four datasets. The first three (Conceptual Captions, GoodNews and Politics) demonstrate the complementarity of image and text within a pair that we describe in Sec. 7.1. To put our method in context, we also evaluate on COCO. **Conceptual Captions** [316] contains of ∼3.3M image-text pairs. The text comes from automatically cleaned alt-text descriptions paired with web images and exhibiting a much wider variety of style and content compared to COCO. **GoodNews** [27] consists of ∼466k images and captions from the New York Times, and **Politics** [341] consists of ∼246k pairs of images with sentences from news articles. **COCO** [211] contains ∼120k images with captions. All of these datasets are large-scale and recent. While COCO and Flickr30K are among the most popular retrieval datasets, both contain heavily overlapping with the image, descriptive captions. We use ConcCap, GoodNews and Politics in place of Flickr30K, to demonstrate the challenge of matching complementary images and text.

Metric. We evaluate top-1 accuracy on image to text, and text to image matching. Upon examination, we find the image-text alignment in GoodNews and Politics the most challenging, hence for these datasets, we use a 5-way multiple-choice task (1 correct option, 4 incorrect ones). For COCO and ConcCap, we found this task to be too simple and all methods easily achieved very high performance due to the literal image-text relationship. To distinguish meaningful performance differences between methods, we used a 20-way task for Conceptual Captions and a 100-way task for COCO.

7.3.1 Main result

We show our main result in Table 27. At the top are four state of the art methods. At the bottom are our three weighting techniques. We observe that the best method per

	GoodNews [27]		**Politics** [341]		**ConcCap** [316]		**COCO** [211]	
Method	**I→T**	**T→I**	**I→T**	**T→I**	**I→T**	**T→I**	**I→T**	**T→I**
PVSE [328]	0.8516	0.8526	0.5919	0.6057	0.7138	0.7168	0.6541	0.6561
HAL [212]	0.8623	0.8579	0.5919	0.5903	0.7638	0.7685	0.6665	0.6845
Amrani [7]	0.8629	0.8678	0.6117	0.6117	0.7376	0.7356	0.6746	0.6756
Mithun [243]	0.8439	0.8463	0.5792	0.5839	0.7523	0.7497	**0.6967**	**0.6950**
Ours-Diversity	0.8499	0.8509	**0.6268**	**0.6366**	**0.7720**	**0.7741**	0.6891	0.6855
Ours-Discrepancy	**0.8730**	**0.8764**	0.6211	0.6228	0.7294	0.7298	0.6863	0.6845
Ours-Density	0.8716	0.8752	0.6209	0.6216	0.7393	0.7416	0.6838	0.6812

Table 27: We show retrieval results (top-1 accuracy) for image to text (**I→T**) and text to image (**T→I**). The best method per task is shown in **bold**, and second-best underlined.

dataset/task is one of our methods on all complementary datasets (**GoodNews, Politics, ConcCap**). On **Politics** and **GoodNews**, the second-best method is also one of ours. On **COCO**, all of our methods essentially tie for second-best. Note that Mithun which outperforms ours on **COCO**, performs much worse on the complementary datasets, often worse than even PVSE. The biggest gain we achieve over PVSE is 8% (in relative terms, or 6% absolute) on **ConcCap**. The relative gain on **Politics** is 5.5%, on **COCO** is 5%, and on **GoodNews** is 2.5%. Our biggest gain over HAL is 7% on **Politics**, over Amrani is 5% on **ConcCap**, and over Mithun is 8.5% on **Politics**. The difference between our best performing method and the best performing baseline per-dataset are all statistically significant ($p < 0.05$).

Among our methods, usually Ours-Diversity performs best, but on **GoodNews**, the best method is Ours-Discrepancy. Overall, Ours-Diversity is best or second-best on six of eight tasks, Ours-Discrepancy on four tasks, and Ours-Density on two. Importantly, all of our methods show benefit, but the choice of weighting method is dataset-dependent, suggesting each dataset exhibits complementarity in a different way which is best captured by a particular method. Ours-Density only uses the structure of the within-

	Diversity (Sec. 7.2.2)		Discrepancy (Sec. 7.2.3)		Density (Sec. 7.2.4)	
Method	**I→T**	**T→I**	**I→T**	**T→I**	**I→T**	**T→I**
$\Gamma = +1$	0.6206	0.6226	0.6158	0.6187	**0.6209**	**0.6216**
$\Gamma = -1$	**0.6268**	**0.6366**	**0.6211**	**0.6228**	0.6106	0.6170
Sum	0.6188	0.6251	0.6130	0.6184	0.6131	0.6181
Diff	**0.6268**	**0.6366**	**0.6211**	**0.6228**	**0.6209**	**0.6216**

Table 28: Ablation on Politics [341]. The first two rows show results for $\Gamma = +1/-1$. The last two show strategies for combining the image/text weight vectors (summing or taking absolute difference).

modality space of the image and text domains, i.e. it does not exploit the paired nature of the multimodal data. In contrast, OURS-DISCREPANCY and OURS-DIVERSITY work cross-modally by exploiting the text space to find semantic text and image neighbors. We also attempted to combine all three methods, but found performance varied across datasets. On **GoodNews**, the combination improved performance to 0.8785 and 0.8811, but on others, combining the three methods was inferior to using any single method.

7.3.2 Ablation results

We first present a result motivating the choice of directionality for our proposed weighting mechanisms. For each of our methods, the weighting could be implemented with the opposite sign (via Γ), e.g. we could prioritize samples that are outliers (as a counterpart to samples that are in dense regions), or we could up-weigh samples that come from homogeneous rather than diverse regions. To test the importance of Γ, we use the **Politics** dataset. In Table 28 (top block), we see that emphasizing samples with high density, low homogeneity (which we term "diversity"), and diffuse, inconsistent semantic neighborhoods (which we term "discrepancy") perform better. We also trained a model which used all equal weights (still scaled by λ) and found it performed even worse than the suboptimal Γ setting.

At the bottom, we explore how to combine the $\boldsymbol{\alpha_X}$ and $\boldsymbol{\alpha_Y}$ scores from the two modalities. We observe taking a difference between the two modalities is better, so weights are larger for samples whose measures differ more across modalities. This underscores the emphasis on weighting complementarity: samples differing across modalities suggest a complementary, rather than overlapping cross-modal alignment, while emphasizing uniformity or overlap (via sum) performs worse.

We next measure how correlated the metrics we propose are. We computed Spearman's rank correlation between the sample weights, and found that our density measure was uncorrelated with discrepancy ($\rho = -0.0032$) and diversity ($\rho = -0.0021$). Discrepancy and diversity were very slightly correlated ($\rho = 0.0520$) because they capture a similar phenomenon (image-text complementarity of the sample), while density measures a different, global property.

Finally, we test the contribution of using s_i'' as opposed to just s_i' in our discrepancy measure, i.e. using neighbors of neighbors, as opposed to just neighbors. On **Politics**, using s_i' (similar to our approach in Chapter 6) dropped results to 0.6034 for **I→T** (vs 0.6214), and 0.6108 for **T→I** (vs 0.6274). Computing neighborhoods in visual rather than semantic space further dropped performance to 0.6005 and 0.6030. We also verified the importance of using weighting on both modalities. If we just used weighting on images, results were 0.6171 and 0.6168; using text only, they were 0.6028 and 0.6043. These are both lower than the 0.6214 and 0.6274 obtained by combining weights from both modalities.

7.3.3 Qualitative results

Samples weighted by measure. In Fig. 28 we show samples receiving the highest or lowest weights. For diversity, high scoring samples concern abstract subjects in which image and text play a complementary role (sad woman-“Great Depression”, American flag-“collusion”), while low-scoring ones are more concrete. For discrepancy, we observe cases where the image-text pairing is more atypical (e.g. football players-“immigration”, pride flags -“Valentine's day” and “flowers”), while low-scoring ones are again more literal (iceberg-“iceberg”, fire-“wildfires”). Finally, the most dense samples had a consistent visual appearance and were

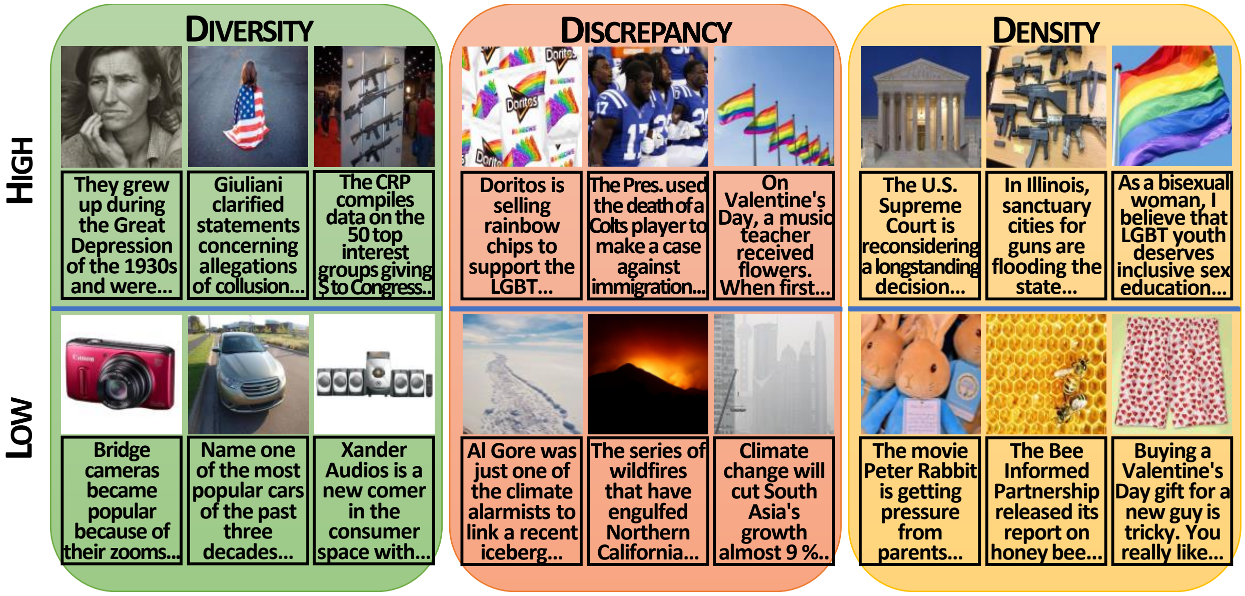

DIVERSITY
DISCREPANCY
DENSITY
HIGH
LOW
They grew up during the Great Depression of the 1930s and were...
Giuliani clarified statements concerning allegations of collusion...
The CRP compiles data on the 50 top interest groups giving $ to Congress...
Doritos is selling rainbow chips to support the LGBT...
The Pres. used the death of a Colts player to make a case against immigration...
On Valentine's Day, a music teacher received flowers. When first...
The U.S. Supreme Court is reconsidering a longstanding decision...
In Illinois, sanctuary cities for guns are flooding the state...
As a bisexual woman, I believe that LGBT youth deserves inclusive sex education...
Bridge cameras became popular because of their zooms...
Name one of the most popular cars of the past three decades...
Xander Audios is a new comer in the consumer space with...
Al Gore was just one of the climate alarmists to link a recent iceberg...
The series of wildfires that have engulfed Northern California...
Climate change will cut South Asia's growth almost 9 %...
The movie Peter Rabbit is getting pressure from parents...
The Bee Informed Partnership released its report on honey bee...
Buying a Valentine's Day gift for a new guy is tricky. You really like...

Figure 28: Example samples from Politics [341] receiving the highest or lowest weights by our measures.

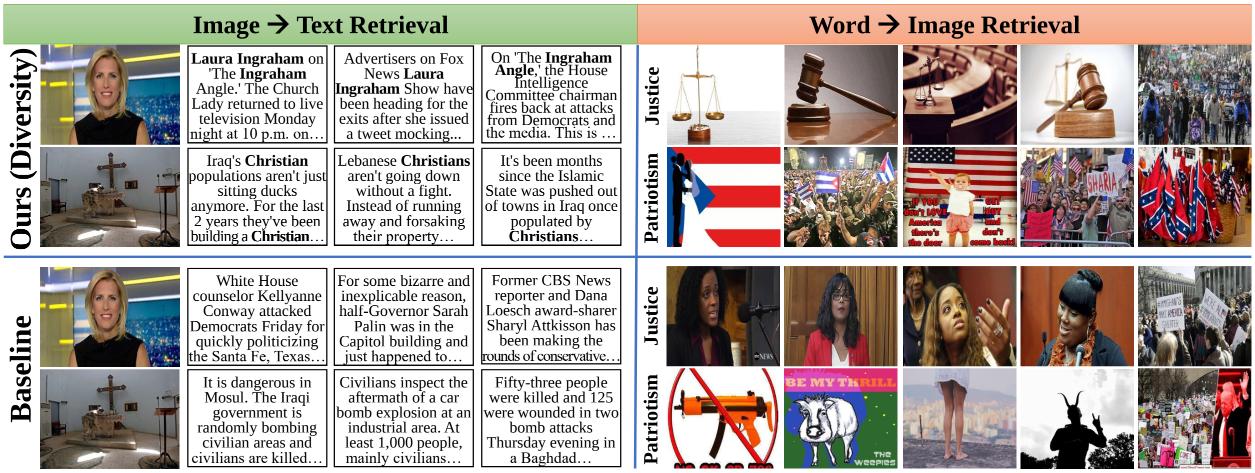


Image → Text Retrieval
Word → Image Retrieval
Ours (Diversity)
Baseline
Justice
Patriotism
Laura Ingraham on 'The Ingraham Angle.' The Church Lady returned to live television Monday night at 10 p.m. on…
Advertisers on Fox News Laura Ingraham Show have been heading for the exits after she issued a tweet mocking...
On 'The Ingraham Angle,' the House Intelligence Committee chairman fires back at attacks from Democrats and the media. This is …
Iraq's Christian populations aren't just sitting ducks anymore. For the last 2 years they've been building a Christian…
Lebanese Christians aren't going down without a fight. Instead of running away and forsaking their property…
It's been months since the Islamic State was pushed out of towns in Iraq once populated by Christians…
White House counselor Kellyanne Conway attacked Democrats Friday for quickly politicizing the Santa Fe, Texas…
For some bizarre and inexplicable reason, half-Governor Sarah Palin was in the Capitol building and just happened to…
Former CBS News reporter and Dana Loesch award-sharer Sharyl Attkisson has been making the rounds of conservative…
It is dangerous in Mosul. The Iraqi government is randomly bombing civilian areas and civilians are killed…
Civilians inspect the aftermath of a car bomb explosion at an industrial area. At least 1,000 people, mainly civilians…
Fifty-three people were killed and 125 were wounded in two bomb attacks Thursday evening in a Baghdad…
SHARIA
BE MY THRILL
THE WEEPIES

Figure 29: Retrieval results for our method and the best performing baseline on Politics [341].

mentioned consistently in the text in the same way (pride flag-"LGBT"), while the low scoring samples mentioned uncommon issues (rabbit-"Peter Rabbit").

Qualitative retrievals. In Fig. 29 we show retrievals using our diversity method vs. Amrani [7] on Politics [341]. We bold words in the text highly aligning with the image. For image to text, we observe that our method, correctly retrieves texts mentioning "Laura Ingraham" for the first image, while the baseline retrieves text mentioning women which aren't shown in the image. For the second image, both methods retrieve text about the Middle East, but the text retrieved by ours mentions Christians (which aligns with the cross in the image). For word to image, we observe that our method performs much better for abstract concepts like "justice" (ours retrieves gavels, balances, and protests, while the baseline retrieves people related to specific court cases). For "patriotism', ours retrieves flags and protests, while the baseline retrieved largely irrelevant images.

7.4 DISCUSSION

In this chapter, we proposed another method for learning visual semantic embeddings which capture abstract semantic concepts. Unlike our prior work, we explicitly encourage our model to learn abstract semantics by weighting abstract samples more in our loss (i.e. emphasizing the impact of semantically informative samples during training). However, like Chapter 6, we again leverage the hypothesized complementarity of communicative multimedia (**H3**). We propose three new techniques for focusing a retrieval method's priorities over individual image-text samples. To ensure that the retrieval method can match images and text that exhibit a complementary, rather than redundant or overlapping relationship, we highly weight examples from diverse neighborhoods and ones where the nearest neighbors of neighbors do not return the original query. We also ensure that samples are representative of the full dataset, through density estimation. We demonstrate our methods generally outperform four very recent methods on four challenging and recent datasets. Collectively, our quantitative and qualitative experimental results demonstrate our hypothesis (**H3**) that communicative multimedia substantially differs from conventional image-text datasets due to the complementary relationship of image and text. In fact, our results establish that de-emphasizing straightforward, highly cross-modally aligned samples during learning works better. Our methods exploit complementarity in order to identify image-text pairs likely containing abstract semantics and to reweight them accordingly. Our approach leverages guided training by measuring visual and semantic diversity as well as controlling for noise, which ultimately allows the model to perform substantially better, proving our hypothesis regarding such training (**H2**).

Learning semantically robust cross-modal embeddings have a variety of applications, including news curation and image captioning (beyond the literal, descriptive level) for visually impaired readers. More broadly, understanding the intricate relationship between images and text has ramifications for understanding persuasion, as well as bias, in multimodal news media sources. In particular, if a system can understand that the image included with a particular text actually contradicts the surface meaning of the text, it may detect cases of irony and mockery, and thus, detect hateful use of conventional or social media. How such detections are used is a matter of policy and not the subject of this dissertation. Instead, the goal of our work is to enable better, more nuanced, modeling of image-text relationships. Eventually, we hope methods like ours will help build more socially aware systems. We believe that in order to ensure AI systems do social good rather than harm, they need to understand subtleties, and our methods are a step in this direction.

8.0 CONCLUSION

In this dissertation, we proposed a new paradigm for studying visual media. We recast the customary computer vision objective of recognizing the contents of visual media, into the problem of computationally understanding visual communication, i.e. how visual is used as a vehicle for communication. Doing so required not only recognition of the content of images for its literal value, but also how such content is used as a type of visual argumentation and rhetoric with respect to some broader concept. These higher-level, broader abstract concepts are what we seek to model throughout this dissertation, rather than just the contents of images themselves. In practice, this involved modeling semantic concepts whose visual expression is inconsistent across different images and often where the overall image collections are noisy, limited, and/or highly visually diverse. In Chapter 3, we studied modeling photographic style in collections of photographs and demonstrated photographic style lives more in semantic than visual space. Next, in Chapter 4, we studied the problem of modeling the persuasive visual rhetoric in advertisement faces which were highly diverse and noisy and demonstrated object appearance differs across ad types. The final problems we studied extended our work by tackling more challenging types of latent semantics. In Chapters 5-7, we tackle the problem of modeling latent semantics in multimedia domains, where samples consist of images and paired text. This setting is particularly challenging because the semantics we seek to capture are latent within both domains and we are primarily interested with leveraging the cross-modal information for solely *visual* modeling. In Chapter 5, we confront the problem of recognizing politically biased images in multimedia news articles and demonstrate that political bias is exhibited visually as well as in text. In Chapters 6-7, we extend our work and provide general, task-agnostic methods for modeling high-level abstract semantic representations in multimedia in a purely data-driven manner.

We show that traditional computer vision techniques fail to model the various latent semantics in the domains we study in this dissertation when applied naïevely (confirming hypothesis **H1**), due to the nature of the problems we study and the datasets in which we study them. The recurring theme of the methods proposed herein is the use of some form of guidance or human-intuition in the model's learning procedure (**H2**). For example, we may restrict the type of learning that can occur by online allowing certain portions of models to learn, provide adjuvant information in the form of features to models to help focus them on the semantics of interest, or impose structural constraints upon the model's architecture and learning procedure to prevent models from seizing upon irrelevant aspects of the data or learning uninteresting representations. Throughout this dissertation, we repeatedly demonstrate that given the highly abstract, latent, and visually incoherent nature of the concepts we seek to model, guided models consistently outperform generic, off-the-shelf models significantly, underscoring the importance of imposing structure on the learning problem as we hypothesized (**H2**). In our later work in multimedia, we show that the paired text domain can itself provide a form of self-supervised guidance, obviating the need for any custom or external interventions (**H3**).

An ancillary theme of this dissertation is the use of generative techniques to visualize what our models have learned. Because our generative models are trained to produce synthetic data containing the latent semantics we seek to model, by viewing the synthetic images produced by the model, we can visualize in a human-understandable way what our feature representations are actually capturing in order to diagnose how well they represent various aspects of the problem. We show that our techniques of guiding the model's training procedure towards relevant features works in the generative setting as well.

We believe a particularly consequential aspect of this dissertation is its contribution towards multimodal latent semantic modeling. We not only study latent visual concepts embedded in visual media, but also their expression in text (e.g. how political bias is exhibited both visually and textually). Unlike traditional computer vision methods which assume that image and paired text provide redundant information, we view each modality as complementary and as contributing to a larger communicative message. We present numerous results showing that multimodal alignment of images and paired text significantly improves performance at modeling latent visual semantics even in the purely visual setting (i.e. even when text is not available). In particular, we show how the text paired with images can act as a sort of semantic fingerprint for the overall message of the image-text pair to guide training. We design methods which exploit the relationship of image and text in real world multimedia, and show that the text can be used to significantly improve purely visual inference by inducing visual models to capture high-level concepts. We also present methods showing how abstract semantics can be better preserved in learned cross-modal representations by exploiting complementarity, by either enforcing semantic cross-over (from text to visual space) or by weighting samples based on their semantic utility. We believe that contributions in this space are particularly apposite for better understanding online communicative multimedia, where image and text relate in indirect and subtle ways.

There are a number of limitations of our work as well as fertile possibilities for future work. Our general approach to handling very limited data was to restrict the learning of the model to pre-defined or transferred feature representations. This had the benefit of preventing overfitting, while allowing the model to reason over semantic feature representations, but this still requires leveraging external semantic resources rather than learning features directly from the dataset. In this dissertation, we used semantic attributes or object representations. However, these features may be inapposite for many problems which require different semantics in order to understand. One idea in this direction is to leverage self-supervision (such as autoencoder networks) to learn dataset-specific features, while simultaneously making use of external resources from which data is plentiful. Additionally, few-shot learning techniques could be applied in these settings. For example, it is possible that photographers of the same "school of thought" exhibit high similarity between their works. Few-shot learning techniques may be able to exploit the representation learned for a photographer with plentiful data and modify it for a similar photographer with limited data. Similarly, in Chapter 3 we proposed a probabilistic generative technique to generate novel photos in the style of particular photographers. We were unable to leverage GANs due to the lack of training data and the high-visual diversity of each photographer's work. However, there have been a number of recent proposals for learning generative networks from even single images [312], which enable one to shuffle the semantic contents of images. These approaches could be extended in order to learn the semantic concept of photographic style of a particular photographer from a collection of images (i.e. object location, position, size, etc.). Then, the model could be trained once again on a single image in order to reshuffle its contents. However, the generated images would be constrainted to be in the style of the photographer, rather than randomly shuffled. This would enable one to generate new photorealistic photographs in the style of a particular photographer. Similarly, such a method could be applied in the context of noisy ads to generate synthetic ads containing objects which obeyed the “style” for the specific ad topic.

Our work in Chapters 5-7 leverages the paired text as a form of semi-supervised learning to guide the model’s purely visual training. However, our work in these sections still have a number of limitations. In particular, though text is used to guide the models towards semantics of interest, it is still left up to our models to discover the precise relationship between images and text. For example, in the image showing a doctor taking money along with text about vaccines (Figure 2, right), the model is left to discover the relationship between the money in the image and the word “promise” in text. We believe explicitly capturing the relationship between image regions and text through structured methods [4, 202, 201] will result in more robust representations capable of capturing the type of abstract arguments made in real-world multimedia. Moreover, these structures can then be directly used to generate text explaining the visual argument made in the image, through language generation techniques. However, existing multimedia knowledge graph extraction [202, 201] methods assume correspondences between image regions and text can be discovered from visual to text parses. This is not necessarily the case in communicative multimedia, where image and text convey different information. One idea is to leverage external knowledge bases (KBs) to complete the inferential steps necessary to connect seemingly disconnected multimodal concepts. For example, we can use KBs to determine money→bribe→dishonesty. The money in the image now can be connected to the word “promise” in text (suggesting the promise will be broken and vaccines are *not* safe), along with a semantic embedding of the inference path from the KB. This approach allows us to capture much richer multimodal semantics in knowledge graphs than is currently possible.

Another limitation of our work in learning cross-modal semantic embeddings is that, despite our addition of constraints and weights to handle complementarity, we still ultimately leverage metric learning losses which assume a redundant image-text relationship. This could be a severe limitation to learning such embeddings on datasets where image and text convey even less redundant information, as the model may never be able to learn embeddings of one modality which can be projected close enough to the other while retaining their semantic representation power. Chapters 6 and 7 both assumed a cross-modal retrieval setting, where the embeddings were used to retrieve images or text from the other modality. While this ensures the embeddings are somewhat discriminative of the other modality's contents, in this dissertation we are ultimately interested in learning embeddings which preserve high-level semantics, rather than just embeddings which are useful for retrieval. For example, in Figure 2, in both cases the text's meaning is in fact the *opposite* when considered with the image. However, our current losses would not encourage each modality to preserve its specific meaning, but instead attempt to preserve their similarities. Thus, another possible extension of our work is the design of custom loss functions to replace the backbone metric learning losses which emphasize the complementarity of image and text and strongly preserve each modality's semantics, while being only secondarily concerned with matching the other modality's representation. To do so, one could impose a constraint which enforces complementarity, i.e. the joint representation of the image AND text should be semantically different than either images OR text. Other possibilities include leveraging feature disentanglement methods [216, 380, 279] for learning visual semantic embeddings which preserve the semantics of each modality, rather than focus on their overlap. Note that this is the opposite of what traditional cross-modal approaches seek to do.

We believe that the work we have undertaken has significance at several different levels. From a scientific point of view, building machines which can understand increasingly abstract semantic phenomena latent within data represents a step towards robust machine intelligence. Many of the problems we study in this dissertation, such as predicting how an image is politically biased, could be thought of as approaching "AI-completeness," i.e. requiring human intelligence to solve. Progress on such problems depends on building robust algorithms which move beyond mere image-level template matching and that are instead capable of semantic inference and generalization about what they see across the entire dataset. Such higher-level reasoning is the "holy-grail" of machine learning and is a hallmark of generalized intelligence. However, despite the inherent complexity of these problems, we show that machines can currently achieve impressive results on them, particularly when they are tailored to exploit human intuitions about the nature of the problem and by rethinking assumptions commonly held in vision (such as literal image-text alignment) which may not hold in real-world data. Our work also has numerous practical applications. For example, technology companies can use the techniques presented in this dissertation in order to better target ads or content to users based on user profiles which take into account semantics inferred from the user's interaction with the site. These could include models of user bias from multimodal data or models that are more aware of what persuasive techniques work for a particular user. Our work on multimodal political bias fulfills a particularly timely need for building more socially-aware systems. Our contributions could be deployed to detect biased content spreading online and to inform users that the content or sites they are viewing may not contain a neutral point of view of the issues. Similarly, our work on learning semantically robust cross-modal representations has numerous applications from detecting disinformation and hate speech on social media to enabling models to produce more nuanced image captions for the visually impaired.

We believe that as the public increasingly engages with machine learning algorithms throughout their daily lives, the desire for personalized content, from search results to news articles, will continue to grow. Machine learning researchers will thus be tasked with solving ever more challenging problems in order to extrapolate latent semantic concepts from noisy and diverse datasets. This may require revisiting standard assumptions in computer vision and instead viewing images as tools of communication, rather than ends in themselves. More broadly, we hope that the contributions and observations made by this dissertation will be useful guideposts for future researchers tasked with modeling such seemingly insurmountable problems, irrespective of what the problem may be. We believe that understanding visual media for its implicit communicative intents, rather than surface-level contents, is an important step to building truly intelligent reasoning systems and that our collective work presented in this dissertation is an important step in that direction.

BIBLIOGRAPHY

[1] S. Agarwal and A. Sureka. Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on tumblr micro-blogging website. *arXiv preprint arXiv:1701.04931*, 2017.

[2] H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, and P. Anderson. nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8948–8957, 2019.

[3] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2927–2936. IEEE, 2015.

[4] H. Akbari, S. Karaman, S. Bhargava, B. Chen, C. Vondrick, and S.-F. Chang. Multilevel multimodal common semantic space for image-phrase grounding. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12476–12486, 2019.

[5] Z. Al-Halah, R. Stiefelhagen, and K. Grauman. Fashion forward: Forecasting visual style in fashion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 388–397, 2017.

[6] X. Alameda-Pineda, A. Pilzer, D. Xu, N. Sebe, and E. Ricci. Viraliency: Pooling local virality. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[7] E. Amrani, R. Ben-Ari, D. Rotman, and A. Bronstein. Noise estimation using density estimation for self-supervised multimodal learning. *arXiv preprint arXiv:2003.03186*, 2020.

[8] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086, June 2018.

[9] J. Aneja, A. Deshpande, and A. G. Schwing. Convolutional image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[10] M. C. Angermeyer and B. Schulze. Reinforcing stereotypes: how the focus on forensic cases in news reporting may influence public attitudes towards the mentally ill. *International Journal of Law and Psychiatry*, 2001.

[11] G. Antipov, M. Baccouche, and J.-L. Dugelay. Face aging with conditional generative adversarial networks. In *2017 IEEE international conference on image processing (ICIP)*, pages 2089–2093. IEEE, 2017.

[12] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, December 2015.

[13] M. Arjovsky and L. Bottou. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.

[14] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

[15] R. S. Arora. *Towards automated classification of fine-art painting style: A comparative study.* PhD thesis, Rutgers University-Graduate School-New Brunswick, 2012.

[16] M. Aubry, S. Paris, S. W. Hasinoff, J. Kautz, and F. Durand. Fast local laplacian filters: Theory and applications. *ACM Transactions on Graphics (TOG)*, 33(5):167, 2014.

[17] S. Bae, S. Paris, and F. Durand. Two-scale tone management for photographic look. *ACM Transactions on Graphics (TOG)*, 25(3):637–645, 2006.

[18] N. Ballas, Y. Yang, Z.-Z. Lan, B. Delezoide, F. Prêteux, and A. Hauptmann. Space-time robust representation for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2704–2711, 2013.

[19] S. Bambach, D. Crandall, L. Smith, and C. Yu. Toddler-inspired visual object learning. In *Advances in neural information processing systems*, pages 1201–1210, 2018.

[20] Y. Bar, N. Levy, and L. Wolf. Classification of artistic styles using binarized features derived from a deep neural network. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 71–84. Springer, 2014.

[21] P. Battaglia, R. Pascanu, M. Lai, D. J. Rezende, et al. Interaction networks for learning about objects, relations and physics. In *Advances in neural information processing systems*, pages 4502–4510, 2016.

[22] E. Baumer, E. Elovic, Y. Qin, F. Polletta, and G. Gay. Testing and comparing computational approaches for identifying the language of framing in political news. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1472–1482, 2015.

[23] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding (CVIU)*, 110(3):346–359, 2008.

[24] Y. Bechavod and K. Ligett. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*, 2017.

[25] C. F. Benitez-Quiroz, R. Srinivasan, A. M. Martinez, et al. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *CVPR*, pages 5562–5570, 2016.

[26] D. Berthelot, T. Schumm, and L. Metz. Began: Boundary equilibrium generative adversarial networks. *CoRR*, abs/1703.10717, 2017.

[27] A. F. Biten, L. Gomez, M. Rusinol, and D. Karatzas. Good news, everyone! context driven entity-aware captioning for news images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12466–12475, 2019.

[28] D. M. Blei, M. I. Jordan, et al. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143, 2006.

[29] A. Blessing and K. Wen. Using machine learning for identification of art paintings. Technical report, Stanford University, 2010.

[30] L. Bogan and R. Limmer. *Journey around my room: The autobiography of Louise Bogan: A mosaic.* Viking Pr, 1981.

[31] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[32] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[33] S. Branson, G. Van Horn, and P. Perona. Lean crowdsourcing: Combining humans and machines in an online system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7474–7483, 2017.

[34] L. Breiman. Random forests. *Machine Learning*, 45(1):5, 2001.

[35] A. Broder. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of Sequences 1997*, page 21. IEEE Computer Society, 1997.

[36] J. P. Brooks. Support vector machines with the ramp loss and the hard margin loss. *Operations research*, 59(2):467–479, 2011.

[37] K. Burns, L. A. Hendricks, T. Darrell, and A. Rohrbach. Women also snowboard: Overcoming bias in captioning models. *arXiv preprint arXiv:1803.09797*, 2018.

[38] J. G. Butler. Visual style. In *The Craft of Criticism*, pages 73–85. Routledge, 2018.

[39] V. Bychkovsky, S. Paris, E. Chan, and F. Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 97–104. IEEE, 2011.

[40] Z. Bylinskii, A. Recasens, A. Borji, A. Oliva, A. Torralba, and F. Durand. Where should saliency models look next? In *European Conference on Computer Vision*, pages 809–824. Springer, 2016.

[41] J. Carpenter, D. Preotiuc-Pietro, L. Flekova, S. Giorgi, C. Hagan, M. L. Kern, A. E. Buffone, L. Ungar, and M. E. Seligman. Real men don't say "cute" using automatic language analysis to isolate inaccurate aspects of stereotypes. *Social Psychological and Personality Science*, 8(3):310–322, 2017.

[42] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4724–4733. IEEE, 2017.

[43] M. Carvalho, R. Cadène, D. Picard, L. Soulier, N. Thome, and M. Cord. Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 35–44, 2018.

[44] M. Cha, Y. Gwon, and H. Kung. Adversarial nets with perceptual losses for text-to-image synthesis. In *Machine Learning for Signal Processing (MLSP), 2017 IEEE 27th International Workshop on*, pages 1–6. IEEE, 2017.

[45] A. Chandrasekaran, A. K. Vijayakumar, S. Antol, M. Bansal, D. Batra, C. Lawrence Zitnick, and D. Parikh. We are humor beings: Understanding and predicting visual humor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4603–4612, 2016.

[46] P. Chattopadhyay, D. Yadav, V. Prabhu, A. Chandrasekaran, A. Das, S. Lee, D. Batra, and D. Parikh. Evaluating visual conversational agents via cooperative human-ai games. *arXiv preprint arXiv:1708.05122*, 2017.

[47] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua. Stylebank: An explicit representation for neural image style transfer. In *CVPR*, 2017.

[48] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *European conference on computer vision*, pages 609–623. Springer, 2012.

[49] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and

fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[50] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, 2016.

[51] X. Chen and A. Gupta. Webly supervised learning of convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1431–1439, 2015.

[52] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8), 2014*, 2014.

[53] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *To appear, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[54] R. G. Cinbis, J. Verbeek, and C. Schmid. Approximate fisher kernels of non-iid image models for image categorization. *IEEE transactions on pattern analysis and machine intelligence*, 38(6):1084–1098, 2016.

[55] R. G. Cinbis, J. Verbeek, and C. Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 39(1):189–203, 2017.

[56] R. Cohen and D. Ruths. Classifying political orientation on twitter: It's not easy! In *Seventh International Association for the Advancement of Artificial Intelligence (AAAI) Conference on Weblogs and Social Media*, 2013.

[57] E. Colleoni, A. Rozza, and A. Arvidsson. Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of communication*, 64(2):317–332, 2014.

[58] M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer. Predicting the political alignment of twitter users. In *IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and IEEE Third International Conference on Social Computing (SocialCom)*, pages 192–199. IEEE, 2011.

[59] B. Cornelis, A. Dooms, I. Daubechies, and P. Schelkens. Report on digital image processing for art historians. In *SAMPTA'09*, 2009.

[60] Y. Cui, F. Zhou, Y. Lin, and S. Belongie. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. In *Proceedings*

of the IEEE conference on computer vision and pattern recognition, pages 1153–1162, 2016.

[61] B. Dai, S. Fidler, R. Urtasun, and D. Lin. Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[62] I. Danihelka, B. Lakshminarayanan, B. Uria, D. Wierstra, and P. Dayan. Comparison of maximum likelihood and gan-based training of real nvps. *arXiv preprint arXiv:1705.05263*, 2017.

[63] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra. Visual dialog. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1080–1089. IEEE, 2017.

[64] T. Davidson, D. Warmsley, M. Macy, and I. Weber. Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*, 2017.

[65] C. Deng, Z. Chen, X. Liu, X. Gao, and D. Tao. Triplet-based deep hashing network for cross-modal retrieval. *IEEE Transactions on Image Processing*, 27(8):3893–3903, 2018.

[66] E. L. Denton, S. Chintala, R. Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pages 1486–1494, 2015.

[67] A. Deza and D. Parikh. Understanding image virality. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1818–1826, 2015.

[68] S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1657–1664. IEEE, 2011.

[69] N. Diamant, D. Zadok, C. Baskin, E. Schwartz, and A. M. Bronstein. Beholder-gan: Generation and beautification of facial images with conditioning on their beauty level. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 739–743. IEEE, 2019.

[70] L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[71] T. L. Dixon and C. L. Williams. The changing misrepresentation of race and crime on network and cable news. *Journal of Communication*, 65(1):24–39, 2014.

[72] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. Efros. What makes paris look like paris? *ACM Transactions on Graphics*, 31(4), 2012.

[73] C. Doherty and R. Weisel. A deep dive into party affiliation: Sharp differences by race, gender, generation, and education. *Pew Research Center*, 2015.

[74] Z. Dong, C. Shi, S. Sen, L. Terveen, and J. Riedl. War versus inspirational in forrest gump: Cultural effects in tagging communities. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[75] A. Dosovitskiy and T. Brox. Generating images with perceptual similarity metrics based on deep networks. In *Advances in Neural Information Processing Systems*, pages 658–666, 2016.

[76] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1110–1118, 2015.

[77] Y. Duan, W. Zheng, X. Lin, J. Lu, and J. Zhou. Deep adversarial metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2780–2789, 2018.

[78] A. Dubey, N. Naik, D. Parikh, R. Raskar, and C. A. Hidalgo. Deep learning the city: Quantifying urban perception at a global scale. In *European Conference on Computer Vision*, pages 196–212. Springer, 2016.

[79] T. B. Edsall. Studies: Conservatives are from mars, liberals are from venus, February 2012. `https://www.theatlantic.com/politics/archive/2012/02/studies-conservatives-are-from-mars-liberals-are-from-venus/252416/`.

[80] A. A. Efros and T. K. Leung. Texture synthesis by non-parametric sampling. In *ICCV*, 1999.

[81] S. Ehrhardt, A. Monszpart, N. J. Mitra, and A. Vedaldi. Learning a physical long-term predictor. *CoRR*, abs/1703.00247, 2017.

[82] C. Eickhoff. Cognitive biases in crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 162–170. ACM, 2018.

[83] M. Elad and P. Milanfar. Style transfer via texture synthesis. *TIP*, 26(5), 2017.

[84] I. A. Essa and A. P. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 19(7):757–763, 1997.

[85] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler. Vse++: Improved visual-semantic embeddings. In *British Machine Vision Conference (BMVC)*, 2018.

[86] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[87] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1778–1785. IEEE, 2009.

[88] H. Farid. Image forgery detection. *Signal Processing Magazine, IEEE*, 26(2):16–25, 2009.

[89] C. Finn, I. Goodfellow, and S. Levine. Unsupervised learning for physical interaction through video prediction. In *NIPS*, 2016.

[90] A. Fokkens, N. Ruigrok, C. J. Beukeboom, G. Sarah, and W. Van Attveldt. Studying muslim stereotyping through microportrait extraction. In *LREC*, 2018.

[91] V. Franzoni, Y. Li, P. Mengoni, and A. Milani. Clustering facebook for biased context extraction. In *International Conference on Computational Science and Its Applications*, pages 717–729. Springer, 2017.

[92] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013.

[93] A. Gaidon, A. Lopez, and F. Perronnin. The reasonable effectiveness of synthetic visual data. *International Journal of Computer Vision*, 126(9):899–901, 2018.

[94] C. Gan, T. Yang, and B. Gong. Learning attributes equals multi-source domain generalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 87–97, 2016.

[95] W. Garbe. Symspell. `https://github.com/wolfgarbe/SymSpell`, 2019.

[96] N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018.

[97] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016.

[98] L. A. Gatys, A. S. Ecker, M. Bethge, A. Hertzmann, and E. Shechtman. Controlling perceptual factors in neural style transfer. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3730–3738. IEEE, 2017.

[99] W. Ge. Deep metric learning with hierarchical triplet loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 269–285, 2018.

[100] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006.

[101] M. Gilens. Race and poverty in americapublic misperceptions and the american news media. *Public Opinion Quarterly*, 60(4):515–541, 1996.

[102] S. Ginosar, K. Rakelly, S. Sachs, B. Yin, and A. A. Efros. A century of portraits: A visual historical record of american high school yearbooks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV-W)*, pages 1–7, 2015.

[103] R. Girdhar, D. Tran, L. Torresani, and D. Ramanan. Distinit: Learning video representations without a single labeled video. *ICCV 2019*, 2019.

[104] R. Girshick. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[105] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, 2014.

[106] G. Gkioxari, R. Girshick, and J. Malik. Contextual action recognition with r* cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1080–1088, 2015.

[107] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 249–256, 2010.

[108] L. Gomez, Y. Patel, M. Rusinol, D. Karatzas, and C. V. Jawahar. Self-supervised learning of visual features through embedding images into text topic spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[109] C. Gong, H. Shi, J. Yang, and J. Yang. Multi-manifold positive and unlabeled learning for visual analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(5):1396–1409, 2019.

[110] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[111] I. J. Goodfellow, M. Mirza, A. C. Da Xiao, and Y. Bengio. An empirical investigation of catastrophic forgeting in gradient-based neural networks. In *In Proceedings of International Conference on Learning Representations (ICLR*. Citeseer, 2014.

[112] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017.

[113] J. Gu, J. Cai, S. R. Joty, L. Niu, and G. Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7181–7189, 2018.

[114] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, et al. Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354–377, 2018.

[115] M. Guerini, J. Staiano, and D. Albanese. Exploring image virality in google plus. In *Social Computing (SocialCom), 2013 International Conference on*, pages 671–678. IEEE, 2013.

[116] C. Guo, M. Zhang, Y. Liu, and S. Ma. A picture is worth a thousand words: Introducing visual similarity into recommendation. In *Intelligent Control and Information Processing (ICICIP), 2016 Seventh International Conference on*, pages 153–160. IEEE, 2016.

[117] Y. Guo, H. Shi, A. Kumar, K. Grauman, T. Rosing, and R. Feris. Spottune: transfer learning through adaptive fine-tuning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4805–4814, 2019.

[118] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016.

[119] S. Gupta, J. Hoffman, and J. Malik. Cross modal distillation for supervision transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2827–2836, June 2016.

[120] S. Gurumurthy, R. Kiran Sarvadevabhatla, and R. Venkatesh Babu. Deligan: Generative adversarial networks for diverse and limited data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 166–174, 2017.

[121] M. Gwilliam and R. Farrell. Intelligent image collection: Building the optimal dataset. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 796–805, 2020.

[122] J. Han, P. Luo, and X. Wang. Deep self-learning from noisy labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5138–5147, 2019.

[123] W. Hao, C. Li, X. Li, L. Carin, and J. Gao. Towards learning a generic agent for vision-and-language navigation via pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13137–13146, 2020.

[124] C. Happer and G. Philo. The role of the media in the construction of public belief and social change. *Journal of Social and Political Psychology*, 1(1):321–336, 2013.

[125] K. He, R. Girshick, and P. Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE international conference on computer vision*, pages 4918–4927, 2019.

[126] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.

[127] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[128] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 558–567, 2019.

[129] X. He, Y. Zhou, Z. Zhou, S. Bai, and X. Bai. Triplet-center loss for multi-view 3d object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1945–1954, 2018.

[130] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[131] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin. Image analogies. In *SIGGRAPH*, 2001.

[132] J. Hessel, L. Lee, and D. Mimno. Quantifying the visual concreteness of words and topics in multimodal datasets. In *North American Association for Computational Linguistics*, 2018.

[133] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[134] E. Hoffer and N. Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer, 2015.

[135] W. Hong, Z. Wang, M. Yang, and J. Yuan. Conditional generative adversarial network for structured domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2018.

[136] M. Honnibal and I. Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 2017.

[137] X. Hou, L. Shen, K. Sun, and G. Qiu. Deep feature consistent variational autoencoder. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1133–1141. IEEE, 2017.

[138] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko. Learning to reason: End-to-end module networks for visual question answering. *CoRR, abs/1704.05526*, 3, 2017.

[139] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko. Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[140] R. Hu, A. Singh, T. Darrell, and M. Rohrbach. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[141] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4555–4564, 2016.

[142] C. Huang, C. C. Loy, and X. Tang. Local similarity-aware deep feature embedding. In *Advances in neural information processing systems*, pages 1262–1270, 2016.

[143] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[144] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[145] X. Huang and A. Kovashka. Inferring visual persuasion via body language, setting, and deep features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 73–79, 2016.

[146] X. Huang, C. Shen, X. Boix, and Q. Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 262–270, 2015.

[147] Y. Huang, Q. Wu, C. Song, and L. Wang. Learning semantic concepts and order for image and sentence matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2018.

[148] M. Huh, P. Agrawal, and A. A. Efros. What makes imagenet good for transfer learning? *NIPS Workshop on Large Scale Computer Vision Systems*, 2016.

[149] Z. Hussain, M. Zhang, X. Zhang, K. Ye, C. Thomas, Z. Agha, N. Ong, and A. Kovashka. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1100–1110. IEEE, July 2017.

[150] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR.

[151] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.

[152] Y. Jae Lee, A. A. Efros, and M. Hebert. Style-aware mid-level representation for discovering visual connections in space and time. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1857–1864, 2013.

[153] M. Jas and D. Parikh. Image specificity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2727–2736, 2015.

[154] D. Jayaraman and K. Grauman. Zero-shot recognition with unreliable attributes. In *Advances in neural information processing systems*, pages 3464–3472, 2014.

[155] S.-H. Jeong. Visual metaphor in advertising: Is the persuasive effect attributable to visual argumentation or metaphorical rhetoric? *Journal of Marketing Communications*, 14(1):59–73, 2008.

[156] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[157] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. G. Hauptmann. Self-paced curriculum learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2694–2700, 2015.

[158] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2309–2318, 2018.

[159] M. Jiang, S. Huang, J. Duan, and Q. Zhao. Salicon: Saliency in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1072–1080, 2015.

[160] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[161] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4565–4574, 2016.

[162] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3668–3678, 2015.

[163] C. R. Johnson Jr, E. Hendriks, I. J. Berezhnoy, E. Brevdo, S. M. Hughes, I. Daubechies, J. Li, E. Postma, and J. Z. Wang. Image processing for artist identification. *Signal Processing Magazine, IEEE*, 25(4):37–48, 2008.

[164] J. Joo, W. Li, F. F. Steen, and S.-C. Zhu. Visual persuasion: Inferring communicative intents of images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 216–223, 2014.

[165] J. Joo, F. F. Steen, and S.-C. Zhu. Automated facial trait judgment and election outcome prediction: Social dimensions of face. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3712–3720, 2015.

[166] T. Kanade, J. F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46–53. IEEE, 2000.

[167] B. Kang, Z. Liu, X. Wang, F. Yu, J. Feng, and T. Darrell. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8420–8429, 2019.

[168] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemoeller. Recognizing image style. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2014.

[169] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137, June 2015.

[170] V. Kaushal, A. Sahoo, K. Doctor, N. R. Uppalapati, S. Shetty, P. Singh, R. K. Iyer, and G. Ramakrishnan. Learning from less data: Diversified subset selection and active learning in image classification tasks. *CoRR*, abs/1805.11191, 2018.

[171] D. Keren. Recognizing image "style" and activities in video using local features and naive bayes. *Pattern Recognition Letters*, 24(16):2913–2922, 2003.

[172] M. Khayatkhoei, M. Singh, and A. Elgammal. Disconnected manifold learning for generative adversarial networks. *arXiv preprint arXiv:1806.00880*, 2018.

[173] M. H. Kiapour, K. Yamaguchi, A. C. Berg, and T. L. Berg. Hipster wars: Discovering elements of fashion styles. In *European conference on computer vision*, pages 472–488. Springer, 2014.

[174] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[175] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[176] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *International Conference on Learning Representations (ICLR)*, 2014.

[177] R. Kiros, R. Salakhutdinov, and R. Zemel. Multimodal neural language models. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 595–603, Bejing, China, 22–24 Jun 2014. PMLR.

[178] J. E. Kjeldsen. Pictorial argumentation in advertising: Visual tropes and figures as a way of creating visual argumentation. In *Topical themes in argumentation theory*, pages 239–255. Springer, 2012.

[179] C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler. What are you talking about? text-to-image coreference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3558–3565, 2014.

[180] S. Kornblith, J. Shlens, and Q. V. Le. Do better imagenet models transfer better? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2661–2671, 2019.

[181] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Image search with relative attribute feedback. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2973–2980. IEEE, 2012.

[182] J. Krause, B. Sapp, A. Howard, H. Zhou, A. Toshev, T. Duerig, J. Philbin, and L. Fei-Fei. The unreasonable effectiveness of noisy data for fine-grained recognition. In *European Conference on Computer Vision*, pages 301–320. Springer, 2016.

[183] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 123(1):32–73, 2017.

[184] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012.

[185] J. Kruk, J. Lubin, K. Sikka, X. Lin, D. Jurafsky, and A. Divakaran. Integrating text and image: Determining multimodal document intent in instagram posts. In *EMNLP*, 2019.

[186] J. Kuen, Z. Wang, and G. Wang. Recurrent attentional networks for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[187] V. Kwatra, I. Essa, A. Bobick, and N. Kwatra. Texture optimization for example-based synthesis. In *SIGGRAPH*, 2005.

[188] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 36(3):453–465, 2014.

[189] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. DENOYER, et al. Fader networks: Manipulating images by sliding attributes. In *Advances in neural information processing systems (NIPS)*, pages 5963–5972, 2017.

[190] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning (ICML)*, pages 1558–1566, 2016.

[191] R. Layne, T. M. Hospedales, and S. Gong. Person re-identification by attributes. In R. Bowden, J. P. Collomosse, and K. Mikolajczyk, editors, *British Machine Vision Conference, BMVC 2012, Surrey, UK, September 3-7, 2012*, pages 1–11. BMVA Press, 2012.

[192] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1188–1196, 2014.

[193] G. Lee, Y.-W. Tai, and J. Kim. Deep saliency with encoded low level distance map and high level features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 660–668, 2016.

[194] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th annual international conference on machine learning*, pages 609–616. ACM, 2009.

[195] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018.

[196] K.-H. Lee, X. He, L. Zhang, and L. Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5447–5456, 2018.

[197] S. Lee, N. Maisonneuve, D. Crandall, A. A. Efros, and J. Sivic. Linking past to present: Discovering style in two centuries of architecture. In *IEEE International Conference on Computational Photography*, 2015.

[198] Y. J. Lee, A. Efros, and M. Hebert. Style-aware mid-level representation for discovering visual connections in space and time. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1857–1864. IEEE, 2013.

[199] H. Li, J. G. Ellis, L. Zhang, and S.-F. Chang. Patternnet: Visual pattern mining with deep neural network. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 291–299. ACM, 2018.

[200] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1378–1386, 2010.

[201] M. Li, A. Zareian, Y. Lin, X. Pan, S. Whitehead, B. Chen, B. Wu, H. zhong Ji, S.-F. Chang, C. R. Voss, D. Napierski, and M. Freedman. Gaia: A fine-grained multimedia knowledge extraction system. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[202] M. Li, A. Zareian, Q. Zeng, S. Whitehead, D. Lu, H. Ji, and S.-F. Chang. Cross-media structured common space for multimedia event extraction. In *Proceedings of The 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[203] T. Li, Z. Liang, S. Zhao, J. Gong, and J. Shen. Self-learning with rectification strategy for human parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9263–9272, 2020.

[204] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang. Universal style transfer via feature transforms. In *NIPS*, 2017.

[205] Y. Li, X. Hou, C. Koch, J. Rehg, and A. Yuille. The secrets of salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 280–287, 2014.

[206] Y. Li, L. Liu, C. Shen, and A. Van Den Hengel. Mining mid-level visual patterns with deep cnn activations. *International Journal of Computer Vision (IJCV)*, 121(3):344–364, 2017.

[207] Z. Li and D. Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

[208] D. Lin, S. Fidler, C. Kong, and R. Urtasun. Visual semantic search: Retrieving videos via complex textual queries. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2657–2664, 2014.

[209] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[210] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[211] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

[212] F. Liu, R. Ye, X. Wang, and S. Li. Hal: Improved text-image matching by mitigating visual semantic hubs. In *Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, 2020.

[213] K. S. Liu, B. Li, and J. Gao. Generative model: Membership attack, generalization and diversity. *arXiv preprint arXiv:1805.09898*, 2018.

[214] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In *Advances in neural information processing systems*, pages 469–477, 2016.

[215] P. Liu, S. Han, Z. Meng, and Y. Tong. Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1805–1812, 2014.

[216] Y. Liu, F. Wei, J. Shao, L. Sheng, J. Yan, and X. Wang. Exploring disentangled feature representation beyond face identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2080–2089, 2018.

[217] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[218] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019.

[219] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[220] M. Long, Y. Cao, and J. Wang. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning (ICML)*, 2015.

[221] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, 2004.

[222] J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.

[223] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020.

[224] J. Lu, C. Xiong, D. Parikh, and R. Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[225] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016.

[226] J. Lu, J. Yang, D. Batra, and D. Parikh. Neural baby talk. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7219–7228, 2018.

[227] F. Luan, S. Paris, E. Shechtman, and K. Bala. Deep photo style transfer. In *CVPR*, 2017.

[228] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet. Are gans created equal? a large-scale study. In *Advances in neural information processing systems*, pages 697–706, 2018.

[229] A. Makhzani, J. Shlens, N. Jaitly, and I. Goodfellow. Adversarial autoencoders. In *International Conference on Learning Representations (ICLR)*, 2016.

[230] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A deep learning approach to visual question answering. *International Journal of Computer Vision*, 125(1-3):110–135, 2017.

[231] Y. A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2016.

[232] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

[233] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley. Least squares generative adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2813–2821. IEEE, 2017.

[234] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1784–1791. IEEE, 2011.

[235] H. Masnadi-Shirazi and N. Vasconcelos. On the design of loss functions for classification: theory, robustness to outliers, and savageboost. In *Advances in neural information processing systems*, pages 1049–1056, 2009.

[236] D. Massiceti, N. Siddharth, P. K. Dokania, and P. H. Torr. Flipdial: A generative model for two-way visual dialogue. *image (referred to as "captioning")*, 13:15, 2018.

[237] D. Massiceti, N. Siddharth, P. K. Dokania, and P. H. Torr. Flipdial: A generative model for two-way visual dialogue. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[238] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. *International Conference on Learning Representations (ICLR)*, 2016.

[239] J. B. Merrill. Liberal, moderate or conservative? see how facebook labels you. *The New York Times*, Aug 2016.

[240] L. Mescheder, S. Nowozin, and A. Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. *arXiv preprint arXiv:1701.04722*, 2017.

[241] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[242] B. S. Minghui Liao and X. Bai. TextBoxes++: A single-shot oriented scene text detector. *IEEE Transactions on Image Processing*, 27(8):3676–3690, 2018.

[243] N. C. Mithun, J. Li, F. Metze, and A. K. Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 19–27, 2018.

[244] N. C. Mithun, R. Panda, E. E. Papalexakis, and A. K. Roy-Chowdhury. Webly supervised joint embedding for cross-modal image-text retrieval. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1856–1864, 2018.

[245] N. C. Mithun, S. Paul, and A. K. Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[246] A.-r. Mohamed, T. N. Sainath, G. Dahl, B. Ramabhadran, G. E. Hinton, and M. A. Picheny. Deep belief networks using discriminative features for phone recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5060–5063. IEEE, 2011.

[247] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[248] F. Morin and Y. Bengio. Hierarchical probabilistic neural network language model. In *Tenth International Workshop on Artificial Intelligence and Statistics (AISTATS)*, volume 5, pages 246–252. Citeseer, 2005.

[249] R. Mottaghi, H. Bagherinezhad, M. Rastegari, and A. Farhadi. Newtonian scene understanding: Unfolding the dynamics of objects in static images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3521–3529, 2016.

[250] C. L. Muñoz and T. L. Towner. The image is the message: Instagram marketing and the 2016 presidential primary season. *Journal of Political Marketing*, 16(3-4):290–318, 2017.

[251] N. Murray, L. Marchesotti, and F. Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2408–2415. IEEE, 2012.

[252] N. Murrugarra-Llerena and A. Kovashka. Cross-modality personalization for retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6429–6438, 2019.

[253] H. Nam, J.-W. Ha, and J. Kim. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 299–307, 2017.

[254] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4467–4477, 2017.

[255] D. Nguyen, D. Trieschnigg, A. S. Doğruöz, R. Gravel, M. Theune, T. Meder, and F. De Jong. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1950–1961, 2014.

[256] S. Nie, Z. Wang, and Q. Ji. A generative restricted boltzmann machine based method for high-dimensional motion data modeling. *Computer Vision and Image Understanding*, 136:14–22, 2015.

[257] L. Niu, X. Dai, J. Zhang, and J. Chen. Topic2vec: learning distributed representations of topics. In *2015 International conference on asian language processing (IALP)*, pages 193–196. IEEE, 2015.

[258] S. U. Noble. *Algorithms of oppression: How search engines reinforce racism.* NYU Press, 2018.

[259] H. Oh Song, S. Jegelka, V. Rathod, and K. Murphy. Deep metric learning via facility location. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5382–5390, 2017.

[260] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4004–4012, 2016.

[261] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision (IJCV)*, 42(3):145–175, 2001.

[262] A. Olteanu, C. Castillo, F. Diaz, and E. Kiciman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13, 2019.

[263] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016.

[264] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Computer Vision and*

Pattern Recognition (CVPR), 2014 IEEE Conference on, pages 1717–1724. IEEE, 2014.

[265] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 685–694, 2015.

[266] G. B. Orr. Removing noise in on-line search using adaptive batch sizes. In *Advances in Neural Information Processing Systems*, pages 232–238, 1997.

[267] J. Otterbacher, A. Checco, G. Demartini, and P. Clough. Investigating user perception of gender bias in image search: the role of sexism. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 933–936. ACM, 2018.

[268] H. Pakdaman. Updating the generator in ppgn-h with gradients flowing through the encoder. *arXiv preprint arXiv:1804.00630*, 2018.

[269] F. Palermo, J. Hays, and A. A. Efros. Dating historical color images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 499–512. Springer, 2012.

[270] Z. Pan, W. Yu, X. Yi, A. Khan, F. Yuan, and Y. Zheng. Recent progress on generative adversarial networks (gans): A survey. *IEEE Access*, 7:36322–36333, 2019.

[271] K. Pang, K. Li, Y. Yang, H. Zhang, T. M. Hospedales, T. Xiang, and Y.-Z. Song. Generalising fine-grained sketch-based image retrieval. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[272] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In X. Xie, M. W. Jones, and G. K. L. Tam, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 41.1–41.12. BMVA Press, September 2015.

[273] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *Advances in Neural Information Processing Systems Workshops (NIPS-W)*, 2017.

[274] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69, 2015.

[275] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L. Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1944–1952, 2017.

[276] T. Peck and N. Boutelier. Big political data. `https://www.isidewith.com/polls`, 2018.

[277] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek. Areas of attention for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[278] O. A. Penatti, S. Avila, E. Valle, and R. d. S. Torres. Semantic diversity versus visual diversity in visual dictionaries. *arXiv preprint arXiv:1511.06704*, 2015.

[279] X. Peng, Z. Huang, X. Sun, and K. Saenko. Domain agnostic learning with disentangled representations. In *International Conference on Machine Learning*, pages 5102–5112, 2019.

[280] Y. Peng. Same candidates, different faces: Uncovering media bias in visual portrayals of presidential candidates with computer vision. *Journal of Communication*, 68(5):920–941, 2018.

[281] M. Pennacchiotti and A.-M. Popescu. A machine learning approach to twitter user classification. In *Fifth International Association for the Advancement of Artificial Intelligence (AAAI) Conference on Weblogs and Social Media*, 2011.

[282] A. Pentina, V. Sharmanska, and C. H. Lampert. Curriculum learning of multiple tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5492–5500, 2015.

[283] M. E. Peters and D. Lecocq. Content extraction using diverse feature sets. In *Proceedings of the 22nd International Conference on World Wide Web (WWW)*, pages 89–90. ACM, 2013.

[284] G. Philo. Active audiences and the construction of public knowledge. *Journalism Studies*, 9(4):535–544, 2008.

[285] H. Pirsiavash, C. Vondrick, and A. Torralba. Inferring the why in images. Technical report, MASSACHUSETTS INST OF TECH CAMBRIDGE, 2014.

[286] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[287] G. Polatkan, S. Jafarpour, A. Brasoveanu, S. Hughes, and I. Daubechies. Detection of forgery in paintings using supervised learning. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 2921–2924. IEEE, 2009.

[288] Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, and L. Carin. Variational autoencoder for deep learning of images, labels and captions. In *Advances in neural information processing systems (NIPS)*, pages 2352–2360, 2016.

[289] S. Qiao, C. Liu, W. Shen, and A. L. Yuille. Few-shot image recognition by predicting parameters from activations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7229–7238, 2018.

[290] M. Recasens, C. Danescu-Niculescu-Mizil, and D. Jurafsky. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1650–1659, 2013.

[291] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[292] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018.

[293] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069, 2016.

[294] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

[295] R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

[296] M. Ren, W. Zeng, B. Yang, and R. Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, pages 4334–4343, 2018.

[297] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 91–99, 2015.

[298] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 1179–1195. IEEE, 2017.

[299] A. Richard, H. Kuehne, and J. Gall. Weakly supervised action learning with rnn based fine-to-coarse modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 754–763, 2017.

[300] K. Ross and C. Carter. Women and news: A long and winding road. *Media, Culture & Society*, 33(8):1148–1165, 2011.

[301] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu. Unsupervised joint object discovery and segmentation in internet images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1939–1946, 2013.

[302] R. Rudinger, C. May, and B. Van Durme. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, 2017.

[303] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[304] H. J. Ryu, M. Mitchell, and H. Adam. Improving smiling detection with race and gender diversity. *arXiv preprint arXiv:1712.00193*, 2017.

[305] B. Saleh and A. Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*, 2015.

[306] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*, 2017.

[307] D. Schill. The visual image and the political image: A review of visual communication research in the field of political communication. *Review of Communication*, 12(2):118–142, 2012.

[308] D. Schreiber, G. Fonzo, A. N. Simmons, C. T. Dawes, T. Flagan, J. H. Fowler, and M. P. Paulus. Red brain, blue brain: Evaluative processes differ in democrats and republicans. *PLOS ONE*, 8(2):1–6, 02 2013.

[309] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[310] S. Sen, M. E. Giesel, R. Gold, B. Hillmann, M. Lesicko, S. Naden, J. Russell, Z. K. Wang, and B. Hecht. Turkers, scholars, arafat and peace: Cultural communities and algorithmic gold standards. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 826–838. ACM, 2015.

[311] P. H. Seo, A. Lehrmann, B. Han, and L. Sigal. Visual reference resolution using attention memory for visual dialog. In *Advances in neural information processing systems*, pages 3719–3729, 2017.

[312] T. R. Shaham, T. Dekel, and T. Michaeli. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4570–4580, 2019.

[313] L. Shamir, T. Macura, N. Orlov, D. M. Eckley, and I. G. Goldberg. Impressionism, expressionism, surrealism: Automated recognition of painters and schools of art. *ACM Transactions on Applied Perception (TAP)*, 7(2):8, 2010.

[314] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009.

[315] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014.

[316] P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[317] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[318] Y. Shih, S. Paris, C. Barnes, W. T. Freeman, and F. Durand. Style transfer for headshot portraits. In *SIGGRAPH*, 2014.

[319] R. Sicre, Y. S. Avrithis, E. Kijak, and F. Jurie. Unsupervised part learning for visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3116–3124, 2017.

[320] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 118–126, 2015.

[321] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*, 2015.

[322] A. Singh, L. Yang, and S. Levine. Gplac: Generalizing vision-based robotic skills using weakly labeled images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5851–5860, 2017.

[323] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 73–86. Springer, 2012.

[324] C. Slade. Seeing reasons: Visual argumentation in advertisements. *Argumentation*, 17(2):145–160, 2003.

[325] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943, 2013.

[326] K. Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, pages 1857–1865, 2016.

[327] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3483–3491, 2015.

[328] Y. Song and M. Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1979–1988, 2019.

[329] H. Su, C. R. Qi, Y. Li, and L. J. Guibas. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2686–2694, 2015.

[330] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852. IEEE, 2017.

[331] J. Sun and J. Ponce. Learning dictionary of discriminative part detectors for image categorization and cosegmentation. *International Journal of Computer Vision*, 120(2):111–133, 2016.

[332] Q. Sun, Y. Liu, T.-S. Chua, and B. Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 403–412, 2019.

[333] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018.

[334] I. Sutskever, G. E. Hinton, and G. W. Taylor. The recurrent temporal restricted boltzmann machine. In *Advances in neural information processing systems*, pages 1601–1608, 2009.

[335] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the Thirty-First Advancement of Artificial Intelligence Conference on Artificial Intelligence (AAAI), February 4-9, 2017, San Francisco, California, USA.*, pages 4278–4284, 2017.

[336] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[337] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[338] D. Teney, P. Anderson, X. He, and A. van den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[339] C. Thomas and A. Kovashka. Seeing behind the camera: Identifying the authorship of a photograph. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3494–3502. IEEE, 2016.

[340] C. Thomas and A. Kovashka. Persuasive faces: Generating faces in advertisements. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[341] C. Thomas and A. Kovashka. Predicting the politics of an image using webly supervised data. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3625–3637, 2019.

[342] C. Thomas and A. Kovashka. Preserving semantic neighborhoods for robust cross-modal retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[343] Z. Thomas. Facebook content moderators paid to work from home, Mar 2020. https://www.bbc.com/news/technology-51954968.

[344] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 776–789. Springer, 2010.

[345] S. Tulkens, L. Hilte, E. Lodewyckx, B. Verhoeven, and W. Daelemans. The automated detection of racist discourse in dutch social media. *Computational Linguistics in the Netherlands Journal*, 6(1):3–20, 2016.

[346] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. Lempitsky. Texture networks: feed-forward synthesis of textures and stylized images. In *ICML*, 2016.

[347] A. van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, pages 4790–4798, 2016.

[348] L. Van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

[349] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, and S. Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 595–604, 2015.

[350] E. van Miltenburg. Stereotyping and bias in the flickr30k dataset. *arXiv preprint arXiv:1605.06083*, 2016.

[351] B. Van Rooyen, A. Menon, and R. C. Williamson. Learning with symmetric label noise: The importance of being unhinged. In *Advances in Neural Information Processing Systems*, pages 10–18, 2015.

[352] V. Veeriah, N. Zhuang, and G.-J. Qi. Differential recurrent neural networks for action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 4041–4049, 2015.

[353] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun. Order-embeddings of images and language. In Y. Bengio and Y. LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[354] S. Venugopalan, L. Anne Hendricks, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko. Captioning images with diverse objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[355] S. Venugopalan, L. A. Hendricks, M. Rohrbach, R. J. Mooney, T. Darrell, and K. Saenko. Captioning images with diverse objects. In *CVPR*, volume 3, page 8, 2017.

[356] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2015.

[357] K. Vogel, S. Shane, and P. Kingsley. How vilification of george soros moved from the fringes to the mainstream. `https://www.nytimes.com/2018/10/31/us/politics/george-soros-bombs-trump.html`, 2018. Accessed: 2020-1-15.

[358] S. Volkova, G. Coppersmith, and B. Van Durme. Inferring user political preferences from streaming communications. In *Proceedings of the 52nd Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 186–196, 2014.

[359] J. Walker, C. Doersch, A. Gupta, and M. Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 835–851. Springer, 2016.

[360] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen. Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 154–162, 2017.

[361] H. Wang, D. Sahoo, C. Liu, E.-p. Lim, and S. C. Hoi. Learning cross-modal embeddings with adversarial networks for cooking recipes and food images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11572–11581, 2019.

[362] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1393, 2014.

[363] J. Wang, F. Zhou, S. Wen, X. Liu, and Y. Lin. Deep metric learning with angular loss. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2593–2601, 2017.

[364] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013, 2016.

[365] L. Wang, Y. Xiong, D. Lin, and L. Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4325–4334, 2017.

[366] P. Wang, Q. Wu, C. Shen, A. Dick, and A. van den Hengel. Fvqa: fact-based visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2017.

[367] X. Wang, G. Oxholm, D. Zhang, and Y.-F. Wang. Multimodal transfer: A hierarchical deep convolutional neural network for fast artistic style transfer. In *CVPR*, 2017.

[368] X. Wang, A. Shrivastava, and A. Gupta. A-fast-rcnn: Hard positive generation via adversary for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[369] Y. Wang, A. Dantcheva, and F. Bremond. From attribute-labels to faces: face generation using a conditional generative adversarial network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[370] Y. Wang, H. Jiang, M. S. Drew, Z.-N. Li, and G. Mori. Unsupervised discovery of action classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1654–1661. IEEE, 2006.

[371] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 155–168. Springer, 2010.

[372] Y.-X. Wang, R. Girshick, M. Hebert, and B. Hariharan. Low-shot learning from imaginary data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7278–7286, 2018.

[373] Z. Wang, X. Tang, W. Luo, and S. Gao. Face aging with identity-preserved conditional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7939–7947, 2018.

[374] D. Wei, B. Zhou, A. Torralba, and W. Freeman. Understanding intra-class knowledge inside cnn. *arXiv preprint arXiv:1507.02379*, 2015.

[375] Y. Wei, Z. Shen, B. Cheng, H. Shi, J. Xiong, J. Feng, and T. Huang. Ts2c: Tight box mining with surrounding segmentation context for weakly supervised object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 434–450, 2018.

[376] F. M. F. Wong, C. W. Tan, S. Sen, and M. Chiang. Quantifying political leaning from tweets, retweets, and retweeters. *IEEE Transactions on Knowledge and Data Engineering*, 28(8):2158–2172, 2016.

[377] H. Wu, J. Mao, Y. Zhang, Y. Jiang, L. Li, W. Sun, and W.-Y. Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6609–6618, 2019.

[378] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163:21–40, 2017.

[379] Q. Wu, P. Wang, C. Shen, A. Dick, and A. van den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[380] F. Xiao, H. Liu, and Y. J. Lee. Identity from here, pose from there: Self-supervised disentanglement and generation of objects using unlabeled videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7013–7022, 2019.

[381] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 842–850, 2015.

[382] C. Xiong, V. Zhong, and R. Socher. Dynamic coattention networks for question answering. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[383] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 451–466. Springer, 2016.

[384] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[385] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[386] X. Yan, J. Yang, K. Sohn, and H. Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision (ECCV)*, pages 776–791. Springer, 2016.

[387] L.-C. Yang, S.-Y. Chou, and Y.-H. Yang. Midinet: A convolutional generative adversarial network for symbolic-domain music generation. *arXiv preprint arXiv:1703.10847*, 2017.

[388] S. Yang, P. Luo, C.-C. Loy, and X. Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5525–5533, 2016.

[389] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[390] K. Ye and A. Kovashka. Advise: Symbolism and external knowledge for decoding advertisements. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 837–855. Springer, 2018.

[391] K. Ye, M. Zhang, A. Kovashka, W. Li, D. Qin, and J. Berent. Cap2det: Learning to amplify weak caption supervision for object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2019.

[392] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do. Semantic image inpainting with deep generative models. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 6882–6890. IEEE, 2017.

[393] K. Yoshida, M. Minoguchi, K. Wani, A. Nakamura, and H. Kataoka. Neural joking machine: Humorous image captioning. *arXiv preprint arXiv:1805.11850*, 2018.

[394] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4651–4659, June 2016.

[395] B. Yu, T. Liu, M. Gong, C. Ding, and D. Tao. Correcting the triplet selection bias for triplet loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 71–87, 2018.

[396] H. Yu and J. M. Siskind. Grounded language learning from video described with sentences. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 53–63, 2013.

[397] Y. Yuan, K. Yang, and C. Zhang. Hard-aware deeply cascaded embedding. In *Proceedings of the IEEE international conference on computer vision*, pages 814–823, 2017.

[398] A. R. Zamir, T.-L. Wu, L. Sun, W. B. Shen, B. E. Shi, J. Malik, and S. Savarese. Feedback networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1808–1817. IEEE, 2017.

[399] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 818–833. Springer, 2014.

[400] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE international conference on computer vision*, pages 1476–1485, 2019.

[401] D. Y. Zhang, L. Shang, B. Geng, S. Lai, K. Li, H. Zhu, M. T. Amin, and D. Wang. Fauxbuster: A content-free fauxtography detector using social media comments. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 891–900. IEEE, 2018.

[402] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5907–5915, 2017.

[403] H. Zhang, J. Zhang, and P. Koniusz. Few-shot learning via saliency-guided hallucination of samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2770–2779, 2019.

[404] J. Zhang, Q. Wu, J. Zhang, C. Shen, and J. Lu. Mind your neighbours: Image annotation with metadata neighbourhood graph co-attention networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2956–2964, 2019.

[405] J. Zhang, T. Zhang, Y. Dai, M. Harandi, and R. Hartley. Deep unsupervised saliency detection: A multiple noisy labeling perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9029–9038, 2018.

[406] M. Zhang, R. Hwa, and A. Kovashka. Equal but not the same: Understanding the implicit relationship between persuasive images and text. In *British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3-6, 2018*, page 8, 2018.

[407] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep networks as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[408] X. Zhang, F. Zhou, Y. Lin, and S. Zhang. Embedding label structures for fine-grained feature representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1114–1123, 2016.

[409] Y. Zhang, P. David, and B. Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2020–2030, 2017.

[410] Z. Zhang and M. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in neural information processing systems*, pages 8778–8788, 2018.

[411] Z. Zhang, Y. Xie, and L. Yang. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6199–6208, 2018.

[412] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[413] L. Zhen, P. Hu, X. Wang, and D. Peng. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10394–10403, 2019.

[414] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016.

[415] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems (NIPS)*, pages 487–495, 2014.

[416] F. Zhou, F. De la Torre, and J. F. Cohn. Unsupervised discovery of facial events. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2581. IEEE, 2010.

[417] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng. Point to set similarity based deep feature learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3741–3750, 2017.

[418] B. Zhu, C.-W. Ngo, J. Chen, and Y. Hao. R2gan: Cross-modal recipe retrieval with generative adversarial network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[419] F. Zhu, L. Shao, J. Xie, and Y. Fang. From handcrafted to learned representations for human action recognition: A survey. *Image and Vision Computing*, 55:42–52, 2016.

[420] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[421] Y. Zhu, M. Elhoseiny, B. Liu, X. Peng, and A. Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1004–1013, 2018.

[422] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[423] Y. Zhu, J. J. Lim, and L. Fei-Fei. Knowledge acquisition for visual question answering via iterative querying. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.