SIN-Seg: A Joint Spatial-Spectral Information Fusion Model for Medical Image Segmentation

by

Siyuan Dai

Bachelor, Anhui University, 2021

Submitted to the Graduate Faculty of

the Swanson School of Engineering in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2024

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This thesis was presented

by

Siyuan Dai

It was defended on

December 1st 2023

and approved by

Wei Gao, Ph.D., Associate Professor, Department of Electrical and Computer Engineering

Peipei Zhou, Ph.D., Assistant Professor, Department of Electrical and Computer Engineering

Thesis, Advisor: Liang Zhan, Ph.D., Associate Professor, Department of Electrical and Computer Engineering

SIN-Seg: A Joint Spatial-Spectral Information Fusion Model for Medical Image Segmentation

Siyuan Dai, M.S.

University of Pittsburgh, 2024

In recent years, the application of deep convolutional neural networks (DCNNs) to medical image segmentation has shown significant promise in computer-aided detection and diagnosis (CAD). Leveraging features from different spaces(i.e. multi-modalities, Euclidean, non-Euclidean, and spectrum spaces) has the potential to enrich the information available to CAD systems, enhancing both effectiveness and efficiency. However, directly acquiring the data across different spaces is often prohibitively expensive and time-consuming. Consequently, most current brain imaging segmentation techniques are confined to the spatial domain, which means just utilizing MRI or CT images. Our research introduces an innovative Joint Spatial-Spectral Information Fusion method that requires no additional data collection. We translate existing MRI data into a new domain to extract features from an alternative space. More precisely, we apply Discrete Cosine Transformation (DCT) to enter the spectrum domain, thereby accessing supplementary feature information from an alternate space. Recognizing that information from different spaces typically necessitates complex alignment modules, we also introduce a contrastive loss function for achieving feature alignment before synchronizing information across different feature spaces. Our empirical results illustrate the effectiveness of our model in harnessing additional information from the spectrum-based space and affirm its superior performance against influential state-of-the-art segmentation baselines.

Table of Contents

List of Tables

List of Figures

Preface

I would like to express my sincere gratitude to the following individuals and organizations for their support throughout the completion of this thesis. First and foremost, I would like to thank my advisor, Dr. Liang Zhan, for his invaluable guidance, feedback, and encouragement throughout the research process within the past years of my MS studies, his dedication, and insights were instrumental in the development of my thesis and in defining the research problem. I also would like to thank my committee members, including Dr. Wei Gao, Dr. Peipei Zhou for their thoughtful feedback and insights on my research studies and my MS thesis.

I would also like to extend my sincere thanks to Dr. Haoteng Tang for his invaluable contributions to this research project. Dr. Tang's help in writing the paper and designing the experiments was crucial in the successful completion of this thesis. His daily input, support, and critical feedback were instrumental in refining my ideas and methodologies. I am truly grateful for his unwavering support and dedication to this research project.

I am sincerely appreciated to my lab mates, including Mr. Kai Ye, Mr. Kun Zhao, Mr. Junyi Li, Mr. Xiyao Fu, Mr. AmirMohammad Mijani, etc., for their assistance and encouragement in my daily works. I really enjoy my life with them during the past few years.

Additionally, I would like to express my gratitude to department of Electrical and Computer Engineering and all the faculty members and staff who have contributed to my education and research during my time in the Master's program. Their guidance, support, and expertise have been invaluable in shaping my research skills and preparing me for the next chapter of my academic journey, which has been instrumental in my success.

Specifically, I am deeply appreciative of my parents, Mr. Tangjun Dai, and Ms Chuanxia Xia, for their attentive education and firm support to my life and education. My lover, Ms Zhusuyi Chen, thank you for your love and unconditional support for more than 6 years, it is never too late to start a new journey with you, I wish the day will come soon.

Finally, I want to thank University of Pittsburgh, National Science Foundation (NSF), National Institutes of Health (NIH) Foundation etc., for their support that allowed me to

conduct the necessary research studies for this thesis.

Thank you all for your valuable contributions and for making my Master's program a fulfilling and enriching experience.

1.0 Introduction

Medical image segmentation is a critical component in the fields of biomedical science research and clinical diagnosis. Its goal is to delineate regions of interest (ROIs) that possess significant diagnostic and therapeutic value for treating physicians and radiologists. The advent of computer-aided detection/diagnosis (CAD) systems has facilitated a unified platform for analyzing vast amounts of medical-specific imaging data. (i.e. MRI, CT, Microscopy, PET, etc) Within this framework, deep neural networks (DNNs) based models have showcased their value, offering precise segmentation outcomes and reducing the time burden traditionally associated with manual analysis.

Despite the impressive achievements of DNNs[25, 33, 12, 7, 17, 41], intrinsic challenges remain to the methodologies currently in medical image segmentation. Medical images, acquired through various specialized devices, are designed to accentuate particular features or abnormalities, often requiring extra interpretative expertise of radiologists to achieve precise diagnosis. A typical CAD system that operates on images from a single type of information, without integrating such expert insight, risks overlooking critical information. Multi-modal learning in medical image analysis[10] can harness the strengths of diverse imaging modalities—such as MRI, CT, and PET—to improve diagnostic accuracy over single-modality data. Yet, the acquisition of multi-modal data for a single subject via different imaging apparatuses is seldom practical. Even though MRI devices can produce images in multiple modalities by capturing different sequence scans in a single session, potentially enhancing diagnostic effectiveness[36]. Such scanning processes require skilled radiologists or technicians and involve setting up various MRI contrast media. This not only is time-intensive but also incurs significant costs. Meanwhile, multiple modalities of imaging require patients to be exposed to radiation from MRI devices, and a typical MRI imaging is diagnosis-oriented, regular MRI images just aim at specific requirements and are captured under specific sequences.

To address this issue, we propose a novel spectrum space-based Joint Spatial-Spectral Information Fusion model (SIN). Prior researchers[46, 40] have illustrated the benefits of spectrum domain learning, particularly in edge detection—a critical element of segmenta-

tion tasks. These studies have established the validity and significance of spectral information from the frequency domain in augmenting image contrast and delineating abnormalities and pathological regions. Spectral information is particularly pivotal in MRI, CT, and Microscopy like frequency sequences-related imaging, where it reveals highly distinctive features of the same segmentation target under varied spectral-related settings during data acquisition[20].

Our SIN model innovatively harnesses both spectral and spatial domain information, synthesizing features from these two spaces. It comprises two primary components: an offline discrete cosine transform (DCT) module and an online trainable feature alignment module, both of which are embeddable and compatible with every encoder-decoder-based segmentation architecture. Meanwhile, since the DCT transformation is color-sensitive and microscopy imaging is somehow captured under RGB space but not like the other two imaging devices are imaged under g*ray-scale* space. So for those three channel-based microscopy images, we implement a space transformation from RGB color space to the $YCbCr$ color space, leveraging the fact that the feature is more sensitive to changes in brightness than color changes, resulting in more efficient form further image processing.

In detail, within the DCT module, spatial space images are partitioned into patches to capture more fine-grained details during the DCT process. A sophisticated DCT workflow is then applied to each patch to generate its spectral representation. However, aligning feature maps from disparate domains, each rooted in different spaces presents a significant challenge, often necessitating complex modules for integration[37, 22]. To overcome this, we implemented a contrastive learning strategy to align the features and fuse within the shared space effectively.

To sum up, our main contributions to this paper can be shown as follows:

- We propose a novel dual information extraction module for fusing the information both from the spectral and the spatial space.
- We introduce a low-dimension flattened strategy for the information from different spaces combined with a contrastive loss for feature alignment.
- We verify our proposed model on multiple datasets from different base imaging types, involving a brain tumor segmentation dataset[27] and a heart segmentation dataset[1]

both from MRI devices, a liver segmentation dataset[18] captured under CT devices, and a cell segmentation dataset[26] from different microscopy imaging methods. According to our comprehensive experiments, our method achieves the state-of-the-art on all UNet-based methods and is also competitive with the Transformer-based model, highlighting its effectiveness and potential for advancing medical image analysis.

2.0 Related Work

In this section, we briefly review the previous works in three different aspects highly related to our works. First, we introduce the improvement of the backbone model. Then, how spectral information shows its significance in the computer vision tasks and the potential for introducing a joint model for medical image segmentation. Finally, some previous works about how to combine and take advantage of different feature spaces, and some alignment strategies are also illustrated.

2.1 UNet-based Medical Image Segmentation Frameworks

Semantic segmentation is always a crucial task for the computer vision domain. FCN[25]] is the first research introduced Convolutional Neural Network(CNN) for segmentation. Then the UNet[33] took advantage of the encoder-decoder-like architecture, initially introduced for biomedical image segmentation, which had revolutionized the field of medical image analysis. Its unique design, characterized by a symmetric expanding path and a contracting path, allows for precise localization and context capture, making it highly effective for tasks like segmentation in medical imaging. Based on such a powerful backbone, more researchers focus on making it better for advanced segmentation frameworks. Attention mechanisms then came to researchers' eyes and made great progress in the computer vision domain[29, 9]. Then Zhang, et. al[44] introduced an attention mechanism to U-Net and utilized the superiority of ResNet, proposing Res-UNet. But such hard attention is not trainable so that limits the efficiency. Att-UNet[30] then designed with a gate module for attention calculation and enhancing the performance of the original UNet. Zhou, et. al[47] noticed that the conventional UNet-based model just use the skip connection under the same level of the features, so they proposed UNet++ for combining the features from different level of features, and a deep-supervision mechanism could promote the supervised learning a lot. In recent years, Vision Tranformer[11, 19] has shown great potential in the computer vision

field. Then TransUNet[2] introduced a vision transformer after a down-sample so that could combine the spatial semantic and the local semantic in consideration in the hidden space. Furthermore, some other research[13, 16] also show their significance in UNet-based medical image segmentation.

2.2 Spectral Information

Conventional computer vision algorithms mainly consider the image analysis in the spatial space, i.e. the *RGB* or *Gray-Scale* images which are easily recognized by human eyes. However the information in such space could obscure lots of detailed features, and some research works[4, 34] have found that when processing a visual scene, animals have more wavebands than humans because of their unique ability to spot the features in the spectral domain. When deep learning and the DCNN frameworks show their power in the computer vision fields, the huge number of parameters and such large models always own redundant information. So, utilizing the DCT transformation and mapping the original images to the spectral space is a good way to compress the images[8, 5, 40] and also the designed networks[3, 24, 38] themselves. Information in the spectral space could utilized for compression because the significant semantic information is easily extracted in such a feature space. The attention mechanism is used to force the networks to focus on the most important part of the feature maps so that it is natural to take advantage to use the spectral information for designing the attention pipelines. Qin, et. al[31] found that many works have used global average pooling (GAP) as an unquestionable preprocessing method for designing the channel attention mechanisms. A potential problem, however, is that different channels may have the same mean value, while their corresponding semantic information may be completely different, which creates the problem of insufficient attention information. So they proved that GAP is a special case of DCT, which is equivalent to the lowest frequency component of DCT and is generalized to the frequency domain, proposing a multi-spectral channel attention framework. What's more, FSDR model[14] tried to achieve domain generalizable for networks, designing a novel attention pipeline based on the information in the spectral space

for forcing the network to learn less domain-related features but more for intrinsic semantic features which are non-related to various domains. Also, spectral space could show important help on some low-level vision scene tasks[23, 45], implementing the DCT or Wavelet transformation.

2.3 Feature Alignment

Like how human beings perceive the world through different organs, it is obvious that more information could also train more powerful neural networks. However, the information in the different spaces always obstacles each other before aligning into the same feature space. For naive feature fuse, some simple operation, e.g. *Concatenation* always be considered, it is widely used in Residual Block, Skip-connection, and related fusion situations. Under this operation, multiple feature maps are spliced together in the depth dimension to obtain a richer representation of features. For example, in encoders and decoders, low-level features in the encoder and high-level features in the decoder are spliced, which improves the perceptual ability of the decoder. As more parameters could learn more feature information, the concatenation-like direct fusion method is not learnable. which will give other learnable blocks more pressure to handle such feature fusion. Research in the multi-modal learning fields always needs to consider such problems more[43, 21]. Autonomous driving[15] is a typical computer vision task that needs multi-modal features. In complex driving environments, a single sensor is not enough to effectively handle changes in the scene. For example, in extreme weather (heavy rain, sandstorms) where visibility is low, the RGB images fed back by the camera alone are not enough to provide feedback on the changes in the environment. In the ordinary road environment, such as traffic lights, color cones, etc., relying only on Lidar's information can not be effectively recognized, and also needs to be combined with the RGB information brought by the camera, to effectively deal with. Tan, et. al[35, 37] tried to introduce multimodal-learning in the medical image segmentation task, they consider using the different medical imaging devices to do the scanning on the same organ and utilize the information from different modalities. They all used extra complex modules to handle such

feature fusion challenges, but they consume too much memory, and capturing such data is impossible in a real clinical situation. Contrastive learning[42, 28] is an efficient and simple way to align and merge the data or feature maps from different feature spaces. So we also introduce a simple contrastive learning strategy to do the alignment.

3.0 Methodology

In this section, we present our proposed SIN model which introduces the spectral information and integrates it with spatial information for segmentation tasks. We first propose a novel off-line DCT transformation module to convert the image from the spatial space to the spectrum space. We then introduce a trainable alignment module with a simple contrastive loss function to align the features yielded from the spectral space and the spatial space as well. Finally, we illustrate the whole segmentation framework (named SIN-Seg) with our proposed SIN model and the loss functions for brain tumor segmentation tasks.

3.1 Off-line DCT Transformation

3.1.1 RGB Images

Microscopy Imaging always generates into the RGB space, which is not suitable for conducting DCT transformation directly on RGB images (denoted as X^{RGB}). Instead, we first transform them to the $YCbCr$ space as $YCbCr$ images (denoted as X^{YCbCr}). This conversion is crucial for two main reasons: Human Visual Sensitivity: YCbCr separates an image into luminance (Y) and chrominance (Cb and Cr). Since human vision is more sensitive to luminance than chrominance, this separation allows for more effective compression. The luminance channel can be preserved with higher fidelity, while the chrominance channels can be compressed more, reducing file size without noticeably impacting image quality. Compression Efficiency: The DCT is more effective in the YCbCr space for compression purposes. It allows for significant data reduction in the chrominance components, which is less perceptible to the human eye while maintaining the crucial details in the luminance component. After such pre-processing, the color information of luminance and chrominance is separated into three channels including Y (i.e., luma or brightness), Cb (i.e., blue-difference chroma), Cr (i.e., red-difference chroma). The $YCbCr$ transformation leverages the fact that the hu-

man visual system is more sensitive to changes in brightness than color changes, resulting in more efficient image processing. To implement the $YCbCr$ transformation, we first normalize image RGB values to the range of $[0, 1]$ with their own min-max values subjects by subjects because of the difference intensity scale for different capturing institutions, and then convert the normalized RGB values to the $YCbCr$ color space as follows:

$$X^{YCbCr} = \begin{cases} Y & = 0.299R + 0.587G + 0.114B \\ Cb & = -0.169R - 0.331G + 0.500B + 0.5 \\ Cr & = 0.500R - 0.419G - 0.081B + 0.5 \end{cases}, \tag{3-1}$$

where R, G, and B represent the intensity values in the three channels (i.e., red, green, blue) of RGB images, respectively, while Y, Cb, and Cr represent the intensity values in the three channels of $YCbCr$ images. The whole pipeline under RGB space for DCT transformation is shown as Figure 2. So that we could get the $YCbCr$ images (i.e., $X_{pc}^{YCbCr} \in \mathcal{R}^{H \times W \times C}$, where H and W denote image size, and C denotes the three channels of such color space) are generated, and the feature map in every channel will be implemented DCT transformation channel by channel.

3.1.2 Gray-Scale Images

MRI or CT images are naturally captured as a 3D volume but not like general RGB images, so we cannot simply presume the intensity scale is all in $[0, 255]$. Since the these types of dataset is collected by different institutions from different patients, we normalize them one patient by one patient with the min-max value of themselves to do subject-wise min-max normalization (SN), mapping them to the same intensity scale in the spatial domain. After normalization, we slice all the 3D MRI volumes into 2D images (i.e., $X_{pc}^{2D} \in \mathcal{R}^{H \times W}$, where H and W denote image size) are generated, we conduct DCT transformation to convert them into the spectrum domain for another modality (see Figure 3).

3.1.3 DCT Transformation in Patches

Particularly, the DCT transformation is conducted on the 8×8 patches of *2D* images, to extract more fine-grained features in the spectrum domain. The DCT transformation (i.e., $\tilde{X}_{pc} \in \mathcal{R}^{1\times 8\times 8}$) on every *2D* image patch is computed as follows:

$$\begin{aligned} \tilde{X}_{pc}\left(i,\ j\right) = & \frac{2}{\sqrt{\left(N_1,\ N_2\right)}} \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} X^{2D}\left(n_1,\ n_2\right) \cdot \\ & a_{n1} a_{n2} cos\left[n_1 \frac{2\pi}{n_1}(n_1+\frac{1}{2})\right] cos\left[n_2 \frac{2\pi}{N_2}(n_2+\frac{1}{2})\right] \\ s.\ t.\ a_{n1},\ a_{n2} = & \begin{cases} \frac{1}{\sqrt{2}}, & k=0 \\ 1, & k \neq 0, \end{cases} \end{aligned} \tag{3-2}$$

where i, j, n_1, n_2 are in range of $[0, 7]$ so that $N_1 = N_2 = 8$, $a_{n1},\ a_{n2}$ are the constant coefficient. To collect the spectrum information along $2D$ images and patches, the $\tilde{X}_{pc}$ is flattened according to the frequency (F^2) from $1 \times 8 \times 8$ to the size of $64 \times 1 \times 1$, while the first number represents the channel and the last two refer as the length and the width, and every channel refers as the feature in different frequency under the spectrum space. We first group the spectrum information from all image patches and generate the channel-wise DCT coefficient cube as $\tilde{X}_c \in \mathcal{R}^{64\times H/8\times W/8}$. Since the intensity value after DCT transformation would be mapped to a high range of scale in different frequency feature representations which is difficult for neural networks to handle and learn, we then implement another frequency-wise normalization (FN) channel by channel for every DCT coefficient cube and let them in the range of $[0, 1]$.

3.2 Segmentation Framework with Feature Alignment

Segmentation Framework. As shown in Figure 1, we utilize *U-Net* as the backbone of our SIN-Seg framework. *U-Net* [33] is a widely used segmentation backbone that has shown convincing and robust performance on a large variety of medical image segmentation

tasks. Here we adopt all default configurations used in the official implementations [1] with the input of 2D *RGB* images. Meanwhile, an extra encoder (i.e., the encoder of *U-Net*) is utilized to embed the DCT coefficient cube simultaneously. A shape alignment (the SA block in Fig. 1, combined with a dimension alignment by the Up block and a channel alignment by the CA block), is operated on the DCT coefficient cube before it goes through the encoder. In the same output scale of the *U-Net* encoder, the feature maps of the original image and DCT coefficient cube are concatenated as a fused feature map.

Feature Alignment. We propose a new contrastive alignment module and conduct the feature alignment after the last down-sample of the *U-Net* encoder. Particularly, we first utilize an MLP layer to flatten the feature maps into a feature band with a size of 1×512. Denote the feature band in the frequency domain and spatial domain as $\tilde{F}$ and F, respectively. An alignment matrix (AM) can then be constructed as $F_{align} = F^{\top}\tilde{F} \in \mathcal{R}^{512\times512}$. Inspired by CLIP[32], they proposed a novel Dual-Modality Learning, which forces their CLIP model to learn from two modalities: images and text. It employs two neural networks, one for processing images and another for processing text. The goal is to map these two different types of data into a shared embedding space where they can be directly compared. This function operates by pulling the embeddings of matching image-text pairs closer together in the shared space while pushing non-matching pairs apart. For instance, an image of a dog and its correct textual description "A dog playing in the park" are pulled closer, whereas mismatches like the same image with the text "A cat sleeping" are pushed apart. This means it can understand and categorize images it has never seen during training, based solely on its learned associations between text and images. For a broader explanation, such a contrastive training strategy could align the correlated features from different spaces to be in a shared new feature space, and those un-correlated features to be pushed away in this new space. In our module, we assume that the corresponding features (i.e., $\tilde{F}_{:,:,i}$ and $F_{:,:,i}$) are more correlated, while the non-corresponding features (i.e., $\tilde{F}_{:,:,i}$ and $F_{:,:,j}$) are less correlated. In other words, the diagonal elements in F_{align} should be dominated. So that all the non-corresponding features would be regarded as negative samples while those corresponding features are positive samples. To this end, a Binary Cross Entropy (BCE) loss is

[1] `https://github.com/milesial/Pytorch-UNet`

proposed to achieve this contrastive alignment process, and the loss function is as follows:

$$L_{Align}(\tilde{X},\ X^{RGB}) = BCE(F_{align},\ E), \tag{3-3}$$

where E is a diagonal matrix (DM) with the size of 512×512.

Loss Function. The loss function for our proposed SIN-Seg framework consists of two parts, including the segmentation loss and the proposed feature alignment loss. Following previous methods [39, 6], we use BCE loss and Dice loss together as the segmentation loss. Therefore, the whole loss function is formulated as:

$$L_{total} \ = \ Ł_{BCE} + L_{Dice} + L_{align}, \tag{3-4}$$

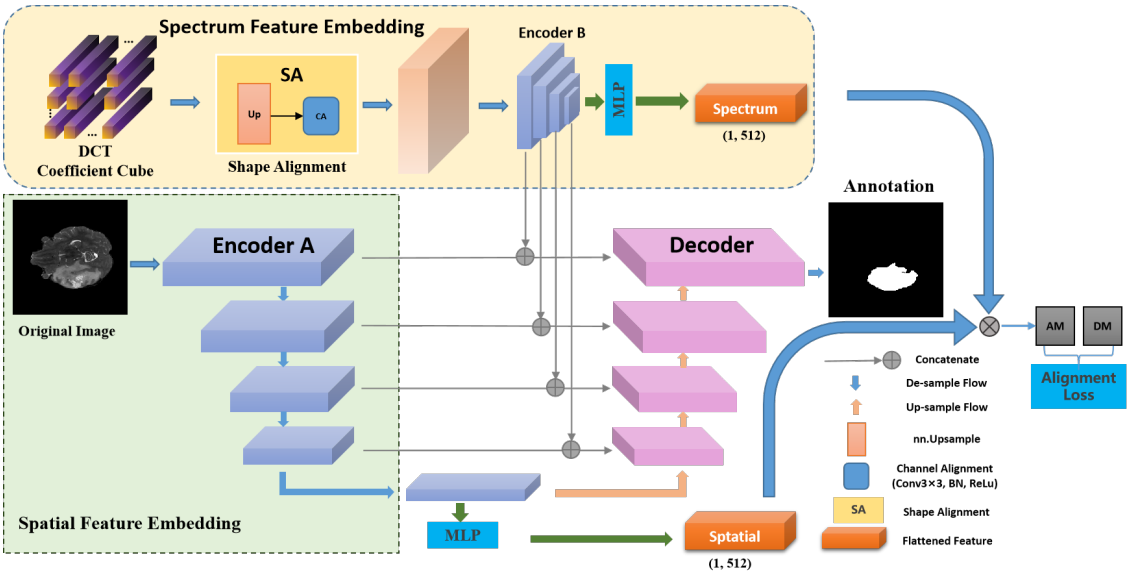

Figure 1: Diagram of the proposed SIN-Seg framework, including two U-Net encoders for the original image in the spatial space and the DCT coefficient cube embedding in the spectrum space, respectively. The coefficient cube is first up-sampled and channel adjusted via the shape-alignment process, to make the input shape aligned to the feature in the spatial space. The features from both encoders are synchronized scale-by-scale. The fused features then are fed forward to the *U-Net* decoder to generate final predicted segmentation masks. Meanwhile, a feature alignment is implemented on the flattened frequency and spatial latent features with the alignment loss.

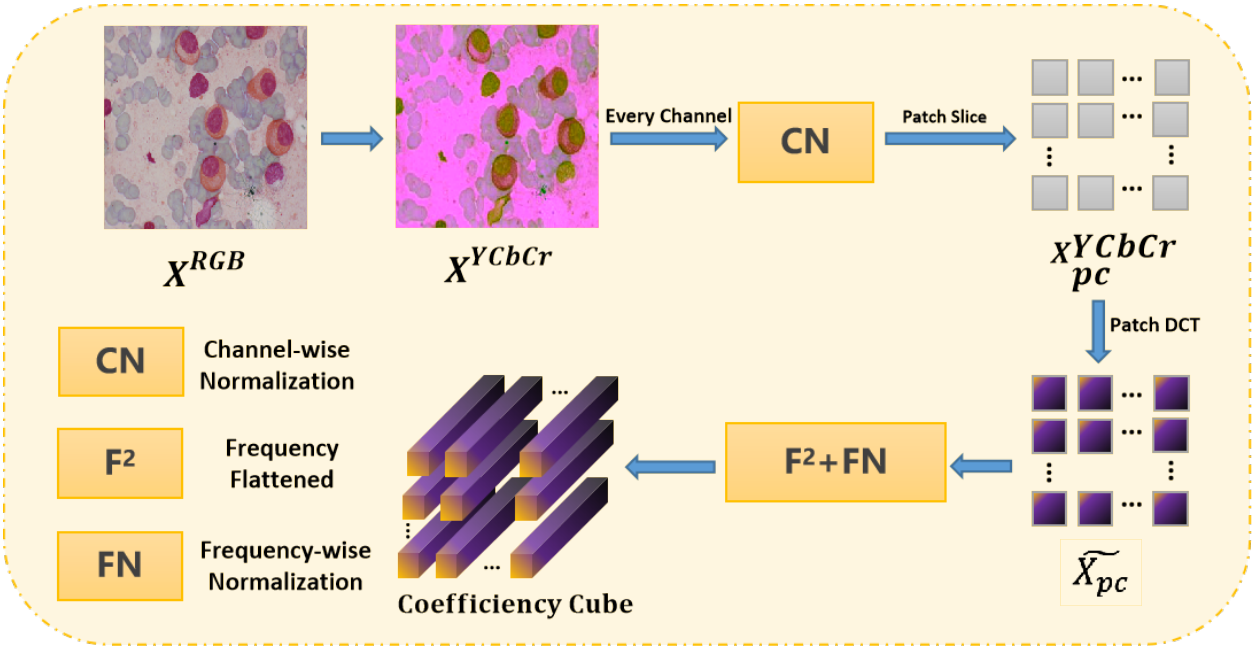

Figure 2: An overview of the off-line DCT transformation module in the SIN model for RGB space. First, a RGB image is converted to the $YCbCr$ domain. Then the $YCbCr$ image is divided into small image patches with a channel-wise normalization (CN). Next, a DCT transformation is implemented on image patches. Finally, the coefficient cube for the whole image is generated from frequency-based flattened (F^2) and frequency-wise normalization (FN) operations.

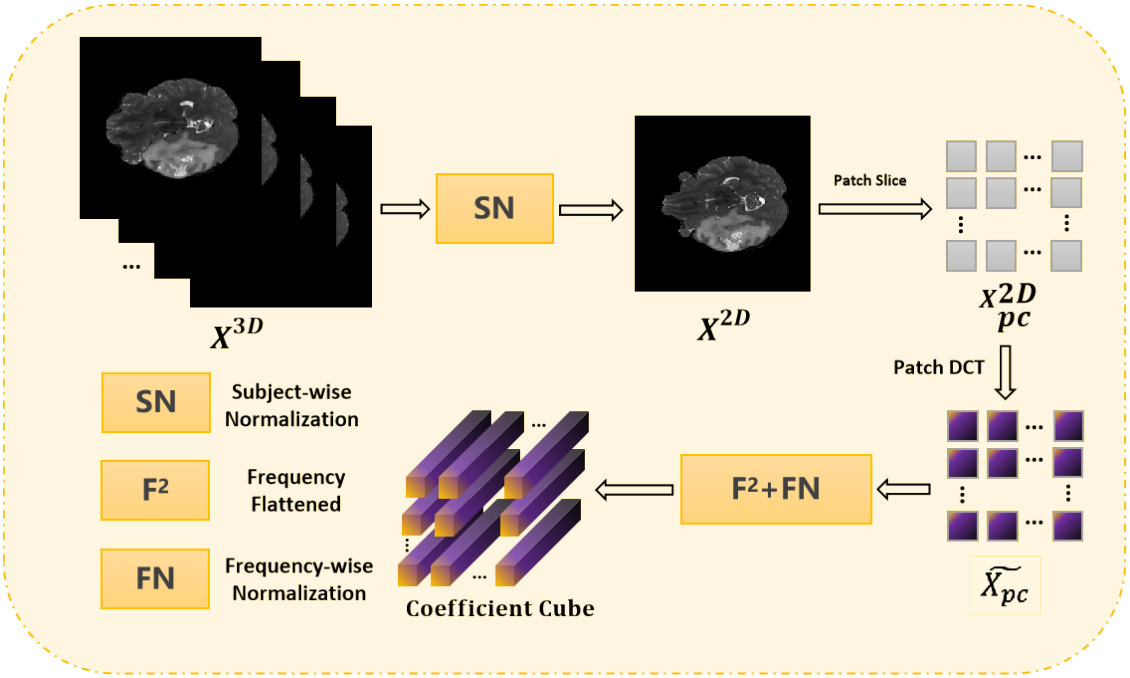

Figure 3: An overview of the off-line DCT transformation module in the SIN model for *gray-scale* space. First, every original 3D-volume brain MRI image is normalized (SN) by the min-max values of itself in the spatial space, and then every 2D slice is partitioned into small image patches. Next, a DCT transformation is implemented on image patches. Finally, the coefficient cube for the whole image is generated from frequency-based flattened (F^2) and frequency-wise normalization (FN) operations.

4.0 Experiments

4.1 Datasets

We use four publically available datasets captured from different commonly used medical imaging devices, including the NeurIPS CellSeg 2022(CellSeg) dataset[26], the CHAOS-CT abdominal organ segmentation (CHAOS-CT) dataset [18], the medical segmentation decathlon heart (MSD-Heart) dataset [1], and a brain tumor segmentation dataset BraTS 2015[27]in this study. In this study, the effects of subjects' age, gender, race, or any other variables on the results are not evaluated since the related information is not provided by the data provider. Details of data description and preprocessing are shown below.

4.1.1 Cell Segmentation

The CellSeg dataset consists of 1000 microscope 2D image slices (i.e., 900 slices training and 100 slices testing) collected from 10 different organizations. It is a specialized dataset designed for advancing research in the field of cellular image analysis, aiding in understanding cellular structures and functions, characterized by its diversity and complexity. It includes a wide range of images capturing various types of cells under different imaging conditions. This variety is essential for developing and testing algorithms that are robust and generalizable across different cell types and imaging modalities. All slices were manually labeled with 11 segmentation regions, such as yeast, adipocyte, brain cell, etc.

4.1.2 Liver Segmentation

The CHAOS CT Liver dataset is a specialized collection of medical images designed for the evaluation and development of computer-aided diagnosis systems, particularly focusing on liver segmentation from CT (Computed Tomography) scans. This dataset is part of the CHAOS challenge (Combined (CT-MR) Healthy Abdominal Organ Segmentation). The dataset comprises a series of abdominal CT scans, providing a comprehensive view of the

liver and surrounding organs. These scans are sourced from different patients, offering a diverse range of liver shapes, sizes, and pathologies. Such diversity is crucial for developing robust segmentation algorithms that can perform accurately across varied clinical scenarios. It consists of 2875 CT slices from 40 different patients collected by the DEU hospital, where the liver regions were manually labeled by expert radiologists, ensuring they can accurately identify and outline the liver in CT images.

4.1.3 Heart Segmentation

The MSD-Heart dataset is part of the Medical Segmentation Decathlon (MSD), a comprehensive collection of datasets aimed at advancing the field of medical image segmentation. Specifically, the MSD-Heart dataset focuses on the segmentation of cardiac structures from MRI (Magnetic Resonance Imaging) scans, This dataset includes a series of MRI scans that capture detailed images of the heart. These scans are sourced from a diverse patient population, encompassing a wide range of heart shapes, sizes, and pathologies. Such diversity is crucial for developing segmentation algorithms that are robust and effective across different patient demographics and clinical conditions. It consists of 2272 MRI slices from 30 subjects, where the experts manually labeled the left atrium.

4.1.4 Brain Tumor Segmentation

The BraTS2015 (Brain Tumor Segmentation 2015) challenge dataset is a significant resource in the field of medical image analysis, particularly for brain tumor segmentation. This dataset was developed for the BraTS challenge, an annual competition that focuses on the segmentation of gliomas, a common type of brain tumor, from multimodal MRI scans. BraTS2015 consists of a collection of MRI scans from multiple patients, featuring various types and stages of gliomas. The dataset includes four different MRI modalities: T1, T1-contrast enhanced, T2, and FLAIR (Fluid Attenuated Inversion Recovery), providing a comprehensive view of the tumor and surrounding brain tissues. This multimodal approach is crucial for accurately identifying and delineating tumor boundaries, as different tumor parts may be more visible in one modality than in others. We used the T2 modality for

experiments, including 35 3D MRI images. We generate 5000 2D image slices from these 3D MRI images for tumor segmentation, where 80% and 20% of image slices are utilized for framework training and validation, respectively.

4.2 Implementation Details

We first resize each image to a size of 128×128 by bilinear interpolation for network training, with training epochs as 200 and 50 epoch for early stop patience. We trained the module by using the Adam optimizer with a batch size of 20 and synchronized batch normalization. The initial learning rate was set to $1e^{-3}$ and decayed by $(1 - \frac{current\_epoch}{max\_epoch})^{0.9}$ with an l_2 weight decay of $5e^{-4}$. We randomly split the dataset into 5 sub-datasets for 5-fold cross-validation. All experiments were conducted based on PyTorch 1.7.1 and were deployed on a workstation with 2× NVIDIA TITAN RTX GPUs.

4.2.1 Baselines and Evaluation Metrics

We compared our proposed SIN-Seg framework with 4 influential U-Net based segmentation baselines, *i.e.*, *U-Net* [33], UNet++ [47], ResUNet [44], and AttUNet [30]. *U-Net* is a cutting-edge backbone framework for medical image segmentation, and UNet++, ResUNet, and AttUNet are three well-performing segmentation frameworks based on the *U-Net* backbone. We adopt two metrics to assess the performance of segmentation methods, including the Dice similarity coefficient (DSC, see as Eq.4-1), which are overlap-based metrics ranging from 0 to 1 and mean intersection over union (IoU, see as Eq.4-2), while X represents the set of pixels in the first segmentation (e.g., the algorithm's output), Y represents the set of pixels in the second segmentation (e.g., the ground truth). $|X \cap Y|$ is the cardinality of the intersection of sets X and Y (i.e., the number of pixels common to both segmentations). $|X| + |Y|$are the cardinalities of sets X and Y, is the cardinality of the union of sets, $|X_i \cup Y_i|$ is the cardinality of the union of sets X_i and Y_i for the i_th class (i.e., the total number of pixels in both the predicted and ground truth segmentations for that class). respectively

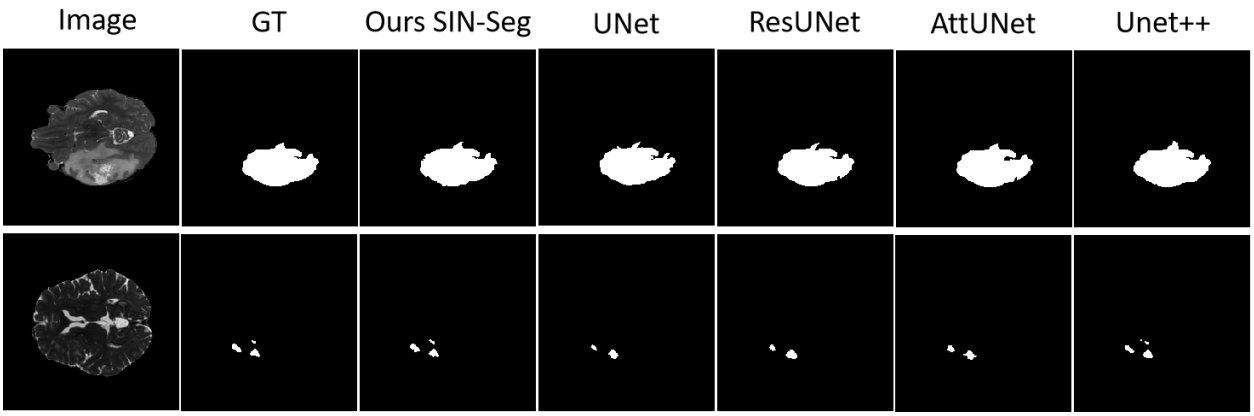

Figure 4: Visualization of the representative segmentation results produced by our frameworks and all competing baselines on the dataset. The first column represents when the lesion is big and the second column illustrates the situation when the tumor is small and discrete distributed.

(i.e., the total number of pixels in each segmentation).

$$\mathrm{DSC} = \frac{2 \times |X \cap Y|}{|X| + |Y|}, \tag{4-1}$$

$$\mathrm{mIoU} = \frac{1}{N} \sum_{i=1}^{N} \frac{|X_i \cap Y_i|}{|X_i \cup Y_i|}, \tag{4-2}$$

4.3 Comparative Experiments

Table. 2 provides the brain tumor segmentation performance of five baseline methods and our SIN-Seg. It shows that our method outperforms all baselines substantially in terms of both metrics on the dataset. Compared to the *U-Net*, our model achieves clearly superior segmentation results, which tends to show the importance of introducing the spectrum information as a complementation of the spatial information in deep neural networks for segmentation tasks. (Details are shown in the ablation study.) The comparison between

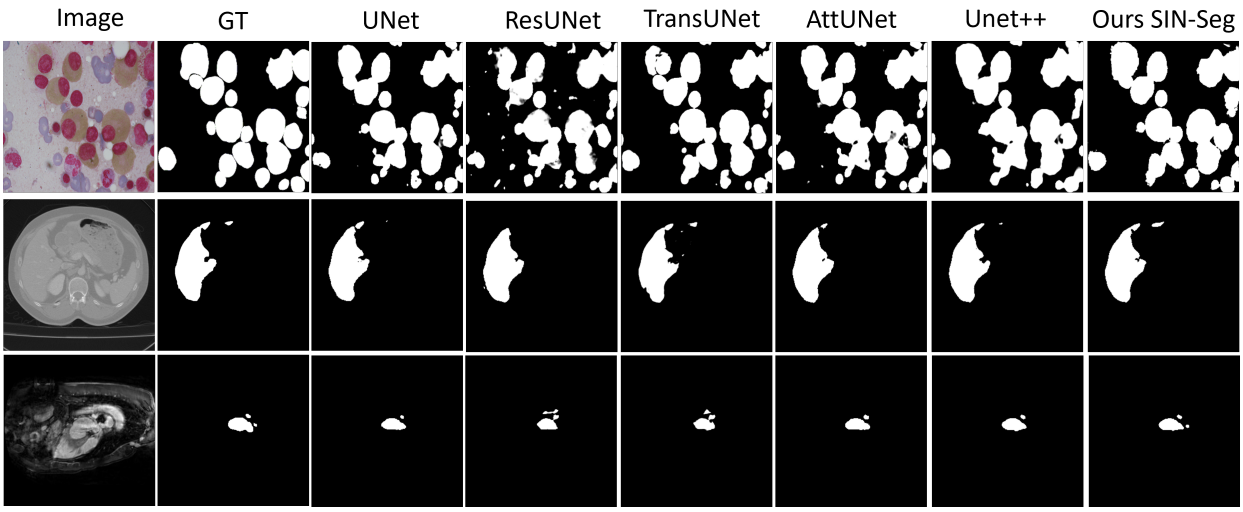

Figure 5: Visualization of the segmentation results produced by our frameworks and all competing baselines on the CellSeg (row 1), CHAOS-CT (row 2), and MSD-Heart (row 3). It better view with colors and zooming in.

SINSeg and SINSeg without feature alignment indicates the contributions provided by the proposed feature alignment loss. We also visualized the qualitative segmentation results for the BraTS dataset in Figure 5. And for the other three dataset, the visualization results is shown in Figure **??** It reveals that the results produced by our SIN-Seg framework are much more similar to the ground truths than those generated by *U-Net*, UNet++, ResUNet, and AttUNet, especially for some detailed edge.

We also implemented further experiments on the other three datasets and also made a comparison with the TransUNet model. The results are shown in Table. 1

4.4 Ablation Study

We conduct an ablation study on the dataset to evaluate the necessity and importance of each component in our framework. Table 4 shows that our SIN-Seg framework improves

Table 1: Quantitative results of different methods the dataset. The best and second best results are shown in red and blue, respectively. The values of DSC and IoU are in percentage terms.

Methods	BraTS			
	DSC/Train	IoU/Train	DSC/Val	IoU/Val
U-Net	74.13	69.29	66.36	50.22
ResUNet	76.02	70.26	65.72	59.94
AttUNet	75.64	72.15	66.31	62.26
UNet++	78.22	74.53	69.11	65.35
SINSeg	80.21	78.42	75.35	72.16

the DSC and IoU substantially compared with U-Net by just using pure spatial or spectral information for the BraTS2015 dataset, which is due to insufficient information. The further ablation experiments are also conducted under the other three datasets, the results is show as Table. 3 Therefore, both spatial and spectral information play important roles in medical image segmentation. But a naive combination of the information from different spaces is also unreasonable. One simple U-Net model obviously cannot handle two types of information space. Alignment of the features and mapping them into a shared space for synchronization is crucial, otherwise, the performance would even be worse.

Table 2: Quantitative results of different methods on the other three datasets. The best and second best results are shown in red and blue, respectively. The values of DSC and IoU are in percentage terms.

Methods	CellSeg				CHAOS-CT		MSD-Heart	
	DSC/Val	IoU/Val	DSC/Test	IoU/Test	DSC	IoU	DSC	IoU
U-Net	81.04	60.29	85.01	71.93	97.47	93.60	91.33	83.92
UNet++	80.18	60.59	83.87	71.47	97.14	92.42	91.74	84.34
ResUNet	78.11	60.02	84.41	71.17	93.48	87.44	87.51	79.69
AttUNet	79.35	58.68	84.94	72.22	96.17	91.70	89.90	82.37
TransUNet	86.62	71.94	86.19	74.35	95.58	90.73	72.53	69.00
Ours-SINSeg	85.44	69.29	86.32	73.26	97.53	94.19	92.29	87.41

Table 3: Ablation studies of our proposed SIN-Seg framework on the other three datasets. The best results are shown in red.

Sttings	CellSeg				CHAOS-CT		MSD-Heart	
	DSC	IoU	DSC/Test	IoU/Test	DSC	IoU	DSC	IoU
U-Net/Pure Spatial	81.04	60.29	85.01	71.93	97.47	93.60	91.33	83.92
U-Net/Pure Spectrum	65.38	42.68	72.72	55.70	96.72	92.24	88.30	81.06
U-Net/Joint wo Aligment	72.84	49.55	78.12	61.38	96.15	91.69	90.16	81.29
U-Net+SINSeg	85.44	69.29	86.32	73.26	97.53	94.19	92.29	87.41

Table 4: Ablation studies of our proposed SIN-Seg framework on the BraTS dataset. The best results are shown in red.

Settings	BraTS	
	DSC/Val	IoU/Val
U-Net/Pure Spatial	66.36	50.22
U-Net/Pure Spectrum	65.47	51.36
U-Net/Joint wo Alignment	52.41	44.37
U-Net+SINSeg	75.35	72.16

5.0 Conclusions

Spectral information plays an important role in brain image segmentation tasks and should be fully considered. The semantic features yielded from the spectrum space should be aligned due to the fact that feature variances, resulting from the inconsistent frequency-related settings of medical imaging modalities, exist on the segmentation ROIs. In this paper, we propose a spectrum information-based feature-enhanced (SIN) model that combines spectrum and spatial information for different segmentation tasks. Experimental results demonstrate the effectiveness and superiority of our proposed model.

6.0 Future Research

This research just used the whole spectral and spatial information for feature fusion. However, as we discussed in the related part of this thesis, Spectral space is always used for image compression, and even the compression of the model itself. After implementing the transformation, both the *RGB* and the *Gray-Scale* images would be mapped to a high-channel space when we do path-wise DCT, and some of the feature maps may be redundant. In the future, we plan to use a advanced feature selection mechanism based on mutual information between these two feature spaces. Furthermore, there are also some domain-unrelated feature could be extracted easier in the spectral space, besides the semantic feature. We consider to develop a novel feature disentangle algorithm to achieve more powerful domain generalization model.

7.0 Data Availability Statement

The Brain Tumor Segmentation dataset is from the BraTS Challenge[27] and is available from https://www.smir.ch/BRATS/Start2015. The MRI heart dataset is from the MSD Challenge[1] and is available from http://medicaldecathlon.com/. The Liver segmentation dataset is from the CHAOS Challenge[18] and is available from https://chaos.grand-challenge.org/. The Cell segmentation dataset is from the NeurIPS2022-cellseg Challenge[26] and is available from https://neurips22-cellseg.grand-challenge.org/.

8.0 Conflict of Interest

The author declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

9.0 Acknowledgement

I would like to appreciate the Pittsburgh Supercomputing Center (PSC) for the computation resources support for part of my work. I would like to thank all the organizers and contributers for BraTS Challenge, CHAOS Challenge, MSD Challenge, and NeurIPS CellSeg challenge, thank you for the dataset you collected and provided.

Bibliography

[1] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

[2] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[3] Wenlin Chen, James Wilson, Stephen Tyree, Kilian Q Weinberger, and Yixin Chen. Compressing convolutional neural networks in the frequency domain. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1475–1484, 2016.

[4] IC Cuthill. Camouflage. *Journal of Zoology*, 308(2):75–92, 2019.

[5] Max Ehrlich and Larry S Davis. Deep residual learning in the jpeg transform domain. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3484–3493, 2019.

[6] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23*, pages 263–273. Springer, 2020.

[7] Xiyao Fu, Zhexian Sun, Haoteng Tang, Eric M Zou, Heng Huang, Yong Wang, and Liang Zhan. 3d bi-directional transformer u-net for medical image segmentation. *Frontiers in Big Data*, 5:1080715, 2023.

[8] Lionel Gueguen, Alex Sergeev, Ben Kadlec, Rosanne Liu, and Jason Yosinski. Faster neural networks straight from jpeg. *Advances in Neural Information Processing Systems*, 31, 2018.

[9] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R Martin, Ming-Ming Cheng, and Shi-Min

Hu. Attention mechanisms in computer vision: A survey. *Computational visual media*, 8(3):331–368, 2022.

[10] Zhe Guo, Xiang Li, Heng Huang, Ning Guo, and Quanzheng Li. Deep learning-based image segmentation on multimodal medical imaging. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 3(2):162–169, 2019.

[11] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1055–1059. IEEE, 2020.

[14] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6891–6902, 2021.

[15] Keli Huang, Botian Shi, Xiang Li, Xin Li, Siyuan Huang, and Yikang Li. Multi-modal sensor fusion for auto driving perception: A survey. *arXiv preprint arXiv:2202.02703*, 2022.

[16] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[17] Haozhe Jia, Haoteng Tang, Guixiang Ma, Weidong Cai, Heng Huang, Liang Zhan, and Yong Xia. A convolutional neural network with pixel-wise sparse graph reasoning for covid-19 lesion segmentation in ct images. *Computers in Biology and Medicine*, page 106698, 2023.

[18] A. Emre Kavur, N. Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan,

Bora Baydar, Dmitry Lachinov, Shuo Han, Josef Pauli, Fabian Isensee, Matthias Perkonigg, Rachana Sathish, Ronnie Rajan, Debdoot Sheet, Gurbandurdy Dovletov, Oliver Speck, Andreas Nürnberger, Klaus H. Maier-Hein, Gözde Bozdağı Akar, Gözde Ünal, Oğuz Dicle, and M. Alper Selver. CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, April 2021.

[19] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

[20] Fred A Kruse, AB Lefkoff, JW Boardman, KB Heidebrecht, AT Shapiro, PJ Barloon, and AFH Goetz. The spectral image processing system (sips)—interactive visualization and analysis of imaging spectrometer data. *Remote sensing of environment*, 44(2-3):145–163, 1993.

[21] Dana Lahat, Tülay Adali, and Christian Jutten. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015.

[22] John Lee, Max Dabagia, Eva Dyer, and Christopher Rozell. Hierarchical optimal transport for multimodal distribution alignment. *Advances in neural information processing systems*, 32, 2019.

[23] Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. Multi-level wavelet-cnn for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 773–782, 2018.

[24] Zhenhua Liu, Jizheng Xu, Xiulian Peng, and Ruiqin Xiong. Frequency-domain dynamic pruning for convolutional neural networks. *Advances in neural information processing systems*, 31, 2018.

[25] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[26] Jun Ma, Ronald Xie, Shamini Ayyadhury, Cheng Ge, Anubha Gupta, Ritu Gupta, Song Gu, Yao Zhang, Gihun Lee, Joonkee Kim, Wei Lou, Haofeng Li, Eric Upschulte, Timo Dickscheid, José Guilherme de Almeida, Yixin Wang, Lin Han, Xin Yang, Marco Labagnara, Sahand Jamal Rahi, Carly Kempster, Alice Pollitt, Leon Espinosa, Tâm Mignot, Jan Moritz Middeke, Jan-Niklas Eckardt, Wangkai Li, Zhaoyang Li, Xiaochen

Cai, Bizhe Bai, Noah F. Greenwald, David Van Valen, Erin Weisbart, Beth A. Cimini, Zhuoshi Li, Chao Zuo, Oscar Brück, Gary D. Bader, and Bo Wang. The multi-modality cell segmentation challenge: Towards universal solutions. *arXiv:2308.05864*, 2023.

[27] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

[28] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576, 2022.

[29] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, 2021.

[30] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. In *Medical Imaging with Deep Learning*, 2022.

[31] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. Fcanet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 783–792, 2021.

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[34] Martin Stevens and Sami Merilaita. Animal camouflage: current issues and new perspectives. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1516):423–427, 2009.

[35] Wei Tan, Prayag Tiwari, Hari Mohan Pandey, Catarina Moreira, and Amit Kumar Jaiswal. Multimodal medical image fusion algorithm in the era of big data. *Neural Computing and Applications*, pages 1–21, 2020.

[36] Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. Transbts: Multimodal brain tumor segmentation using transformer. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 109–119. Springer, 2021.

[37] Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12186–12195, 2022.

[38] Yunhe Wang, Chang Xu, Shan You, Dacheng Tao, and Chao Xu. Cnnpack: Packing convolutional neural networks in the frequency domain. *Advances in neural information processing systems*, 29, 2016.

[39] Jun Wei, Yiwen Hu, Guanbin Li, Shuguang Cui, S Kevin Zhou, and Zhen Li. Boxpolyp: Boost generalized polyp segmentation using extra coarse bounding box annotations. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III*, pages 67–77. Springer, 2022.

[40] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. Learning in the frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1740–1749, 2020.

[41] Kai Ye, Haoteng Tang, Siyuan Dai, Lei Guo, Johnny Yuehan Liu, Yalin Wang, Alex Leow, Paul M Thompson, Heng Huang, and Liang Zhan. Bidirectional mapping with contrastive learning on multimodal neuroimaging data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 138–148. Springer, 2023.

[42] Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6995–7004, 2021.

[43] Yifei Zhang, Désiré Sidibé, Olivier Morel, and Fabrice Mériaudeau. Deep multimodal fusion for semantic image segmentation: A survey. *Image and Vision Computing*, 105:104042, 2021.

[44] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018.

[45] Bolun Zheng, Shanxin Yuan, Chenggang Yan, Xiang Tian, Jiyong Zhang, Yaoqi Sun, Lin Liu, Aleš Leonardis, and Gregory Slabaugh. Learning frequency domain priors for image demoireing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7705–7717, 2021.

[46] Yijie Zhong, Bo Li, Lv Tang, Senyun Kuang, Shuang Wu, and Shouhong Ding. Detecting camouflaged object in frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4504–4513, 2022.

[47] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018.