Multimodal Representation Distribution Learning for Medical Image Segmentation

Chao Huang[1], **Weichao Cai**[1*], **Qiuping Jiang**[2*] and **Zhihua Wang**[3]
[1]School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University
[2]School of Information Science and Engineering, Ningbo University
[3]Department of Engineering, Shenzhen MSU-BIT University
huangch253@mail.sysu.edu.cn, 21wccai@stu.edu.cn, jiangqiuping@nbu.edu.cn,
zhihua.wang@my.cityu.edu.hk

Abstract

Medical image segmentation is one of the most critical tasks in medical image analysis. However, the performance of typical methods is limited by the lack of high-quality labeled data due to the expensive spending of annotation data. To alleviate this limitation, we propose a novel multi-modal learning method for medical image segmentation. In our method, medical text annotation is adopted to compensate for the quality deficiency in image data. Moreover, previous multi-modal fusion methods ignore the redundant information between different modalities. In this paper, we propose a novel multi-modal feature distribution learning method to reduce redundancy by capturing the discriminate information between text and image. Additionally, medical image segmentation needs to predict detailed segmentation boundaries. Thus, a prompt encoder is designed to achieve fine-grained segmentation. Experimental results on three datasets show that our method has superior performance. Source codes will be available at https://github.com/GPIOX/Multimodal.git.

1 Introduction

Medical image segmentation, as a pivotal component of auxiliary disease diagnosis, holds a crucial role in medical image applications. Enormous advances in medical image segmentation benefit from the more and more annotated datasets. However, it is still commonly impression that medical imaging datasets are too small to develop robust deep learning models [Liu *et al.*, 2023]. One reason for this is the expensive spending of high-quality pixel-level annotations. Expert annotators typically spend several hours to annotate a medical image. In addition, traditional methods only adopt vision modality to train the model, which limits the ability of encoded knowledge [Zhao *et al.*, 2023; Wang *et al.*, 2021c; Wang *et al.*, 2021b]. In contrast, medical text annotation is naturally rich in semantics. These annotations are much easier to obtain. Hence, it is intuitive to integrate text annotation into medical image segmentation.

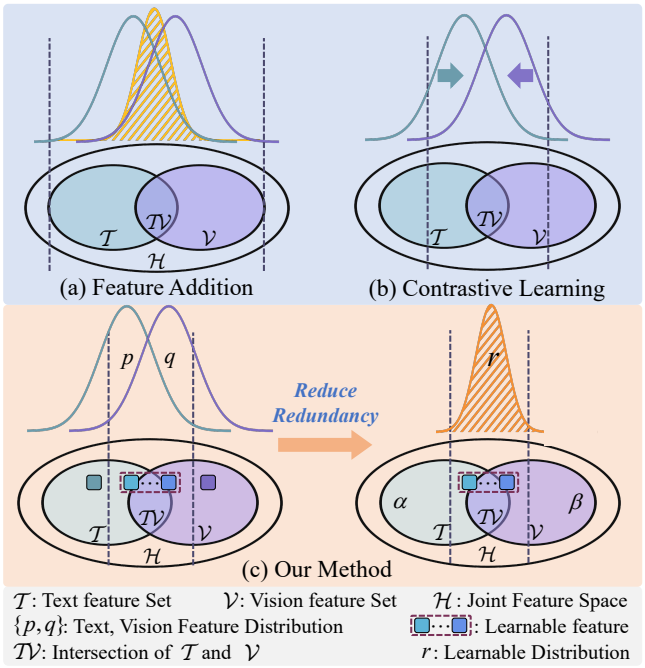

Figure 1: Comparison of different multi-modal fusion methods and the proposed method.

Previous multi-modal methods adopt element addition to fuse text and vision features. As shown in Fig. 1 (a), the element addition method can be seen to fuse two distributions into one in joint feature space $\mathcal{H}$. The fused distribution contains all the information of text and vision features. However, both task-related and irrelevant information is included in the fused distribution, which results in information redundancy in the fused features [Hosseini *et al.*, 2024]. Thus, the performance improvement is limited. Recently, contrastive learning has been commonly used to align text and image features [Li *et al.*, 2023a]. Specifically, it considers text caption as a linguistic view of the image under consideration as shown in Fig. 1 (b). It then pulls the pair of image and text feature distribution (p and q, respectively) close in semantic space. Nevertheless, contrastive learning only models the common information between text and image. It ignores the modality-specific information. For example, the

*Corresponding authors.

edge information in the vision modal. Thus, the value of distinctive perspectives is neglected due to the suppression of modality-specific information [Zhang *et al.*, 2024]. In addition, a large number of image-text pairs data are necessary to train a contrastive learning model. Thus, contrastive learning is possibly unsuitable for medical image segmentation due to high-quality image-text pairs lacking in many medical domains [Thawkar *et al.*, 2023; Huang *et al.*, 2022a; Huang *et al.*, 2022b].

Furthermore, recent studies [Li *et al.*, 2023a; Wang *et al.*, 2023] show that text-image models perform poorly in spatial understanding tasks but excel in semantic understanding tasks. We also observe this phenomenon in experiments. Remarkable spatial perception is essential for accurate segmentation of object boundaries, which is important in medical image segmentation. The delineation of boundaries can signify the transitions between different human tissues or anatomical structures, thus providing essential information for accurately separating these instances [Wei *et al.*, 2023; Huang *et al.*, 2021b; Huang *et al.*, 2021a].

To address these issues, we proposed a novel multi-modal feature distribution method for medical image segmentation. Firstly, medical text annotation is incorporated into our method to compensate for the lack of quality in image data. It also enhances the ability of the model to encode knowledge. Secondly, we propose a novel multi-modal feature fusion method, aiming to reduce redundant information. Modality-specific information is also retained in the proposed method. Specifically, as shown in Fig 1 (c), a set of learnable features is adopted to extract the modality-specific information of two modalities by maximizing the similarity of corresponding distributions. Then a distribution is adopted to model these learnable features. It can maximize discriminative information about tasks. Finally, the boundary regions in images demonstrate a strong correlation with features in the frequency domain. Consequently, we propose a frequency prompt encoder to improve the spatial perception of the model. It can fully leverage high-frequency information. These features serve as spatial information prompts, which along with the multi-modal features are decoded into segmentation results.

Our main contributions are summarized as follows:

- We propose a novel multi-modal method that leverages the text and image features for medical image segmentation. The medical text annotation serves as additional information, which can enhance the ability of model to encode knowledge.
- We propose a novel multi-modal fusion paradigm, aiming to reduce redundancy while retaining the modality-specific information.
- A novel frequency prompt encoder is proposed to leverage high-frequency information for accurately segmenting boundaries.

2 Related Work

2.1 Medical Image Segmentation

Medical image segmentation refers to the identification of organ or lesion pixels from medical images [Yao *et al.*, 2023]. U-Net [Ronneberger *et al.*, 2015] is commonly considered as a typical segmentation model for medical images. TransUNet [Chen *et al.*, 2021] and Swin-UNet [Cao *et al.*, 2022] combine Transformer and UNet for achieving better performance. However, the performance of traditional segmentation methods depend on the high-quality annotated data. Additionally, these methods only are trained with visual modality data, which potentially constrains model to encoded knowledge. Several approaches [Cherti *et al.*, 2023; Sun *et al.*, 2023] explore leveraging medical text annotation to reduce reliance on high-quality annotations while maintaining performance. Inspired by these methods, we propose a novel method that adopts medical text annotation to strengthen the ability of model to encoded knowledge. In this way, it can achieve solid performance on the dataset with limited high-quality annotations.

2.2 Multi-Modal Learning

Recent studies have shown that text information is useful for medical image analysis. Thus, it is essential to design an effective multimodal fusion paradigm to fuse text and vision features. LAVT [Yang *et al.*, 2022] and VLT [Ding *et al.*, 2022] integrate text feature into vision feature by a language encoder. LViT [Li *et al.*, 2023b] directly adopts element-wise addition to fuse text and vision features. The problem with these methods is that containing redundant information in the fused features. Several methods such as MedCLIP [Wang *et al.*, 2022b], TPRO [Zhang *et al.*, 2023], and others [Wu *et al.*, 2023; Cao *et al.*, 2023] base CLIP [Radford *et al.*, 2021] adopt contrastive learning to align text and image features. However, these methods only model the commonalities between different modalities. It ignores the wealth of model-specific information. In addition, the contrastive learning model requires a large amount of image-text pair data for training. This paper proposes a novel multi-modal fusion method. It can effectively reduce redundancy while retaining modality-specific information.

2.3 Spectral Representation

Deep neural networks are biased towards learning low-frequency representations [Mildenhall *et al.*, 2021]. To better leverage high-frequency information, several methods [Mao *et al.*, 2023] investigate the Fast Fourier Transform (FFT)-based frequency representation. FFT is a powerful tool, and it can leverage the strengths of both spectral and spatial representations [Tang *et al.*, 2021]. Recent research [Zhu *et al.*, 2023] has focused on the parallel extraction of spatial and frequency information within the encoder block. However, the spatial and computational complexity is increased due to this parallel extraction for the high-dimension tensor. In this paper, we propose a novel frequency prompt encoder to fully aggregate high-frequency information from the frequency domain. The proposed method avoids parallel processing in the vision encoder, which requires fewer parameters and lower computational cost.

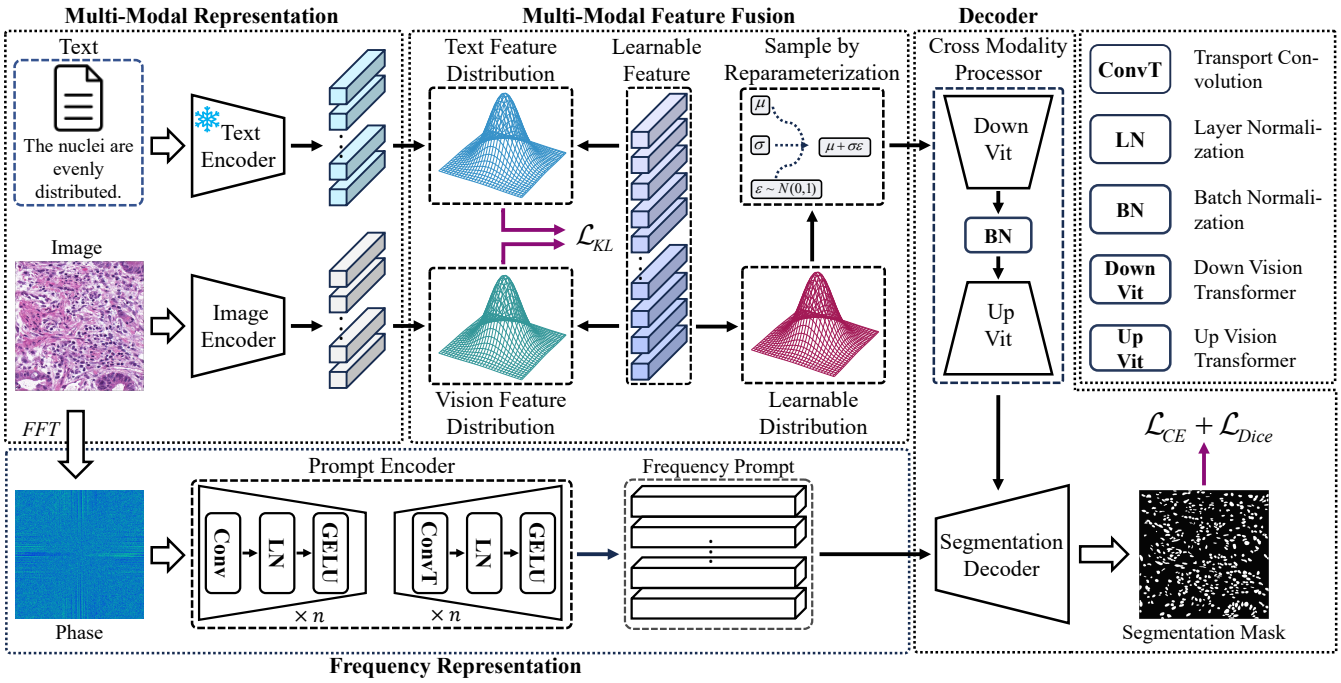

Figure 2: Overview of the proposed method. First, given the medical image and corresponding medical text annotation, Text encoder and vision encoder to generate text and vision features. Then a set of learnable features is adopted to reduce redundancy. A distribution is adopted to model these learnable features. The Fused feature is sampled from this distribution and a cross-modality processer is adopted to further process it. Finally, the segmentation output is obtained by decoding the post-processed features and the frequency prompt. Additionally, 'snow' means freezing the parameters of the text encoder.

3 Methodology

3.1 Overview

As shown in Fig. 2, the proposed method consists of four components. The first part integrates a vision encoder Enc_I and a text encoder Enc_T to extract the vision and text features. Enc_T is a pre-trained CLIP text encoder. The second part is a multi-modal feature fusion module. It leverages a distribution to reduce redundant information while retaining the modality-specific information. The third part is a frequency prompt encoder Enc_P. Specifically, considering the frequency information is significant for boundary segmentation, we designed Enc_P to fully aggregate high-frequency information from the frequency domain. Detailed explanations of these two parts are provided in the following sections. The last part is a segmentation decoder Dec. It leverages the fused features from a cross-modality processor P_{CM} and frequency prompt from Enc_P to generate segmentation results.

3.2 Multi-Modal Feature Fusion

Previous multi-modal fusion methods usually ignore the redundancy in the fused features. Consequently, it is hard to capture the discriminative information. To address these issues, we propose a novel multi-modal feature fusion method with a learnable distribution. Specifically, given the medical text annotation T and medical image I, Enc_T and Enc are adopted to extract the text and vision features $\mathbf{F}_T$ and $\mathbf{F}_V$:

$$\mathbf{F}_T^{'} = Enc_T(T), \mathbf{F}_V^{'} = Enc_I(I). \tag{1}$$

Moreover, a convolution block $ConvB$ is used to map visual and text features to the joint feature space:

$$\mathbf{F}_V = Embedding(\mathbf{F}_V^{'}), \mathbf{F}_T = ConvB(\mathbf{F}_T^{'}), \tag{2}$$

where $ConvB$ consists of a convolution layer and a batch normalization layer. ReLU is activation function.

A set of learnable features $\mathbf{F}_L$ is defined in the joint feature space $\mathcal{H}$. It is adopted to capture the discriminative information about task-related in $\mathbf{F}_T$ and $\mathbf{F}_V$. Specifically, texts with similar semantics have high cosine similarity, which means they are close to each other in the feature space [Nlong Zhao *et al.*, 2023]. Previous research [Lu *et al.*, 2022; Huang *et al.*, 2020] further demonstrates that text describing identical categories cluster together in the feature space. Inspired by this phenomenon, a prompt T_P that is similar in semantics to T is defined. T_P is encoded by Enc_T and $ConvB$, which generates a prompt feature $\mathbf{F}_P$ in $\mathcal{H}$, defined as:

$$\mathbf{F}_P = ConvB(Enc_T(T_P)). \tag{3}$$

$\mathbf{F}_P$ is close to T_P in $\mathcal{H}$. Then $\mathbf{F}_P$ is repeated K times to obtain $\mathbf{F}_P = \{\mathbf{F}_P^1, \mathbf{F}_P^2, \cdots, \mathbf{F}_P^K\}$. However, these features are identical, directly initializing them as learnable features would cause them to converge to the same vector. To avoid this trouble, Gaussian noise $\mathcal{N}(\mathbf{0}, \mathbf{I})$ is added to $\mathbf{F}_P^i$,

$$\mathbf{F}_L^i = \mathbf{F}_P^i + 0.1 \times \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{4}$$

This operation can introduce variation to the N features while ensuring that every $\mathbf{F}_L^i$ is still close to $\mathbf{F}_T$ in $\mathcal{H}$. Consequently, $\mathbf{F}_L = \{\mathbf{F}_L^1, \mathbf{F}_L^2, \cdots, \mathbf{F}_L^K\}$ is initialized as learnable

features. To model the text and vision feature distributions (p, q, respectively) in $\mathcal{H}$, we assume that p, q follow Gaussian distributions $p \sim \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\sigma}_p^2)$ and $p \sim \mathcal{N}(\boldsymbol{\mu}_q, \boldsymbol{\sigma}_q^2)$. With N learnable features $\mathbf{F}_L$, we can estimate the $\boldsymbol{\mu}_p$ and $\boldsymbol{\sigma}_p$ of p as:

$$\boldsymbol{\mu}_p = \frac{1}{K+1}(\mathbf{F}_T + \sum_{i=1}^{K} \mathbf{F}_L^i), \tag{5}$$

$$\boldsymbol{\sigma}_p = \frac{1}{K+1}(\mathbf{F}_T + \sum_{i=1}^{K} \mathbf{F}_L^i - \boldsymbol{\mu}_p)^T(\mathbf{F}_T + \sum_{i=1}^{K} \mathbf{F}_L^i - \boldsymbol{\mu}_p). \tag{6}$$

Similarity, $\boldsymbol{\mu}_q$ and $\boldsymbol{\sigma}_q$ of q can be estimated as:

$$\boldsymbol{\mu}_q = \frac{1}{K+1}(\mathbf{F}_V + \sum_{i=1}^{K} \mathbf{F}_L^i), \tag{7}$$

$$\boldsymbol{\sigma}_q = \frac{1}{K+1}(\mathbf{F}_V + \sum_{i=1}^{K} \mathbf{F}_L^i - \boldsymbol{\mu}_q)^T(\mathbf{F}_V + \sum_{i=1}^{K} \mathbf{F}_L^i - \boldsymbol{\mu}_q). \tag{8}$$

As a result, $\mathbf{F}_L$ can extract the modality-specific information in $\mathbf{F}_T$ and $\mathbf{F}_V$.

Then p and q are adopted to reduce the redundant information by modeling the discriminative information Specifically, we leverage Kullback-Leibler divergence $\mathcal{L}_{KL}$ to measure the similarity between p and q. $\mathbf{F}_L$ can extract modality-common information by minimizing $\mathcal{L}_{KL}$. $\mathcal{L}_{KL}$ is defined as :

$$\mathcal{L}_{KL} = \int_{-\infty}^{\infty} p(\mathbf{x}) \ln(\frac{p(\mathbf{x})}{q(\mathbf{x})}) d\mathbf{x} \tag{9}$$

In the proposed method, $\mathcal{L}_{KL}$ can be represented as:

$$\mathcal{L}_{KL} = log\frac{\boldsymbol{\sigma}_q}{\boldsymbol{\sigma}_p} + \frac{1}{2\boldsymbol{\sigma}_p^2}(\boldsymbol{\sigma}_p^2 + (\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)^2) - \frac{1}{2} \tag{10}$$

To further maximize discriminative information and reduce redundant information, we use a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_r, \boldsymbol{\sigma}_r)$, where $\boldsymbol{\mu}_r = \boldsymbol{\mu}(\mathbf{F}_L)$ and $\boldsymbol{\sigma}_r = \boldsymbol{\sigma}(\mathbf{F}_L)$ are the mean and covariance of r, respectively. The fused feature $\mathbf{F}$ is sampled from r. $\mathbf{F}$ effectively fuses modality-common and modality-specific information. Nevertheless, sampling $\mathbf{F}$ from r is not differentiable for optimization. Thus, the reparameterization trick is adopted, which is similar to VAE [Kingma and Welling, 2013]. Formally, the process of sampling $\mathbf{F}$ is rewrite as:

$$\mathbf{F} = \boldsymbol{\mu}_r + \boldsymbol{\omega}\boldsymbol{\sigma}_r, \tag{11}$$

where $\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ has same dimension as $\boldsymbol{\mu}_r$ and $\boldsymbol{\sigma}_r$.

3.3 Frequency Prompt Encoder

To improve the spatial perception ability of the text-image model, we design a frequency prompt encoder Enc_P. It is designed to fully utilize high-frequency information, which aims to prompt the boundaries information for the segmentation decoder. In contrast to previous methods, the proposed Enc_P is not integrated into basic blocks. Thus it can avoid parallel processing in the vision encoder, which requires fewer parameters and lower computational cost.

The Discrete Fourier Transform (DFT) is a powerful tool for converting images to the frequency domain. In practice, the FFT is used to compute the DFT efficiently, defined as:

$$\mathcal{F}_{u,v} = \sum_{h=1}^{H}\sum_{w=1}^{W} f_{h,w} \cdot \mathrm{e}^{-j2\pi\left(\frac{h}{H}u + \frac{w}{W}v\right)}, \tag{12}$$

where $f_{h,w}$ dnotes the spectrum represention of I.

Furthermore, the amplitude and phase spectrum of $\mathcal{F}_{u,v}$ can be computed as $|\mathcal{F}_{u,v}|$ and $arg(\mathcal{F}_{u,v})$. The phase spectrum contains information about the edges and overall structure [Ghiglia and Pritt, 1998]. Consequently, the phase spectrum outperforms the amplitude spectrum in terms of representation ability in ablation studies. Therefore, we adopt the phase spectrum $\mathbf{F}_{pha}$ as the default in the proposed method. As shown in Fig. 2, Enc_P consists of two components. The first component Enc_{P1} is consists of sequences of convolution layers, layer normalization, and activations. It can be represented as:

$$\mathcal{Z}_P^i = GeLU^i(LN^i(Conv^i(\mathcal{Z}_P^{i-1}))), \tag{13}$$

where $\mathcal{Z}_P^i$ is the output of layer i and $i = \{1, 2, \cdots, n\}$. $Conv^i$, LN^i, and $GeLU^i$ are the i-th convolution layer, GeLU activation function, and layer normalization, respectively. Similarity, the second component Enc_{P2} can be defined as:

$$\mathbf{F}_{pha}^i = GeLU^i(LN^i(ConvT^i(\mathcal{Z}_P^{i-1}))), \tag{14}$$

$$\mathbf{F}_{pha} = \mathbf{F}_{pha}^n, \tag{15}$$

where $\mathbf{F}_{pha}$ is the frequency prompt of the I, $ConvT^i$ is the i-th transposed convolution layer. In experiments, we set $N = 2$.

3.4 Segmentation Decoder

Like LViT [Li *et al.*, 2023b], the Enc_I and Dec are U-shape architecture. Moreover, the cross-modality processor P_{CM} is also a U-shape architecture. More details about the network architectures of Enc_I, Dec, and P_{CM} are provided in the supplementary materials. Last decoder bolck outputs the segmentation result $\mathbf{O}$, formulated as:

$$\mathbf{O} = Dec(\mathbf{F}_V^M, \mathbf{F}_{pha} + P_{CM}(\mathbf{F})), \tag{16}$$

where $\mathbf{F}_V^M$ is the output of the last layer of Enc_I.

3.5 Loss Function

The loss function used in this work are Dice loss $\mathcal{L}_{Dice}$, cross-entropy loss $\mathcal{L}_{CE}$, and $\mathcal{L}_{KL}$, formulated as:

$$L_{Dice} = 1 - \sum_{i=1}^{S}\sum_{j=1}^{C} \frac{1}{SC} \cdot \frac{2|d_{ij} \cap o_{ij}|}{(|d_{ij}| + |o_{ij}|)}, \tag{17}$$

$$L_{CE} = -\sum_{i=1}^{S}\sum_{j=1}^{C} \frac{1}{S} \cdot o_{ij} \log(d_{ij}), \tag{18}$$

Method	Venue	Text	MoNuSeg		GlaS		MosMedData+		Param (M)	FLOPs (G)
			Dice	MIoU	Dice	MIoU	Dice	MIoU		
UNet	MICCAI'2015	✗	76.45	62.86	85.45	74.78	64.60	50.73	**19.1**	412.7
DCA	EAAI'2023	✗	79.50	65.97	89.90	81.68	72.05	59.78	30.58	87.28
AttUNet	MICCAI'2021	✗	76.67	63.47	88.80	80.69	66.34	52.82	34.9	101.9
Swin-UNet	ECCV'2022	✗	77.69	63.77	89.58	82.06	63.29	50.19	82.3	67.3
TransuNet	arXiv'2021	✗	78.53	65.05	88.40	80.40	71.24	58.44	105	56.7
UCTransNet	AAAI'2022	✗	79.09	66.68	89.76	81.91	65.90	52.69	65.6	63.2
MedSAM	Nat. Commun'2024	✗	—	21.40	—	54.79	—	—	91.0	371.9
SAM	ICCV'2023	✗	—	32.95	—	67.91	—	—	91.0	371.9
Our		✗	80.24	67.15	89.16	81.22	74.26	61.18	27.5	57.4
ViLT	PMLR'2021	✓	—	—	—	—	72.36	60.15	87.4	55.9
CLIP	PMLR'2021	✓	—	—	—	—	71.97	59.64	87.0	105.3
TGANet	MICCAI'2022	✓	70.09	61.63	87.96	82.74	71.81	59.28	19.8	41.9
LAVT	CVPR'2022	✓	65.36	52.12	76.28	67.52	73.29	60.41	118.6	83.8
LViT-T	TMI'2023	✓	80.15	67.00	90.02	82.68	74.57	61.33	29.7	**54.1**
Our		✓	**80.96**	**68.12**	**91.08**	**84.00**	**76.02**	**63.03**	29.2	57.5

Table 1: Quantitative comparison of the proposed method and other SOTA methods on MoNuSeg, GlaS, and MosMedData+ datasets.

where S is the number of pixels in the image, C is the number of classes, o_{ij} and d_{ij} are the ground truth and predicted segmentation output, respectively.

The final loss function $\mathcal{L}$ is defined as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{Dice} + \lambda_2 \mathcal{L}_{CE} + \lambda_3 \mathcal{L}_{KL}. \tag{19}$$

Base on experiments, we set $\lambda_1 = 0.5$, $\lambda_2 = 0.5$, and $\lambda_3 = 1$.

4 Experiments

4.1 Datasets

The proposed method is evaluated on three medical image segmentation datasets: MoNuSeg [Kumar *et al.*, 2017], MosMedData+ [Li *et al.*, 2023b], and GlaS [Sirinukunwattana *et al.*, 2017]. The first two datasets are the same benchmark datasets used in [Li *et al.*, 2023b]. MoNuSeg contains 30 digital microscopic tissue images of several patients, while MosMedData+ [Morozov *et al.*, 2020; Hofmanninger *et al.*, 2020] contains 2729 CT scan slices of lung infections. The ratio of training, validation, and test sets are the same as in [Li *et al.*, 2023b]. GlaS has 85 images for training and 80 for testing.

4.2 Evaluation Metrics

Dice score and IoU [Li *et al.*, 2023b] are used to evaluate the performance of the proposed method and other SOTA methods. Moreover, we have reported the time complexity and space complexity of the proposed method. Specifically, the time complexity is evaluated in Floating Point Operations per second (FLOPs), and the space complexity is measured by the number of parameters.

4.3 Implementation Details

The proposed method is optimized by AdamW [Loshchilov and Hutter, 2017] and cosine annealing learning rate scheduler is adopted. The initial learning rate is set to 1e-3 for all datasets. Image input sizes are 224×224 both for MoNuSeg, GlaS, and MosMedData+. An early stop mechanism is adopted until the performance of the model does not increase for 50 epochs. The batch size is 2 for MoNuSeg and GlaS and 24 for MosMedData+. The default number of learnable features K is set to 32. All experiments are conducted on a single NVIDIA RTX 3090 GPU with 24GB memory.

4.4 Baseline

Baselines are divided into two sets: single-modal approaches and multi-modal approaches. Single-modal approaches include U-Net [Ronneberger *et al.*, 2015], DoubleUnet [Ates *et al.*, 2023], AttUNet [Wang *et al.*, 2021a], Swin-UNet [Cao *et al.*, 2022], TransUNet [Chen *et al.*, 2021], UCTransNet [Wang *et al.*, 2022a] MedSAM [Han *et al.*,] and SAM [Han *et al.*,]. Multi-modal approaches include ViLT [Kim *et al.*, 2021], CLIP [Radford *et al.*, 2021], TGANet [Tomar *et al.*, 2022], LAVT [Yang *et al.*, 2022], and LViT [Li *et al.*, 2023b].

4.5 Comparison with State-of-the-Art Methods

The quantitative comparison results are shown in Tab. 1. From this table, we can clearly observe that the proposed method achieves the best performance on all datasets. In particular, the proposed method achieves an 80.24% Dice score and 67.15% IoU on MoNuSeg, which is outperformance in the single modal set. The proposed method maintains a similar performance to DCA on the GlaS dataset with a smaller number of parameters and lower computational complexity.

The proposed method achieves an 80.96% Dice score and 68.12% IoU on MoNuSeg, which respectively increases by 0.81% and 0.65% compared to the suboptimal method LViT-T. Similarly, the proposed method obtains 91.08% Dice score and 84.00% MIoU on GlaS. It outperforms the sec-

Backbone	Text	Frequency Prompt	MoNuSeg		GlaS	
			Dice	MIoU	Dice	MIoU
✓			79.16	65.84	88.91	80.77
✓	✓		80.62	67.63	89.43	81.49
✓		✓	80.24	67.15	89.16	81.22
✓	✓	✓	**80.96**	**68.12**	**91.08**	**84.00**

Table 2: Ablation study on the effectiveness of supervised components on MoNuSeg and GlaS datasets.

Fusion Method	MoNuSeg		GlaS	
	Dice	MIoU	Dice	MIoU
Addition	80.49	67.47	90.21	83.03
Distribution	**80.96**	**68.12**	**91.08**	**84.00**

Table 3: Ablation study of the different multi-modal fusion methods on MoNuSeg and GlaS datasets.

$\mathbf{FP}_{amp}$	$\mathbf{FP}_{pha}$	MoNuSeg		GlaS	
		Dice	MIoU	Dice	MIoU
✓		80.40	67.36	90.37	83.14
	✓	**80.96**	**68.12**	**91.08**	**84.00**
✓	✓	80.71	67.78	89.20	81.22

Table 4: Ablation study on frequency prompt encoder. $\mathbf{FP}_{pha}$ and $\mathbf{FP}_{amp}$ denote the phase and amplitude spectrum of frequency prompt, respectively.

Value Setting	MoNuSeg		GlaS	
	Dice	MIoU	Dice	MIoU
0.1	**80.96**	**68.12**	90.37	83.14
0.5	80.89	68.03	90.77	83.73
1.0	80.84	67.94	**91.08**	**84.00**
1.5	80.74	67.83	89.98	82.36

Table 5: Ablation study of different λ_3 value.

ond best method by 1.06% and 1.32%, respectively. Surprisingly, we observe notable improvements in the MosMedData+ dataset. In particular, the performance respectively increases by 1.45% and 1.70% in Dice score and MIoU. This performance improvement benefits from reducing redundant information and maximizing the discriminative information.

It is worth noting that the proposed method with text annotation achieves better performance than the proposed method without text annotation. In detail, the proposed method with text annotation respectively increased by 0.72% and 0.97% in Dice score and MIoU on MoNuSeg. Similarly, the performance improvements are 0.92% and 1.36% on GlaS, and 0.45% and 0.70% on MosMedData+. It demonstrates that medical text annotation can enhance the robustness of encoded semantics. Moreover, as shown in Fig. 3, the proposed method achieves more higher-quality segmentation results.

4.6 Ablation Study

Effectiveness of Supervised Components

Tab. 2 reports the ablation experimental results of the effectiveness of supervised components. These results indicate that all of these components are effective for the proposed method. It is worth noting that incorporating text annotation significantly improves segmentation performance.

Specifically, the Dice score and MIoU are increased by 1.46% and 1.29% on MoNuSeg, respectively. The performance improvements are 0.52% in Dice score and 0.72% in MIoU on GlaS. The frequency prompt encoder is also effective for the proposed method. It respectively increases by 1.08% and 1.13% in Dice score and MIoU on MoNuSeg. Similarly, performance improvements are 0.25% and 0.97% on GlaS in Dice score and MIoU, respectively. The proposed method with both text annotation and frequency prompt encoder achieves the best performance.

Impact of Different Multi-modal Fusion Methods

Tab. 3 reports the ablation experimental results of the different multi-modal fusion methods. These results indicate that the distribution fusion method is more effective than the addition fusion method. In details, the Dice score and MIoU are increased by 0.47% and 0.65% on MoNuSeg, respectively. The similar trends are observed on GlaS dataset, which respectively increased by 0.87% and 0.97% in Dice score and MIoU. This significant improvement benefits from minimizing redundant information while retaining the modality-specific information. Consequently, multi-modal features can be effectively fused and more valuable in improving segmentation results.

Impact of Amplitude and Phase Spectrum

Tab. 4 indicates the effectiveness of the frequency adapter, and it can be observed that phase spectrum information is more helpful for spectrum representation compared to phase information. In other words, the proposed frequency adapter can extract more valuable edge and overall information from the phase spectrum. Thus our proposed frequency adapters fully take advantage of amplitude information, which is more related to segmentation boundaries.

Impact of Different Balance Factors

We conduct the ablation study of different λ_3 values. In particular, λ_3 is respectively set as 0.1, 0.5, 1.0, and 1.5. The experimental results are shown in Tab. 5. It can be observed that the proposed method achieves the best performance when λ_3 is set to 1.0 on MoNuSeg. And the best performance is achieved on GlaS when λ_3 is 0.1. This is because of the size of the dataset.

On the one hand, the model would remove more information that it considers to be redundant if the size of the dataset is too small in the large lambda. Smaller datasets mean less information. In other words, excessive information compression within modalities may discard useful knowledge while λ_3 are further increased in small datasets. On the one hand, the fused feature would contain more redundant information while λ_3 are small in GlaS. It results in multi-modal features that could be effective to adopt and have a negative impact

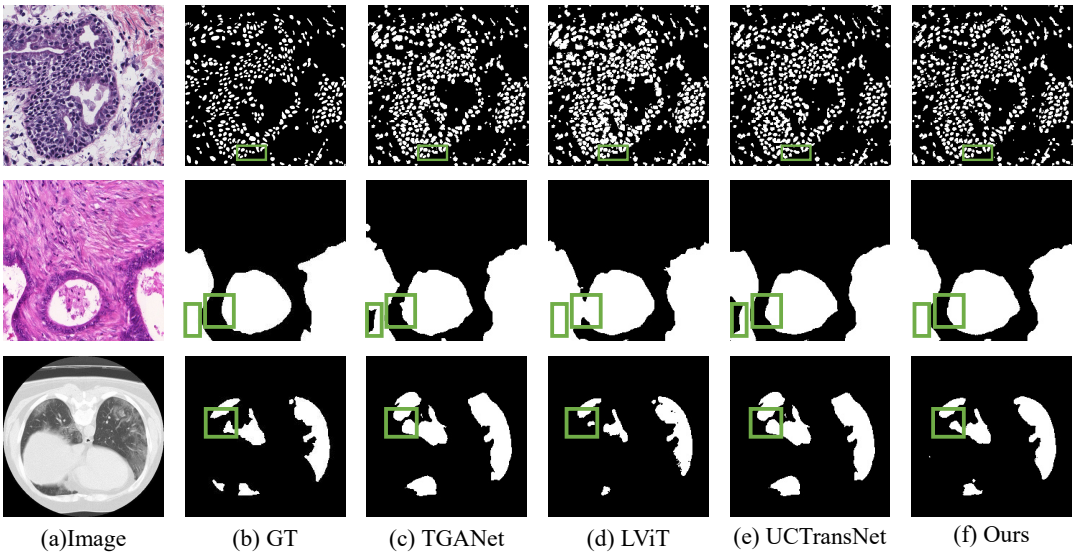

Figure 3: Qualitative comparison of the proposed method and other SOTA methods on MoNuSeg, GlaS, and MosMedData+ datasets. The green boxes highlight regions where the proposed method performs better than the other methods.

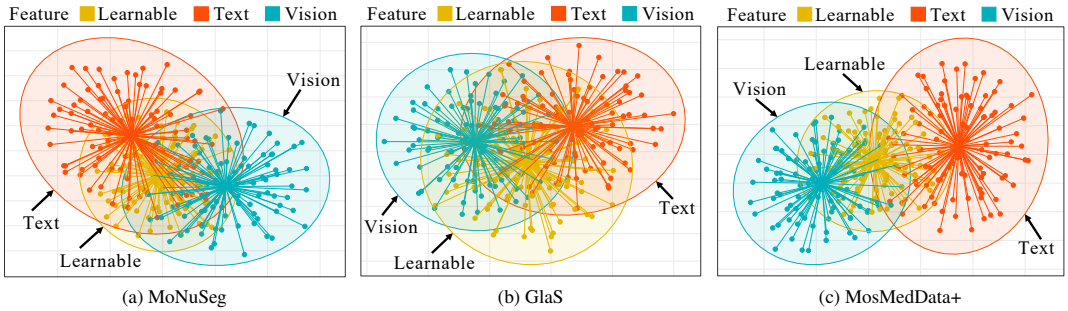

Figure 4: The visualization results of text, vision, and learnable features on three datasets. Specifically, red ellipses, yellow ellipses, and blue ellipses represent the distributions of text features, learnable features, and visual features, respectively. Different colored points represent the samples drawn from the corresponding distributions.

for decoder. Summarized, this variation could be attributed to differences in multimodal redundant levels in the disparity size of datasets, which conversely affects the optimal weighting to minimize redundant information in the training stage.

Interpretability of Proposed Fusion Method

To validate that the distribution of learnable features have modeled the discriminative information within multimodal data, we conduct random sampling from each distribution on three datasets. Subsequently, we reduce the obtained high-dimensional vectors to a two-dimensional plane using t-SNE. As illustrated in Fig. 4, the distribution of learnable features lies between the text and visual feature distribution, indicating the effectiveness of the proposed approach. Specifically, the distribution of learnable features can reduce the redundancies (modality-specific task-irrelated infirmation) multi-modal data by only modeling the most discriminative information. Meanwhile, the discriminative information extracted from text and vision feature contains both modality-common and modality-specific task-related information.

5 Conclusion

In this paper, a novel multi-modal method that leverages the text and image features for medical image segmentation. Specifically, we propose a novel multi-modal fusion paradigm, aiming to reduce the redundant information while retaining the modality-specific information. And the discriminative information can be extracted from the fused feature, which significantly improves the performance. Moreover, a novel frequency prompt encoder is designed to leverage high-frequency information for accurately segmenting boundaries. Experimental results demonstrate the effectiveness of the proposed method.

Acknowledgments

This work was supported by National Natural Science Foundation of China (62301621, 62301323) and Shenzhen Science and Technology Program (20231121172359002).

References

[Ates *et al.*, 2023] Gorkem Can Ates, Prasoon Mohan, and Emrah Celik. Dual cross-attention for medical image segmentation. *EAAI*, 126:107139, 2023.

[Cao *et al.*, 2022] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *ECCV*, pages 205–218. Springer, 2022.

[Cao *et al.*, 2023] Yunkang Cao, Xiaohao Xu, Chen Sun, Yuqi Cheng, Zongwei Du, Liang Gao, and Weiming Shen. Segment any anomaly without training via hybrid prompt regularization. *arXiv preprint arXiv:2305.10724*, 2023.

[Chen *et al.*, 2021] 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.

[Cherti *et al.*, 2023] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, pages 2818–2829, 2023.

[Ding *et al.*, 2022] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vlt: Vision-language transformer and query generation for referring segmentation. *IEEE TPAMI*, 2022.

[Ghiglia and Pritt, 1998] Dennis C Ghiglia and Mark D Pritt. Two-dimensional phase unwrapping: theory, algorithms, and software. *Wiely-Interscience, first ed.(April 1998)*, 1998.

[Han *et al.*,] Xianjun Han, Tiantian Li, and Hongyu Yang. Integrating prior knowledge into bi-branch pyramid network for medical image segmentation. *Available at SSRN 4564024*.

[Hofmanninger *et al.*, 2020] Johannes Hofmanninger, Forian Prayer, Jeanny Pan, Sebastian Röhrich, Helmut Prosch, and Georg Langs. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *European Radiology Experimental*, 4(1):1–13, 2020.

[Hosseini *et al.*, 2024] Mahdi S Hosseini, Babak Ehteshami Bejnordi, Vincent Quoc-Huy Trinh, Lyndon Chan, Danial Hasan, Xingwen Li, Stephen Yang, Taehyo Kim, Haochen Zhang, Theodore Wu, et al. Computational pathology: a survey review and the way forward. *Journal of Pathology Informatics*, page 100357, 2024.

[Huang *et al.*, 2020] Chao Huang, Zongju Peng, Yong Xu, Fen Chen, Qiuping Jiang, Yun Zhang, Gangyi Jiang, and Yo-Sung Ho. Online learning-based multi-stage complexity control for live video coding. *IEEE TIP*, 30:641–656, 2020.

[Huang *et al.*, 2021a] Chao Huang, Zhihao Wu, Jie Wen, Yong Xu, Qiuping Jiang, and Yaowei Wang. Abnormal event detection using deep contrastive learning for intelligent video surveillance system. *IEEE TII*, 18(8):5171–5179, 2021.

[Huang *et al.*, 2021b] Chao Huang, Zehua Yang, Jie Wen, Yong Xu, Qiuping Jiang, Jian Yang, and Yaowei Wang. Self-supervision-augmented deep autoencoder for unsupervised visual anomaly detection. *IEEE Trans Cybern*, 52(12):13834–13847, 2021.

[Huang *et al.*, 2022a] Chao Huang, Chengliang Liu, Jie Wen, Lian Wu, Yong Xu, Qiuping Jiang, and Yaowei Wang. Weakly supervised video anomaly detection via self-guided temporal discriminative transformer. *IEEE Trans Cybern*, 2022.

[Huang *et al.*, 2022b] Chao Huang, Jie Wen, Yong Xu, Qiuping Jiang, Jian Yang, Yaowei Wang, and David Zhang. Self-supervised attentive generative adversarial networks for video anomaly detection. *IEEE TNNLS*, 2022.

[Kim *et al.*, 2021] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, pages 5583–5594. PMLR, 2021.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Kumar *et al.*, 2017] Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE TMI*, 36(7):1550–1560, 2017.

[Li *et al.*, 2023a] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*, 2023.

[Li *et al.*, 2023b] Zihan Li, Yunxiang Li, Qingde Li, Puyang Wang, Dazhou Guo, Le Lu, Dakai Jin, You Zhang, and Qingqi Hong. Lvit: language meets vision transformer in medical image segmentation. *IEEE TMI*, 2023.

[Liu *et al.*, 2023] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *ICCV*, pages 21152–21164, 2023.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[Lu *et al.*, 2022] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *CVPR*, pages 5206–5215, 2022.

[Mao *et al.*, 2023] Xintian Mao, Yiming Liu, Fengze Liu, Qingli Li, Wei Shen, and Yan Wang. Intriguing findings of frequency selection for image deblurring. In *AAAI*, volume 37, pages 1905–1913, 2023.

[Mildenhall *et al.*, 2021] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as

neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[Morozov *et al.*, 2020] Sergey P Morozov, AE Andreychenko, NA Pavlov, AV Vladzymyrskyy, NV Ledikhova, VA Gombolevskiy, Ivan A Blokhin, PB Gelezhe, AV Gonchar, and V Yu Chernina. Mosmeddata: Chest ct scans with covid-19 related findings dataset. *arXiv preprint arXiv:2005.06465*, 2020.

[Nlong Zhao *et al.*, 2023] Brian Nlong Zhao, Yuhang Xiao, Jiashu Xu, Xinyang Jiang, Yifan Yang, Dongsheng Li, Laurent Itti, Vibhav Vineet, and Yunhao Ge. Dreamdistribution: Prompt distribution learning for text-to-image diffusion models. *arXiv e-prints*, pages arXiv–2312, 2023.

[Radford *et al.*, 2021] 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 *ICML*, pages 8748–8763. PMLR, 2021.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.

[Sirinukunwattana *et al.*, 2017] Korsuk Sirinukunwattana, Josien PW Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J Matuszewski, Elia Bruni, Urko Sanchez, et al. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, 35:489–502, 2017.

[Sun *et al.*, 2023] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[Tang *et al.*, 2021] Xianlun Tang, Jiangping Peng, Bing Zhong, Jie Li, and Zhenfu Yan. Introducing frequency representation into convolution neural networks for medical image segmentation via twin-kernel fourier convolution. *Computer Methods and Programs in Biomedicine*, 205:106110, 2021.

[Thawkar *et al.*, 2023] Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*, 2023.

[Tomar *et al.*, 2022] Nikhil Kumar Tomar, Debesh Jha, Ulas Bagci, and Sharib Ali. Tganet: Text-guided attention for improved polyp segmentation. In *MICCAI*, pages 151–160. Springer, 2022.

[Wang *et al.*, 2021a] Sihan Wang, Lei Li, and Xiahai Zhuang. Attu-net: attention u-net for brain tumor segmentation. In *MICCAI Workshop*, pages 302–311. Springer, 2021.

[Wang *et al.*, 2021b] Wei Wang, Baopu Li, Mengzhu Wang, Feiping Nie, Zhihui Wang, and Haojie Li. Confidence regularized label propagation based domain adaptation. *IEEE TCSVT*, 32(6):3319–3333, 2021.

[Wang *et al.*, 2021c] Wei Wang, Haojie Li, Zhengming Ding, Feiping Nie, Junyang Chen, Xiao Dong, and Zhihui Wang. Rethinking maximum mean discrepancy for visual domain adaptation. *IEEE TNNLS*, 34(1):264–277, 2021.

[Wang *et al.*, 2022a] Haonan Wang, Peng Cao, Jiaqi Wang, and Osmar R Zaiane. Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In *AAAI*, volume 36, pages 2441–2449, 2022.

[Wang *et al.*, 2022b] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022.

[Wang *et al.*, 2023] Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. Sam-clip: Merging vision foundation models towards semantic and spatial understanding. *arXiv preprint arXiv:2310.15308*, 2023.

[Wei *et al.*, 2023] Xiaobao Wei, Jiajun Cao, Yizhu Jin, Ming Lu, Guangyu Wang, and Shanghang Zhang. I-medsam: Implicit medical image segmentation with segment anything. *arXiv preprint arXiv:2311.17081*, 2023.

[Wu *et al.*, 2023] Yongjian Wu, Yang Zhou, Jiya Saiyin, Bingzheng Wei, Maode Lai, Jianzhong Shou, Yubo Fan, and Yan Xu. Zero-shot nuclei detection via visual-language pre-trained models. In *MICCAI*, pages 693–703. Springer, 2023.

[Yang *et al.*, 2022] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, pages 18155–18165, 2022.

[Yao *et al.*, 2023] Wenjian Yao, Jiajun Bai, Wei Liao, Yuheng Chen, Mengjuan Liu, and Yao Xie. From cnn to transformer: A review of medical image segmentation models. *arXiv preprint arXiv:2308.05305*, 2023.

[Zhang *et al.*, 2023] Shaoteng Zhang, Jianpeng Zhang, Yutong Xie, and Yong Xia. Tpro: Text-prompting-based weakly supervised histopathology tissue segmentation. In *MICCAI*, pages 109–118. Springer, 2023.

[Zhang *et al.*, 2024] Yilan Zhang, Yingxue Xu, Jianqi Chen, Fengying Xie, and Hao Chen. Prototypical information bottlenecking and disentangling for multimodal cancer survival prediction. *arXiv preprint arXiv:2401.01646*, 2024.

[Zhao *et al.*, 2023] Zihao Zhao, Yuxiao Liu, Han Wu, Yonghao Li, Sheng Wang, Lin Teng, Disheng Liu, Xiang Li, Zhiming Cui, Qian Wang, et al. Clip in medical imaging: A comprehensive survey. *arXiv preprint arXiv:2312.07353*, 2023.

[Zhu *et al.*, 2023] Qiang Zhu, Pengfei Li, and Qianhui Li. Attention retractable frequency fusion transformer for image super resolution. In *CVPR*, pages 1756–1763, 2023.