

# Frequency-domain Multi-modal Fusion for Language-guided Medical Image Segmentation

Bo Yu<sup>1†</sup>, Jianhua Yang<sup>2†</sup>, Zetao Du<sup>3</sup>, Yan Huang<sup>2,4</sup>, Chenglong Li<sup>5</sup>, and Liang Wang<sup>2,4✉</sup>

<sup>1</sup> School of Computer Science and Technology, Anhui University

<sup>2</sup> NLPR, MAIS, Institute of Automation, Chinese Academy of Sciences

<sup>3</sup> School of Information Science and Technology, ShanghaiTech University

<sup>4</sup> School of Artificial Intelligence, University of Chinese Academy of Sciences

<sup>5</sup> School of Artificial Intelligence, Anhui University

wangliang@nlpr.ia.ac.cn

**Abstract.** Automatically segmenting infected areas in radiological images is essential for diagnosing pulmonary infectious diseases. Recent studies have demonstrated that the accuracy of the medical image segmentation can be improved by incorporating clinical text reports as semantic guidance. However, the complex morphological changes of lesions and the inherent semantic gap between vision-language modalities prevent existing methods from effectively enhancing the representation of visual features and eliminating semantically irrelevant information, ultimately resulting in suboptimal segmentation performance. To address these problems, we propose a Frequency-domain Multi-modal Interaction model (FMISeg) for language-guided medical image segmentation. FMISeg is a late fusion model that establishes interaction between linguistic features and frequency-domain visual features in the decoder. Specifically, to enhance the visual representation, our method introduces a Frequency-domain Feature Bidirectional Interaction (FFBI) module to effectively fuse frequency-domain features. Furthermore, a Language-guided Frequency-domain Feature Interaction (LFFI) module is incorporated within the decoder to suppress semantically irrelevant visual features under the guidance of linguistic information. Experiments on QaTa-COV19 and MosMedData+ demonstrated that our method outperforms the state-of-the-art methods qualitatively and quantitatively.

**Keywords:** Medical Image Segmentation · Frequency-domain Features · Multi-modal Fusion.

## 1 Introduction

The technology of medical image segmentation (MIS) is crucial for delineating pathological areas of pulmonary infectious diseases, such as COVID-19. It facilitates the precise identification of lesions and greatly aids in diagnosis, treatment

---

<sup>†</sup> These authors contributed equally to this work.

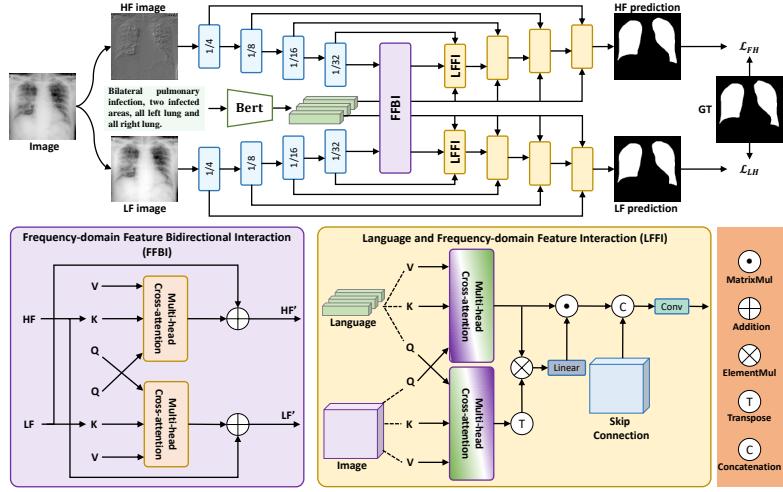
The code is available at <https://github.com/demoyu123/FMISeg>.

planning, and disease monitoring. With the rapid developments of deep learning, numerous MIS methods based on CNN (e.g., U-Net [1] and U-Net++ [2]) and hybrid CNN-Transformer (e.g., TransUNet [3] and SwinUnet [4]) architectures have been proposed for the segmentation of pulmonary lesions from radiological images. These methods substantially assist physicians in identifying and evaluating pulmonary structures and pathological anomalies. Despite the remarkable progress of these methods, the intricate morphological characteristics of lesion regions (e.g., shape, size, and blurry boundaries) still pose critical challenges to effectively boosting the segmentation accuracy of pulmonary lesions.

In clinical practice, pulmonary imaging is typically accompanied by clinical text reports, which provide detailed descriptions of lesion regions, including their position, shape, size, number, and other relevant characteristics. Inspired by the significant performance improvement achieved by integrating textual information with visual information in MedCLIP [5], the task of language-guided medical image segmentation (LMIS) has attracted increasing attention [6,7,8,9,10,11,12,13]. This task involves providing a medical image along with its corresponding text and predicting segmentation masks of pulmonary lesions. By leveraging the semantic guidance from the text, LMIS approaches significantly improve segmentation performance compared to uni-modal methods [1,2,3,4]. To bridge the semantic gap between the medical image and the text in multi-modal frameworks, various fusion strategies have been explored, where linguistic and visual features are integrated either within visual encoder blocks (early fusion) [6,7,11] or decoder blocks (late fusion) [8,9,10,12,13]. Among these methods, both unidirectional interaction [8,9,10] and bidirectional interaction [12,13] between two modalities through attention mechanisms, as well as language-guided adapters in SAM [14] have been investigated to improve semantic alignment.

However, two key issues still hinder the accurate localization and segmentation of target lesions specified by textual descriptions. Firstly, *the visual representation of the medical image lacks sufficient discriminative ability*. The complex morphological changes of lesions indicate that the extraction of distinguished features is important for accurate segmentation, especially for the small or subtle lesion regions. Compared to spatial-domain features, frequency-domain features can enrich visual representations by providing complementary structural and textural information, which is beneficial for MIS [15]. Nevertheless, the integration of frequency-domain features into LMIS task has not been explored to date. Secondly, *semantically irrelevant visual information cannot be effectively suppressed*. Although existing LMIS methods [8,9,10,12,13] leverage attention mechanisms to integrate visual and linguistic features, they often struggle to distinguish the lesion areas described by the text from complex anatomical backgrounds. For instance, the cross-modal attention in LanGuideSeg [8] is insufficient to eliminate irrelevant visual information within the decoder, leading to suboptimal segmentation performance.

In this study, we propose a Frequency-domain Multimodal Interaction framework (FMISeg) to address the aforementioned issues in LMIS task. FMISeg follows a late fusion strategy to integrate textual information into high fre-



**Fig. 1.** Overview of the proposed frequency-domain multimodal interaction method.

quency (HF) and low frequency (LF) features within decoder blocks. Specifically, FMISeg first utilizes wavelet transform to generate HF and LF images from the raw image. Then, HF and LF images are fed into a dual-branch encoder to extract corresponding HF and LF features. Since HF features contain fine-grained textural details, while LF features encode high-level semantic contexts, their combination can enhance the representation of the raw image and contributes to more accurate lesion segmentation. To this end, our method introduces a Frequency-domain Feature Bidirectional Interaction (FFBI) module to fuse HF and LF features before feeding them into a dual-branch decoder. Furthermore, to effectively model the interaction between linguistic and visual features, we propose a Language and Frequency-domain Feature Interaction (LFFI) module within the decoder. This module first establishes bidirectional interaction between linguistic and visual features through a cross-attention mechanism. The LFFI module subsequently utilizes the generated filter weights to reweight the output visual features, thereby suppressing semantically irrelevant visual information. We conduct experiments on the QaTa-COV19 [16] and MosMedData+[17] datasets to validate the effectiveness and state-of-the-art performance of our method.

## 2 Methodology

An overview of our proposed method is illustrated in Fig. 1. The framework is comprised of four key components, including a dual-branch visual encoder to extract HF and LF features, a language encoder to extract linguistic features, the FFBI module to exchange supplementary information between HF and LF features at the final stage of the encoder, and a language-guided dual-branch decoder with LFFI modules to effectively align linguistic and visual modalities.

## 2.1 Visual and Linguistic Feature Extraction

**Vision Encoder:** For an input raw image  $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ , we follow XNet [15] to apply wavelet transform to decompose it into the LF image  $\mathbf{I}_L$  and the HF image  $\mathbf{I}_H$ . These LF and HF images are subsequently fed into dual-branch visual encoders (i.e., ConvNeXt-Tiny [18]) to extract LF and HF visual features at different stages. Following previous works [10,11,12,13], we extract multi-stage LF and HF features with downsampling rates of 4, 8, 16, and 32. The extracted features are denoted as  $\mathbf{F}_m^1 \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C_1}$ ,  $\mathbf{F}_m^2 \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C_2}$ ,  $\mathbf{F}_m^3 \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C_3}$ , and  $\mathbf{F}_m^4 \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times C_4}$ . Here,  $H$  and  $W$  represent the height and width of the input image, and  $m \in \{LF, HF\}$ , where  $LF$  and  $HF$  indicate low frequency and high frequency, respectively.

**Language Encoder:** For the input clinical text  $\mathbf{T} \in \mathbb{R}^L$ , we follow prior works [8,12,13] to adopt CXR-BERT [19] as the language encoder to extract word-level linguistic features  $\mathbf{F}_T \in \mathbb{R}^{L \times C}$ . Here,  $L$  and  $C$  denote the number of words in the text and the channel dimension of features, respectively. Benefiting from its domain-specific optimization for chest X-ray reports, the linguistic features extracted from CXR-BERT can effectively facilitate semantic alignment between textual prompts and medical images.

## 2.2 Frequency-domain Feature Bidirectional Interaction (FFBI)

The LF features can encode high-level semantic contexts (e.g., organ morphology and lesion localization) and suppress noise interference. In contrast, HF features retain textural details (e.g., lesion and tissue boundaries) but are susceptible to HF artifacts and stochastic noise. To improve the segmentation of pulmonary lesions, we propose a Frequency-domain Feature Bidirectional Interaction (FFBI) module at the final stage of visual encoders, as illustrated in the bottom-left of Fig. 1. This module dynamically recalibrates HF features using LF semantic guidance, while simultaneously refines LF features with HF boundary details. Specifically, we leverage the cross-attention mechanism to model the bidirectional interactions between LF features  $\mathbf{F}_{LF}^4 \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times C_4}$  and HF features  $\mathbf{F}_{HF}^4 \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times C_4}$ . This process can be formulated as:

$$\hat{\mathbf{F}}_{HF}^4 = LN(\mathbf{F}_{HF}^4 + MHCA(\mathbf{F}_{HF}^4, \mathbf{F}_{LF}^4, \mathbf{F}_{LF}^4)), \quad (1)$$

$$\hat{\mathbf{F}}_{LF}^4 = LN(\mathbf{F}_{LF}^4 + MHCA(\mathbf{F}_{LF}^4, \mathbf{F}_{HF}^4, \mathbf{F}_{HF}^4)), \quad (2)$$

where  $LN(\cdot)$  denotes layer normalization,  $MHCA(Q, K, V)$  represents multi-head cross-attention with inputs of query ( $Q$ ), key ( $K$ ), and value ( $V$ ). Through bi-directional interaction, LF and HF features are respectively enhanced with local textural details and global semantic information. As a result, the proposed module yields more robust visual representations across diverse medical images.

## 2.3 Language and Frequency-domain Feature Interaction (LFFI)

To effectively model cross-modal interactions between linguistic features and LF (or HF) features, we propose a Language and Frequency-domain Feature Interaction (LFFI) module, as illustrated in the bottom-right of Fig. 1. Specifically,

the LFFI module first establishes interactions between linguistic features and LF (or HF) features using two multi-head cross-attention structures. For clarity, we take LF features  $\mathbf{F}_{LF}$  from arbitrary stages of the decoder as an example and omit the specific stage index. The process can be represented as:

$$\mathbf{F}'_{LF} = MHCA(\mathbf{F}_{LF}, \mathbf{F}_T, \mathbf{F}_T), \quad (3)$$

$$\mathbf{F}'_T = MHCA(\mathbf{F}_T, \mathbf{F}_{LF}, \mathbf{F}_{LF}). \quad (4)$$

The interactions between linguistic and LF features via cross-attention are inadequate to bridge the semantic gap between modalities, as standard cross-attention may introduce semantically irrelevant noise and often struggles to capture fine-grained positional dependencies. To address this, we further design a semantically irrelevant filter to fuse  $\mathbf{F}'_{LF} \in \mathbb{R}^{h \times w \times C}$  and  $\mathbf{F}'_T \in \mathbb{R}^{L \times C}$ , where  $h$  and  $w$  denote height and width of the feature map, respectively. Specifically,  $\mathbf{F}'_{LF}$  and  $\mathbf{F}'_T$  first perform the matrix multiplication and obtain the feature vector  $\mathbf{F}_M \in \mathbb{R}^{h \times w \times L}$ . Then, a linear projection layer followed by a sigmoid function is applied to generate the filter weights with the same dimensions as  $\mathbf{F}'_{LF}$ . Finally, the weighted features resulting from the element-wise multiplication between the weights and  $\mathbf{F}'_{LF}$  are combined with encoder features via a skip connection to produce the output features  $\mathbf{F}_{LF}^o$ . This process can be described as:

$$\mathbf{F}_M = \mathbf{F}'_{LF} \otimes (\mathbf{F}'_T)^\top, \quad (5)$$

$$\mathbf{F}_{LH}^o = Conv(\mathbf{F}_{LF} + \mathbf{F}'_{LF} \odot \delta(Linear(\mathbf{F}_M))), \quad (6)$$

where  $\otimes$  denotes matrix multiplication,  $\odot$  means element-wise multiplication,  $\delta(\cdot)$  represents the sigmoid function,  $Linear(\cdot)$  indicates linear projection,  $Conv(\cdot)$  refers to convolution operation. Our dual-branch decoders progressively upsample the text-injected LF and HF features from high-level to low-level stages. Each branch independently predicts masks using two separate segmentation heads, which helps reduce learning bias. During training, we follow prior works [6,13] to adopt Dice loss and cross-entropy loss to optimize dense predictions.

### 3 Experiments

#### 3.1 Datasets and Evaluation Metrics

To evaluate our method, we conducted experiments on two public medical image segmentation datasets with text prompts, including MosMeData+ [17] and QaTa-COV19 [16]. MosMedData+ contains 2,729 COVID-19 lung CT scan slices annotated with binary segmentation masks. It's divided into 2,183 training, 273 validation, and 273 test samples. [6]. The QaTa-COV19 dataset contains 9,258 chest X-ray images, annotated with COVID-19 lesion details and textual descriptions of infection and location. It's split into 5,716 training, 1,429 validation, and 2,113 test samples. Following existing studies [6,7], we adopt two widely used evaluation metrics, namely the Dice coefficient (*Dice*) and mean intersection-over-union (*mIoU*), to evaluate the performance of the proposed method.

**Table 1.** Performance comparison of our FMISeg with existing medical image segmentation methods on the QaTa-COV19 and MosMedData+ datasets.

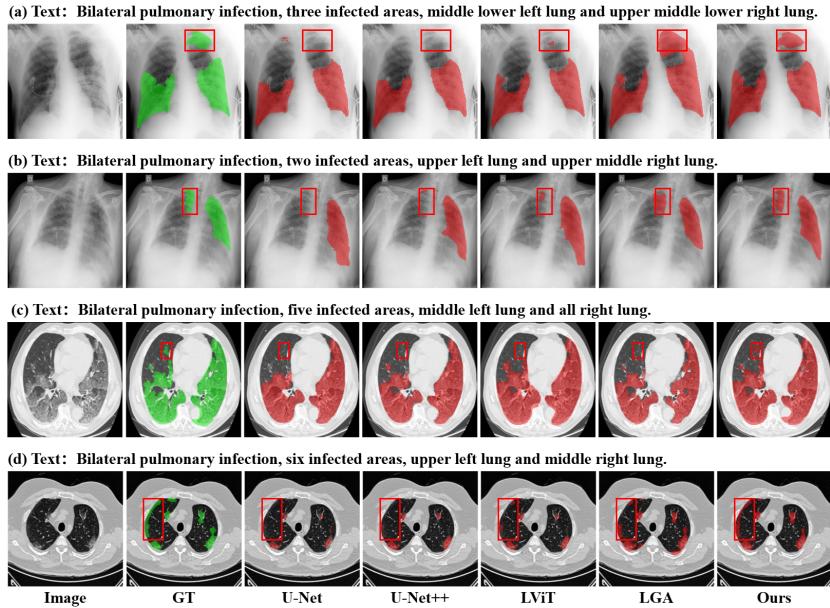
Methods	Backbone	Text	QaTa-COV19		MosMedData+	
			Dice(%)	mIoU(%)	Dice(%)	mIoU(%)
U-Net [1]	CNN	✗	79.02	69.46	64.60	50.73
UNet++ [2]	CNN	✗	79.62	70.25	71.75	58.39
nnUnet [22]	CNN	✗	80.42	70.81	72.59	60.36
TransUNet [3]	Hybrid	✗	78.63	69.13	71.24	58.44
Swin-Unet [4]	Hybrid	✗	78.07	68.34	63.29	50.19
UCTransNet [23]	Hybrid	✗	79.15	69.60	65.90	52.69
LViT-T [6]	Hybrid	✓	83.66	75.11	74.57	61.33
LGA [7]	Transformer	✓	84.65	76.23	75.63	62.52
CausalCLIPSeg [9]	Hybrid	✓	85.21	76.90	-	-
ReclMIS [11]	CNN	✓	85.22	77.00	77.48	65.07
SGSeg [10]	CNN	✓	87.41	77.82	-	-
LanGuideSeg [8]	CNN	✓	89.78	81.45	-	-
MAdapter [12]	CNN	✓	90.22	82.16	78.62	64.78
TGCAM [13]	CNN	✓	90.60	82.81	77.82	63.69
<b>FMISeg (ours)</b>	CNN	✓	<b>91.21</b>	<b>83.84</b>	<b>79.30</b>	<b>65.71</b>

### 3.2 Implementation Details

We implemented our method in PyTorch [20] with an NVIDIA RTX 3090 GPU. We adopted ConvNeXt-Tiny [18] as the backbone for the dual-branch visual encoder. The model was optimized by the AdamW [21] with an initial learning rate of 3e-4, which is eventually reduced to 1e-6, in conjunction with a cosine annealing learning rate strategy. For fair comparisons, the input image resolution was set to  $224 \times 224$ , and the default batch size was 32. The hidden dimension of the interaction module was set to 768.

### 3.3 Comparison With State-of-the-Art Methods

We compared the segmentation performance of our method with a series of state-of-the-art methods, including some classic uni-modal models (i.e., UNet [1], UNet++ [2], nnUnet [22], TransUNet [3], Swin-Unet [4], and UCTransNet [23]) and all multi-modal models (i.e., LViT-T [6], LGA [7], CausalCLIPSeg [9], SGSeg [10], LanGuideSeg [8], ReclMIS [11], MAdapter [12], and TGCAM [13]). The comparison results are shown in Table 1. Note that the results of multi-modal methods are directly cited from their original papers, while the performance of uni-modal methods is taken from the reproduction provided by LViT [6]. It can be observed that all methods without text show a significant performance gap compared to methods with text. For example, our method outperforms the best uni-modal nnUnet by 10.79% in *Dice* score and 13.03% in *mIoU* on the QaTa-COV19 dataset. When compared with LGA, which is based on foundation model



**Fig. 2.** Qualitative comparison of our method with uni-modal methods and multi-modal methods on QaTa-COV19 (a-b) and MosMedData+ (c-d).

SAM, our method has 6.56% and 7.61% improvement in terms of *Dice* score and *mIoU* on QaTa-COV19 dataset. Compared to methods using the same backbone, such as TGCAM, MAdapter, LanGuideSeg, and SGSeg, our method achieves superior performance on both datasets. Specifically, our method outperforms the best method (i.e., TGCAM) by 0.61% in *Dice* score and 1.03% in *mIoU* on QaTa-COV19 dataset, and by 1.48% in *Dice* score and 2.02% in *mIoU* on MosMedData+ dataset.

We also provide a qualitative comparison with uni-modal methods (i.e., U-Net and U-Net++) and multi-modal methods (i.e., LViT and LGA) in Fig. 2. It can be observed that uni-modal methods exhibit noticeable segmentation errors, particularly in boundary regions and detailed areas. While multi-modal methods show improved performance, they still suffer from missing or inaccurate segmentations in some complex and small areas. In contrast, our proposed method demonstrates higher accuracy, with more precise segmentation results.

### 3.4 Ablation Study

**Effectiveness of FFBI Module:** Our proposed method first decomposes the input raw image into LF and HF images, then establishes bidirectional interactions between LF and HF features in FFBI module. We conducted ablation studies to demonstrate the effectiveness of FFBI module, the experimental results are shown in Table 2. The models of #1, #2, and #3 are single-branch

**Table 2.** The effectiveness of the FFBI module.

No.	Model	QaTa-COV19		MosMedData+	
		Dice(%)	mIoU(%)	Dice(%)	mIoU(%)
#1	Raw Image	89.86	81.72	78.21	64.17
#2	HF Image	88.75	80.15	77.16	63.04
#3	LF Image	89.54	81.23	77.89	63.69
#4	Cat(HF, LF)	90.61	82.93	78.65	64.88
#5	FFBI(HF, LF)	<b>91.21</b>	<b>83.84</b>	<b>79.30</b>	<b>65.71</b>

**Table 3.** The impact of the numbers of LFFI layers. The LFFI layers in each branch of the decoder range from 1 to 4. “No Text” means the absence of text in our method.

No.	Model	QaTa-COV19		MosMedData+	
		Dice(%)	mIoU(%)	Dice(%)	mIoU(%)
#6	No Text	87.63	78.13	76.45	62.87
#7	1 layer	90.64	82.98	78.71	64.93
#8	2 layers	90.86	83.26	78.97	65.25
#9	3 layers	91.06	83.63	79.16	65.48
#10	4 layers	<b>91.21</b>	<b>83.84</b>	<b>79.30</b>	<b>65.71</b>

encoder-decoder models with inputs of raw, HF, and LF images, respectively. The models of #4 and #5 are two-branch encoder-decoder models, they adopt concatenation (Cat) and FFBI to fuse two modalities of visual features, respectively. The experimental results show that the performance of models fusing HF and LF visual features outperforms models using a single modality of visual features. This indicates that HF and LF modalities contain complementary information for segmentation. Furthermore, the FFBI module performs better than simply concatenating HF and LF features, since the bidirectional interaction with cross-attention is more conducive for mining complementary information from each other.

**The Impact of LFFI Layers:** In our two-branch decoder, we progressively inject textual information using the LFFI module and upsample the fused visual features from high-level to low-level for mask prediction. Here, we conducted ablation studies to investigate the impact of LFFI layers on segmentation performance, experimental results are shown in Table 3. These results show that Model #6 performs the worst on two datasets when textual information is absent. However, it still significantly outperforms all uni-modal methods in Table 1. As the layers of LFFI increase from high-level to low-level, the segmentation performance improves gradually. These experimental results demonstrate that the LFFI module effectively interacts with visual features and aligns the lesion region with textual semantics in our decoder.

## 4 Conclusion

In this paper, we propose a novel method FMISeg for language-guided medical image segmentation. FMISeg improves the segmentation accuracy of lesion regions from two aspects. Firstly, FMISeg extracts frequency-domain features and fuses different components of features with a frequency-domain feature bidirectional interaction module, resulting in the discriminative representation of visual features. Secondly, a language and frequency-domain feature interaction module is devised to effectively integrate linguistic features into frequency-domain features. Experiments show that FMISeg achieves state-of-the-art performance on the QaTa-COV19 and MosMedData+ datasets.

**Acknowledgments.** This work was jointly supported by the National Natural Science Foundation of China (62236010, 62322607 and 62276261), and Youth Innovation Promotion Association of Chinese Academy of Sciences under Grant 2021128.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this paper.

## References

1. Ronneberger, O., Fischer, P. and Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234-241 (2015)
2. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N. and Liang, J.: 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, pp. 3-11 (2018)
3. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L. and Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
4. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q. and Wang, M.: Swinunet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision, pp. 205-218 (2022)
5. Azad, R., Heidari, M., Shariatnia, M., Aghdam, E.K., Karimjafarbigloo, S., Adeli, E. and Merhof, D.: Wang, Z., Wu, Z., Agarwal, D. and Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. In Proceedings of the Conference on Empirical Methods in Natural Language Conference, Vol. 2022, pp. 3876 (2022)
6. Li, Z., Li, Y., Li, Q., Wang, P., Guo, D., Lu, L., Jin, D., Zhang, Y. and Hong, Q.: Lvit: language meets vision transformer in medical image segmentation. IEEE Transactions on Medical Imaging, 43(1), pp.96-107 (2023)
7. Hu, J., Li, Y., Sun, H., Song, Y., Zhang, C., Lin, L. and Chen, Y.W.: LGA: A Language Guide Adapter for Advancing the SAM Model's Capabilities in Medical Image Segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention pp. 610-620 (2024)

8. Zhong, Y., Xu, M., Liang, K., Chen, K., Wu, M.: Ariadne’s Thread: Using Text Prompts to Improve Segmentation of Infected Areas from Chest X-ray Images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 724-733 (2023)
9. Chen, Y., Wei, M., Zheng, Z., Hu, J., Shi, Y., Xiong, S., Mou, L.: Causal-CLIPSeg: Unlocking CLIP’s Potential in Referring Medical Image Segmentation with Causal Intervention. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 77-87 (2024)
10. Ye, S., Meng, M., Li, M., Feng, D., Kim, J.: Enabling Text-free Inference in Language-guided Segmentation of Chest X-rays via Self-guidance. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 242-252 (2024)
11. Huang, X., Li, H., Cao, M., Chen, L., You, C. and An, D.: Cross-modal conditioned reconstruction for language-guided medical image segmentation. IEEE Transactions on Medical Imaging, (2024)
12. Zhang, X., Ni, B., Yang, Y. and Zhang, L.: MAdapter: A Better Interaction Between Image and Language for Medical Image Segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 425-434 (2024)
13. Guo, Y., Zeng, X., Zeng, P., Fei, Y., Wen, L., Zhou, J. and Wang, Y.: Common Vision-Language Attention for Text-Guided Medical Image Segmentation of Pneumonia. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 192-201 (2024)
14. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y. and Dollár, P.: Segment anything. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 4015-4026 (2023)
15. Zhou, Y., Huang, J., Wang, C., Song, L. and Yang, G.: Xnet: Wavelet-based low and high frequency fusion networks for fully-and semi-supervised semantic segmentation of biomedical images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 21085-21096 (2023)
16. Degerli, A., Kiranyaz, S., Chowdhury, M.E., Gabbouj, M.: Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 2306–2310 (2022)
17. Andreychenko, A., Pavlov, N., Vladzimyrskyy, A., Ledikhova, N., Gombolevskiy, V., Gelezhe, P., Gonchar, A., Chernina, V.Y.: Mosmeddata: Chest ct scans with covid-19 related findings dataset. arXiv preprint arXiv:2005.06465 (2020)
18. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986 (2022)
19. Boecking, B., Usuyama, N., Bannur, S., Castro, D. C., Schwaighofer, A., Hyland, S.: Making the most of text semantics to improve biomedical vision–language processing. In: European conference on computer vision pp. 1-21 (2022)
20. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32 (2019)
21. Loshchilov, I.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
22. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J. and Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods, 18(2), pp.203-211 (2021)

23. Wang, H., Cao, P., Wang, J. and Zaiane, O.R.: Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In: Proceedings of the AAAI conference on artificial intelligence, pp. 2441-2449 (2022)