Self-adaptive Vision-Language Model for 3D Segmentation of Pulmonary Artery and Vein

Xiaotong Guo[⋆2], **Deqian Yang**[⋆1,3], **Dan Wang**[3], **Ying Zhu**[3], **Haochen Zhao**[5], **Yuan Li**[8], **Zhilin Sui**[2], **Tao Zhou**[4], **Lijun Zhang**[1], **Hui Meng**[†3], **Yanda Meng**[†6,7]

[1] Key Laboratory of System Software (Chinese Academy of Sciences) and State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, China

[2] National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Shenzhen, 518116, China

[3] School of Intelligent Science and Technology, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, 310024, China

[4] R&D Center, Guangxi Huayi Artificial Intelligence Medical Technology Co., Ltd

[5] School of Computer Science and Engineering, Beihang University, Beijing, China

[6] Department of Computer Science, University of Exeter, Exeter, UK

[7] Department of Cardiovascular & Metabolic Medicine, University of Liverpool, Liverpool, UK

[8] Guangzhou Jiayi Software Technology Co., Ltd

[†] *Corresponding authors:* `huimeng@ucas.ac.cn, Y.m.meng@exeter.ac.uk`

Abstract. Accurate segmentation of pulmonary structures is crucial in clinical diagnosis, disease study, and treatment planning. Significant progress has been made in deep learning-based segmentation techniques, but most require large amount of labeled data for training. Consequently, developing precise segmentation methods that demand fewer labeled datasets is paramount in medical image analysis. We constructed PAV-Seg3D, the largest Pulmonary Arteriovenous 3D Segmentation Dataset to date (718 scans).The emergence of pre-trained vision-language foundation models, such as CLIP, recently opened the door for universal computer vision tasks. However, exploring these models for pulmonary artery-vein segmentation is still limited. This paper proposes a novel framework called LA-CAF, which adopts pre-trained CLIP as a strong feature extractor for generating the segmentation of 3D CT scans, while adaptively aggregating the cross-modality of text and image representations. We propose a specially designed adapter module to fine-tune pre-trained CLIP with a self-adaptive learning strategy to effectively fuse the two modalities of embeddings. We validate LA-CAF on two datasets: PAV-Seg3D and the public PARSE2022 dataset. The experiments show that our method outperformed other state-of-the-art methods by a large margin. The dataset and code is made publicly available on https://github.com/zhuji423/LA-CAF-MICCAI2025.

Keywords: Vision-language model · Pulmonary A/V segmentation · CLIP.

1 Introduction

In recent years, pulmonary vascular diseases, including pulmonary embolism and pulmonary hypertension, have emerged as conditions with elevated morbidity and mortality rates. Computed tomography (CT) has been widely adopted as a diagnostic tool to elucidate tomographic patterns of pulmonary diseases [25]. Therefore, implementing automated pulmonary vascular segmentation is of significant clinical importance for achieving a three-dimensional reconstruction of the pulmonary vascular architectures. However, the manual delineation process remains labour-intensive due to the complexity of tubular structures. Segmentation methods for lung vessels have primarily focused on Convolutional Neural Networks (CNNs), particularly the U-Net architecture and its variants. These approaches have effectively maximized the potential of limited labeled data, especially from CT scans. Many semi-supervised and weakly supervised learning approaches are proposed based on pseudo labeling of the partially labeled data [13–18,21,23,26,27]. However, they often suffer significantly from the incorrectness of pseudo labels associated with unlabeled parts of the CT data [22,28].

The emerging paradigm of Vision-Language Model (VLM) pre-training with zero-shot transferability has attracted considerable interest for leveraging web-scale image-text pairs. Exemplified by CLIP [19], these models align modalities via a symmetric contrastive loss, minimizing cosine similarity between matched image-text pairs while maximizing it for negatives. This mechanism enables direct deployment to downstream tasks without task-specific fine-tuning. However, significant domain gaps persist when applying VLMs to specialized domains containing unseen data modalities (e.g., 3D medical imaging data absent from CLIP's training corpus). To address these adaptation challenges, numerous efforts are being made to adapt VLMs to specific task domains. For example, some approaches [4,19,24] modify the contrastive objectives to generative or alignment objectives to retrain a VLM. On the other hand, other methods fine-tune existing VLMs at a lower cost, including techniques such as prompt tuning [9] and feature adapters [3]. Thus, it raises a question: **How can we effectively utilize CLIP for 3D pulmonary artery/vein segmentation tasks?**

This work introduces an efficient Language-Guided self-adaptive Cross Attention Fusion framework called LA-CAF that integrates adaptive modules designed explicitly for pulmonary artery/vein (A/V) segmentation tasks. Our model not only preserves the performance of the pre-trained model to a great extent but also leverages the unique characteristics of PAV-Seg3D settings more effectively. By incorporating these adaptive modules, LA-CAF achieved an average DSC score of 77.26% on PAV-Seg3D, significantly surpassing the performance of other methods, such as nnU-Net [6] by an average DSC score of 9.74%, and nnFormer [29] by an average DSC score of 12.8%. Additionally, LA-CAF achieved an average DSC score of 84.71% on PARSE, surpassing other methods by 2.79% -9.5%. The primary contributions of this study are as follows:

(1) We exploit a large pre-trained vision-language model to segment pulmonary arteries and veins using a substantial local dataset PAV-Seg3D compris-

ing 718 annotated CT scans. PAV-Seg3D dataset will be made publicly available upon acceptance.

(2) We propose LA-CAF, an adaptive module incorporating attention mechanisms and data augmentation methods that are specially designed for PAV-Seg3D to highlight the vascular characteristics of pulmonary arteries and veins. These mechanisms significantly enhance the fusion of features between the language and visual models, yielding awe-inspiring results on the test dataset.

(3) Extensive experiments on PAV-Seg3D and PARSE2022 datasets have been conducted to validate the effectiveness of our proposed methods. The results of these experiments have demonstrated significant performance improvement over other state-of-the-art methods.

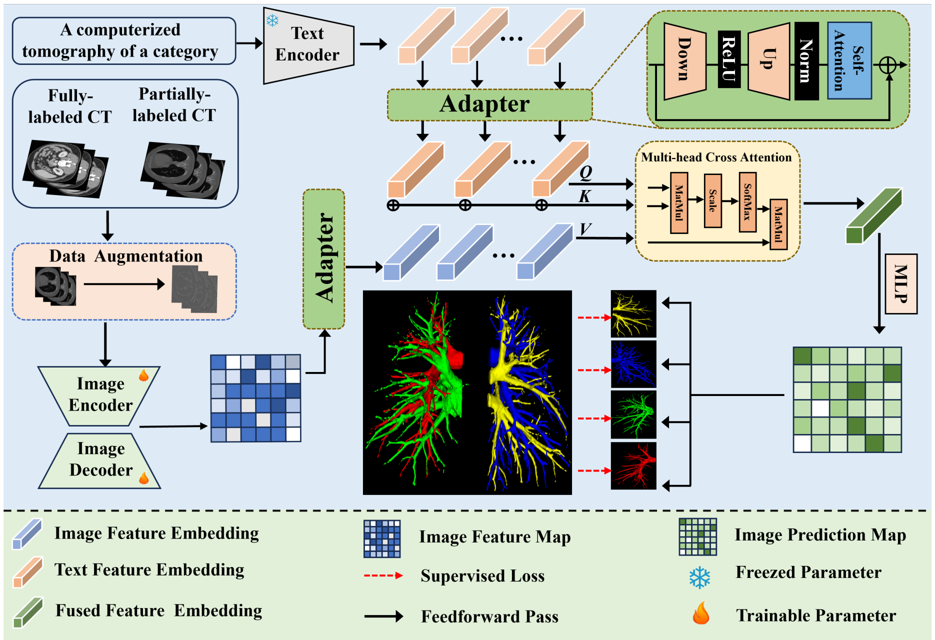

Fig. 1. Overview of the proposed LA-CAF framework, which comprises a text encoder and an image segmentation model. Best viewed in color.

2 Methods

CLIP (Contrastive Language-Image Pre-training) is a pretraining method developed by OpenAI [19]. Built upon the methodology of contrastive pretraining [10], it jointly optimizes a vision encoder and a text encoder, where the vision encoder is based on either ResNet [5] or Vision Transformer(ViT) [2]. The language encoder is rooted in a transformer-based model like BERT [1], forcing the paired image-text information to be as close as possible to the joint image-text latent space after encoding. We adopt the original CLIP model as our text embedding extractor. Trained on a vast collection of image-text pairs, CLIP learns visual representation through text supervision, known as prompt. We design a specialized prompt for our pulmonary vessel segmentation task.

2.1 Pretrained Text Encoder and Vision Model

Text Encoder: We use the original pre-trained CLIP encoder E_{text} with a specially designed medical prompt x_{text} (*i.e.* 'A computerized tomography of a category with small branches') to generate text embeddings $H_t \in \mathbb{R}^{K*D}$, where K represents the number of classes, and D represents the length of the embedding. The pre-trained encoder consists of a 12-layer 512-wide transformer with eight attention heads . The 512-wide output of the transformer is used as text embedding. To enhance the CLIP architecture's medical capability for medical image segmentation tasks, we use K text adapters A_{text} to fine-tune E_{text}.

$$H_t = E_{text}(x_{text}), H_t^a = A_{text}(H_t). \tag{1}$$

Vision Model: The CLIP-Driven Universal Model [8] introduces the first effective integration of CLIP's visual-language representations into medical 3D semantic understanding at the voxel level. Accounting for its strong ability to segment organs, the pre-trained model minimizes the time cost of training a model and inherits the weights that are suitable for organ segmentation. Therefore, we adopt a pre-trained U-Net model as the backbone for segmentation. Specifically, in our model, the 3D CT images $x_{img} \in \mathbb{R}^{H*W*L}$ are encoded into a feature map $H_v \in \mathbb{R}^{B*C*H*W*L}$ through the U-Net encoder E_{img}, where B represents batch and C represents channels. An image adapter A_{img} is used to map every batch B of raw high-level features to the embedding $H_v^a \in \mathbb{R}^{B*D}$.

$$H_v = E_{img}(x_{img}), H_v^a = A_{img}(H_v), \tag{2}$$

To match the shape of H_t, we duplicate H_v^a according to the class number K. We define:

$$\text{rep}(H, k) = concat[\underbrace{H, H, \ldots, H}_{k \text{ times}}], \tag{3}$$

then, we obtain the result $H_v^a \in \mathbb{R}^{B*K*D}$ by inputting $A_{img}(H_v)$,

$$H_v^a = rep(A_{img}(H_v), K). \tag{4}$$

2.2 Attention-based Self-Adaptive Learning Pipeline

The vision-language models have shown promising results across various tasks, attributable to their generalizability and interpretability. However, they often face the image and text distribution gap when applied to downstream tasks. For example, a medical segmentation dataset may have task-specific image styles and text formats that are not included in the pre-trained data sources. Therefore, how to fine-tune the pre-trained model at a lower cost and how to fuse different modalities of embeddings can be a noteworthy problem. We propose an attention-based self-adaptive learning pipeline to address the problem effectively.

In detail, the capability of CLIP is rooted in the natural image-text pairs. We enhance it for medical image segmentation tasks through fine-tuning. During training, the pre-trained CLIP encoder maintained frozen instead of fully adjusting all parameters to reduce the computing workload. We devise an adapter module and integrate it into designated positions shown in Fig 1. The adapter consists of a down-projection, ReLU activation, and up-projection with batch normalization and self-attention sequentially. The down-projection compresses the given embedding into a lower dimension using an MLP layer. At the same time, the up-projection expands the compressed embedding back to its original dimension using another MLP layer. Self-attention calculation captures the correlation of each class. The adapter trains CLIP embedding at a low cost with frozen parameters while intensely learning the attention of each class, guiding the segmentation model through the fusion method. Additionally, a trainable adapter is introduced into the vision model component, which is based on a pre-trained U-Net due to its manageable parameter size, making it an efficient starting point for training. This adapter facilitates the transition from image features to embeddings, enhancing the segmentation process.

In terms of fusing text and image embeddings after adopting the vision-language model as the backbone, many researchers [7,30] adopt simple strategies, such as direct plus or concatenation ignoring domain gap between text and image. Differently, we adopt the cross-attention(CA) module to integrate the two domain embeddings adaptively. The attention function serves as the operation to discover inner relationships from one modality to another. We have used the aforementioned adapters to get text embedding $H_t^a \in \mathbb{R}^{B*K*D}$ and image embedding $H_v^a \in \mathbb{R}^{B*K*D}$, for every batch $H_t^a(b)$, we calculate the attention scores H_f:

$$f_{\mathrm{CA}}(H) = \mathrm{softmax}\left(\frac{q(H_t^a)^T k(H_t^a + H_v^a(b))}{\sqrt{d_k}}\right) v(H_v^a(b)). \tag{5}$$

The input sequences of these two modalities are identically ordered in our input. Based on this, contextual clues can be propagated between modalities. As for the cross-attention module's detail, we choose the text embedding to be query (Q), image embedding to be value (V), and the plus of them to be the key (K). With language embedding's guidance, more precise features can be automatically selected rather than concatenated or plus in a hand-crafted way. The fusion of image and text embedding H_f uses a multi-layer perceptron (MLP) to generate parameters (θ_k). Three sequential convolutional layers with $1 \times 1 \times 1$ kernels filling with (θ_k) convert vision decoder output features F into k predictions, where $P_k = Sigmoid((F * \theta_{k_1}) * \theta_{k_2}) * \theta_{k_3}), \theta_k = \{\theta_{k_1}, \theta_{k_2}, \theta_{k_3}\}$. $*$ represents convolution operation. For each class k, we get every foreground class $P_k \in \mathbb{R}^{1 \times H \times W \times L}$. After that, we merge k classes of prediction into one prediction P, shown in Fig 1. P_k is supervised by label Y_k, where the overall loss is represented as:

$$\mathcal{L}_{sup} = \frac{1}{|B|} \sum_{i=1}^{|B|} \left[\mathcal{L}_S(P_k, Y_k) \right], \tag{6}$$

where $\mathcal{L}_S = \frac{1}{2} \left[\mathcal{L}_{Dice} + \mathcal{L}_{ce} \right]$; $\mathcal{L}_{Dice}$ and $\mathcal{L}_{ce}$ represent the Dice and cross-entropy losses, respectively.

3 Dataset and Implementation Detail

3.1 Dataset

We conduct extensive experiments on two datasets, including private PAV-Seg3D and public PARSE2022 [12] dataset. PAV-Seg3D collects a large-scale Pulmonary ArterioVenous Segmentation 3D Dataset from a real-world local hospital, comprising a total of 718 3D CT volumes provided in compressed NIFTI format. Among these, the pulmonary arteries and veins are manually annotated, where 79 CT scans are fully labeled and 639 CT scans are half-labeled, indicating either the left lung or the right lung are labeled. The sizes of these CT volumes range from 512×512×169 to 512×512×985, with varying slice thicknesses from 0.62 to 1.25 mm. Annotations are obtained from five junior clinicians (with one to five years of experience) who used MIMICS to manually refine the segmentation results under the supervision of two board-certified radiologists. PARSE2022 consists of 100 CT scans which are annotated at the pixel level for artery. In our experiments, we divide the dataset into training, validation, and test sets at a ratio of 7:1:2. Results are shown in Table 2.

3.2 Implementation Detail

Our model is implemented with U-Net as the backbone and optimizing the parameters via AdamW [11]. The training utilizes a batch size of 4 and a patch size of 96 × 96 × 96. The default initial learning rate is set to 8e-4, with a momentum of 0.9 and a decay of 1e-5. The framework is implemented in MONAI version 0.9.05. The Dice Similarity Coefficient (DSC), Normalized Surface Distance (NSD), Jaccard and 95% Hausdorff distance (HD95) are used to evaluate vessel segmentation performance in this work. We use specialized data augmentation to our dataset. Firstly, we adjusted the window width and level ranging from -700 to 300. Then, we calculate the Hessian matrix of the CT and obtain the eigenvalue to fill the Z-axis, which strengthens the CT's tubular structures, as shown in Fig 1.

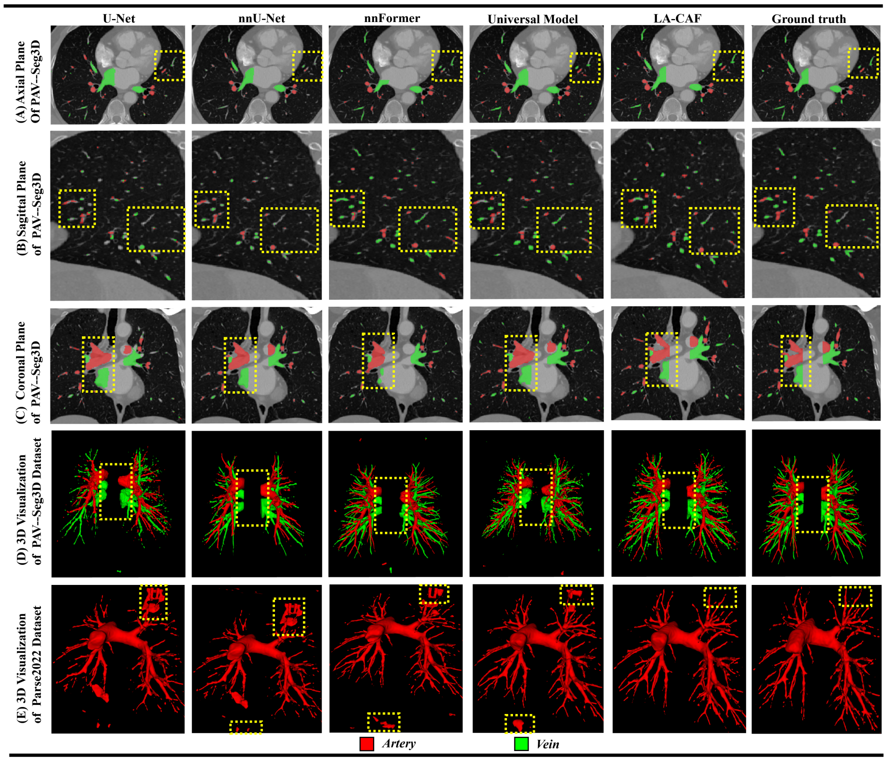
U-Net
nnU-Net
nnFormer
Universal Model
LA-CAF
Ground truth
(A) Axial Plane Of PAV–Seg3D
(B) Sagittal Plane of PAV–Seg3D
(C) Coronal Plane of PAV–Seg3D
(D) 3D Visualization of PAV–Seg3D Dataset
(E) 3D Visualization of Parse2022 Dataset
Artery
Vein

Fig. 2. Visualization of segmentation results on PAV-Seg3D(A-D) and Parse2022(E). The regions enclosed by the dashed yellow boxes indicate misclassification executed by other models; Mask colors have been set to red for arteries and green for veins.

Table 1. Ablation study of every component of our framework. UM indicates Universal Model, DA indicates Data Augmentation, AAP indicates Attention-based self-Adaptive learning Pipeline

UM	DA	AAP	DSC(%) ↑	Jaccard(%)↑	NSD↓	HD95↓
✓			$64.49_{0.45}$	$56.32_{0.15}$	$0.98_{0.03}$	$47.34_{0.35}$
✓	✓		$71.24_{0.25}$	$58.33_{0.02}$	$0.91_{0.06}$	$46.43_{0.01}$
✓	✓	✓	$76.22_{0.76}$	$62.74_{0.24}$	$0.86_{0.43}$	$14.48_{0.22}$

Table 2. Quantitative results of comparison experiment. Metrics are presented in the form of $mean_{std}$, where each method is evaluated over three trials for averaging.

Methods Dataset		U-Net	nnU-Net	nnFormer	Universal Model	LA-CAF
PAV-Seg3D	DSC(%) ↑	$61.23_{0.48}$	$67.52_{0.63}$	$64.46_{0.51}$	$70.24_{0.57}$	$77.26_{0.64}$
	Jaccard(%)↑	$48.34_{0.38}$	$56.45_{0.56}$	$51.32_{0.12}$	$57.25_{0.57}$	$64.13_{0.61}$
	NSD↓	$1.94_{0.19}$	$0.81_{0.05}$	$0.89_{0.06}$	$0.79_{0.02}$	$0.86_{0.05}$
	HD95↓	$132.23_{1.32}$	$43.76_{0.51}$	$86.61_{0.87}$	$25.34_{1.78}$	$13.72_{0.14}$
PARSE2022	DSC(%) ↑	$75.21_{0.32}$	$80.54_{0.36}$	$80.06_{0.23}$	$81.92_{0.33}$	$84.71_{0.32}$
	Jaccard(%)↑	$60.51_{0.25}$	$67.61_{0.22}$	$66.91_{0.48}$	$67.32_{0.28}$	$73.57_{0.18}$
	NSD↓	$0.77_{0.31}$	$0.85_{0.03}$	$0.84_{0.34}$	$0.84_{0.09}$	$0.92_{0.04}$
	HD95↓	$70.81_{2.56}$	$54.35_{0.46}$	$54.11_{0.56}$	$58.45_{3.69}$	$14.43_{0.26}$

4 Experiments

4.1 Ablation Studies

We conduct ablation studies to evaluate every component of LA-CAF using on PAV-Seg3D using 79 fully labeled dataset. The quantitative results of the different methods are presented in Table 1. The pre-trained Universal Model(UM) [8] is used as our baseline. We first use a specialized Data Augmentation(DA) to effectively use our half-labeled data, contributing a performance gain of over 6.75% DSC, 2.01% Jaccard over baseline. Subsequently, our proposed attention-based self-adaptive learning pipeline is introduced to fine-tune the pre-trained model and align text representations with image representations with an adaptive attention mechanism. We observe a further increment of 4.98% in DSC, 4.41% in Jaccard and a significant decrement of 31.95 in HD95. Each component significantly enhanced our method.

4.2 Comparison of Quantitative Results on Test Dataset

Table 2 presents a qualitative comparison of PAV-Seg3D and PARSE2022 test dataset against other state-of-the-art methods [6,8,20,29], which consists of 143 CT volumes and 20 CT volumes. Compared to the vanilla 3D U-Net, all other methods outperformed it in terms of DSC and Jaccard. Our method achieved DSC of 77.26% on PAV-Seg3D and 84.71% on PARSE2022, outperforming the baseline universal model [8] by 2.79% - 7.02% in DSC, as well as surpassing the supervised self-configuring model nnU-Net by 4.17% - 9.74% in DSC. Additionally, it outperformed the attention-based self-configuring model nnFormer by 4.54% - 12.8% in DSC. Overall, our model significantly surpasses all other compared methods on two datasets with integrating more semantic information from CLIP, achieving superior state-of-the-art performance. Furthermore, Figure 2 displays the visualization of segmentation results for our method and others, illustrating that our method segments both veins and arteries closer to the ground truth.

5 Conclusion

This work introduces a novel segmentation framework LA-CAF, integrating vision-language models with a self-adaptive feature learning pipeline and a designated data augmentation strategy. We leverage our partially annotated dataset to adhere to the best practices from large vision-language models. The framework incorporates our proposed adapter for fine-tuning CLIP embeddings, enhanced with self-attention to capture inter-class relationships. Furthermore, a cross-attention mechanism is seamlessly integrated to promote the effective fusion of the vision model with the segmentation model. We present PAV-Seg3D as the most extensive clinical dataset to date for pulmonary artery vein segmentation. The experiment results on two datasets affirm the superiority of our framework in the challenging task of pulmonary vessel segmentation against current state-of-the-art methods.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under grant numbers 62303127,62273009.

Disclosure of Interests.

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Devlin, J.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
3. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision **132**(2), 581–595 (2024)
4. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
6. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)

7. Lin, Y., Chen, M., Wang, W., Wu, B., Li, K., Lin, B., Liu, H., He, X.: Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15305–15314 (2023)
8. Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., A Landman, B., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z.: Clip-driven universal model for organ segmentation and tumor detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21152–21164 (2023)
9. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys **55**(9), 1–35 (2023)
10. Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., Tang, J.: Self-supervised learning: Generative or contrastive. IEEE transactions on knowledge and data engineering **35**(1), 857–876 (2021)
11. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
12. Luo, G., Wang, K., Liu, J., Li, S., Liang, X., Li, X., Gan, S., Wang, W., Dong, S., Wang, W., Yu, P., Liu, E., Wei, H., Wang, N., Guo, J., Li, H., Zhang, Z., Zhao, Z., Gao, N., An, N., Pakzad, A., Rangelov, B., Dou, J., Tian, S., Liu, Z., Wang, Y., Sivalingam, A., Punithakumar, K., Qiu, Z., Gao, X.: Efficient automatic segmentation for multi-level pulmonary arteries: The parse challenge (2024), `https://arxiv.org/abs/2304.03708`
13. Meng, H., Zhao, H., Yang, D., Wang, S., Li, Z.: Coarse to fine segmentation method enables accurate and efficient segmentation of organs and tumor in abdominal ct. In: MICCAI Challenge on Fast and Low-Resource Semi-supervised Abdominal Organ Segmentation, pp. 115–129. Springer (2023)
14. Meng, H., Zhao, H., Yu, Z., Li, Q., Niu, J.: Uncertainty-aware mean teacher framework with inception and squeeze-and-excitation block for miccai flare22 challenge. In: MICCAI Challenge on Fast and Low-Resource Semi-supervised Abdominal Organ Segmentation, pp. 245–259. Springer (2022)
15. Meng, Y., Chen, X., Zhang, H., Zhao, Y., Gao, D., Hamill, B., Patri, G., Peto, T., Madhusudhan, S., Zheng, Y.: Shape-aware weakly/semi-supervised optic disc and cup segmentation with regional/marginal consistency. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 524–534. Springer (2022)
16. Meng, Y., Zhang, H., Zhao, Y., Gao, D., Hamill, B., Patri, G., Peto, T., Madhusudhan, S., Zheng, Y.: Dual consistency enabled weakly and semi-supervised optic disc and cup segmentation with dual adaptive graph convolutional networks. IEEE transactions on medical imaging **42**(2), 416–429 (2022)
17. Meng, Y., Zhang, Y., Xie, J., Duan, J., Joddrell, M., Madhusudhan, S., Peto, T., Zhao, Y., Zheng, Y.: Multi-granularity learning of explicit geometric constraint and contrast for label-efficient medical image segmentation and differentiable clinical function assessment. Medical Image Analysis **95**, 103183 (2024)
18. Meng, Y., Zhang, Y., Xie, J., Duan, J., Zhao, Y., Zheng, Y.: Weakly/semi-supervised left ventricle segmentation in 2d echocardiography with uncertain region-aware contrastive learning. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). pp. 98–109. Springer (2023)
19. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)

20. Ronneberger, O., Fischer, P., Brox, T.: 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. pp. 234–241. Springer (2015)
21. Wang, Y., Zhao, T., Wang, X.: Fine-grained heartbeat waveform monitoring with rfid: A latent diffusion model. In: Proceedings of the 3rd International Workshop on Human-Centered Sensing, Modeling, and Intelligent Systems. pp. 86–91 (2025)
22. Wang, Y., Zhong, J., Kumar, R.: A systematic review of machine learning applications in infectious disease prediction, diagnosis, and outbreak forecasting (2025)
23. Yang, D., Zhao, H., Jin, G., Meng, H., Zhang, L.: Class-aware cross pseudo supervision framework for semi-supervised multi-organ segmentation in abdominal ct scans. In: Lin, Z., Cheng, M.M., He, R., Ubul, K., Silamu, W., Zha, H., Zhou, J., Liu, C.L. (eds.) Pattern Recognition and Computer Vision. pp. 148–162. Springer Nature Singapore, Singapore (2025)
24. Yang, J., Li, C., Zhang, P., Xiao, B., Liu, C., Yuan, L., Gao, J.: Unified contrastive learning in image-text-label space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19163–19173 (2022)
25. Yuan, C., Song, S., Yang, J., Sun, Y., Yang, B., Xu, L.: Pulmonary arteries segmentation from ct images using pa-net with attention module and contour loss. Medical Physics **50**(8), 4887–4898 (2023)
26. Zhao, H., Meng, H., Yang, D., Wu, X., Li, Q., Niu, J., et al.: Guidednet: Semi-supervised multi-organ segmentation via labeled data guide unlabeled data. In: ACM Multimedia 2024
27. Zhao, H., Niu, J., Meng, H., Wang, Y., Li, Q., Yu, Z.: Focal u-net: A focal self-attention based u-net for breast lesion segmentation in ultrasound images. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 1506–1511. IEEE (2022)
28. Zhong, J., Wang, Y.: Enhancing thyroid disease prediction using machine learning: A comparative study of ensemble models and class balancing techniques (2025)
29. Zhou, H.Y., Guo, J., Zhang, Y., Yu, L., Wang, L., Yu, Y.: nnformer: Interleaved transformer for volumetric segmentation. arXiv preprint arXiv:2109.03201 (2021)
30. Zhou, Z., Lei, Y., Zhang, B., Liu, L., Liu, Y.: Zegclip: Towards adapting clip for zero-shot semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11175–11185 (2023)