Exploring Text-enhanced Mixture-of-Experts for Semi-supervised Medical Image Segmentation with Composite Data

Qingjie Zeng[1,2] ⋆, Huan Luo[1*], Xinke Ma[1], Zilin Lu[1], Yang Hu[1], and Yong Xia[1,2,3(✉)]

[1] National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China
[2] Ningbo Institute of Northwestern Polytechnical University, Ningbo 315048, China
[3] Research & Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen 518057, China
yxia@nwpu.edu.cn

Abstract. Semi-supervised learning (SSL) has emerged as an effective approach to reduce reliance on expensive labeled data by leveraging large amounts of unlabeled data. However, existing SSL methods predominantly focus on visual data in isolation. Although text-enhanced SSL approaches integrate supplementary textual information, they still treat image-text pairs independently. In this paper, we explore the potential of jointly learning from related text-image datasets to further advance the capabilities of SSL. To this end, we introduce a novel text-enhanced Mixture-of-Experts (MoE) model, augmented with textual information, for semi-supervised medical image segmentation (TextMoE). TextMoE incorporates a universal vision encoder and a text-assisted MoE (TMoE) decoder, enabling it to simultaneously process CT-text and X-Ray-text data within a unified framework. To achieve effective knowledge integration from heterogeneous unlabeled data, a content regularization with frequency space exchange is designed, guiding TextMoE to learn modality-invariant representations. Additionally, the proposed TMoE decoder is enhanced by modality indicators, securing the effective fusion of visual and textual features. Finally, a differential loss is introduced to diversify the semantic understanding between visual experts, ensuring complementary insights to the overall interpretation. Experiments conducted on two public datasets indicate that TextMoE outperforms SSL and text-assisted SSL methods, achieving superior performance. Code is available at: https://github.com/jgfiuuuu/TextMoE.

Keywords: Semi-supervised learning · Medical image segmentation · Mixture-of expert · Textual knowledge.

⋆ Q. Zeng and H. Luo contributed equally. Corresponding author: Y. Xia.

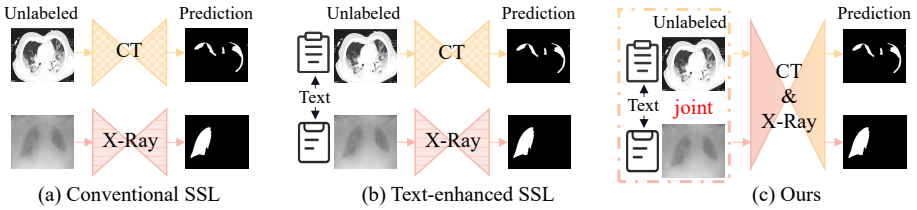

Fig. 1. Illustration of existing SSL paradigms: (a) Conventional SSL using only visual data; (b) Text-enhanced SSL; (c) Proposed SSL approach using joint CT-text and X-Ray-text datasets within a unified framework.

1 Introduction

Medical image segmentation plays a critical role in healthcare, assisting clinicians in accurately diagnosing and treating various conditions [28,4,27]. Traditional fully-supervised methods, while effective, face significant challenges in medical imaging due to the high cost and scarcity of labeled data. To address this, semi-supervised learning (SSL) has emerged as a promising approach [19,14,26]. SSL combines a small amount of labeled data with abundant unlabeled data, improving model performance and generalization. This approach is particularly relevant in medical imaging, where expert annotations are both time-consuming and expensive to obtain [5,6]. Established SSL techniques, such as consistency regularization [1,25,22] and pseudo-labeling [16,23], have been shown to enhance the segmentation accuracy of models across diverse clinical applications.

Despite these advancements, traditional SSL methods face several limitations. First, they typically rely on visual data alone [16,25] (see Fig. 1 (a)), which can miss crucial semantic details necessary for accurate segmentation. Second, they often struggle to generalize across different medical imaging modalities, resulting in suboptimal performance when applied to multi-source datasets [24]. Third, they underutilize the contextual information available in unlabeled data. Recent studies have explored text-enhanced SSL approaches [12,21,9], which integrate clinical notes and reports, providing richer semantic guidance that improves outcomes, particularly when dealing with diverse medical data.

The focus of existing text-enhanced SSL methods has been on two main objectives: (1) effectively combining textual and visual features to generate more accurate pseudo-labels, and (2) aligning image and text representations in a shared latent space for efficient feature integration. Key approaches include fine-grained pixel-word attention modules [8,20], which establish precise mappings between image regions and corresponding text segments, and cross-modal contrastive learning [17,21], which enhances the alignment of visual and textual features by maximizing the similarity of shared representations. Additionally, text-guided pseudo-label generation [12,9] has been used to generate more reliable pseudo-labels, improving the quality of the training signals for segmentation tasks from textual data. However, these methods typically treat paired image-text datasets independently, *e.g.*, training separate SSL models for each modality,

such as one for CT images and another for X-ray images (see Fig. 1 (b)). Given the correlation in imaging principles across modalities, such as CT and X-ray, we propose a novel approach that leverages these related text-image datasets within a single model to further enhance text-assisted SSL (see Fig. 1 (c)). To achieve this, we address two main challenges: (1) how to represent visual features from heterogeneous unlabeled data with textual guidance, requiring SSL models to handle variability in both visual and textual data and learn modality-invariant representations that capture essential semantic information across different data sources, and (2) how to resolve conflicts between modalities to ensure effective knowledge sharing. Although CT and X-ray share similar imaging principles, they can yield conflicting information due to differences in characteristics such as contrast and noise.

In this paper, we propose a novel text-enhanced Mixture-of-Experts (MoE) model, TextMoE, which jointly processes mixed CT-text and X-ray-text data within a unified framework for semi-supervised medical image segmentation. The TextMoE framework consists of a teacher and a student model, each comprising a universal vision encoder and a text-assisted MoE (TMoE) decoder. To mitigate modality conflict between CT and X-ray images, we introduce a content regularization loss that improves visual content understanding. This is achieved by swapping the frequency space of randomly selected unlabeled data, thereby enhancing robustness to modality-specific variations and refining feature representations. Furthermore, the TMoE decoder, aided by modality indicators, enables the seamless integration of visual and textual features. Finally, a differential loss is introduced to encourage diverse interpretations among the experts, further enriching semantic understanding.

The key contributions of this work are three-fold: (1) we propose TextMoE, a unified text-enhanced SSL framework capable of concurrently processing CT-text and X-ray-text data; (2) we address modality conflicts by devising a content-based loss and a modality indicator-based MoE architecture; and (3) extensive experiments and ablation studies demonstrate the superiority of our framework.

2 Method

2.1 Preliminaries

We construct a labeled dataset $D_l = \{(x_i^l, t_i^l, y_i^l)\}_{i=1}^{N_l}$ and an unlabeled dataset $D_u = \{(x_i^u, t_i^u)\}_{i=1}^{N_u}$, where x_i^l and x_i^u represent the i-th labeled and unlabeled images, respectively. The corresponding textual descriptions are denoted by t_i^l and t_i^u, while y_i^l is the ground truth label for x_i^l. Note that N_l and N_u indicate the total number of labeled and unlabeled data.

As shown in Fig. 2, the proposed TextMoE model is built upon the mean-teacher framework [18], with both the teacher and student models following a U-Net-like architecture [15]. Unlike conventional SSL and text-enhanced approaches, our TextMoE collaboratively learns from heterogeneous datasets using a unified model. The model consists of a universal vision encoder $f(\cdot;\theta_v)$ and

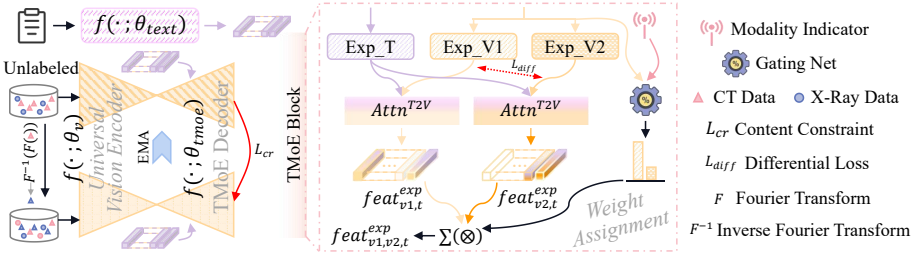

Fig. 2. Illustration of our TextMoE framework. To mitigate modality conflict, we introduce a content regularization loss $\mathcal{L}_{\text{cr}}$ based on frequency space exchange when processing unlabeled data. To effectively adapt feature representations, we design a TMoE decoder with a modality indicator to discern the type of data being processed. Our TextMoE is built upon the mean-teacher framework.

a TMoE decoder $f(\cdot;\theta_{\text{tmoe}})$. Additionally, a text encoder $f(\cdot;\theta_{\text{text}})$ is used to extract textual knowledge. We now delve into the details of TextMoE.

2.2 Architecture of TextMoE

The core design of TextMoE lies in its TMoE decoder $f(\cdot;\theta_{\text{tmoe}})$, which comprises three experts: $f(\cdot;\theta_{\text{exp\_t}})$, $f(\cdot;\theta_{\text{exp\_v1}})$ and $f(\cdot;\theta_{\text{exp\_v2}})$, each specializing in textual knowledge and visual semantic understanding. A gating network, augmented by a modality indicator, dynamically integrates the outputs of these experts by assigning adaptive weights for fusion. Since the forward process for labeled and unlabeled data is identical, we omit subscripts i, l, and u for clarity.

Text-guided Visual Feature Integration. Given an image x (from CT or X-Ray modalities), the general visual features feat_v are produced by the vision encoder $f(x;\theta_v)$, while the associated textual features feat_t are obtained from the text encoder $f(t;\theta_{\text{text}})$. Unlike typical MoE models, which integrate both visual and textual features uniformly, *i.e.*, integrating concurrently both visual and textual features during the forward pass to produce the outcomes [11], we propose employing textual features to guide the visual semantic understanding, and then fusing the enhanced visual features only. This design is motivated by the observation that segmentation outcomes are primarily determined by visual input, with textual data providing supplementary context. Accordingly, the above process can be formalized as:

$$\text{feat}^{\text{exp}}_{v_1} = f(\text{feat}_v;\theta_{\text{exp\_v1}}), \quad \text{feat}^{\text{exp}}_{v_2} = f(\text{feat}_v;\theta_{\text{exp\_v2}}), \tag{1}$$

where $\text{feat}^{\text{exp}}_{v_1}$ and $\text{feat}^{\text{exp}}_{v_2}$ are visual features processed by the respective visual experts. The textual features feat_t are passed through the text expert $f(\text{feat}_t;\theta_{\text{exp\_t}})$ to generate $\text{feat}^{\text{exp}}_t$. Two groups of text-enhanced visual features

are then computed as:

$$\begin{aligned} \text{feat}_{v_1,t}^{\text{exp}} &= \text{Attn}^{T2V}(\text{feat}_{v_1}^{\text{exp}}, \text{feat}_t^{\text{exp}}, \text{feat}_t^{\text{exp}}), \\ \text{feat}_{v_2,t}^{\text{exp}} &= \text{Attn}^{T2V}(\text{feat}_{v_2}^{\text{exp}}, \text{feat}_t^{\text{exp}}, \text{feat}_t^{\text{exp}}), \\ \text{Attn}^{T2V}(Q,K,V) &= \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \end{aligned} \tag{2}$$

where $\text{feat}_{v_1,t}^{\text{exp}}$ and $\text{feat}_{v_2,t}^{\text{exp}}$ are the improved visual features. $\text{Attn}^{T2V}(\cdot)$ follows a standard cross-attention operation mechanism. The two sets of enhanced visual features are then fused using a gating network $f(\cdot;\theta_g)$.

Weights Assignment with Modality Indicator. To address the modality-specific variations between CT and X-Ray images, we introduce a one-hot modality indicator, which is concatenated with the visual features to generate gating weights. These weights determine the contribution of each expert to the final feature representation. The weight computation is given by:

$$\{\text{weight}_{v1}\ , \text{weight}_{v2}\} = f(\text{feat}_v, [\text{Indicator}]; \theta_g), \tag{3}$$

where the gating weights $\{\text{weight}_{v1}\ , \text{weight}_{v2}\}$ are normalized using the SoftMax function. The final output within a TMoE block is a weighted combination of the enhanced visual features:

$$\text{feat}_{v_1,v_2,t}^{\text{exp}} = \text{weight}_{v1} \times \text{feat}_{v_1,t}^{\text{exp}} + \text{weight}_{v2} \times \text{feat}_{v_2,t}^{\text{exp}}. \tag{4}$$

Similarly, the refined visual features $\text{feat}_{v_1,v_2,t}^{\text{exp}}$ and txtual features $\text{feat}_t^{\text{exp}}$ are then passed to the next block for further processing.

2.3 Content Regularization based on Frequency Space Exchange

In this section, we describe the learning process for labeled and unlabeled data. For labeled data, the standard supervised loss is computed as:

$$\mathcal{L}_{\text{sup}} = \frac{1}{N_l}\sum_{i=1}^{N_l}\mathcal{L}(f(\text{feat}_{i,v}^l, \text{feat}_{i,t}^l; \theta_{\text{tmoe}}), y_i^l) \tag{5}$$

where $\text{feat}_{i,v}^l = f(x_i^l;\theta_v)$, $\text{feat}_{i,t}^l = f(t_i^l;\theta_{\text{text}})$, and $\mathcal{L}(\cdot)$ represents a combination of Dice and Cross-Entropy losses. To address the challenges associated with heterogeneous data sources, we propose exchanging the frequency space between randomly selected unlabeled images. The original image is processed by the teacher model to generate pseudo-labels, while the exchanged image is processed by the student model. This strategy improves robustness to modality-specific variations and enhances the model's understanding of both CT and X-ray content. The frequency exchange operation is given by:

$$x_{i,f}^u = \mathcal{F}^{-1}(\mathcal{F}(x_j^u)), \tag{6}$$

where $\mathcal{F}$ denotes the Fourier Transformation, $\mathcal{F}^{-1}$ stands for the Inverse Fourier Transformation, and $x^u_{i,f}$ represents the exchanged image, which combines the content of x^u_i with the style of x^u_j from a different modality. The content regularization loss $\mathcal{L}_{\text{cr}}$ is applied to the unlabeled data, as follows:

$$\mathcal{L}_{\text{cr}} = \frac{1}{N_u} \sum_{i=1}^{N_u} \mathcal{L}(f(\text{feat}^u_{i,v,f}, \text{feat}^u_{i,t}; \theta_{\text{tmoe}}), \hat{y}^u_i) \tag{7}$$

where $\text{feat}^u_{i,v,f} = f(x^u_{i,f}; \theta_v)$, $\text{feat}^u_{i,t} = f(t^u_i; \theta_{\text{text}})$, and $\hat{y}^u_i$ is the pseudo-label generated by the teacher model. The teacher is updated using an exponential moving average of the student with a momentum of 0.99.

2.4 Differentiated Semantic Understanding between Experts

To ensure diverse interpretations of the input image, we introduce a differentiated loss $\mathcal{L}_{\text{diff}}$, which is applied to the outputs of the visual experts. The loss is formulated as:

$$\mathcal{L}_{\text{diff}} = \frac{1}{N} \sum_{i=1}^{N} \frac{\|\text{feat}^{\text{exp}}_{v_1} \cdot \text{feat}^{\text{exp}}_{v_2}\|_2}{\|\text{feat}^{\text{exp}}_{v_1}\|_2 \cdot \|\text{feat}^{\text{exp}}_{v_2}\|_2} \tag{8}$$

where $N = N_l + N_u$. This loss encourages the visual experts to learn distinct, complementary representations of the input data. So far, the objective function of our TextMoE consists of three losses, formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sup}} + \mathcal{L}_{\text{cr}} + \mathcal{L}_{\text{diff}}. \tag{9}$$

As for inference, given a test image with associated text data, only the student model is used, with a text encoder providing supplemental textual knowledge.

3 Experiments and Results

3.1 Datasets and Implementation Details

Datasets and Metrics. We evaluated our method using two public datasets. The MosMedData+ dataset [7] consists of 2,729 CT scan slices for lung infections, while the QaTa-COV19 dataset [3] includes 9,258 COVID-19 Chest X-Ray images. Both datasets provide text descriptions detailing the number and location of infected areas, as outlined in [9]. To ensure fair comparison with existing methods, we used the same data splits as in [9,12]. Specifically, MosMedData+ was divided into 2,183 training, 273 validation, and 273 test images, QaTa-COV19 was split into 5,716 training, 1,429 validation, and 2,113 test images. 25% and 50% labels were considered, with Dice and mIoU evaluation metrics.
Implementation Details. Our model was trained using the AdamW optimizer with an input size of 224×224. A cosine annealing learning rate schedule was employed, starting at 3e-4 and gradually decreasing to a minimum of 1e-6. ConvNeXt [10] and CXR-BERT [2] were employed as the vision and text encoders.

Table 1. Comparisons on the QaTa-COV19 and MosMedData+ datasets, under label percentages of 25% and 50%. The most right presets the averaged Dice and mIoU scores. The best and second-best results are highlighted in **bold** and underline, respectively.

Dateset	Method	Modality	25%		50%		Avg. Results	
			Dice (%)	mIoU (%)	Dice (%)	mIoU (%)	Dice (%)	mIoU (%)
QaTa -COV19	UCMT	I	76.09	64.13	77.81	68.65	76.95	66.39
	BCP	I	74.79	65.26	75.57	65.31	75.18	65.29
	LeFeD	I	78.15	67.12	78.19	69.21	78.17	68.16
	LAVT	I+T	77.08	67.21	79.10	70.88	78.09	69.05
	LViT	I+T	78.12	66.75	80.32	72.16	79.22	69.46
	CPAM	I+T	80.21	70.59	82.08	74.12	81.15	72.36
	DuSSS	I+T	79.00	68.21	82.52	75.87	80.76	72.04
	Ours	I+T	**88.61**	**79.55**	**89.55**	**81.08**	**89.08**	**80.32**
MosMed Data+	UCMT	I	68.73	52.86	70.80	55.09	69.76	53.97
	BCP	I	69.56	53.42	71.38	55.76	70.47	54.59
	LeFeD	I	70.72	54.50	72.06	56.21	71.39	55.36
	LAVT	I+T	71.02	54.73	72.11	56.29	71.57	55.51
	LViT	I+T	71.38	54.82	72.25	56.33	71.82	55.58
	CPAM	I+T	71.62	55.08	72.37	56.47	72.01	55.78
	DuSSS	I+T	72.39	55.60	73.18	57.32	72.79	56.46
	Ours	I+T	**74.15**	**58.92**	**75.17**	**60.22**	**74.66**	**59.57**

Data augmentation techniques, including random resizing and cropping, were applied. To prevent overfitting, we incorporated an early stopping mechanism. To address the disparity in data sizes between QaTa-COV19 and MosMedData+, over-sampling was performed to balance the data. All experiments were conducted using PyTorch [13] on a single NVIDIA 3080Ti GPU.

3.2 Comparisons and Ablations

Results Analysis. Table 1 presents the model performance at label percentages of 25% and 50%, including comparisons with advanced SSL approaches (UCMT [16], BCP [1], LeFeD [25]) and text-enhanced SSL methods (LAVT [20], LViT [9], CPAM [8], DuSSS [12]). Based on the results, three primary observations can be made: (1) Our TextMoE consistently outperforms all other approaches in both Dice and mIoU scores across both datasets, particularly at lower labeled data percentages. This highlights the superiority of our method in concurrently leveraging limited image-text data from heterogeneous sources, making it available for clinical applications. (2) TextMoE shows higher performance gains on the QaTa-COV19 dataset (X-Ray) compared to the MosMedData+ dataset (CT). We attribute this to the more homogeneous nature of X-Ray images, which may make them more conducive to multimodal integration. In contrast, the CT images in MosMedData+ are more complex, requiring more nuanced modeling to handle their diverse and detailed structures. (3) Visualizations in Fig. 3 further demonstrate the efficacy of TextMoE in establishing seamless connections between image and text, highlighting its ability to integrate multimodal infor.

Table 2. Ablation studies on two datasets with label percentages of 25% and 50%. The right column displays the average gains compared to the baseline.

Setup	$\mathcal{L}_{\text{sup}}$	Plain-MoE [11]	TMoE	$\mathcal{L}_{\text{cr}}$	$\mathcal{L}_{\text{diff}}$	QaTa-COV19		MosMedData+		Avg. gain Δ	
						Dice (%)	mIoU (%)	Dice (%)	mIoU (%)	Dice (%)	mIoU (%)
25%	✓	×	×	×	×	77.88	67.53	67.55	50.99	-	-
	✓	✓	×	×	×	84.54	74.74	68.48	52.07	+ 3.80	+ 4.15
	✓	×	✓	×	×	85.63	74.87	70.72	54.70	+ 5.46	+ 5.53
	✓	×	✓	✓	×	87.59	77.92	72.60	56.99	+ 7.38	+ 8.20
	✓	×	✓	✓	✓	88.61	79.55	74.15	58.92	+ 8.85	+ 9.98
50%	✓	×	×	×	×	79.64	71.88	69.27	52.99	-	-
	✓	✓	×	×	×	87.23	77.35	70.91	54.94	+ 4.62	+ 3.71
	✓	×	✓	×	×	87.66	78.02	71.86	56.08	+ 5.31	+ 4.62
	✓	×	✓	✓	×	88.60	79.53	73.00	57.49	+ 6.35	+ 6.08
	✓	×	✓	✓	✓	89.55	81.08	75.17	60.22	+ 7.91	+ 7.07

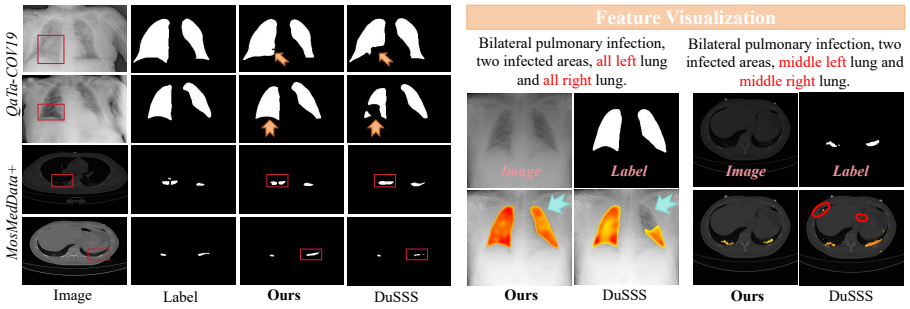

Fig. 3. Visualization analysis. The left side displays the segmentation results, while the right side highlights key features along with their corresponding textual descriptions.

Ablation Study. Table 2 presents the results for five configurations: (1) the baseline model with labeled supervision $\mathcal{L}_{\text{sup}}$ only; (2) using both labeled and unlabeled data with Plain-MoE, which fuses visual and textual features element-wise; (3) replacing Plain-MoE with our TMoE, which enhances visual features using text cues and fuses the enhanced visual features solely; (4) substituting the naive unsupervised teacher-student loss with our content regularization loss $\mathcal{L}_{\text{cr}}$; and (5) incorporating the differentiated semantic understanding loss $\mathcal{L}_{\text{diff}}$ to further diversify the knowledge learned from experts. As an example, with 25% labeled data, incorporating Plain-MoE already yields results on the QaTa-COV19 dataset that outperform both competing SSL and text-enhanced SSL methods, highlighting the importance of integrating data sources, especially when labels are limited. Replacing Plain-MoE with our TMoE leads to significant improvements on the MosMedData+ dataset, validating the efficacy of our MoE design for handling composite data. Furthermore, introducing $\mathcal{L}_{\text{cr}}$ to mitigate modality conflicts between CT and X-Ray results in notable improvements on both datasets, demonstrating the effectiveness of our content understanding approach. Finally, by diversifying semantics between visual experts using $\mathcal{L}_{\text{diff}}$, the gains achieved by TextMoE are further enhanced.

4 Conclusion

In this paper, we introduce a novel text-enhanced SSL model, TextMoE, designed to jointly process multiple image-text datasets within a unified framework. To effectively extract information from unlabeled data and resolve modality conflicts between CT and X-Ray images, we propose a content regularization technique based on frequency space exchange. To integrate complementary textual knowledge, we design a TMoE decoder with modality indicators, enhancing the fusion of visual and textual features for improving the visual semantics. Furthermore, we introduce a differential loss to promote diverse interpretations among visual experts, securing variability in the learned representations.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China under Grants 62171377 and 92470101, in part by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang, China, under Grant 2025C01201(SD2), in part by the Ningbo Clinical Research Center for Medical Imaging under Grant 2021L003 (Open Project 2022LYKFZD06), in part by the Shenzhen Science and Technology Program under Grants JCYJ20220530161616036, and in part by the Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University under Grant CX2025019.

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

References

1. Bai, Y., Chen, D., Li, Q., Shen, W., Wang, Y.: Bidirectional copy-paste for semi-supervised medical image segmentation. In: CVPR. pp. 11514–11524 (2023) 2, 7
2. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al.: Making the most of text semantics to improve biomedical vision–language processing. In: ECCV. pp. 1–21. Springer (2022) 6
3. Degerli, A., Kiranyaz, S., Chowdhury, M.E., Gabbouj, M.: Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images. In: ICIP. pp. 2306–2310. IEEE (2022) 6
4. Feuerriegel, S., Frauen, D., Melnychuk, V., Schweisthal, J., Hess, K., Curth, A., Bauer, S., Kilbertus, N., Kohane, I.S., van der Schaar, M.: Causal machine learning for predicting treatment outcomes. Nature Medicine **30**(4), 958–968 (2024) 2
5. Huang, Z., Jiang, R., Aeron, S., Hughes, M.C.: Systematic comparison of semi-supervised and self-supervised learning for medical image classification. In: CVPR. pp. 22282–22293 (2024) 2
6. Jin, C., Guo, Z., Lin, Y., Luo, L., Chen, H.: Label-efficient deep learning in medical image analysis: Challenges and future directions. arXiv preprint arXiv:2303.12484 (2023) 2
7. Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., Sethi, A.: A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE Transactions on Medical Imaging **36**(7), 1550–1560 (2017) 6

8. Lee, G.E., Kim, S.H., Cho, J., Choi, S.T., Choi, S.I.: Text-guided cross-position attention for segmentation: Case of medical image. In: MICCAI. pp. 537–546. Springer (2023) 2, 7
9. Li, Z., Li, Y., Li, Q., Wang, P., Guo, D., Lu, L., Jin, D., Zhang, Y., Hong, Q.: Lvit: language meets vision transformer in medical image segmentation. IEEE Transactions on Medical Imaging (2023) 2, 6, 7
10. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: CVPR. pp. 11976–11986 (2022) 6
11. Mustafa, B., Riquelme, C., Puigcerver, J., Jenatton, R., Houlsby, N.: Multimodal contrastive learning with limoe: the language-image mixture of experts. NeurIPS **35**, 9564–9576 (2022) 4, 8
12. Pan, Q., Qiao, W., Lou, J., Ji, B., Li, S.: Dusss: Dual semantic similarity-supervised vision-language model for semi-supervised medical image segmentation. In: AAAI (2025) 2, 6, 7
13. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. NeurIPS **32** (2019) 7
14. Ran, L., Li, Y., Liang, G., Zhang, Y.: Pseudo labeling methods for semi-supervised semantic segmentation: A review and future perspectives. IEEE Transactions on Circuits and Systems for Video Technology (2024) 2
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015) 3
16. Shen, Z., Cao, P., Yang, H., Liu, X., Yang, J., Zaiane, O.R.: Co-training with high-confidence pseudo labels for semi-supervised medical image segmentation. In: IJCAI. pp. 4199–4207 (2023) 2, 7
17. Song, X., Zhang, X., Ji, J., Liu, Y., Wei, P.: Cross-modal contrastive attention model for medical report generation. In: COLING. pp. 2388–2397 (2022) 2
18. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. NeurIPS **30** (2017) 3
19. Weng, Y., Zhang, Y., Wang, W., Dening, T.: Semi-supervised information fusion for medical image analysis: Recent progress and future perspectives. Information Fusion **106**, 102263 (2024) 2
20. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: Lavt: Language-aware vision transformer for referring image segmentation. In: CVPR. pp. 18155–18165 (2022) 2, 7
21. You, K., Gu, J., Ham, J., Park, B., Kim, J., Hong, E.K., Baek, W., Roh, B.: Cxr-clip: Toward large scale chest x-ray language-image pre-training. In: MICCAI. pp. 101–111. Springer (2023) 2
22. Zeng, Q., Lu, Z., Xie, Y., Lu, M., Ma, X., Xia, Y.: Reciprocal collaboration for semi-supervised medical image classification. In: MICCAI. pp. 522–532. Springer (2024) 2
23. Zeng, Q., Lu, Z., Xie, Y., Xia, Y.: Pick: Predict and mask for semi-supervised medical image segmentation. International Journal of Computer Vision pp. 1–16 (2025) 2
24. Zeng, Q., Xie, Y., Lu, Z., Lu, M., Wu, Y., Xia, Y.: Segment together: A versatile paradigm for semi-supervised medical image segmentation. IEEE Transactions on Medical Imaging (2025) 2
25. Zeng, Q., Xie, Y., Lu, Z., Lu, M., Zhang, J., Zhou, Y., Xia, Y.: Consistency-guided differential decoding for enhancing semi-supervised medical image segmentation. IEEE Transactions on Medical Imaging (2024) 2, 7

26. Zeng, Q., Xie, Y., Lu, Z., Xia, Y.: Pefat: Boosting semi-supervised medical image classification via pseudo-loss estimation and feature adversarial training. In: CVPR. pp. 15671–15680 (2023) 2
27. Zeng, Q., Xie, Y., Lu, Z., Xia, Y.: A human-in-the-loop method for pulmonary nodule detection in ct scans. Visual Intelligence **2**(1), 19 (2024) 2
28. Zhang, K., Zhou, R., Adhikarla, E., Yan, Z., Liu, Y., Yu, J., Liu, Z., Chen, X., Davison, B.D., Ren, H., et al.: A generalist vision–language foundation model for diverse biomedical tasks. Nature Medicine pp. 1–13 (2024) 2