This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Textmatch: Using Text Prompts to Improve Semi-supervised Medical Image Segmentation

Aibing Li[1], Xinyi Zeng[1], Pinxian Zeng[1], Sixian Ding[1], Peng Wang[1], Chengdi Wang[2(✉)], and Yan Wang[1(✉)]

[1]School of Computer Science, Sichuan University, China
wangyanscu@hotmail.com
[2] Department of Respiratory and Critical Care Medicine, West China Hospital, Chengdu, China
chengdi_wang@scu.edu.cn

Abstract. Semi-supervised learning, a paradigm involving training models with limited labeled data alongside abundant unlabeled images, has significantly advanced medical image segmentation. However, the absence of label supervision introduces noise during training, posing a challenge in achieving a well-clustered feature space essential for acquiring discriminative representations in segmentation tasks. In this context, the emergence of vision-language (VL) models in natural image processing has showcased promising capabilities in aiding object localization through the utilization of text prompts, demonstrating potential as an effective solution for addressing annotation scarcity. Building upon this insight, we present Textmatch, a novel framework that leverages text prompts to enhance segmentation performance in semi-supervised medical image segmentation. Specifically, our approach introduces a Bilateral Prompt Decoder (BPD) to address modal discrepancies between visual and linguistic features, facilitating the extraction of complementary information from multi-modal data. Then, we propose the Multi-views Consistency Regularization (MCR) strategy to ensure consistency among multiple views derived from perturbations in both image and text domains, reducing the impact of noise and generating more reliable pseudo-labels. Furthermore, we leverage these pseudo-labels and conduct Pseudo-Label Guided Contrastive Learning (PGCL) in the feature space to encourage intra-class aggregation and inter-class separation between features and prototypes, thus enhancing the generation of more discriminative representations for segmentation. Extensive experiments on two publicly available datasets demonstrate that our framework outperforms previous methods employing image-only and multi-modal approaches, establishing a new state-of-the-art performance.

Keywords: Medical image segmentation, Semi-supervised learning, Bilateral Prompt, Multi-views Consistency, Contrastive learning.

1 Introduction

Accurate results of medical image segmentation provide salient and insightful information for clinicians, facilitating clinical diagnosis, disease progression, and treatment planning. With the recent advancements in deep learning, numerous approaches have leveraged various types of deep neural networks trained on large annotated datasets, leading to outstanding performance across various medical image segmentation tasks [1-3]. Nevertheless, obtaining extensive pixel-level annotations is often time-consuming and labor-intensive, necessitating expertise and incurring significant costs. Hence, it is imperative to devise methods to alleviate the aforementioned constraints.

To tackle these issues, semi-supervised medical image segmentation has emerged as a promising technique, leveraging unlabeled data to enhance performance with only a limited number of labeled samples. According to [4], current semi-supervised approaches applied to medical image segmentation can be categorized into pseudo-labeling, consistency regularization, and hybrid methods. Pseudo-labeling methods [5-8] involve generating pseudo-labels for a large portion of unlabeled data to expand training data, which are further utilized to train the segmentation network in a self-training manner. Consistency regularization methods [9,10] encourage the similarity among predictions from the perturbed inputs to enhance the generalization capability of the model. Hybrid methods [11,12] combine the aforementioned ideas to achieve better performance.

Despite advancements in previous studies, semi-supervised medical image segmentation still faces challenges. As introduced, during the process of generating pseudo-labels, the lack of supervised data contributes to the accumulation of significant noise in these labels, which adversely affects the efficacy of model training. Moreover, current semi-supervised learning methods [9,10] provide supervision solely within the logit(pixel) space, yet they lack explicit guidance within the feature space. This absence of guidance impedes the acquisition of a well-clustered feature space, compromising the attainment of discriminative representations beneficial for segmentation tasks.

Recently, vision language (VL) models [13-15] have garnered considerable attention in natural image processing and have found applications in medical image analysis [16-19], showing promising results in fully supervised scenarios. As for medical image segmentation, numerous studies [18-20] have shown that the introduction of text prompts can improve the models' object localization ability and reduce the impact of noise, thus enhancing the segmentation performance. Nonetheless, in semi-supervised scenarios, the utilization of VL learning remains largely unexplored, as current methods [5-10] predominantly rely solely on image-level data, overlooking the potential of valuable complementary information provided by text prompts, which can act as an effective solution for addressing the scarcity of pixel-level annotations.

In this paper, we propose Textmatch, a novel semi-supervised medical image segmentation framework that explores the potential of text prompts for better segmentation performance. Specifically, we first introduce a Bilateral Prompt Decoder (BPD) to harmonize the discrepancy between visual and linguistic features and mine additional information, taking advantage of mutual complementarity among multiple modalities. Furthermore, we propose a Multi-view Consistency Regularization (MCR) strategy. This strategy incorporates both image and text perturbations to derive various augmented views with distinct image appearances and different text prompts of similar semantics. By incorporating a consistency regularization constraint to these views, we significantly reduce the impact of noise and generate more robust pseudo-labels.

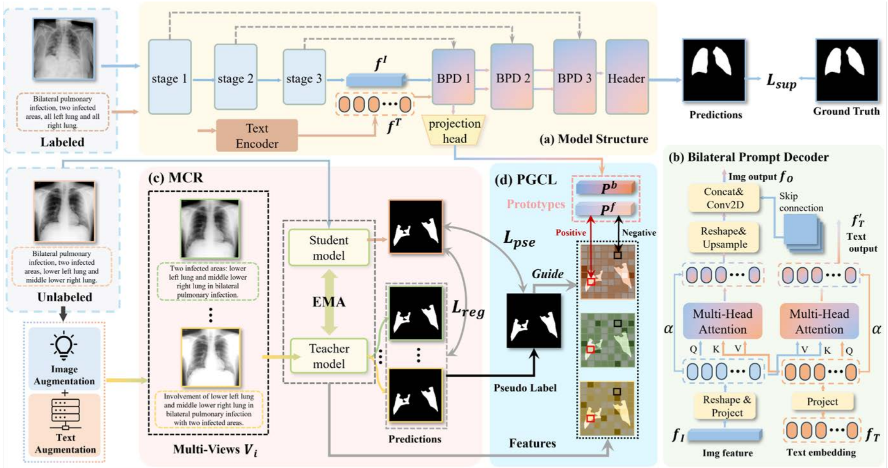

Fig. 1. The overview of our model for Semi-supervised Medical Image Segmentation.

Moreover, to address the challenge of insufficient guidance in feature space, we introduce a novel Pseudo-label Guided Contrastive Learning (PGCL) strategy. This strategy encourages pixels belonging to the same class to converge towards their respective class prototypes in feature space while pushing those from different classes apart, thereby further facilitating the exploration of class-discriminative features. In summary, the contributions of this work consist of the following aspects:

- We design a Bilateral Prompt Decoder to harmonize the modal discrepancy between visual and linguistic features and comprehensively extract mutually complementary multi-modal feature representations.
- We introduce a Multi-views Consistency Regularization strategy that incorporates both image and text perturbations to reduce the influence of noise and generate high-quality pseudo labels.
- We propose a Pseudo-label Guided Contrastive Learning strategy to supervise feature space and explore class-discriminative features.
- Extensive experiments on two public datasets demonstrate the significant advancement and superiority of the proposed framework.

2 Method

The overview of our proposed Textmatch is illustrated in Fig. 1. It takes the Mean Teacher structure [21] as the backbone, where both the student and teacher models comprise a visual encoder, a text encoder, several proposed bilateral prompt decoders, and a segmentation head. The visual and text encoders extract features independently. Subsequently, the bilateral prompt decoders perform multi-modal feature fusion, with the segmentation head producing the final segmentation mask. The teacher model is updated from the student model using the Exponential Moving Average (EMA). For labeled data, we employ ground truth for supervised learning. For unlabeled data, we first ensure consistency regularization by generating augmented multiple views. Then, we derive high-quality pseudo-labels from the multi-views predictions to supervise the student model. Finally, we utilize pseudo-labels to guide the contrastive learning of the feature space. The details of our model and the objective functions will be introduced in the following sub-sections.

Problem Setting and Feature Extraction. In our framework, the training dataset consists of a small labeled subset containing N labeled data and a large unlabeled subset containing M unlabeled data, where $M \gg N$. In addition, both labeled and unlabeled data have their corresponding text prompts. We denote the labeled subset as $D^l = \left\{\left(x_i^l, t_i^l, y_i^l\right)\right\}_{i=1}^{N}$ and the unlabeled subset as $D^u = \{(x_i^u, t_i^u)\}_{i=1}^{M}$, where $x_i \in \mathbb{R}^{C\times H\times W}$ denotes the training image, $y_i \in \mathbb{R}^{H\times W}$ denotes the label, and $t_i \in \mathbb{R}^{L}$ denotes the corresponding text prompt containing L words. For an input image $x_i \in \mathbb{R}^{H\times W\times D}$, we extract multiple visual features $f^I = \left\{f_i^I \in \mathbb{R}^{\frac{H}{d_i}\times\frac{W}{d_i}\times C_i}\right\}_{i=1}^{4}$ from the different stages of the visual encoder, where d_i denotes the downsampling rate and C_i denotes feature dimension. For an input text prompt $t_i \in \mathbb{R}^{L}$, we adopt the text encoder to extract the linguistic features $f^T \in \mathbb{R}^{L\times C}$, where L is the number of words in the text prompt and C denotes the feature dimension.

Bilateral Prompt Decoder (BPD). In contrast to previous multi-modal fusion methods that utilize features from one modality to refine the other [14,20,22], our bilateral prompt method concurrently enhances the features of both modalities by mutual prompting, as shown in Fig. 1(b). Given input visual feature $f_I \in \mathbb{R}^{H\times W\times C_I}$ and linguistic feature $f_T \in \mathbb{R}^{L\times C_T}$, the visual feature is reshaped into token sequences $f_I \in \mathbb{R}^{(H\times W)\times C_I}$, and the dimensionality of image tokens and text tokens are aligned as:

$$f_I^d \in \mathbb{R}^{(H\times W)\times C_d} = \sigma\big(Conv(f_I)\big), f_T^d \in \mathbb{R}^{L\times C_d} = \sigma\big(Conv(f_T)\big), \quad (1)$$

where $Conv(\cdot)$ denotes a 1×1 convolution layer, σ denotes the activation function and C_d denotes the aligned feature dimension. Then, the bilateral prompt can be formulated as:

$$f_I' \in \mathbb{R}^{(H\times W)\times C_d} = f_I^d + \alpha\left(MHSA(f_I^d, f_T^d, f_T^d)\right), \quad (2)$$

$$f_T' \in \mathbb{R}^{L\times C_d} = f_T^d + \alpha\left(MHSA(f_T^d, f_I^d, f_I^d)\right), \quad (3)$$

where $MHSA(\cdot)$ denotes the Multi-Head Self-Attention layer and α is a learnable parameter that controls the weight of the residual connection. The refined visual feature is reshaped and upsampled to obtain $f_I'' \in \mathbb{R}^{H'\times W'\times C_d}$. Subsequently, f_I'' is concatenated with $f_C \in \mathbb{R}^{H'\times W'\times C_d}$ (the low-level visual feature from skip connection) on the channel dimension. Finally, the concatenated features are processed through a convolution layer and an activation function to obtain the final fused feature f_O and $f_T^{'}$. The process can be expressed as:

$$f_I'' = Upsample\big(Reshape(f_I')\big), \quad (4)$$

$$f_O = \sigma\big(Conv([f_I''; f_C])\big), \quad (5)$$

where [· ; ·] represents the concatenate operation on the channel dimension.

Multi-views Consistency Regularization (MCR). As shown in Fig. 1(c), the MCR strategy integrates both image and text perturbations, generating varied augmented image-text pairs for unlabeled data, denoted as distinct views. Specifically, for an unlabeled sample $D_i^u = (x_i^u, t_i^u)$, the image is augmented following [12], denoted as $\mathcal{A}(\cdot)$. The corresponding augmented texts are generated through the generative model [23] with different expressions, denoted as $\mathcal{T}(\cdot)$. The multiple views are formulated as:

$$V_i = \left\{(\mathcal{A}(x_i^u), \mathcal{T}(t_i^u))_j\right\}_{j=0}^{n}, \tag{6}$$

where n denotes the number of augmented views and V_{i0} denotes the original input. Subsequently, V_{i0} is forwarded to the student model, while $\left\{V_{ij}\right\}_{j=1}^{n}$ is transmitted to the teacher model to obtain the prediction. Then, the ultimate regularization loss is formulated as:

$$L_{reg} = \frac{1}{B \times n} \sum_{i=1}^{B} \sum_{j=1}^{n} MSE\left(S(V_{i0}), T\left(V_{ij}\right)\right), \tag{7}$$

where B denotes the batch size, $MSE(\cdot)$ denotes the mean squared error, S denotes the student model, and T denotes the teacher model. Leveraging the complementary characteristics of multiple views, we employ the average prediction results from the teacher model as pseudo-labels to mitigate label noise before guiding the student model. Finally, we use $Dice(.)$ to denote the Dice loss [3], and the pseudo-label supervision loss can be formulated as:

$$L_{ps} = \frac{1}{B} \sum_{i=1}^{B} Dice\left(S(V_{i0}), \frac{1}{n} \sum_{j=1}^{n} T\left(V_{ij}\right)\right). \tag{8}$$

Pseudo-label Guided Contrastive Learning (PGCL). To explore class-discriminative features, we propose the contrastive learning strategy, as shown in Fig. 1(d). For labeled data, where $y \in \mathbb{R}^{H \times W}$ denotes the ground truth indicating foreground and background and $f^l \in \mathbb{R}^{H \times W \times C}$ represents the projected and upsampled visual feature from the student model, we compute the prototypes of foreground (P^f) and background (P^b) as follows:

$$P^f = \frac{\Sigma_{i,j} y_{ij} \cdot f_{ij}^l}{\Sigma_{i,j} y_{ij}}, \qquad P^b = \frac{\Sigma_{i,j}(1 - y_{ij}) \cdot f_{ij}^l}{\Sigma_{i,j}(1 - y_{ij})}, \tag{9}$$

The formulas calculate the mean feature vectors of pixels for the foreground and background regions based on the ground truth label, serving as the prototypes that can be updated through EMA when training. For unlabeled data, with the pseudo-label $\hat{y} \in \mathbb{R}^{H \times W}$ derived from predictions of multiple views and the visual feature $f^u \in \mathbb{R}^{H \times W \times C}$ of unlabeled data, our objective is to minimize the distance between the pixel-level feature and its corresponding prototype while simultaneously maximizing the distance between the pixel feature and the non-corresponding prototype guided by the pseudo-label. The contrastive learning loss based on InfoNCE [24] is formulated as:

$$\mathcal{H}\left(f_{ij}^u, P^x, P^y\right) = \frac{\exp\left(\text{sim}\left(f_{ij}^u, P^x\right)/\tau\right)}{\exp\left(\text{sim}\left(f_{ij}^u, P^x\right)/\tau\right) + \exp\left(\text{sim}\left(f_{ij}^u, P^y\right)/\tau\right)}, \tag{10}$$

$$L_c = -\frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \left[\widehat{y_{ij}} \cdot log\left(\mathcal{H}\left(f_{ij}^u, P^f, P^b\right)\right) + \left(1 - \widehat{y_{ij}}\right) \cdot log\left(\mathcal{H}\left(f_{ij}^u, P^b, P^f\right)\right)\right], \tag{11}$$

where $\text{sim}(.,.)$ denotes cosine similarity and τ denotes temperature coefficient.

Overall Learning Objective. For labeled data, we integrate the cross-entropy loss and Dice loss to supervise the model training (i.e., L_{sup}). For the unlabeled data, we calculate the multi-views consistency loss L_{reg} in Eq. 7, pseudo-label supervision loss L_{ps} in Eq. 8 and contrastive learning loss L_c in Eq. 11. Finally, the total loss function can be formulated as below:

$$L = L_{sup} + \lambda_1 L_{reg} + \lambda_2 L_{ps} + \lambda_3 L_c, \tag{12}$$

where λ_1, λ_2 and λ_3 are hyperparameters to trade off the importance of corresponding terms.

3 Experiments and Results

Datasets and Metrics. We evaluate the performance of our proposed framework on two publicly available datasets: 1) QaTa-COV19 [25] consists of 9258 COVID-19 chest X-ray radiographs with manual annotations of COVID-19 lesions. 2) MosMedData+ [26] consists of 2729 CT scan slices of lung infections. Li et al. [19] extended the text prompts for these datasets. The text prompts focus on whether both lungs are infected, the number of lesion regions, and the approximate location of the infected areas. Following previous works [8,19], we use $Dice$ and $mIoU$ coefficient metrics to evaluate the segmentation results objectively. Both of them calculate the intersection regions over the union regions of the given predicted mask and ground truth.

Implementation Details. Our model is implemented using the PyTorch framework and executed using four NVIDIA GeForce RTX 2080Ti GPUs, each with a memory of 10GB. To be specific, ConvNeXt-Tiny [27] is selected as the image encoder and BERT [28] is adopted as the text encoder. The projection head is basically a shallow FC layer [29]. For image augmentations, we use random scaling, morphological, and brightness changes following [12]. For text augmentations, we utilize the generative pre-trained transformer model [23] to generate similar text prompts. For a fair comparison, we follow the previous works [8,19] and use 5%, 15%, and 25% labeled data for training the model. The model is converged using an Adam optimizer with a batch size of 48 and a learning rate of $3e-4$. For hyperparameter settings, we set the number of views $n =$ 3 and temperature coefficient $\tau = 0.9$. For both datasets, λ_1 and λ_3 in Eq. 12 are set to 0.1 to maintain balanced gradient scales. For the QaTa-COV19 dataset, λ_2 in Eq. 12 is set to 0.1, while for the MosMedData+ dataset, λ_2 is set to 0.5 due to its higher segmentation difficulty, requiring a larger weight to learn more from the unlabeled data.

Comparison with other state-of-the-art (SOTA) methods. We experiment with different percentages of labeled data and compare them with previous image-only and multi-modal methods. The quantitative results of QaTa-COV19 and MosMedData+ datasets are presented in Table 1. Compared to the best image-only method, our framework significantly improves segmentation performance by an average of 10.19% on QaTa-COV19 and 5.03% on MosMedData+, which demonstrates using text prompts can significantly improve the model’s ability of object localization in the case of limited

Table 1. Result comparison (%) of image-only and multi-modal methods on two available datasets, with the different labeled percentages. $\mathcal{I}$ denotes image and $\mathcal{T}$denotes text.

Dataset	Method	Sup	5%		15%		25%	
			$Dice$	$mIoU$	$Dice$	$mIoU$	$Dice$	$mIoU$
QaTa-COV19	MC-Net+ (2022) [10]	$\mathcal{I}$	73.15	61.48	75.93	65.81	76.93	67.02
	PLCT (2023) [8]	$\mathcal{I}$	73.01	61.32	75.42	65.12	76.65	66.71
	LAVT (2022) [15]	$\mathcal{I}+\mathcal{T}$	74.45	64.69	77.14	65.86	77.08	67.21
	TGCPA (2023) [18]	$\mathcal{I}+\mathcal{T}$	76.32	65.36	79.18	69.14	80.21	70.59
	LViT-T (2023) [19]	$\mathcal{I}+\mathcal{T}$	77.24	66.31	79.98	70.04	80.95	71.31
	Ours	$\mathcal{I}+\mathcal{T}$	**83.56**	**71.67**	**86.21**	**75.64**	**87.26**	**77.12**
MosMed Data+	MC-Net+ (2022) [10]	$\mathcal{I}$	66.75	54.38	68.92	56.57	70.32	57.94
	PLCT (2023) [8]	$\mathcal{I}$	66.97	54.65	69.14	56.81	70.54	58.13
	LAVT (2022) [15]	$\mathcal{I}+\mathcal{T}$	67.54	55.39	69.75	57.64	71.18	58.84
	TGCPA (2023) [18]	$\mathcal{I}+\mathcal{T}$	67.92	55.87	70.42	58.89	71.94	59.78
	LViT-T (2023) [19]	$\mathcal{I}+\mathcal{T}$	68.61	56.49	71.45	59.64	72.48	60.31
	Ours	$\mathcal{I}+\mathcal{T}$	**72.46**	**58.12**	**75.73**	**60.93**	**76.46**	**62.69**

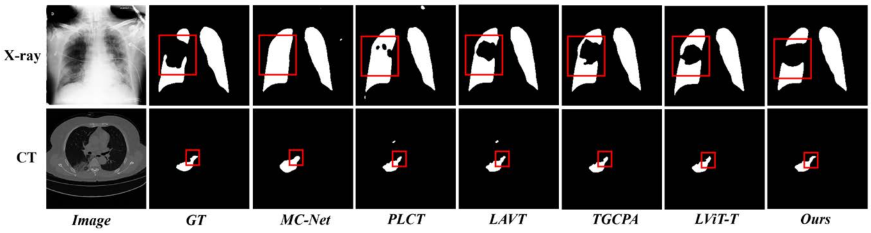

Fig. 2. Visual comparison of segmentation results on two datasets.

annotated data. Besides, our framework still achieves better performance by an average of 5.94% and 2.92% on two datasets respectively than existing multi-modal methods. Particularly, our framework in the case of 5% labeled data can achieve or even exceed the performance of multi-modal methods in the case of 25% labeled data, which shows that our model can learn more class-discriminative feature from unlabeled data and reduce the effect of noise in semi-supervised learning.

Besides, the visual comparison results are shown in Fig. 2. It shows that our framework can create a segmentation mask with more accurate regions and distinctive borders than other methods whether X-rays or CT images with the text prompts.

Ablation Studies. We conduct ablation studies to validate the effectiveness of our proposed modules and report the quantitative results on the QaTA-COV19 test set in Table 2. As is shown in Table 2, each part we proposed contributes to the final performance improvement. Specifically, the bilateral prompt decoder (BPD) significantly improved segmentation performance by an average of 3.88% in $Dice$ and 3.63% in $MIou$ due to

Table 2. Ablation Studies on QaTa-COV19 test set, with different labeled rates.

L_{sup}	BPD	L_{reg}	L_{ps}	PGCL	5%		15%		25%	
					$Dice$	$mIoU$	$Dice$	$mIoU$	$Dice$	$mIoU$
✓					75.14	64.04	78.35	67.74	79.64	69.47
✓	✓				79.32	67.42	82.13	71.52	83.32	73.21
✓	✓	✓			81.79	69.91	84.53	73.98	85.27	75.23
✓	✓	✓	✓		82.76	70.84	85.64	75.09	86.58	76.41
✓	✓	✓	✓	✓	83.56	71.67	86.21	75.64	87.26	77.12

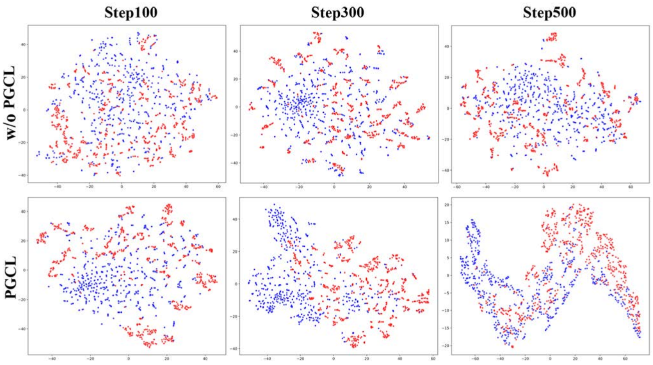

Fig. 3. T-SNE decomposition of feature space produced by encoder and projection head at different training stages on QaTa-COV19 dataset (15% labeled) w/wo the proposed PGCL.

the multi-modal complementarity. With the proposed multi-views consistency and pseudo-labels supervision, the metrics increase by an average of 2.29% and 1.14% respectively, which indicate that the multi-views incorporating both image and text perturbations can reduce the influence of noise and encourage the model to learn generalized representations.

Besides, we also analyze the T-SNE decomposition of representation space with and without PGCL, as shown in Figure 3. Despite some boundary confusion, PGCL notably enhances the clustering of feature embeddings during training, fostering improved inter-class separability and intra-class compactness. Conversely, without PGCL, embeddings from various classes become entangled in the feature space. This effectively demonstrates the capability of the proposed strategy to address the lack of explicit guidance within the feature space.

4 Conclusions

In this paper, we propose Textmatch, a novel semi-supervised medical image segmentation framework that explores the potential of text prompts for better segmentation results. Specifically, we design a Bilateral Prompt Decoder (BPD) to mine information from visual and linguistic features. Furthermore, we introduce a Multi-views Consistency Regularization (MCR) strategy that incorporates both image and text perturbations to reduce the influence of noise. Finally, we propose a Pseudo-label Guided Contrastive Learning (PGCL) strategy to explore class-discriminative features. Extensive experiments on two available datasets demonstrate the significant advantage of our framework compared to previous image-only and multi-modal methods.

Acknowledgments. This work is supported by National Natural Science Foundation of China (NSFC 62371325, 62071314), Sichuan Science and Technology Program 2023YFG0025, 2023YFG0101, and 2023 Science and Technology Project of Sichuan Health Commission 23LCYJ002.

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

Reference

1. Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence 40(4), 834-848 (2017)
2. Ronneberger O, Fischer P, Brox T, et al.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) MICCAI 2015, Part III 18, pp. 234-241. Springer, Cham (2015)
3. Milletari, F., Navab, N., Ahmadi, S. A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV), pp. 565-571. (2016)
4. Yang, X., Song, Z., King, I., Xu, Z.: A survey on deep semi-supervised learning. IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 9, 8934-8954 (2022)
5. Bai W, Oktay O, Sinclair M, et al.: Semi-supervised learning for network-based cardiac MR image segmentation. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D., Duchesne, S. (eds) MICCAI 2017, Part II 20, pp. 253-260. Springer, Cham (2017)
6. Lyu, F., Ye, M., Carlsen, J. F., Erleben, K., Darkner, S., Yuen, P. C.: Pseudo-label guided image synthesis for semi-supervised covid-19 pneumonia infection segmentation. IEEE Transactions on Medical Imaging 42(3), 797-809 (2022)
7. Seibold, C. M., Reiß, S., Kleesiek, J., Stiefelhagen, R.: Reference-guided pseudo-label generation for medical semantic segmentation. In: Proceedings of the AAAI conference on artificial intelligence, Vol. 36, No. 2, pp. 2171-2179. (2022)
8. Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E.: Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. Medical Image Analysis 87, 102792 (2023)
9. Wang, K., Zhan, B., Zu, C., Wu, X., Zhou, J., Zhou, L., Wang, Y.: Semi- supervised Medical Image Segmentation via a Tripled-uncertainty Guided Mean Teacher Model with Contrastive Learning. Medical Image Analysis 79, 102447. (2022)
10. Tang, C., Zeng, X., Zhou, L., Zhou, Q., Wang, P., Wu, X., Ren, H., Zhou, J., Wang, Y.: Semi-supervised medical image segmentation via hard positives oriented contrastive learning. Pattern Recognition 146, 110020. (2024)
11. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2613-2622. (2021)

12. Sohn K, Berthelot D, Carlini N, et al.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems 33, 596-608 (2020)
13. Radford A, Kim J W, Hallacy C, et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning, pp. 8748-8763. PMLR (2021)
14. Rao Y, Zhao W, Chen G, et al.: Denseclip: Language-guided dense prediction with context-aware prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18082-18091. (2022)
15. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P. H.: Lavt: Language-aware vision transformer for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18155-18165. (2022)
16. Tomar N K, Jha D, Bagci U, et al.: TGANet: Text-guided attention for improved polyp segmentation. In: Wang, L. et al. (eds.) MICCAI 2022, pp. 151-160. Springer, Cham (2022)
17. Müller, P., Kaissis, G., Zou, C., Rueckert, D.: Radiological reports improve pre-training for localized imaging tasks on chest x-rays. In: Wang, L. et al. (eds.) MICCAI 2022, pp. 647-657. Springer, Cham (2022)
18. 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: Greenspan, H., et al. (eds.) MICCAI 2023, pp. 537-546. Springer, Cham (2023)
19. Li Z, Li Y, Li Q, et al.: Lvit: language meets vision transformer in medical image segmentation. IEEE Transactions on Medical Imaging, vol. 43, no. 1, 96-107 (2023)
20. 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: Greenspan, H., et al. (eds.) MICCAI 2023, pp. 724-733. Springer, Cham (2023)
21. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems, 30 (2017)
22. Zhou, K., Yang, J., Loy, C. C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16816-16825. (2022)
23. Brown T, Mann B, Ryder N, et al.: Language models are few-shot learners. Advances in neural information processing systems 33, 1877-1901 (2020)
24. 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)
25. Degerli, A., Kiranyaz, S., Chowdhury, M. E., Gabbouj, M.: Osegnet: Operational segmentation network for COVID-19 detection using chest X-ray images. In: IEEE International Conference on Image Processing (ICIP), pp. 2306-2310. (2022)
26. Morozov S P, Andreychenko A E, Pavlov N A, et al.: Mosmeddata: Chest ct scans with covid-19 related findings dataset. arXiv preprint arXiv:2005.06465. (2020)
27. 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)
28. Devlin, J., Chang, M. W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. (2018)
29. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning, pp. 1597-1607. PMLR (2020)