PromptSSL: Prompt-guided Semi-Supervised Spine Segmentation with Global-Local Semantic Constrained Vision-Language Model

Anonymous submission

Abstract

Vision-language model (VLM) has great potential to enhance the confidence of pseudo labels in semi-supervised learning (SSL) for spine segmentation, but no one has studied it yet. However, existing VLMs suffer from the cross-modal uncertainty problem (i.e., multiple correspondences between modalities), which limits the performance of image-text alignment. In this paper, we propose a global-local semantic constrained VLM (GLsc) driven SSL, PromptSSL, to integrate semantic priors from GLsc into semi-supervised spine segmentation, improving the quality of pseudo labels. Specifically, the proposed GLsc addresses the cross-modal uncertainty with two novel designs. 1) The local semantic constraint learns the semantic relationships between each image patch and text for uncertainty metric of local views. 2) The global semantic constraint measures the Wasserstein-2 distance between cross-modal distribution representations to restrain the corresponding global semantic similarity. Moreover, a prompt-guided semi-supervised spine segmentation framework is designed to enhance the quality of pseudo labels in the training process through the pre-trained GLsc. Experimental results show that our PromptSSL achieves superior spine segmentation performance with Dice of 79.35%, only using 5% labeled data on a public spine segmentation dataset, surpassing existing SSL and VLM methods. Our code will be available.

Introduction

Automatic spine segmentation plays a significant role in various orthopedic applications, including spinal disease diagnosis (Han et al. 2018), surgical treatment planning (Chen et al. 2015), and spine pathology identification (Han et al. 2018). Fully supervised segmentation methods have made significant progress relying on a significant amount of labeled data (Chen et al. 2018). However, labeling pixel-level annotations is laborious and demands expertise, particularly in spine images with multiple categories. Semi-supervised learning (SSL) methods has the capability to mitigate the label scarcity problem by using a limited amount labeled data and any quantity of unlabeled data (Van Engelen 2020).

Existing SSL methods are usually based on consistency regularization (Tarvainen and Valpola 2017) and pseudo labeling (Lee et al. 2013). Conducting consistency regularization and pseudo labeling between the sub-networks (Fig. 1 (left)), has achieved favorable performance in semi-supervised segmentation (Liu et al. 2022; Ouali, Hudelot, and Tami 2020). However, it may lead to inconsistent predictions for the same sample if there is insufficient complementarity between models, thereby affecting the quality of pseudo-labels. The question that comes to mind is: how to effectively improve the quality of pseudo labels for SSL.

Current Vision-Language Models (VLM) are hard to apply to semi-supervised spine segmentation due to facing the challenge of cross-modal uncertainty (Chun et al. 2021; Yang et al. 2021), even though VLM has great potential to enhance pseudo labels. The challenge manifests as: multiple different images may correspond to the same text, while multiple ambiguous yet similar texts (i.e., synonym) tend to correspond to the same image (Fig. 2). Distribution modeling methods are usually used for cross-modal uncertainty awareness (Chun et al. 2021; Yang et al. 2021; Yu et al. 2019). Nevertheless, relying solely on distribution representations cannot be fully aware of complex semantic association.

In this work, we propose a prompt-guided SSL (PromptSSL) with global-local semantic constrained VLM (GLsc) for spine segmentation (Fig. 1 (right)). The GLsc estimates the uncertainty relationships between local image semantics and text via local semantic constraint, and simultaneously restrains the uncertainty of semantic similarity via global semantic constraint. Furthermore, a novel semi-supervised spine segmentation framework is proposed. It learns better semantic decision boundaries via the benefit of text prompts, enhancing the quality of pseudo labels. Our contributions include:

(1) We are the first to migrate the prompt-guided mask to SSL for spine segmentation. It enhances the quality of pseudo labels in SSL training by leveraging the semantic priors of the pre-trained VLM.

(2) Our GLsc pre-trains a powerful uncertainty aware VLM. It estimates the uncertainty relationship between local image semantics and text, and constrains the global semantic similarity via the distribution-based uncertainty level, thus restraining the adverse effects of uncertain local regions and improving the robustness of image-text alignment during VLM pre-training.

(3) The experimental results on spine MR images demonstrate that our method exhibits superior spine segmentation

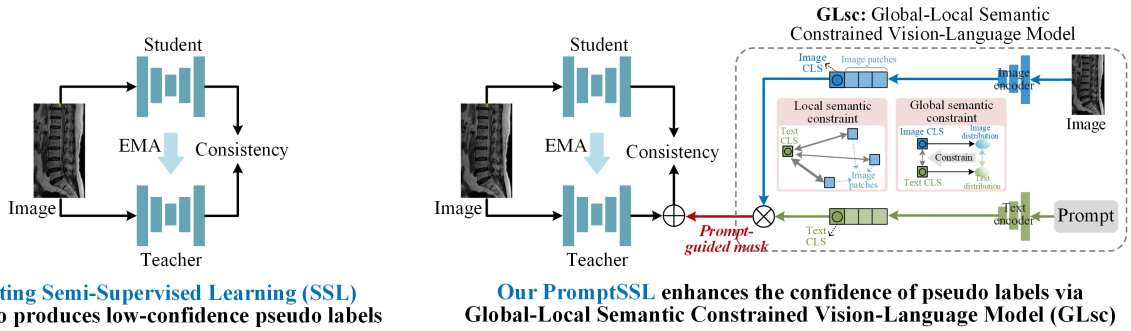

Figure 1: Our PromptSSL innovatively proposes GLsc for prompt-guided mask generation, enhancing the quality of pseudo labels in SSL.

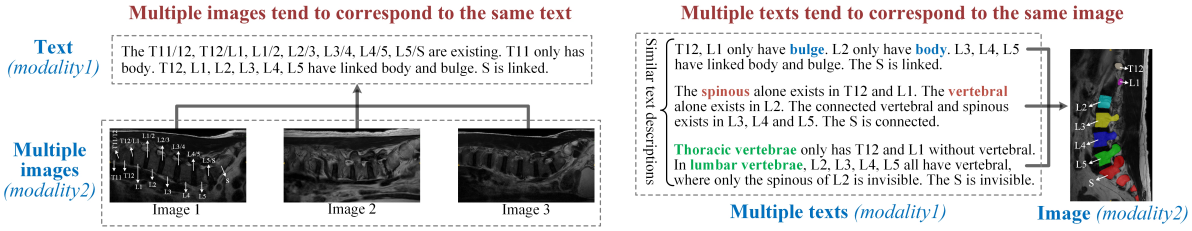

Figure 2: Our GLsc addresses the challenge of cross-modal (image-text) semantic uncertainty in existing VLM, where the uncertainty manifests: the same object often has different descriptions from different modalities.

performance and uncertainty handling capability compared to other SSL and VLM methods.

Related Work

Spine Segmentation

Traditional methods (Lessmann et al. 2019; Tao, Liu, and Zheng 2022), have been widely researched for spine segmentation. Following the segmentation objects of the spine, these algorithms can be divided into sole segmentation of VBs or IVDs and simultaneous segmentation of multiple spinal structures. For the sole segmentation of VBs, InteractiveFCN (Lessmann et al. 2019) labels and segments VB by referring to the prior knowledge of VB continuity. For the sole segmentation of IVDs, Li et al. (2018) proposed an integrated multi-scale fully convolutional network with random modality voxel dropout learning for IVD localization and segmentation. Although good segmentation performance is achieved, these methods may lose some useful information on disease diagnosis of the spine. In this paper, we introduce text-prompt to spine segmentation, guiding the model to learn segmentation regions.

Semi-Supervise Medical Image Segmentation

In the field of Semi-Supervise Medical Image Segmentation, the main methods includes self-training methods (Zhu et al. 2021), co-training methods (Wang et al. 2021; Xia et al. 2020), and consistency regularization methods (Wang et al. 2020; Zhang et al. 2023). Consistency regularization focuses on maintaining consistent model predictions under different perturbations. The state-of-the-art technique is Mean Teacher (MT) (Tarvainen and Valpola 2017). In MT, the teacher model is employed to generate pseudo-labels for unlabeled data while maintaining prediction consistency between the teacher and student models through various regularization methods. Afterward, the teacher model is the exponential moving average (EMA) of the student model's weights. This method enables the teacher model to continually aggregate historical prediction information from unlabeled data. Subsequent improvements use different consistency regularization strategies to improve the prediction quality of unlabeled data (Li et al. 2020; Chen et al. 2021). However, these methods are still based on the single-modal approaches, leading to poor pseudo labels. In this paper, we introducing VLM into SSMIS to enhancing the quality of pseudo labels by generating text prompt guided multimodal supervision information.

Vision-Language Model

Although existing VLMs learns generic visual-textual representations by aligning image-text, they are limited by uncertainty awareness of cross-modal and intra-modal. CLIP (Wang et al. 2022) is a representative work of VLM, using the contrastive loss to calculate similarity scores between images and texts. Accordingly, numerous works use text information to improve the image segmentation capabilities (Yang et al. 2022; Ding et al. 2022). For instance, Yang et al. (Yang et al. 2022) conducted early fusion of image-text features in intermediate layers of a transformer network, achieving significantly cross-modal alignment. Ding et al. (Ding et al. 2022) built an encoder decoder attention

network with transformer and multi-head attention to provide the language expression "queries" for the given image. Inspired by VLM in natural images, a few works have started utilizing text information for medical image analysis (Bhalodia et al. 2021; Müller et al. 2021). Li et al. (Li et al. 2023) proposed a Language meets Vision Transformer model (LViT) to incorporated text annotations with images in down-sampling and up-sampling processes, compensating for the quality deficiencies in image data. Boecking et al. (Boecking et al. 2022) used the attention weights learned during local alignment to conduct medical semantic segmentation. In this work, we propose GLsc to address cross-model uncertainty problem for better image-text alignment.

Methodology

Our GLsc boosts vision-language alignment through local semantic uncertainty metric and global semantic similarity restraint, driving the PromptSSL to enhance the quality of pseudo labels in semi-supervised spine segmentation (Fig. 3). It contains two stages. **1) VLM pre-training:** The GLsc estimates the uncertainty relationship between each image patch and text via local semantic constraint, and constrains the semantic similarity between image and text via global semantic constraint, addressing the uncertainty of cross-modal alignment during the VLM pre-training. **2) Downstream task:** The prompt-guided SSL integrates the semantic priors of the pre-trained GLsc into semi-supervised spine segmentation, improving the quality of pseudo labels.

GLsc for Robust Cross-Modal Alignment

The GLsc adaptively models the uncertainty in each patch-text correspondence for local semantic understanding and measures distribution-based uncertainty representations for semantic similarity restraint, improving the robustness for cross-modal alignment.

Local Semantic Constraint: The goal of local semantic constraint is to guide the model adaptively focus on various semantic information through uncertainty metric between each image patch and text. The mismatched patch-text pairs' similarities are regarded as a distribution. The variance σ_k^2 of the distribution represents the uncertainty weight for patch-text contrastive learning.

$$\sigma_k^2 = var(\frac{x_{I_{i,k}}^\top x_{T_j}}{||x_{I_{i,k}}||||x_{T_j}||}|j=1,...,T_n, j\neq i), \quad (1)$$

where k is the k-th patch of an image. σ_k^2 is applied in contrastive learning loss between k-th patch and text. The loss of local semantic constraint is:

$$\mathcal{L}_{LS} = -\frac{1}{N}\frac{1}{K}\sum_{i=1}^{N}\sum_{k=1}^{K}\left[log\frac{exp(\frac{1}{\sigma_k^2}sim((x_{I_1}^k, x_{T_+})/\tau)}{\sum_{j=1}^{N}exp(\frac{1}{\sigma_k^2}sim(x_{I_1}^k, x_{T_j})/\tau)}\right], \quad (2)$$

where σ_k^2 is regarded as local uncertainty. A smaller σ_k^2 corresponds to lower uncertainty, bringing higher attention for the corresponding patch-text loss.

Global Semantic Constraint: The goal of global semantic constraint is to restrain the global semantic similarity based on the corresponding uncertainty level. It models the semantic embeddings (f_I, f_T) of both images and texts as distribution representations (d_I, d_T) via the Gaussian distribution (Fig. 3). The distance of distribution representations is formulated as the uncertainty level.

Given N image-text pairs (x_{I_n}, x_{T_n}), $n \in \{1, 2, ..., N\}$, the semantic distance is defined as $D_s(x_{I_n}, x_{T_n}) = \|f_{I_n} - f_{T_n}\|_2$, where (f_{I_n}, f_{T_n}) is the semantic embeddings of n-th image-text pair. The Gaussian distribution is used to model the semantic embeddings as distribution representations $\mathcal{N}(\mu_{I_n}, \sigma_{I_n})$ and $\mathcal{N}(\mu_{T_n}, \sigma_{T_n})$. The 2-Wasserstein (Kantorovich 1960; Kim, Son, and Kim 2021) is utilized to measure the difference between image-text distribution representations. The 2-Wasserstein distance is defined as:

$$D_{2W}(x_{I_n}, x_{T_n}) = ||\mu_{I_n} - \mu_{T_n}||_2^2 + ||\sigma_{I_n} - \sigma_{T_n}||_2^2, \quad (3)$$

where μ represents the mean vector and σ is the standard deviation vector. The uncertainty level is given by:

$$D_u(x_{I_n}, x_{T_n}) = a \cdot D_{2W}(x_{I_n[\texttt{CLS}]}, x_{T_n[\texttt{CLS}]}) + b, \quad (4)$$

where a is a positive scale factor and b is a deviation value. The ratio of the uncertainty level $D_u(x_{I_n}, x_{T_n})$ to the semantic discrepancy $D_s(x_{I_n}, x_{T_n})$ is defined as the relative uncertainty $\widehat{D}_u$,

$$\widehat{D}_u(x_{I_n}, x_{T_n}) = e^{-\lambda\frac{D_u(x_{I_n}, x_{T_n})}{D_s(x_{I_n}, x_{T_n})}}, \quad (5)$$

where λ is a positive parameter for controlling the constraining degree.

To avoid semantic uncertainty of image-text alignment, the cosine similarity $sim(x_{I_n}, x_{T_n})$ between image and text is constrained via the relative uncertainty $\widehat{D}_u$, thus obtaining the uncertainty cosine similarity $\widehat{sim}(x_{I_n}, x_{T_n})$. The InfoNCE loss (Oord, Li, and Vinyals 2018) is used to optimize $\widehat{sim}(x_{I_n}, x_{T_n})$. $\widehat{sim}(x_{I_n}, x_{T_n})$ and the global semantic constraint loss are denoted as:

$$\widehat{sim}(x_{I_n}, x_{T_n}) = 1 - (1 - sim(x_{I_n}, x_{T_n})) \cdot \widehat{D}_u(x_{I_n}, x_{T_n}), \quad (6)$$

$$\begin{aligned}\mathcal{L}_{GS} = &- \mathbb{E}_{(I,T)}\left[log\frac{exp(\widehat{sim}((x_{I_1}, x_{T_+})/\tau)}{\sum_{n=1}^{N}exp(\widehat{sim}(x_{I_1}, x_{T_n})/\tau)}\right] \\ &- \mathbb{E}_{(T,I)}\left[log\frac{exp(\widehat{sim}(x_{T_1}, x_{I_+})/\tau)}{\sum_{n=1}^{N}exp(\widehat{sim}(x_{T_1}, x_{I_n})/\tau)}\right],\end{aligned} \quad (7)$$

where τ is a learned temperature hyper-parameter, T_+ denotes the positive text matched with I_1, and I_+ is the positive image that matches to T_1.

Summarized advantages: Our GLsc tackles the cross-modal uncertainty problem with two novel semantic constraint strategies. 1) The local semantic constraint strategy learns the local uncertainty in each patch-text correspondence for powerful local semantic understanding. 2)

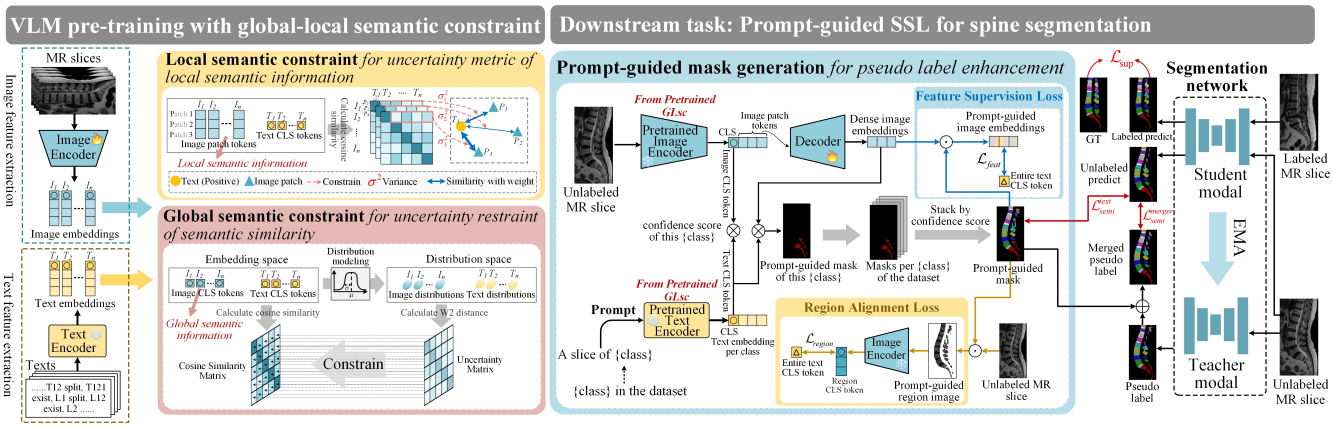

Figure 3: The framework of our PromptSSL. **VLM pre-training:** Our GLsc conducts local semantic uncertainty metric and global semantic similarity restraint. **Downstream task:** Our prompt-guided SSL generates text prompt-guided mask based on the pre-trained GLsc for enhancing the quality of pseudo labels.

The global semantic constraint strategy models distribution-based representations to constrain global semantic similarity. Therefore, our GLsc has strong robustness for image-text alignment.

PromptSSL for Pseudo Label Quality Improvement

Our PromptSSL integrates the semantic priors from the pre-trained GLsc into semi-supervised spine segmentation. It enhances the quality of pseudo labels via the prompt-guided mask. It includes two paths: 1) Prompt-guided mask generation; 2) Teacher-Student Network guided pseudo-label generation.

Prompt-guided Mask Generation: The spine image embeddings are merged with textual prompts for each category, generating masks corresponding to each category's textual prompt. The masks per category are stacked based on the corresponding confidence scores, forming the final prompt-guided mask. Specifically, the image patch tokens v_f^{patch} encoded by the image encoder are decoded into dense image embeddings v_f through a decoder $f_d(\cdot)$. The masks per category are generated by dot product between v_f and $t_{f,c}$ (text embedding for each category).

$$y_c^{text} = \sigma(t_{f,c}^{\top} f_d(v_f^{patch})) \quad y^{text} = stack(y_c^{text} \cdot s_c), \quad (8)$$

where $\sigma(\cdot)$ is the softmax function. The $stack(\cdot)$ operation stacks each y_c^{text} according to the similarity score s_c of each y_c^{text}. To obtain reliable prompt-guided mask, the region alignment loss and feature supervision loss are introduced.

Region Alignment Loss: To directly learn region-text alignment for better semantic correspondence, the region image embeddings V_I are extracted from the prompt-guided region image to align with the entire text embedding T_f. The region alignment loss is defined as:

$$\mathcal{L}_{region} = -\mathbb{E}_{(I,T)}\left[log\frac{exp(sim((V_{I,1}, T_{f,+})/\tau)}{\sum_{n=1}^{N} exp(sim(V_{I,1}, T_{f,n})/\tau)}\right] - \mathbb{E}_{(T,I)}\left[log\frac{exp(sim(T_{f,1}, V_{I,+})/\tau)}{\sum_{n=1}^{N} exp(sim(T_{f,1}, V_{f,n})/\tau)}\right], \quad (9)$$

Feature Supervision Loss: To suppress the model from generating masks for regions undescribed in the text, a feature supervision loss is introduced to restrain negative masks obtained from unrelated texts, which is denoted as:

$$\mathcal{L}_{feat} = -\mathbb{E}_{(I,T)}\left[log\frac{exp(sim((V_{F,1}, T_{f,+})/\tau)}{\sum_{n=1}^{N} exp(sim(V_{F,1}, T_{f,n})/\tau)}\right] - \mathbb{E}_{(T,I)}\left[log\frac{exp(sim(T_{f,1}, V_{F,+})/\tau)}{\sum_{n=1}^{N} exp(sim(T_{f,1}, V_{F,n})/\tau)}\right], \quad (10)$$

where V_F is prompt-guided image embeddings.

Teacher-Student Guided Pseudo Label Generation: The teacher network f_{θ_t} is updated by the student network f_{θ_s} via the Exponential Moving Average (EMA). f_{θ_s} predicts the labeled image x_l and the unlabeled image x_u. f_{θ_t} generates the pseudo-label y_u^t of the unlabeled image.

$$y_l = f_{\theta_s}(x_l), \quad y_u^s = f_{\theta_s}(x_u), \quad y_u^t = f_{\theta_t}(x_u), \quad (11)$$

where y_l is the labeled prediction, and y_u^s is the unlabeled prediction.

Summarized advantages: This is the first attempt to integrate prompt-guided mask into SSL, which improves the quality of pseudo labels. The designed region alignment loss and feature supervision loss drive the decoder produce reliable prompt-guided mask. Therefore, our PromptSSL achieves powerful spine segmentation performance through the advantages of text descriptions for locating the target segmentation area.

Table 1: The comparative experiments on MRSpineSeg dataset demonstrate that our powerful segmentation performance. The best results are highlighted in bold.

Method	Text	Metrics							
		5% labeled		10% labeled		25% labeled		50% labeled	
		mDice (%)	mIoU (%)	mDice (%)	mIoU (%)	mDice (%)	mIoU (%)	mDice (%)	mIoU (%)
MT	×	74.25	64.14	75.54	64.22	77.11	67.65	79.31	69.05
BCP	×	49.86	37.09	66.24	53.81	70.35	58.97	70.68	59.68
MC-Net	×	70.57	59.89	71.30	59.54	73.34	62.27	75.09	64.18
SS-Net	×	71.89	61.26	70.01	58.65	73.71	63.22	74.60	64.30
UCMT	×	74.51	63.22	74.88	63.17	76.42	65.78	77.09	66.81
CLIP	✓	76.17	66.12	76.15	66.08	77.19	66.83	77.25	67.14
MedCLIP	✓	76.47	66.84	76.72	67.05	77.01	66.59	78.30	67.99
MAP	✓	77.84	68.65	78.30	67.99	**80.45**	**70.18**	81.04	71.93
Ours	✓	**79.35**	**69.11**	**79.56**	**70.76**	80.05	69.49	**81.46**	**72.28**

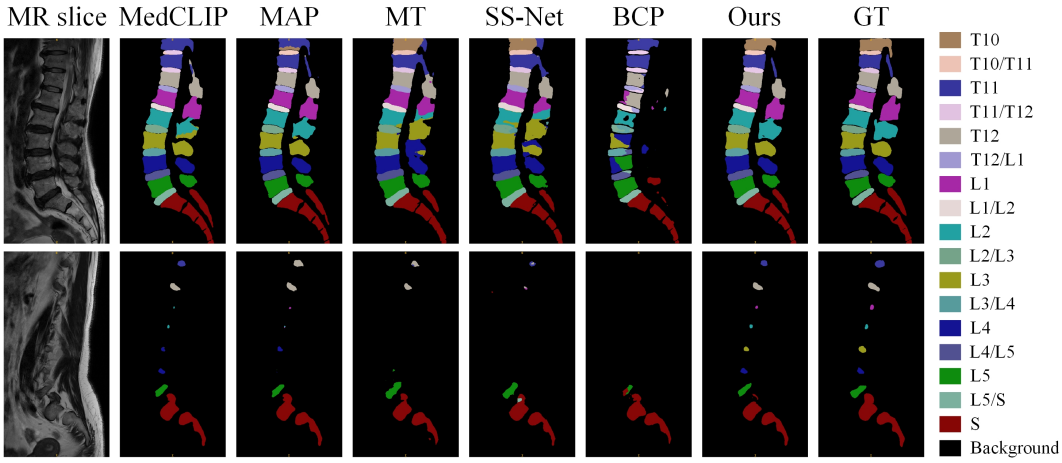

Figure 4: Our method shows high-quality segmentation. The spine segmentation superiority of the proposed method with 5% labeled data.

Training Paradigm of PromptSSL

For labeled images, $\mathcal{L}_{sup}$ is computed between the ground truth y_{gt} and the labeled prediction y_l. For unlabeled images x_u, $\mathcal{L}_{semi}$ is calculated between the unlabeled prediction y_u^s and incorporated pseudo label, where the incorporated pseudo label is the combination of both the teacher network-generated pseudo-label and the text prompt-guided mask, as well as between the unlabeled prediction y_u^s and the text prompt-guided mask y_u^{text}. $\mathcal{L}_{sup}$ and $\mathcal{L}_{semi}$ are defined as:

$$\mathcal{L}_{sup} = -\frac{1}{N_l}\frac{1}{HW}\sum_{i=1}^{N_l}\sum_{j=1}^{HW}\ell_{ce}(y_{l_{i,j}}, y_{gt_{i,j}}), \tag{12}$$

$$\mathcal{L}_{semi}^{merged} = -\frac{1}{N_u}\frac{1}{HW}\sum_{i=1}^{N_u}\sum_{j=1}^{HW}\ell_{ce}(y_{u_{i,j}}^s, \sigma(y_{u_{i,j}}^t + y_{u_{i,j}}^{text}), \tag{13}$$

$$\mathcal{L}_{semi}^{text} = -\frac{1}{N_u}\frac{1}{HW}\sum_{i=1}^{N_u}\sum_{j=1}^{HW}\ell_{ce}(y_{u_{i,j}}^s, y_{u_{i,j}}^{text}), \tag{14}$$

$$\mathcal{L}_{semi} = (\mathcal{L}_{semi}^{merged} + \mathcal{L}_{semi}^{text})/2 \tag{15}$$

where $\ell(\cdot)$ is the cross-entropy loss. (i,j) represents the j-th pixel in i-th mask. N_l and N_u are the batch size of labeled images and unlabeled images. W and H represent the width and height of an image. The overall loss is $\mathcal{L}_{overall} = \mathcal{L}_{sup} + \mathcal{L}_{semi} + \lambda_i\mathcal{L}_{region} + \lambda_f\mathcal{L}_{feat}$

Theoretical Analysis

Our theoretical proofs have proven that the Global Semantic Constraint can address the alignment uncertainty between image and text.

Proposition For the proposed uncertainty cosine similarity $sim(I,T)$ between image and text, if the uncertainty $D_u(\cdot)$ increase, $\widehat{sim}(I,T)$ will increase.

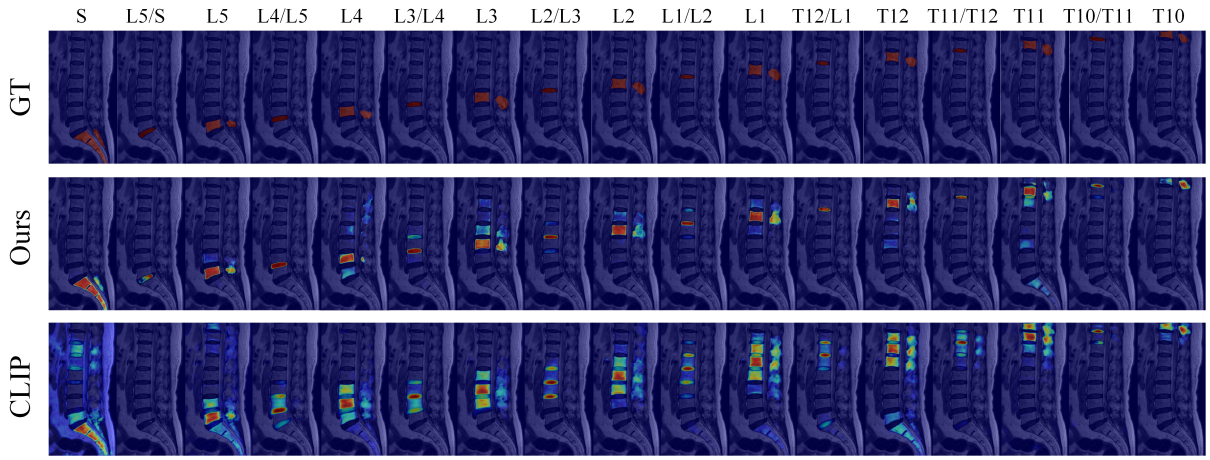

Figure 5: Our GLsc driven SSL shows great target region activation effects of each class in spine. The class activate map (CAM) upon last layer of the teacher network generated by AblationCAM (Ramaswamy et al. 2020).

Proof.

$$\begin{aligned}\widehat{sim}(I,T) &= 1-(1-sim(I,T))\cdot \widehat{D}_u(x_{I_n},x_{T_n}) \\ &= 1-(1-sim(I,T))\cdot e^{-\lambda\frac{D_u(x_{I_n},x_{T_n})}{D_s(x_{I_n},x_{T_n})}} \\ &\propto D_u(x_{I_n},x_{T_n}).\end{aligned} \tag{16}$$

For the image-text pair with high uncertainty $D_u(x_{I_n}, x_{T_n})$, i.e. high relative uncertainty $\widehat{D}_u(x_{I_n}, x_{T_n})$, the proposed uncertainty cosine similarity make such image-text pair more similar, addressing the alignment uncertainty.

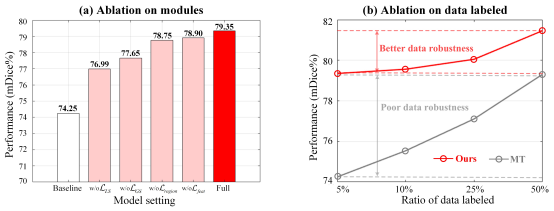

Figure 6: (a) Ablation on modules demonstrates the effectiveness of each component of our method. (b) Ablation on data labeled shows that our method has better data robustness with a subtle decrease when the labeled training data reduce to 5%.

Experiment

Dataset

The proposed method was evaluated on MRSpineSeg dataset (Pang et al. 2020). It contains 215 T2-weighted MR volumetric images. There are 20 categories including 19 spinal structures and the background in the dataset. Not all images contain 19 spinal structures. The in-plane resolutions range from 512 × 512 to 1024 × 1024 and the number of slices ranges from 12 to 18.

Implementation Details

Our method is implemented by Pytorch. The operating system is Ubuntu 20.04.4 LTS with 24GB V100 GPU. For VLM pre-training, Vit (Dosovitskiy et al. 2020) is image encoder, BERT Base is text encoder. Adam optimizer with the learning rate of 1e-5 and ReduceLR scheduler are applied. For PromptSSL, the teacher-student network architecture is 2D ResUNet. The PromptSSL is trained for 300 epochs using Adam optimizer with a weight decay of 1e-4 and a batch size of 8. The learning rate is set to 1e-4 initially and is lowered by 5 times at epoch 100 and 200.

Comparison Study

Fig. 4 shows our powerful spine segmentation performance, compared with existing VLM and SSL methods, including MT (Tarvainen and Valpola 2017), BCP (Bai et al. 2023), MC-Net (Wu et al. 2022a), SS-Net (Wu et al. 2022b), UCMT (Shen et al. 2023), CLIP (Radford et al. 2021), MedCLIP (Wang et al. 2022), MAP (Ji et al. 2023). Our method accurately localizes target regions and generates coherent boundaries, even in small object circumstances. MT, SS-Net, and BCP all have more severe mis-segmentation, which indicates that the introduction of the prompt-guided mask can help to produce more accurate segmentation. In addition, our method generates more delicate segmentation boundary compared with CLIP, MedCLIP, and MAP. It is attributed to the semantic uncertainty understanding during cross-modal alignment. For quantitative experimental results (Table 1), our method outperforms the second best method by 1.51% and 0.46% in mDice and mIoU in 5% labeled data, respectively, indicating that the proposed method can exploit the advantages of text description to enhance the quality of pseudo label in SSL.

Uncertainty Analysis: Fig. 5 demonstrates the reliability of our GLsc driven SSL. For each class in spinal image, our method generates robust target region activation effects. However, the CLIP driven SSL shows unstable activation effects. Therefore, our method can aid the model to address

semantic uncertainty, thus enhancing its performance for pseudo label enhancement.

Ablation Study: Ablation studies on modules and labeled data demonstrate the effectiveness of each component and powerful data robustness. Fig. 5 (a) shows that using baseline (i.e., teacher-student network) achieves poor performance in spine segmentation. The local semantic constraint loss $\mathcal{L}_{LS}$ and global semantic constraint loss $\mathcal{L}_{GS}$ all bring benefits for pseudo label enhancement. $\mathcal{L}_{LS}$ improves the mDice by 2.36%, and adding $\mathcal{L}_{GS}$ could improve 1.70%. This demonstrates that our GLsc is reliable for alleviating semantic uncertainty. The region alignment loss $\mathcal{L}_{region}$ and feature supervision loss $\mathcal{L}_{feat}$ improves the mDice by 0.60% and 0.45%, respectively. This is attributed to the superiority of their region-text alignment and suppression of irrelevant text, optimizing the decoder to generate high-confidence pseudo labels. Fig. 5 (b) demonstrates that our method has better data robustness with a subtle decrease when the labeled training data reduce to 5%.

Conclusion

In this work, we proposed a GLsc driven SSL for spine segmentation. Specifically, the GLsc conducts local semantic uncertainty metric and global semantic similarity restraint to address cross-modal semantic uncertainty. Furthermore, the PromptSSL is proposed to improve the quality of pseudo labels via the pre-trained GLsc. The experiments verify the effectiveness of our framework.

References

Bai, Y.; Chen, D.; Li, Q.; Shen, W.; and Wang, Y. 2023. Bidirectional copy-paste for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11514–11524.

Bhalodia, R.; Hatamizadeh, A.; Tam, L.; Xu, Z.; Wang, X.; Turkbey, E.; and Xu, D. 2021. Localization via Cross-Attention on Medical Images and Reports. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 571–581.

Boecking, B.; Usuyama, N.; Bannur, S.; Castro, D. C.; Schwaighofer, A.; Hyland, S.; Wetscherek, M.; Naumann, T.; Nori, A.; Alvarez-Valle, J.; Poon, H.; and Oktay, O. 2022. Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 1–21.

Chen, C.; Belavy, D.; Yu, W.; Chu, C.; Armbrecht, G.; Bansmann, M.; Felsenberg, D.; and Zheng, G. 2015. Localization and Segmentation of 3D Intervertebral Discs in MR Images by Data Driven Estimation. *IEEE Transactions on Medical Imaging*, 34(8): 1719–1729.

Chen, L. C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *European Conference on Computer Vision*.

Chen, X.; Yuan, Y.; Zeng, G.; and Wang, J. 2021. Semi-supervised semantic segmentation with cross pseudo supervision. *In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2613–2622.

Chun, S.; Oh, S. J.; De Rezende, R. S.; Kalantidis, Y.; and Larlus, D. 2021. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8415–8424.

Ding, H.; Liu, C.; Wang, S.; and Jiang, X. 2022. VLT: Vision-Language Transformer and Query Generation for Referring Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6): 1–16.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Han, Z.; Wei, B.; Mercado, A.; Leung, S.; and Li, S. 2018. Spine-GAN: Semantic segmentation of multiple spinal structures. *Medical image analysis*, 50: 23–35.

Ji, Y.; Wang, J.; Gong, Y.; Zhang, L.; Zhu, Y.; Wang, H.; Zhang, J.; Sakai, T.; and Yang, Y. 2023. Map: Multimodal uncertainty-aware vision-language pre-training model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23262–23271.

Kantorovich, L. V. 1960. Mathematical methods of organizing and planning production. *Management science*, 6(4): 366–422.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, 5583–5594.

Lee, D.-H.; et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 896.

Lessmann, N.; Van Ginneken, B.; De Jong, P. A.; and Išgum, I. 2019. Iterative fully convolutional neural networks for automatic vertebra segmentation and identification. *Medical Image Analysis*, 53: 142–155.

Li, X.; Dou, Q.; Chen, H.; Fu, C. W.; Qi, X.; Belavy, D. L.; Armbrecht, G.; Felsenberg, D.; Zheng, G.; and Heng, P. A. 2018. 3D multi-scale FCN with random modality voxel dropout learning for Intervertebral Disc Localization and Segmentation from Multi-modality MR Images. *Medical Image Analysis*, 45: 41–54.

Li, X.; Yu, L.; Chen, H.; Fu, C.-W.; Xing, L.; and Heng, P.-A. 2020. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Trans. Neural Netw. Learn. Syst.*, 32(2): 523–534.

Li, Z.; Li, Y.; Li, Q.; Wang, P.; Guo, D.; Lu, L.; Jin, D.; Zhang, Y.; and Hong, Q. 2023. LViT: Language meets Vision Transformer in Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, 1–1.

Liu, Y.; Tian, Y.; Chen, Y.; Liu, F.; Belagiannis, V.; and Carneiro, G. 2022. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4258–4267.

Müller, P.; Kaissis, G.; Zou, C.; and Rueckert, D. 2021. Joint Learning of Localized Representations from Medical Images and Reports. *arXiv preprint*.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Ouali, Y.; Hudelot, C.; and Tami, M. 2020. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12674–12684.

Pang, S.; Pang, C.; Zhao, L.; Chen, Y.; Su, Z.; Zhou, Y.; Huang, M.; Yang, W.; Lu, H.; and Feng, Q. 2020. SpineParseNet: spine parsing for volumetric MR image by a two-stage segmentation framework with semantic image representation. *IEEE Transactions on Medical Imaging*, 40(1): 262–273.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763.

Ramaswamy, H. G.; et al. 2020. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *proceedings of the IEEE/CVF winter conference on applications of computer vision*, 983–991.

Shen, Z.; Cao, P.; Yang, H.; Liu, X.; Yang, J.; and Zaiane, O. R. 2023. Co-training with high-confidence pseudo labels for semi-supervised medical image segmentation. *arXiv preprint arXiv:2301.04465*.

Tao, R.; Liu, W.; and Zheng, G. 2022. Spine-transformers: Vertebra labeling and segmentation in arbitrary field-of-view spine CTs via 3D transformers. *Medical image analysis*, 75: 102258.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

Van Engelen, H. H., J.E. 2020. A survey on semi-supervised learning. Machine learning. *Mach. Learn*, 109(2): 373–440.

Wang, P.; Peng, J.; Pedersoli, M.; Zhou, Y.; Zhang, C.; and Desrosiers, C. 2021. Self-paced and self-consistent co-training for semi-supervised image segmentation. *Med. Image Anal.*, 73: 102146.

Wang, Y.; Zhang, Y.; Tian, J.; Zhong, C.; Shi, Z.; Zhang, Y.; and He, Z. 2020. Double-uncertainty weighted method for semi-supervised learning. 542–551.

Wang, Z.; Wu, Z.; Agarwal, D.; and Sun, J. 2022. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*.

Wu, Y.; Ge, Z.; Zhang, D.; Xu, M.; Zhang, L.; Xia, Y.; and Cai, J. 2022a. Mutual consistency learning for semi-supervised medical image segmentation. *Medical Image Analysis*, 81: 102530.

Wu, Y.; Wu, Z.; Wu, Q.; Ge, Z.; and Cai, J. 2022b. Exploring smoothness and class-separation for semi-supervised medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, 34–43.

Xia, Y.; Yang, D.; Yu, Z.; Liu, F.; Cai, J.; Yu, L.; Zhu, Z.; Xu, D.; Yuille, A.; and Roth, H. 2020. Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. *Med. Image Anal.*, 65: 101766.

Yang, G.; Zhang, J.; Zhang, Y.; Wu, B.; and Yang, Y. 2021. Probabilistic modeling of semantic ambiguity for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12527–12536.

Yang, Z.; Wang, J.; Tang, Y.; Chen, K.; Zhao, H.; and Torr, P. H. 2022. LAVT: Language-Aware Vision Transformer for Referring Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18134–18144.

Yu, T.; Li, D.; Yang, Y.; Hospedales, T. M.; and Xiang, T. 2019. Robust person re-identification by modelling feature uncertainty. In *Proceedings of the IEEE/CVF international conference on computer vision*, 552–561.

Zhang, S.; Zhang, J.; Tian, B.; Lukasiewicz, T.; and Xu, Z. 2023. Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation. *Med. Image Anal.*, 83: 102656.

Zhu, Y.; Zhang, Z.; Wu, C.; Zhang, Z.; He, T.; Zhang, H.; Manmatha, R.; Li, M.; and Smola, A. 2021. Improving semantic segmentation via self-training. arXiv 2020. *arXiv preprint arXiv:2004.14960*.

Reproducibility Checklist

Unless specified otherwise, please answer "yes" to each question if the relevant information is described either in the paper itself or in a technical appendix with an explicit reference from the main paper. If you wish to explain an answer further, please do so in a section titled "Reproducibility Checklist" at the end of the technical appendix.

This paper:

1. Includes a conceptual outline and/or pseudocode description of AI methods introduced. **yes**
2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results. **yes**
3. Provides well marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper. **yes**

Does this paper make theoretical contributions? yes

If yes, please complete the list below.

1. All assumptions and restrictions are stated clearly and formally. **yes**
2. All novel claims are stated formally (e.g., in theorem statements). **yes**
3. Proofs of all novel claims are included. **yes**

4. Proof sketches or intuitions are given for complex and/or novel results. **yes**
5. Appropriate citations to theoretical tools used are given. **yes**
6. All theoretical claims are demonstrated empirically to hold. **yes**
7. All experimental code used to eliminate or disprove claims is included. **yes**

Does this paper rely on one or more datasets? one

If yes, please complete the list below.

1. A motivation is given for why the experiments are conducted on the selected datasets. **yes**
2. All novel datasets introduced in this paper are included in a data appendix. **yes**
3. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. **yes**
4. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations. **yes**
5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available. **yes**
6. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing. **N/A**

Does this paper include computational experiments? yes

If yes, please complete the list below.

1. Any code required for pre-processing data is included in the appendix. **yes**
2. All source code required for conducting and analyzing the experiments is included in a code appendix. **yes**
3. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. **yes**
4. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from. **yes**
5. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results. **yes**
6. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks. **yes**
7. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics. **yes**
8. This paper states the number of algorithm runs used to compute each reported result. **yes**
9. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information. **yes**
10. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank). **yes**
11. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments. **yes**
12. This paper states the number and range of values tried per (hyper-)parameter during development of the paper, along with the criterion used for selecting the final parameter setting. **no**