Low-Rank Text-Guided Spectral Learning for Semi-Supervised MHSI Segmentation

Siqi Zhang, Qing Zhang(✉), Yan Wang, and Qingli Li

Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai 200241, China
HangzhouHyperspectral Imaging Technology Co., Ltd, Hangzhou 310024, China
qzhang@cee.ecnu.edu.cn

Abstract. Microscopic hyperspectral image (MHSI) provides rich spectral information that enables the detection of subtle biochemical variations in tissue, making it a powerful modality for computational pathology. However, hyperspectral data exhibits substantial spectral redundancy, which may obscure critical discriminative cues. Current approaches rarely focus on identifying informative spectral components, thus failing to leverage the spectral invariance of pathological samples. Moreover, large-scale precise annotations are difficult to obtain, and textual description for MHSI remains largely unavailable, limiting supervised learning. To address these challenges, we propose a novel **Lo**w-rank **T**ext-guided **S**pectral learning **Net**work for semi-supervised MHSI segmentation, named as **LoTS-Net**. Specifically, we introduce a text-guided spectral channel selection mechanism to extract refined low-rank spectral representations and incorporate textual semantics to enable interpretable and redundancy-aware channel selection. Then, we leverage intrinsic spectral similarity across samples by constructing a spectral feature container that retrieves correlated prototypes to improve spectral-spatial fusion. This container mitigates noise during the interaction between labeled and unlabeled samples. To support future research, we construct the first microscopic hyperspectral vision-language benchmark and conduct comprehensive evaluations. Experimental results demonstrate that our method significantly outperforms the state-of-the-art methods, particularly under the challenging 5% labeled data setting. Code is available at https://github.com/ECNU-MultiDimLab/LoTS-Net.

Keywords: Hyperspectral Image Segmentation · Semi-Supervised Learning · Spectral Selection · Spectral-Spatial Fusion.

1 Introduction

Microscopic hyperspectral image (MHSI) has emerged as a promising modality in computational pathology [16]. MHSI is a three-dimensional data cube, capturing rich spectral signatures across hundreds of contiguous wavelengths. The fine-grained spectral representation enables the characterization of subtle biochemical and structural alterations within tissues, making it a critical tool to

precisely delineate lesion boundaries and identify malignant regions that are difficult to distinguish using RGB image. However, the spectral information is not directly interpretable in a visual sense, making large-scale pixel-level annotation particularly difficult and labor-intensive.

To alleviate this limitation, semi-supervised learning (SSL) has been a promising solution by leveraging limited labeled data together with abundant unlabeled data. Existing SSL methods for medical image segmentation rely on two paradigms: pseudo-labeling [23,24,29] and consistency regularization [10,3]. The former generates artificial labels based on high-confidence predictions, while the latter enforces prediction invariance under perturbations. Recent studies have extended these strategies to MHSI segmentation. For example, ACCL-CINet [17] proposes an adversarial consistency constraint learning method by enforcing distribution alignment between labeled and unlabeled data. These methods rely exclusively on visual information to generate pseudo labels for unlabeled data. This visual-only strategy is suboptimal for MHSI segmentation because their pathological tissues often exhibit minimal spatial contrast but distinctive spectral signatures. Consequently, pseudo-labels generated solely from spatial-visual cues are prone to confirmation bias and error accumulation.

The spectral signature of a specific tissue type remains inherently stable and physically grounded, suggesting that it can serve as a reliable prior to guide pseudo-label generation. However, MHSI data contains substantial redundancy due to strong spectral correlation among adjacent bands [26,30]. Although traditional dimensionality reduction methods like principal component analysis (PCA) have been widely adopted to remove redundancy, they are data-driven linear projections that may not preserve task-relevant spectral components [22]. Recent studies indicate that the low-rank structure of hyperspectral data encodes essential discriminative information for tissue characterization [13]. Nevertheless, most existing methods fail to explicitly exploit this low-rank prior in SSL settings. Meanwhile, vision-language models (VLMs) [32,11] have demonstrated remarkable success in image understanding by introducing textual guidance. In medical image analysis, text prompt derived from clinical knowledge has shown effectiveness in enhancing feature discrimination and suppressing noise [31,14]. However, it has not yet been explored in semi-supervised MHSI segmentation.

To solve these challenges, we propose a novel **Lo**w-rank **T**ext-guided **S**pectral learning **Net**work (**LoTS-Net**) for semi-supervised MHSI segmentation. To effectively reduce spectral redundancy, we design a *text-guided spectral channel selection* that enhances discriminative spectral features, while explicitly modeling the low-rank structure of MHSI data. We then introduce a *spectral feature container* to store high-value spectral feature as the dataset-level spectral knowledge. This historical spectral prior knowledge guides the precise segmentation of labeled data and reliable pseudo labeling of unlabeled data. By integrating the spectral priors and semantic guidance into the SSL paradigm, LoTS-Net mitigates confirmation bias and enhances segmentation robustness under limited supervision. We construct the first vision-language benchmark for microscopic hyperspectral pathology by extending two public datasets with text prompts.

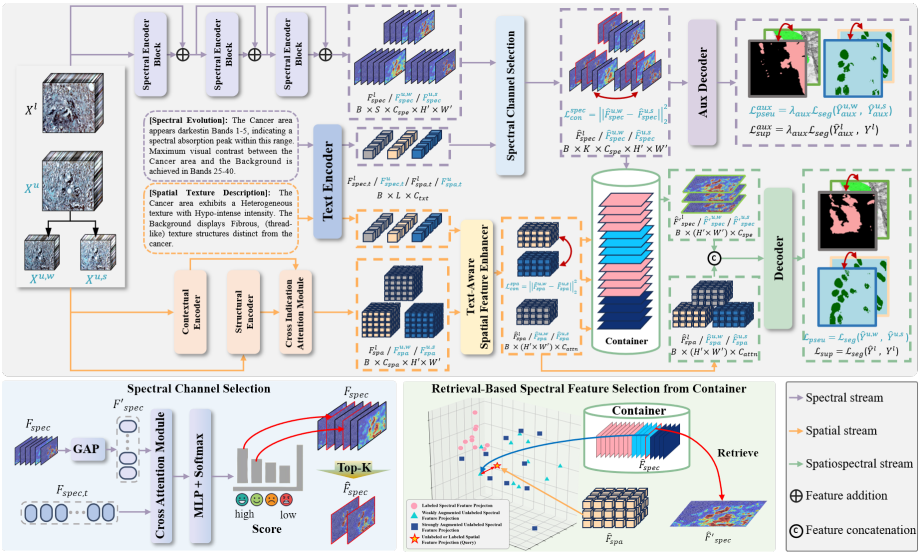

Fig. 1. Overview of LoTS-Net. The texts are leveraged for spectral-spatial feature enhancement. Key spectral channels are selected and stored in container, guiding the spatial-spectral feature fusion and the pseudo-label generation for unlabeled data by retrieving the most matched spectral feature.

Extensive experiments on them demonstrate the superiority of the proposed LoTS-Net compared to existing MHSI segmentation methods.

2 Method

2.1 Problem Definition and Network Structure

Given labeled dataset $\mathcal{D}_{\mathrm{L}} = \{(X_i^l, T_i^l, Y_i^l)\}_{i=1}^{N_{\mathrm{L}}}$, our goal is to capture precise semi-supervised MHSI segmentation. Here, $X_i \in \mathbb{R}^{S\times H\times W}$ represents the input hyperspectral image with S spectral bands and spatial resolution $H \times W$, $T_i^l = \{T_{\mathrm{spa}_i}^{\ l}, T_{\mathrm{spec}_i}^{\ l}\}$ represents the textual prompts of both spatial and spectral modalities, and $Y_i \in \{0,1\}^{H\times W}$ is the ground truth binary mask. Similarly, let $\mathcal{D}_{\mathrm{U}} = \{(X_i^u, T_{spa_i}^{\ u}, T_{spec_i}^{\ u}\}_{i=1}^{N_{\mathrm{U}}}$ denote the unlabeled dataset. Fig. 1 illustrates the overall architecture of the proposed LoTS-Net. Specifically, it leverages spectral text prompt T_{spec} to identify key spectral channels, and spatial text prompt T_{spa} to enhance the texture information. These modalities are fused by retrieving relevant spectral prototypes, followed by a unified decoder for segmentation. For semi-supervised learning, we apply weak-to-strong augmentation on unlabeled data and enforce consistency constraints with pseudo-labels by the intrinsic spectral consistency to guide model training.

2.2 Text-Guided Spatial-Spectral Feature Enhancement

Text-Guided Spectral Channel Selection. To address the spectral redundancy inherent in MHSI and physically discriminative spectral features, we design a text-guided spectral channel selection module. Inspired by [26], we employ 3 spectral encoder blocks to extract low-rank spectral feature map $F_{\text{spec}} \in \mathbb{R}^{B\times S\times C_{\text{spe}}\times H'\times W'}$, where B is the batch size, $H' \times W'$ represents the spatial dimensions. Then, we capture the global spectral fingerprint $F'_{\text{spec}} \in \mathbb{R}^{B\times S\times C_{\text{spe}}}$ via global average pooling (GAP). Simultaneously, spectral text prompt T_{spec} is encoded by a frozen text encoder to obtain the semantic feature $F_{\text{spec,t}}$. We inject the textual semantics into the spectral features through a cross-attention mechanism, producing an enhanced spectral feature. Based on this feature, we employ a multi-layer perceptron (MLP) block to generate importance scores for each channel, and then select the top-K channels. To preserve the physical purity of the features, we sample directly from the original spectral feature F_{spec} using the selected channel indexes. The refined feature $\hat{F}_{\text{spec}} \in \mathbb{R}^{B\times K\times C_{\text{spe}}\times H'\times W'}$ serve as high-value spectral prototypes for subsequent interactions.

Text-Aware Spatial Feature Enhancement. We employ the feature extraction module in CINet [17], combining contextual and structural encoders, to capture spatial feature F_{spa} with both local textural details and global structure. Similar to the spectral stream, we encode the spatial text prompts T_{spa} by a frozen text encoder to obtain the semantic embedding $F_{\text{spa,t}}$. We then fuse the textual feature with the spatial feature by a cross-attention mechanism to obtain the enhanced spatial feature $\hat{F}_{\text{spa}} \in \mathbb{R}^{B\times(H'\times W')\times C_{\text{attn}}}$.

2.3 Retrieval-based Spatial-Spectral Fusion

Spectral Feature Container. Each pathological area has its own inherent spectral attributes. Therefore, we construct a first-in-first-out global queue $\mathcal{Q}$ as a spectral feature container to store the high-value spectral feature $\hat{F}_{\text{spec}}$ as the dataset-level spectral knowledge. This container ensures that the proposed model can retrieve the best-matching historical spectral prior knowledge. By enabling cross-sample spectral prototype storage and interaction, this container provides stable spectral feature support for the spatial and spectral feature fusion and the reliable pseudo label generation of unlabeled data in semi-supervised learning.

Spectral-Guided Decoder. Leveraging the stored spectral knowledge, we introduce a spectral-guided decoder by a retrieval-based interaction mechanism. We utilize $\hat{F}_{\text{spa}}$ as a query to retrieve the most relevant spectral feature $\hat{F}'_{\text{spec}}$ from the container $\mathcal{Q}$ within each epoch with batch size of B.

$$\mathrm{I}_{\text{top}} = \texttt{argmax}^{(-1)}\left(\texttt{Softmax}\left(\frac{\hat{F}_{\text{spa}}(\texttt{MLP}(\mathcal{Q}))^T}{\sqrt{d}}\right)\right), \quad \hat{F}'_{\text{spec}} = \mathcal{Q}[\mathrm{I}_{\text{top}}]. \tag{1}$$

where $\mathrm{I}_{\text{top}} \in \mathbb{R}^{B\times 1}$ denotes the indices of the spectral feature maps within the container. The resulting $\hat{F}'_{\text{spec}} \in \mathbb{R}^{B\times(H'\times W')\times C_{\text{spe}}}$ represents a spectral feature

map that is spatially aligned with the spatial feature $\hat{F}_{\text{spa}}$. We then concatenate them along the channel dimension and feed them into a unified decoder equipped with skip connections. This decoder progressively upsamples the fused features to restore resolution and generates the final segmentation prediction $\hat{Y}$.

2.4 Semi-Supervised Learning Strategy

For the semi-supervised segmentation task, we employ a hybrid training strategy that integrates a fully supervised loss for labeled data with a weak-to-strong consistency regularization mechanism for unlabeled data.

Hybrid Supervision. For labeled data, we compute a combination of Dice and Cross-Entropy losses as the segmentation loss $\mathcal{L}_{\text{seg}}$ between the main prediction $\hat{Y}^l$ and the ground truth Y^l. Additionally, to ensure effective training of the spectral stream, we introduce an auxiliary decoder with the selected features $\hat{F}^l_{\text{spec}}$ as input to generate the segmentation mask $\hat{Y}^l_{\text{aux}}$. The supervision loss is:

$$\mathcal{L}_{\text{sup}} = \mathcal{L}_{\text{seg}}(\hat{Y}^l, Y^l) + \lambda_{\text{aux}}\mathcal{L}_{\text{seg}}(\hat{Y}^l_{\text{aux}}, Y^l), \tag{2}$$

where λ_{aux} is a weight hyperparameter.

Pseudo-Labeling and Spatiospectral Consistency Regularization. For unlabeled data, we generate two perturbed views: a weak augmentation view $X^{u,w}$ with salt-and-pepper noise and a strong augmentation view $X^{u,s}$ with semantic perturbations (random brightness, contrast jitter, and Gaussian noise) [6]. Subsequently, we generate pseudo labels $\hat{Y}^{u,w}$ generated from $X^{u,w}$ by the shared network of the labeled data, supervising the prediction $\hat{Y}^{u,s}$ of $X^{u,s}$. The pseudo-labeling supervision loss $\mathcal{L}_{\text{pseu}}$ is defined as:

$$\mathcal{L}_{\text{pseu}} = \mathcal{L}_{\text{seg}}(\hat{Y}^{u,s}, \hat{Y}^{u,w}), \tag{3}$$

where $\hat{Y}^{u,w}$ serves as the pseudo-label target generated from the weak augmentation view.

To enhance the robustness of the model's internal representations, we propose a spatiospectral consistency regularization by enforcing invariance within both the spatial and spectral feature space of one unlabeled sample. Specifically, we minimize the discrepancy between $\hat{F}^{u,s}_{\text{spa}}$ and $\hat{F}^{u,w}_{\text{spa}}$, the spatial features extracted from its strong and weak views, as well as the discrepancy between $\hat{F}^{u,s}_{\text{spec}}$ and $\hat{F}^{u,w}_{\text{spec}}$, the selected spectral features of strong and weak views. The consistency regularization is optimized by the loss:

$$\mathcal{L}_{\text{con}} = \|\hat{F}^{u,s}_{\text{spa}} - \hat{F}^{u,w}_{\text{spa}}\|_2^2 + \|\hat{F}^{u,s}_{\text{spec}} - \hat{F}^{u,w}_{\text{spec}}\|_2^2, \tag{4}$$

where $\|\cdot\|_2^2$ represents Euclidean Distance.

By enforcing feature-level invariance across augmentation views in both spatial and spectral domains, this consistency regularization significantly improves the robustness of the semi-supervised segmentation. The total unsupervised loss is a sum of the above two loss functions:

$$\mathcal{L}_{\text{unsup}} = \mathcal{L}_{\text{pseu}} + \mathcal{L}_{\text{con}}. \tag{5}$$

Table 1. Performance comparison of fully-supervised methods on two datasets. Modality I and T separately denote Image and Text. Best results are highlighted in **bold**.

Method	Mod.	MDC			GPCC		
		DICE↑	IoU↑	HD95↓	DICE↑	IoU↑	HD95↓
nnU-Net [9]	I	78.34	65.67	47.81	91.20	84.69	45.73
Swin-Unet [4]	I	71.52	57.19	57.60	81.72	70.83	63.55
Spec-Tr [27]	I	77.06	65.23	42.01	89.21	82.41	43.43
Swin-UM [15]	I	78.58	65.64	50.52	89.51	81.84	55.15
VMUNet [20]	I	77.15	63.61	50.96	83.54	73.38	62.14
Omni-Fuse [30]	I	84.12	69.39	54.84	79.78	66.84	72.96
MK-UNET [19]	I	81.31	71.69	59.79	91.01	84.19	48.96
EffiDec3D [18]	I	84.02	74.80	46.56	85.76	76.87	66.89
LViT-T [14]	$I\&T$	65.74	49.73	86.78	89.59	82.76	43.48
RecLMIS [8]	$I\&T$	72.40	57.53	52.13	91.58	85.06	47.20
ViTexNet [2]	$I\&T$	82.99	74.06	63.56	82.84	73.40	57.32
ARSeg [21]	$I\&T$	81.71	71.75	54.41	84.50	74.86	64.80
LoTS-Net	$I\&T$	**85.62**	**77.05**	**55.02**	**91.65**	**85.33**	**41.81**

Dynamic Weighting for Unsupervised Loss. To prevent low-quality pseudo-labels from destabilizing the optimization process during the early stages of training, we introduce a time-dependent Gaussian ramp-up weight $\lambda_{\text{ramp}}(t)$ that dynamically adjusts the importance of the unsupervised loss. This weight of current training epoch t is defined as:

$$\lambda_{\text{ramp}}(t) = \lambda_{\max} \cdot e^{-0.5\left(1-\frac{t}{\mathcal{T}_{\max}}\right)^2}, \tag{6}$$

where $\lambda_{\max}$ is the maximum consistency weight, and $\mathcal{T}_{\max}$ is the total length of the ramp-up phase. This function ensures a minimal unsupervised weight at the beginning and gradually increasing as model predictions improve. The total loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sup}} + \lambda_{\text{ramp}}(t) \cdot \mathcal{L}_{\text{unsup}}. \tag{7}$$

3 Experiments and Results

3.1 Experimental Setup

Datasets. We evaluate the performance of LoTS-Net on two public MHSI datasets: MDC and GPCC datasets [30]. For **MDC**, we cut 2,276 patches with resolution of $256\times256\times60$ from 684 cancerous ROIs captured under a $20\times$ objective lens. For **GPCC**, we extract 1,366 patches with resolution of $256\times256\times40$ from 414 ROIs captured under a $10\times$ objective lens. To enable the proposed multimodal learning framework, we extend both datasets by generating both spatial and spectral text prompts for each patch using BioBERT [12]. We adopt a random split ratio of 6:2:2 for training, validation, and testing under the fully supervised setting. To evaluate semi-supervised performance, we further randomly select 5%, 10%, and 20% of the training set as labeled data, treating the remaining training samples as unlabeled data.

Table 2. Performance comparison of semi-supervised methods (5%, 10%, 20% labeled data). Modality I and T denote Image and Text. Best results are highlighted in **bold**.

Data	Method	Mod.	5% Label			10% Label			20% Label		
			DICE↑	IoU↑	HD95↓	DICE↑	IoU↑	HD95↓	DICE↑	IoU↑	HD95↓
MDC	CPS [5]	I	69.25	54.59	64.77	70.34	55.55	62.98	76.98	63.91	54.42
	BCP [1]	I	69.90	55.19	65.22	72.02	58.32	54.60	75.42	62.28	55.99
	ACCL-CINet [17]	I	69.12	54.98	60.90	75.60	62.45	50.88	77.19	64.06	56.12
	SimTxt [25]	$I\&T$	74.87	61.51	50.57	79.31	67.03	49.82	80.51	68.26	43.93
	KnowSAM [7]	$I\&T$	73.07	58.66	60.84	77.74	64.37	59.16	78.82	65.85	59.08
	TextMoE [28]	$I\&T$	73.30	59.22	60.87	79.37	66.79	47.66	80.24	67.94	42.45
	LoTS-Net	$I\&T$	**77.59**	**64.83**	**52.20**	**80.19**	**67.79**	**52.34**	**81.37**	**69.43**	**40.14**
GPCC	CPS [5]	I	81.62	70.27	60.09	85.81	76.17	54.59	86.74	77.49	54.22
	BCP [1]	I	81.99	70.80	62.47	83.17	72.50	61.69	84.24	74.05	58.60
	ACCL-CINet [17]	I	83.47	73.27	55.08	86.74	77.68	49.01	87.80	79.15	46.90
	SimTxt [25]	$I\&T$	82.60	72.41	53.49	86.68	77.60	49.22	88.82	80.69	49.01
	KnowSAM [7]	$I\&T$	79.28	66.68	47.65	81.70	69.99	43.88	82.65	71.25	44.05
	TextMoE [28]	$I\&T$	80.57	68.32	47.04	86.78	77.80	49.86	90.40	83.10	49.69
	LoTS-Net	$I\&T$	**83.76**	**73.84**	**48.59**	**86.90**	**77.96**	**50.94**	**90.59**	**83.33**	**47.45**

Table 3. Evaluation of the main components on two datasets with 20% labeled data.

$\mathcal{L}_{\text{sup}}$	SFC	SCS-M	SCS-T	$\mathcal{L}_{\text{con}}$	$\mathcal{L}_{\text{pseu}}$	MDC (20%)			GPCC (20%)		
						DICE↑	IoU↑	HD95↓	DICE↑	IoU↑	HD95↓
✓	×	×	×	×	×	49.58	35.02	88.48	55.40	39.77	72.45
✓	✓	×	×	×	×	53.42	38.11	89.54	65.53	51.68	74.81
✓	✓	✓	×	×	×	65.27	51.72	63.98	74.82	63.55	57.73
✓	✓	✓	✓	×	×	73.03	59.83	52.89	80.84	69.95	55.34
✓	✓	✓	✓	✓	×	78.79	65.72	42.58	87.86	79.27	52.59
✓	✓	✓	✓	✓	✓	**81.37**	**69.43**	**40.14**	**90.59**	**83.33**	**47.45**

Implementation Details. The model is optimized using the AdamW optimizer for 200 epochs, an initial learning rate of 8×10^{-4} and a batch size of 2. We set the number of selected spectral channels K to 15. For the semi-supervised setting, we set the auxiliary loss weight $\lambda_{\text{aux}} = 0.8$, the maximum consistency weight $\lambda_{\max} = 0.8$, and the ramp-up period $\mathcal{T}_{\max} = 50$ epochs. We employ the Dice Similarity Coefficient (Dice, %), Intersection over Union (IoU, %), and the 95th percentile Hausdorff Distance (HD95, px) to quantitatively evaluate the segmentation performance. All experiments are conducted using the PyTorch 2.4.1 library and trained on one NVIDIA RTX 3090 GPU.

3.2 Performance Evaluation

We compare LoTS-Net with SOTA unimodal and dual-modal segmentation methods on the MDC and GPCC datasets under both fully supervised and semi-supervised settings. As shown in Table 1, under fully supervised setting, LoTS-Net demonstrates highly competitive performance across all metrics. Specifically,

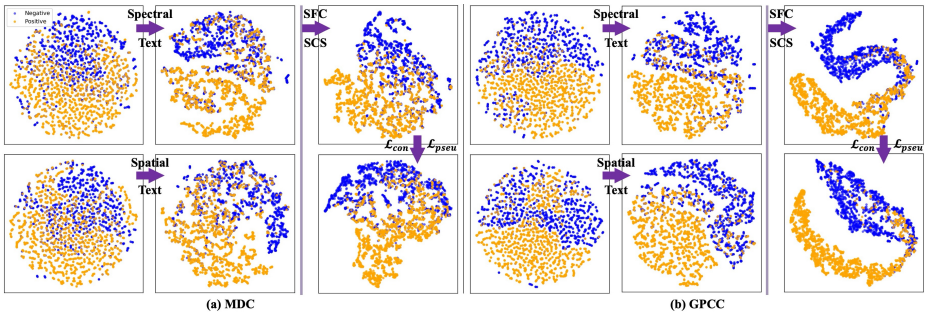

Fig. 2. t-SNE visualization of feature distributions across different modules. "Spectral/Spatial Text" are text prompts, "SFC" is spectral feature container, "SCS" is spectral channel selection, and $\{\mathcal{L}_{\text{con}}, \mathcal{L}_{\text{pseu}}\}$ are spatiospectral consistency losses.

it achieves a Dice score of 85.62% on MDC, surpassing the top-performing unimodal method EffiDec3D (84.02%) and the latest dual-modal method baseline ViTexNet (82.99%), highlighting LoTS-Net's superior ability to leverage both spectral and spatial features. Additionally, LoTS-Net achieves the lowest HD95 error on both datasets, demonstrating its robustness in capturing fine-grained segmentation boundaries. We also evaluate its semi-supervised learning ability, as shown in Table 2. The results reveal that LoTS-Net consistently outperforms SOTA methods across all label ratios (5%, 10%, and 20%) in terms of all metrics. As the amoun of labeled data increases, its DICE score improves, reaching 81.37% on MDC and 90.59% on GPCC at 20% ratio. All these results validate the effectiveness of the proposed semi-supervised learning strategy and the guidance provided by the selected key spectral features in LoTS-Net.

3.3 Ablation Study

To validate the effectiveness of each proposed component in LoTS-Net, we conduct comprehensive ablation studies on both datasets with 20% labeled data. The quantitative results are reported in Table 3, and the corresponding feature space evolution is visualized using t-SNE in Fig. 2. The baseline model utilizes only $\mathcal{L}_{\text{sup}}$ on a naive dual-stream architecture, performing poorly on both MDC and GPCC. Integrating the spectral feature container (SFC) alone brings a modest improvement, as historical spectral knowledge provides broader data distributions. A substantial leap in performance is observed when spectral channel selection (SCS-M) is conducted by using an MLP to assign importance scores to each spectral channel, demonstrating that the spectral redundancy negatively impacts the model performance. Specifically, the inclusion of text-guided selection strategy (SCS-T) boosts the DICE score on MDC to 73.03% and GPCC to 80.84%. Finally, the designed semi-supervised learning strategy with $\{\mathcal{L}_{\text{con}}, \mathcal{L}_{\text{pseu}}\}$ further improve the segmentation performance of LoTS-Net.

4 Conclusion

In this paper, we propose LoTS-Net, a novel language-aware dual-stream framework for semi-supervised microscopic hyperspectral image (MHSI) segmentation. We introduce a text-guided spectral channel selection mechanism to extract representative dataset-level spectral knowledge, with high-quality spectral feature maps stored in a spectral feature container to enable robust spatial-spectral fusion and reliable pseudo-label generation through a retrieval-based decoder. Experiments on the two extended public datasets demonstrate that LoTS-Net substantially outperforms SOTA semi-supervised approaches, particularly under the highly constrained 5% labeled data regime.

Acknowledgments. This work is supported by National Natural Science Foundation of China (Grant No. 62475072, 42530110), Fundamental Research Funds for the Central Universities, the Science and Technology Commission of Shanghai Municipality (Grant No. 25xtcx00600, 22DZ2229004), Joint Laboratory of Hyperspectral Big Data and Artificial Intelligence (No. MIP202602), the AI project from the economic and information commission of Shanghai (Grant No. 2024-GZL-RGZN-01038).

Disclosure of Interests. The authors have no competing interests to declare in relation to this paper.

References

1. Bai, Y., Chen, D., Li, Q., Shen, W., Wang, Y.: Bidirectional copy-paste for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11514–11524 (2023)
2. Bhardwaj, R., Tambe, U.Y., Neog, D.R.: Vitexnet: Vision-text guided dynamic convolution network for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 690–699. Springer (2025)
3. Bortsova, G., Dubost, F., Hogeweg, L., Katramados, I., De Bruijne, M.: Semi-supervised medical image segmentation via learning consistency under transformations. In: International conference on medical image computing and computer-assisted intervention. pp. 810–818. Springer (2019)
4. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. pp. 205–218. Springer (2022)
5. 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)
6. Huang, K., Zhang, Y., Zhou, Y., Xu, T., Zhou, T.: Bidirectional channel-selective semantic interaction for semi-supervised medical segmentation. arXiv preprint arXiv:2601.05855 (2026)
7. Huang, K., Zhou, T., Fu, H., Zhang, Y., Zhou, Y., Gong, C., Liang, D.: Learnable prompting sam-induced knowledge distillation for semi-supervised medical image segmentation. IEEE Transactions on Medical Imaging **44**(5), 2295–2306 (2025)

8. Huang, X., Li, H., Cao, M., Chen, L., You, C., An, D.: Cross-modal conditioned reconstruction for language-guided medical image segmentation. IEEE Transactions on Medical Imaging **44**(4), 1821–1835 (2024)
9. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
10. Jiao, R., Zhang, Y., Ding, L., Xue, B., Zhang, J., Cai, R., Jin, C.: Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation. Computers in Biology and Medicine **169**, 107840 (2024)
11. Kumar, G.M.K., Chadha, A., Mendola, J.D., Shmuel, A.: Medvisionllama: Leveraging pre-trained large language model layers to enhance medical image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1114–1124 (2025)
12. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)
13. Li, L., Li, W., Du, Q., Tao, R.: Low-rank and sparse decomposition with mixture of gaussian for hyperspectral anomaly detection. IEEE Transactions on Cybernetics **51**(9), 4363–4372 (2020)
14. 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 **43**(1), 96–107 (2023)
15. Liu, J., Yang, H., Zhou, H.Y., Xi, Y., Yu, L., Li, C., Liang, Y., Shi, G., Yu, Y., Zhang, S., et al.: Swin-umamba: Mamba-based unet with imagenet-based pretraining. In: International conference on medical image computing and computer-assisted intervention. pp. 615–625. Springer (2024)
16. Lu, G., Fei, B.: Medical hyperspectral imaging: a review. Journal of biomedical optics **19**(1), 010901–010901 (2014)
17. Qin, G., Liu, H., Zhang, X., Li, W., Guo, Y., Guo, C.: Semi-supervised medical hyperspectral image segmentation using adversarial consistency constraint learning and cross indication network. IEEE Transactions on Image Processing (2025)
18. Rahman, M.M., Marculescu, R.: Effidec3d: an optimized decoder for high-performance and efficient 3d medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10435–10444 (2025)
19. Rahman, M.M., Marculescu, R.: Mk-unet: Multi-kernel lightweight cnn for medical image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1042–1051 (2025)
20. Ruan, J., Li, J., Xiang, S.: Vm-unet: Vision mamba unet for medical image segmentation. ACM Transactions on Multimedia Computing, Communications and Applications (2024)
21. Wang, Q., Lin, X., Yan, Z.: Towards robust medical image referring segmentation with incomplete textual prompts. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 636–646. Springer (2025)
22. Wang, Y., Xie, X., Gao, L., Zhang, B., Zhou, C., Zou, D., Lu, L., Li, Q.: S 4 r: Separated self-supervised spectral regression for hyperspectral histopathology image diagnosis. IEEE Transactions on Image Processing (2025)
23. Wu, H., Wang, C., Cui, Z.: Dual cross-image semantic consistency with self-aware pseudo labeling for semi-supervised medical image segmentation. IEEE Transactions on Medical Imaging (2025)

24. Wu, H., Zhang, B., Chen, C., Qin, J.: Federated semi-supervised medical image segmentation via prototype-based pseudo-labeling and contrastive learning. IEEE Transactions on Medical Imaging **43**(2), 649–661 (2023)
25. Xie, Y., Zhou, T., Zhou, Y., Chen, G.: Simtxtseg: Weakly-supervised medical image segmentation with simple text cues. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 634–644. Springer (2024)
26. Yun, B., Li, Q., Mitrofanova, L., Zhou, C., Wang, Y.: Factor space and spectrum for medical hyperspectral image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 152–162. Springer (2023)
27. Yun, B., Wang, Y., Chen, J., Wang, H., Shen, W., Li, Q.: Spectr: Spectral transformer for hyperspectral pathology image segmentation. arXiv preprint arXiv:2103.03604 (2021)
28. Zeng, Q., Luo, H., Ma, X., Lu, Z., Hu, Y., Xia, Y.: Exploring text-enhanced mixture-of-experts for semi-supervised medical image segmentation with composite data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 226–236. Springer (2025)
29. Zhang, Q., Li, L., Cai, R., Li, Q., Dissanayake, B., Wang, Y., Kot, A.: Ps-net: high-frequency attention and bayesian analysis based facial pore segmentation with no human annotation. Visual Intelligence **3**(1), 20 (2025)
30. Zhang, Q., Pei, G., Wang, Y.: Omni-fusion of spatial and spectral for hyperspectral image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 471–481. Springer (2025)
31. Zhao, Y., Zhong, E., Yuan, C., Li, Y., Zhao, M., Li, C., Hu, J., Liu, W., Liu, C.: Med-vlm: Enhancing medical image segmentation accuracy through vision-language model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7283–7293 (2025)
32. Zhong, L., Huang, Z., Liu, Y., Liao, W., Zhang, S., Wang, G., Zhang, S.: Vlm-cpl: Consensus pseudo-labels from vision-language models for annotation-free pathological image classification. IEEE Transactions on Medical Imaging (2025)