Multi-stage Multi-granularity Focus-tuned Learning Paradigm for Medical HSI Segmentation

Haichuan Dong[1], Runjie Zhou[1], Boxiang Yun[1], Huihui Zhou[1], Benyan Zhang[2], Qingli Li[1], and Yan Wang[1](✉)

[1] Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai, China
[2] Department of Pathology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China
10212140417@stu.ecnu.edu.cn, 10212140428@stu.ecnu.edu.cn, 52265904012@stu.ecnu.edu.cn, hhZhou0724@163.com, benyan_zhang07@163.com, qlli@cs.ecnu.edu.cn, ywang@cee.ecnu.edu.cn

Abstract. Despite significant breakthrough in computational pathology that **M**edical **H**yperspectral **I**maging (MHSI) has brought, the asymmetric information in spectral and spatial dimensions pose a primary challenge. In this study, we propose a multi-stage multi-granularity Focus-tuned Learning paradigm for Medical HSI Segmentation. To learn subtle spectral differences while equalizing the spatiospectral feature learning, we design a quadruplet learning pre-training and focus-tuned fine-tuning stages for capturing both disease-level and image-level subtle spectral differences while integrating spatially and spectrally dominant features. We propose an intensifying and weakening strategy throughout all stages. Our method significantly outperforms all competitors in MHSI segmentation, with over 3.5% improvement in DSC. Ablation study further shows our method learns compact spatiospectral features while capturing various levels of spectral differences. Code will be released at https://github.com/DHC233/FL.

Keywords: Medical hyperspectral images · self-supervised learning

1 Introduction

Hyper**s**pectral **i**maging (HSI) heralds a pivotal advancement in the field of medical technology, offering deep insights into the physiological and biochemical properties of tissues, especially in cancer detection and management [9]. Despite the fact that extensive spectral information and pixel-level spatial resolution of HSI (Fig. 1(a)) facilitate the precise identification, its inherent high dimensionality results in complex data processing, *i.e.*, spectral redundancy. Spectral redundancy exhibits low-rank properties of the dataset with overlapping or repetitive information across spectral bands. By calculating the spectral angles, **S**pectral **A**ngle **M**apper (SAM) heatmap [10] quantifies the similarity between spectral

H. Dong and R. Zhou—Contributed equally to this work.

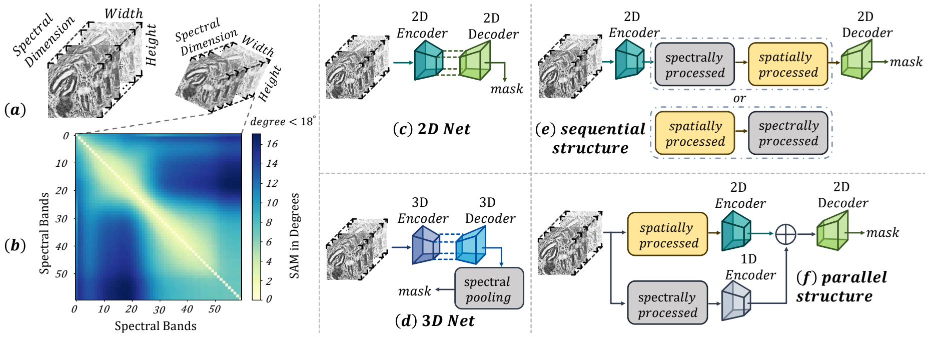


Fig. 1. (a): an example of MHSI. (b): Spectral Angle Mapper (SAM) heatmap. (c)-(f): MHSI segmentation structures.

signatures. It reveals common low spectral angles (< 18° in Fig. 1(b)) among adjacent spectra, indicating high spectral redundancy and underscoring the challenge in discerning subtle spectral variations.

A key challenge in **m**edical **h**yper**s**pectral **i**mage (MHSI) representation learning is how to balance the asymmetric information of spectral and spatial dimensions. Three types of structures are adopted for MHSI segmentation, including (1) 2D/3D networks, (2) spatiospectral sequential learning and (3) spatiospectral parallel learning. It is not suitable to directly employ a 2D or 3D network with the shortage of the spectral analysis [7] for 2D (Fig. 1(c)) and the surplus of network parameters [6] for 3D (Fig. 1(d)). Sequential operation (Fig. 1(e)) merely processes one dimension (spectral or spatial) with absence of the other at each stage. Though it attempts to extract subtle spectral differences [16, 18], focusing solely on spectral dimensions may suppress previously learnt spatial information, leading to insufficient learning of spatiospectral representations. Spatiospectral parallel learning structure (Fig. 1(f)) achieves speedup by learning MHSI features separately, and fusing them afterwards [19]. This post-fusion discourages the network to learn effective spatiospectral features and parallel learning disables the network to learn spectral differences across spatial locations. Thus, an ideal structure ought to focus on subtle spectral differences as well as equalize the relationship between spatiospectral features.

Subtle spectral differences exists at both disease-level (raw MHSIs from dataset) and image-level (extracted spatial and spectral features from images). For a specific disease, its spectral curve (Fig. 4(a)) is an important prior. But no methods considers the spectral differences from the disease level. Despite a strong feature extractor, pure deformable attention [22] neglects the spatial information in extracting the spectral discriminative feature. Simply designing a network to learn subtle spectral differences while maintaining spatial information is not easy. A self-supervised model which forcefully diverges spectral differences but gently separates spatial disparities can offer essential initialization information. Actively diverging spectral differences improves the accuracy of downstream tasks by detecting subtle spectral variations among different substances.

To this end, we propose a multi-stage multi-granularity **F**ocus-tuned **L**earning paradigm (FL) for MHSI segmentation. With the focuses on subtle spectral differences as well as the balance between spectral and spatial dimensions (focus-tuned), our FL includes both pre-training and fine-tuning stages (multi-stage) for addressing both disease-level and image-level subtle spectral differences (multi-granularity). An "intensifying and weakening" strategy is launched for different spectral bands and different (spatial and spectral) dimensions. In the mainstream, we first design **S**pectral **F**ocus **F**orge module (SFF) to intensify dominant bands and weaken redundant ones to focus on subtle spectral differences at the disease level. Then, we design a Bi-Scale Extractor and an Indicative Spatiospctral Transformer to further process and re-integrate spatially and spectrally dominant features for a better spatiospectral balance at the image level. Moreover, with the jointly pre-trained architectures (SFF and feature encoder), we tailor Self-Supervised spatiospectral Quadruplet Learning (**Q**uad-**S** **Q**uadruplet **L**earning or QSQL) to highlight subtle spectral differences while equalizing the spatiospectral features for the downstream segmentation task. Experiments show our architecture secures a notable 3.5% improvement in Dice coefficient metrics over the second best competitor.

2 Methodology

2.1 Mainstream Focus-tuned Learning Architecture

The overall Focus-tuned Learning architecture (Fig. 2) is proposed to learn subtle spectral differences from two granularities. At the disease level, Spectral Focus Forge module is devised for focusing on subtle spectral differences from dataset. At the image level, Bi-scale Extractor and Indicative Spatiospectral Transformer is designed for balancing the spatiospectral features with focus on subtle spectral differences. As the input, an MHSI is denoted by $\mathbf{Z} \in \mathbb{R}^{H \times W \times B}$, where $H \times W$ refers to spatial resolution and B refers to the number of the spectral bands.

Spectral Focus Forge Module (SFF) In SFF module, each band $z_i \in \mathbf{Z}(1 \leq i \leq B)$ goes through Global Average Pooling, following with two fully connected layers and one activation function (sigmod). After layer normalization, the focal coefficients $\omega_i \in \mathbb{R}^1$ are acquired. The smaller focal coefficients mirror the similarity among adjacent or near spectral bands, that is, existing spectral redundancy. Assigned with focal coefficients on MHSI from dataset directly, z_i is transformed into $z_i' \in \mathbf{Z}'$,where $\mathbf{Z}'$ is the output feature of the SFF module, with the enhancement of subtle spectral differences and the alleviation of spectral redundancy from disease level.The disease-level processed $\mathbf{Z}'$ is divided into $\mathbf{G}_j(1 \leq j \leq \frac{B}{3})$ spectral groups with each three bands.

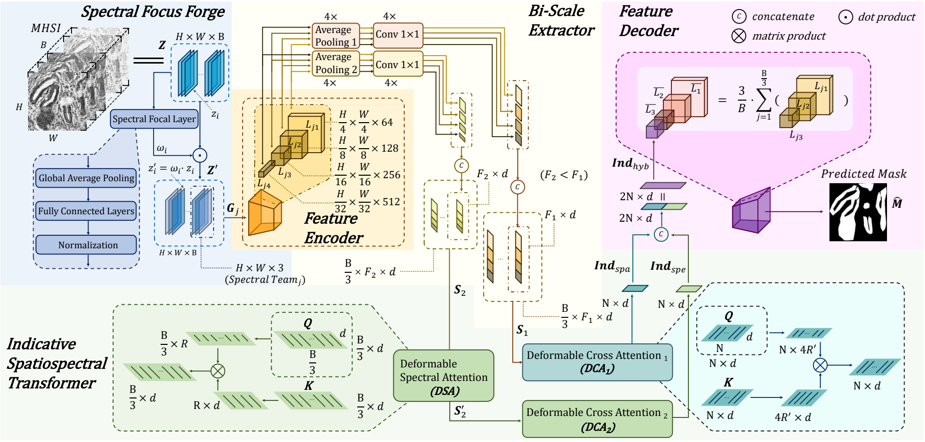


Fig. 2. Mainstream Focus-tuned Learning architecture. The MHSI cube **Z**, is processed through the Spectral Focus Forge and fed into the Feature Encoder (orange region) with each three bands to construct feature pyramids. Then, these feature pyramids undergo bifurcated processing in spatially dominant and spectrally dominant streams via Bi-scale Extractor and Indicative Spatiospectral Transformer for the spatiospectral hybrid indicator Ind_{hyb}. Ultimately, with the Ind_{hyb}, a standard Feature Decoder generates the predicted mask.

Bi-Scale Extractor Bi-Scale Extractor contains one spatially intensified stream and another spatially weakened stream. Spectral groups $\mathbf{G}_j$ are first fed into the feature extractor in parallel for the feature pyramid $\{L_{j1}, L_{j2}, L_{j3}, L_{j4}\}$ with four-level hierarchy. The feature pyramid is then fed into two streams, each of which is equipped with four independent adaptive average pooling layers connected with four independent convolutional layers with d output channels. Due to different scales in two adaptive average pooling layers, the number of spatial feature dimension($F_1 > F_2$) varies in two streams. Greater spatial feature dimension represents more intensified focus on the spatial feature, and vice versa. Concatenated with processed feature pyramid, the tensor S_1 and S_2 with respective weakened and intensified spatial focus are generated.

Indicative spatiospectral Transformer Indicative spatiospectral Transformer intensifies inter-spectral features by **d**eformable **s**pectral **a**ttention (DSA) and re-integrates spatiospectral features by **d**eformable **c**ross **a**ttention (DCA$_1$/DCA$_2$ in Fig. 2). Tensor S_2', generated from tensor S_2 via DSA, and tensor S_1 separately go through DCA$_2$ and DCA$_1$. Inspired by [22], DSA and DCA are both based on the deformable attention mechanism, which can be summarized as:

$$DeformAttn(q, p, x) = \sum_{i=1}^{N_{\text{head}}} \mathcal{W}_i \sum_{j=1}^{N_{\text{key}}} \mathcal{A}_{ij} \cdot \mathcal{W}_i' x(p + \Delta p_{ij}) \tag{1}$$

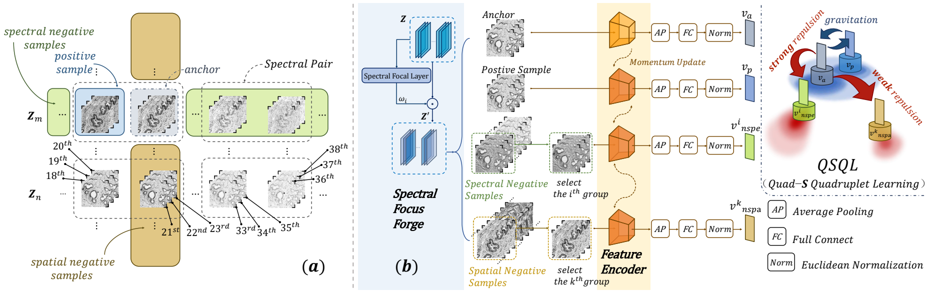


Fig. 3. (a): An example of sample selection in Quadruplet Learning. (b): Self-supervised Spatiospectral Quadruplet Learning architecture. Parameters of the SFF module and the feature encoder are trained during this process.

where q, p, x respectively represent the query, reference points and input features. N_{head} is the total number of attention heads and N_{key} is the total sampled key number for each head. $\mathcal{W}_i$ and $\mathcal{W}'_i$ are the learnable weights. $\mathcal{A}_{ij}$ is the predicted attention weight. Δp_{ij} are the predicted offsets to the reference points p.

For DSA, in the f^{th} position of S_2, it exists $\frac{B}{3}$ spectral vectors with d spatial feature dimensions. Each spectral vector is regarded as a query vector. The query vector is fed into a linear projection head to encode the sampled offset and the attention weight. The number of sampled points R is derived from output channels d (Fig. 2) to efficiently learn subtle spectral differences. The updated spatiospectral features S'_2 are generated after inter-group interactions by DSA.

For DCA, a set of learnable tokens $Ind \in \mathbb{R}^{N\times d}$, as the queries of DCA, extract more representative high-level pathological information in the feature map. The number of sampled points R' is set to $8 \times \frac{B}{3}$ at each level of the feature pyramid with four-level hierarchy. Ind_{spe} and Ind_{spa}, constructed respectively by S'_2 and S_1 from spectral and spatial dominant streams, are concatenated into Ind_{hyb} as the substitute for the top level of the average feature pyramid that represents the high semantic information. The refined feature pyramid $\{\overline{L_1}, \overline{L_2}, \overline{L_3}, Ind_{hyb}\}$ is finally fed into the feature decoder for the predicted mask $\widehat{M} \in \mathbb{R}^{H\times W}$.

2.2 Self-supervised Spatiospectral Quadruplet Learning (QSQL)

To achieve better weight initialization which can balance spatiospectral features with actively diverging subtle spectral differences and gently diverging spatial disparities, we develop a self-supervised model called QSQL. We define four elements: anchor A, positive sample S_p, spectral negative samples S_{nspe} and spatial negative samples S_{nspa} (Fig. 3 (a)). Metric-learning-based quadruplet loss is utilized as the objective function. In the process, the weights of the SFF module and the feature encoder for the anchor are trained, while those for positive and negative samples are updated in a momentum manner (Fig. 3 (b)). Let N_{MHSI} denote the number of MHSI. Each spectral group $G_{n,j}$ in certain MHSI $\mathbf{Z}_n$ is coupled with another spectral group, thus generating the spectral pairs $P_{n,k}$, where k and n refer to the k^{th} spectral pair in the n^{th} MHSI. Each pair consists of an anchor and a positive sample. The other spectral groups in the MHSI $\mathbf{Z}_n$ are regarded as spectral negative samples. Spectral groups in other MHSIs (except $\mathbf{Z}_n$) with the same bands of the anchor are classified as spatial negative samples. According to the set theory, the description mentioned above is formulated as:

$$\begin{aligned} &P_{n,k} = \{G_{n,2k-1}, G_{n,2k}\} \quad (1 \leq k \leq \frac{B}{6}); A = G_{n,2k}; S_p = G_{n,2k-1}; \\ &S_{nspa} = \cup_{i=1,i\neq n}^{N_{MHSI}} G_{i,2k}; S_{nspe} = \cup_{i=1,i\neq k}^{\frac{B}{6}} P_{n,i} = \cup_{i=1,i\neq\{2k-1,2k\}}^{\frac{B}{3}} G_{n,i} \end{aligned} \tag{2}$$

Fig. 3 (b) illustrates the generation process of four features v_a, v_p, v_{nspe}^i, v_{nspa}^k, and the quadruplet loss $\mathcal{L}_Q$ is proposed as follows:

$$\begin{aligned} \mathcal{L}_Q = &\frac{1}{\frac{B}{3}-2} \cdot \sum_{i=1}^{\frac{B}{3}-2} \max\{d(v_a, v_p) - d(v_a, v_{nspe}^i) + \alpha, 0\} \\ &+ \frac{1}{N_{MHSI}-1} \cdot \sum_{k=1}^{N_{MHSl}} \max\{d(v_a, v_p) - d(v_a, v_{nspa}^k) + \beta, 0\} \end{aligned} \tag{3}$$

where $d(\cdot,\cdot)$ is Euclidean distance between two samples, $d(x, y) = \|x-y\|_2$. α and β are two margin parameters that α is greater than β in terms of the stronger repulsion in spectral dimension and weaker repulsion in spatial dimension.

3 Experimental Results

3.1 Experimental Setup

We use the public **M**ulti-**D**imensional **C**holedoch (MDC) Dataset [21] with 538 scenes, and the private **G**astric **P**oorly-**C**ohesive **C**arcinoma (GPCC) Dataset with 600 scenes. Both are high-quality annotated for binary MHSI segmentation tasks. The MDC dataset encompasses 60 spectral bands, whereas the GPCC dataset contains 40 spectral bands per scene. Each scene's single band image is resized to 256 × 256 spatial resolution. Consistent with [17], datasets are divided into training, validation, and test sets following a patient-centric hard split method, maintaining a ratio of 3:1:1. This strategy ensures that data from the same patient are exclusively assigned to one of the three subsets, preventing any overlap of patient data across different sets.

Initialized with ImageNet-1K [12] pre-trained weights, QSQL adopts data augmentation techniques (rotation, flipping, and brightness/contrast adjustments). We employ the cosine learning rate decay scheduler with a peak rate of 0.03 and a 10-epoch warm-up for pre-training, capping at 300 epochs. For fine-tuning in semantic segmentation tasks, a AdamW [5] optimizer and a cosine decay scheduler are used, starting at a learning rate of 4×10^{-4} for 100 epochs. The loss function in fine-tuning combines dice loss and cross-entropy loss. Experiments are conducted using PyTorch 1.11.0 on an NVIDIA GeForce RTX 4090 GPU.

Table 1. Performance comparison with SOTA methods on MDC dataset and GPCC dataset. The best-performing configurations are **highlighted** for clarity.

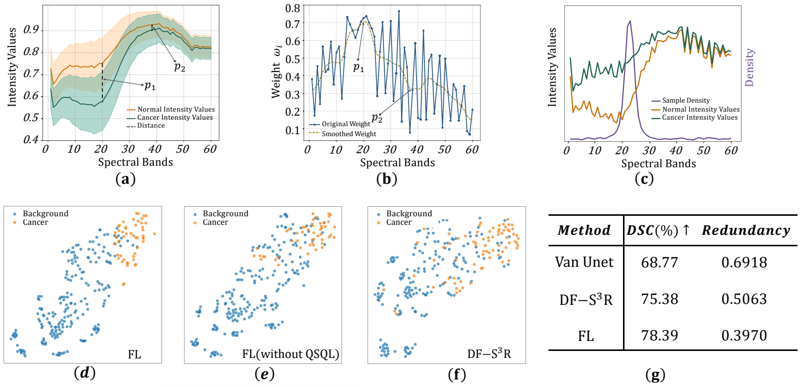
Method	MDC			GPCC		
	DSC(%)↑	IoU(%)↑	HD(px)↓	DSC(%)↑	IoU(%)↑	HD(px)↓
HyperNet [14]	72.31	59.54	75.36	69.52	54.83	92.99
nnUNet [4]	74.06	61.33	71.69	71.87	55.32	86.37
FSS [19]	74.62	61.61	68.67	71.45	54.99	87.48
Spec-Tr [18]	73.06	60.36	67.96	71.60	55.01	86.84
Swin-UNETR [13]	72.04	58.69	72.80	70.93	54.93	89.66
3DUNet [2]	72.48	58.92	73.99	70.76	53.79	92.26
FL(without QSQL)(ours)	76.80	64.24	66.40	74.27	58.18	83.92
DMVL [8]	71.97	59.49	72.75	70.36	54.37	90.07
SimSiam [1]	73.42	60.78	68.61	71.34	55.48	85.89
BYOL [3]	72.83	60.18	68.54	71.73	55.64	86.70
DF-S^{3}R [16]	75.38	63.39	69.09	72.29	56.52	86.13
FL(ours)	**78.39**	**65.78**	**63.01**	**75.76**	**60.44**	**80.92**

3.2 Evaluation of the FL

Comparisons with State-of-the-art MHSI Segmentation Methods Table 1 shows that our models, FL (without QSQL) and FL, outperform current state-of-the-art approaches on the MDC and GPCC datasets compared to four main competitors: (1) popular architectures for medical image segmentation (nnUNet [4], 3D-UNet [2] and SwinUNETR [13]), (2) architectures for HSI segmentation (Spec-Tr [18], FSS [19] and HyperNet [14]), (3) architectures for natural image segmentation with pre-training (DMVL [8], SimSiam [1] and BYOL [3]), (4) architectures for HSI segmentation with pre-training (DF-S^{3}R [16]). Results show that our architecture secures a notable 3.5% improvement in Dice coefficient metrics over the second best in the GPCC dataset. Even FL (without QSQL) maintains a leading position in metrics across both datasets. The segmentation visualizations are shown in the supplementary material.

We further explore the effectiveness of FL by insightful interpretation of the proposed method with t-SNE visualization (Fig. 4 (d)-(f)). Compared to the self-supervised methods DF-S^{3}R [16], FL significantly increases the inter-class distance and reduces the intra-class distance, due to the ability of diverging subtle spectral differences and spatial disparities provided by Quadruplet learning.

Ablation Study We conduct an ablation study to elucidate the efficacy of each component integrated into our architecture. The learned weights at the disease level (Fig. 4 (b)) explain how SFF module function. The significant value differences of focal coefficients for adjacent bands indicate that our SFF can reduce the bands redundancy. After applying a Gaussian filter to focal coefficients, the golden dashed line reveals a clear correlation between the weight distribution trend and the distance between foreground and background Intensity Values in Fig. 4 (a). This correlation highlights the strength of SFF, *i.e.*, spectral bands crucial for segmentation are assigned with higher value to emphasize their roles for segmentation, leading to 1.21% improvement in DSC (Table 2). Table 2 shows



Method	***DSC***(%) ↑	***Redundancy***
Van Unet	68.77	0.6918
DF−S^3R	75.38	0.5063
FL	78.39	0.3970

Fig. 4. (a): Spectral intensity in MDC dataset. (b): Spectral focal coefficients distribution. (c): Sampled points distribution in the Indicative Spatiospectral Transformer module of one image in MDC dataset and its corresponding spectral intensity plot. (d)-(f): t-SNE visualization of the first layer features from the Decoder across different methods. (g): Feature redundancy in MDC dataset.

Table 2. Ablation study on MDC dataset using RegNet34. CrossAtt1 and CrossAtt2 are Deformable Cross Attention mechanisms for spectral and spatial streams, respectively. "Van" represents vanilla self-attention and cross-attention operations. "SFF" and "BSE" refer to **S**pectral **F**ocus **F**orge module and **B**i-**s**cale **E**xtractor. QSQL indicates the use of self-supervised pre-training.

SpeAttn	CrossAttn		BSE	SFF	QSQL	Metrics		
	1	2				DSC(%)↑	IoU(%)↑	HD(px)↓
✗	✗	✗	✗	✗	✗	72.82	59.77	72.69
✓	✗	✗	✗	✗	✗	73.42	60.60	71.94
✓	✓	✗	✗	✗	✗	74.63	61.79	70.70
✗	✓	✓	✗	✗	✗	73.94	60.46	70.65
✓	✓	✓	✗	✗	✗	75.38	62.07	68.82
Van	*Van*	*Van*	✗	✗	✗	73.28	61.17	72.72
✓	✓	✓	✓	✗	✗	75.59	62.15	67.59
✓	✓	✓	✓	✓	✗	76.80	64.24	66.40
✓	✓	✓	✓	✓	✓	**78.39**	**65.78**	**63.01**

applying Indicative Spatiospectral Transformer (DCA_1 and DSA+DCA_2) yields a 2.56% improvement in DSC. We show the sampled points by DCA_2 in Fig. 4 (c). The band where the most sampled points gather is in immediate vicinity to the most discriminative bands with the vast intensity value gap from cancer to normal tissues. With the initialized weights provided by our QSQL, the DSC is further boosted from 76.80% to 78.39%.

FL Reduces Feature Redundancy Since high feature redundancy limits the generalization of neural networks [20], we demonstrate that FL effectively reduces the redundancy of high-level features. Following [15], we calculated the Pearson correlation coefficient among features in the feature pyramid's last layer to assess feature redundancy. As indicated in Fig. 4 (g), FL exhibits lower feature redundancy compared to DF-S^{3}R [16] and the vanilla UNet [11] baseline.

4 Conclusion

In this paper, a multi-stage multi-granularity Focus-tuned Learning paradigm for MHSI segmentation is proposed with intensifying and weakening strategy. The pre-training stage, adopting self-supervised spatiospectral quadruplet learning, can well initialize the downstream with a appreciable balance in spatiospectral feature learning. The fine-tuning stage can better focus on both disease-level and image-level subtle spectral differences and balance the spatiospectral feature learning. Experiments show our method achieves much better segmentation performance than other state-of-the-arts with over 3.5% improvement in DSC.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant No. 62101191), Shanghai Natural Science Foundation (Grant No. 21ZR1420800), and the Science and Technology Commission of Shanghai Municipality (Grant No. 22DZ2229004).

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

References

1. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15750–15758 (2021)
2. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19. pp. 424–432. Springer (2016)
3. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)
4. 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)
5. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
6. Li, Q., Wang, Q., Li, X.: Mixed 2d/3d convolutional network for hyperspectral image super-resolution. Remote sensing **12**(10), 1660 (2020)
7. Li, Q., Wang, Q., Li, X.: Exploring the relationship between 2d/3d convolution for hyperspectral image super-resolution. IEEE Transactions on Geoscience and Remote Sensing **59**(10), 8693–8703 (2021)

8. Liu, B., Yu, A., Yu, X., Wang, R., Gao, K., Guo, W.: Deep multiview learning for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **59**(9), 7758–7772 (2020)
9. Lu, G., Little, J.V., Wang, X., Zhang, H., Patel, M.R., Griffith, C.C., El-Deiry, M.W., Chen, A.Y., Fei, B.: Detection of head and neck cancer in surgical specimens using quantitative hyperspectral imaging. Clinical Cancer Research **23**(18), 5426–5436 (2017)
10. Myasnikov, E.: Embedding spatial context into spectral angle based nonlinear mapping for hyperspectral image analysis. In: International Conference on Computer Vision and Graphics. pp. 263–274. Springer (2018)
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
12. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**, 211–252 (2015)
13. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20730–20740 (2022)
14. Wang, Q., Sun, L., Wang, Y., Zhou, M., Hu, M., Chen, J., Wen, Y., Li, Q.: Identification of melanoma from hyperspectral pathology image using 3d convolutional networks. IEEE Transactions on Medical Imaging **40**(1), 218–227 (2020)
15. Wang, Y., Tang, S., Zhu, F., Bai, L., Zhao, R., Qi, D., Ouyang, W.: Revisiting the transferability of supervised pretraining: an mlp perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9183–9193 (2022)
16. Xie, X., Jin, T., Yun, B., Li, Q., Wang, Y.: Exploring hyperspectral histopathology image segmentation from a deformable perspective. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 242–251 (2023)
17. Xie, X., Wang, Y., Li, Q.: S^3r: Self-supervised spectral regression for hyperspectral histopathology image classification. In: Proc. MICCAI (2022)
18. Yun, B., Lei, B., Chen, J., Wang, H., Qiu, S., Shen, W., Li, Q., Wang, Y.: Spectr: Spectral transformer for microscopic hyperspectral pathology image segmentation. IEEE Transactions on Circuits and Systems for Video Technology (2023)
19. 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)
20. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: International Conference on Machine Learning. pp. 12310–12320. PMLR (2021)
21. Zhang, Q., Li, Q., Yu, G., Sun, L., Zhou, M., Chu, J.: A multidimensional choledoch database and benchmarks for cholangiocarcinoma diagnosis. IEEE access **7**, 149414–149421 (2019)
22. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)