Factor Space and Spectrum for Medical Hyperspectral Image Segmentation

Boxiang Yun[1], Qingli Li[1], Lubov Mitrofanova[2], Chunhua Zhou[3], and Yan Wang[1(✉)]

[1] Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai, China
52265904012@stu.ecnu.edu.cn, qlli@cs.ecnu.edu.cn, ywang@cee.ecnu.edu.cn
[2] Almazov National Medical Research Centre, St. Petersburg, Russia
[3] Rui Jin Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China

Abstract. Medical Hyperspectral Imaging (MHSI) brings opportunities for computational pathology and precision medicine. Since MHSI is a 3D hypercube, building a 3D segmentation network is the most intuitive way for MHSI segmentation. But, high spatiospectral dimensions make it difficult to perform efficient and effective segmentation. In this study, in light of information correlation in MHSIs, we present a computationally efficient, plug-and-play space and spectrum factorization strategy based on 2D architectures. Drawing inspiration from the low-rank prior of MHSIs, we propose spectral matrix decomposition and low-rank decomposition modules for removing redundant spatiospectral information. By plugging our dual-stream strategy into 2D backbones, we can achieve state-of-the-art MHSI segmentation performances with 3–13 times faster compared with existing 3D networks in terms of inference speed. Experiments show our strategy leads to remarkable performance gains in different 2D architectures, reporting an improvement up to 7.7% compared with its 2D counterpart in terms of DSC on a public Multi-Dimensional Choledoch dataset. Code is publicly available at https://github.com/boxiangyun/Dual-Stream-MHSI.

Keywords: Medical hyperspectral images · MHSI segmentation

1 Introduction

Medical Hyperspectral Imaging (MHSI) is an emerging imaging modality which acquires two-dimensional medical images across a wide range of electromagnetic spectrum. It brings opportunities for disease diagnosis, and computational pathology [16]. Typically, an MHSI is presented as a hypercube, with hundreds of narrow and contiguous spectral bands in spectral dimension, and thousands of pixels in spatial dimension (Fig. 1(a)).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43901-8_15.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14223, pp. 152–162, 2023.
https://doi.org/10.1007/978-3-031-43901-8_15

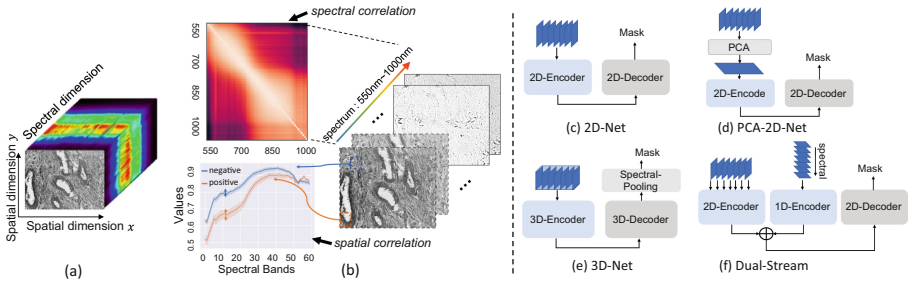


Fig. 1. (a): An example of MHSI. (b): Illustration of two types of correlation in MHSI. (c)–(f): HSI classification and segmentation backbones. 2D decoder can be changed into classification head for classification tasks.

Due to the success of 2-Dimensional (2D) deep neural network in natural images, the simplest way to classify/segment an MHSI is to treat its two spatial dimensions as input spatial dimension, and treat its spectral dimension as input channel dimension [25] (Fig. 1(c)). Dimensionality reduction [12] and recurrent approaches [1] are usually adopted to aggregate spectral information before feeding the HSI into 2D networks (Fig. 1(d)). These methods are not suitable for high spatial resolution MHSI, and they may bring noises in spatial features while reducing spectral dimension. The 2D networks are computationally efficient, usually much faster than 3D networks. But, they mix spectral information after the first convolutional layer, making the interband correlations of MHSIs underutilized. Building a 3D network usually suffers from high computational complexity, but it is the most straightforward way to learn interpixel and interband correlations of MHSIs [23] (Fig. 1(e)). Since spatiospectral orientations are not equally likely, there is no need to treat space and spectrum symmetrically, as is implicit in 3D networks. We might instead design a dual-stream strategy to "factor" the architecture. A few HSI classification backbones try to design dual-stream architectures that treat spatial structures and spectral intensities separately [2, 20, 30] (Fig. 1(f)). But, these methods simply adopt convolutional or MLP layers to extract spectral features. SpecTr [28], learning spectral and spatial features alternatively, utilizes Transformer to capture the global spectral feature. They overlook the low rankness in the spectral domain, which contains discriminative information for differentiating targets from the background.

High spatiospectral dimensions make it difficult to perform a thorough analysis of MHSI. In MHSIs, there exist two types of correlation. One is a spectral correlation in adjacent pixels. As shown in Fig. 1(b), the intensity values vs. spectral bands for the local positive (cancer) area and negative (normal) area are highly correlated. The other is spatial correlation between adjacent bands. Figure 1(b) plots the spatial similarity among all bands, and shows large cosine similarity scores among nearby bands (error band of line chart in the light color area) and small scores between bands in a long distance. The correlation implies spectral redundancy when representing spatial features, and spatial redundancy

when learning spectral features. The low-rank structure in MHSIs holds significant discriminatory and characterizing information [11]. Exploring MHSI's low-rank prior can promote the segmentation performance.

In this paper, we consider treating spatiospectral dimensions separately and propose an effective and efficient dual-stream strategy to "factor" the architecture, by exploiting the correlation information of MHSIs. Our dual-stream strategy is designed based on 2D CNNs with U-shaped [16] architecture. For the spatial feature extraction stream, inspired from spatial redundancy between adjacent bands, we group adjacent bands into a spectral agent. Different spectral agents are fed into a 2D CNN backbone as a batch. For the spectral feature extraction stream, inspired by the low-rank prior on the spectral space, we propose a matrix factorization-based method to capture global spectral information. To remove the redundancy in the spatiospectral features and promote the capability of representing the low-rank prior of MHSI, we further design Low-rank Decomposition modules, and employ the Canonical-Polyadic decomposition method [9,32]. Our space and spectrum factorization strategy is plug-and-play. The effectiveness of the proposed strategy is compared and verified by plugging in different 2D architectures. We also show that with our proposed strategy, U-Net model using ResNet-34 can achieve state-of-the-art MHSI segmentation with 3–13 faster than other 3D architectures.

2 Methodology

Mathematically, let $\mathbf{Z} \in \mathbb{R}^{S\times H\times W}$ denote the 3D volume of a pathology MHSI, where $H \times W$ is the spatial resolution, and S is the number of spectral bands. The goal of MHSI segmentation is to predict the per-pixel annotation mask $\hat{\mathbf{Y}} \in \{0,1\}^{H\times W}$. Our training set is $\mathcal{D} = \{(\mathbf{Z}_i, \mathbf{Y}_i)\}_{i=1}^{N}$, where $\mathbf{Y}_i$ denotes the per-pixel groundtruth for MHSI $\mathbf{Z}_i$.

The overall architecture of our proposed method is shown in Fig. 2, where the 2D CNN in the figure is a proxy which may represent all widely-used 2D architectures. It represents a spatial stream, which focuses on extracting spatial features from spectral agents (Sect. 2.1). The lightweight spectral stream learns multi-granular spectral features, and it consists of three key modules: Depthwise Convolution (DwConv), Spectral Matrix Decomposition (SMD) and Feed Forward Network, where SMD module effectively leverages low-rank prior from spectral features (Sect. 2.1). Besides, the Low-rank Decomposition module (LD) represents high-level low-rank spatiospectral features (Sect. 2.2). The input MHSI $\mathbf{Z}$ is decomposed into a spatial input $\mathbf{Z}_{spa} \in \mathbb{R}^{G\times(S/G)\times H\times W}$ and a spectral input $\mathbf{Z}_{spe} \in \mathbb{R}^{S\times C_0^{spe}\times H\times W}$, where G indicates evenly dividing spectral bands into G groups, *i.e.*, spectral agents. S/G and $C_0^{spe} = 1$ are the input feature dimensions for two streams respectively.

2.1 Dual-Stream Architecture with SpatioSpectral Representation

As mentioned above, for the spatial stream, we first reshape MHSI $\mathbf{Z} \in \mathbb{R}^{S\times H\times W}$ into $\mathbf{Z}_{spa} \in \mathbb{R}^{G\times(S/G)\times H\times W}$, which has G spectral agents. Each spectral agent is

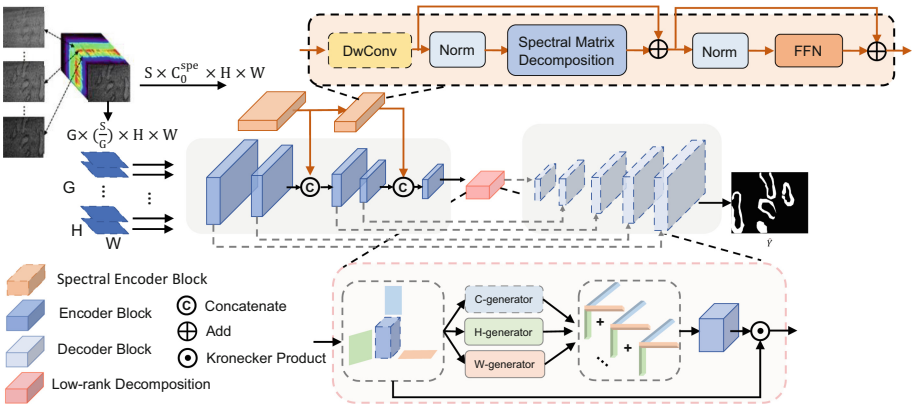


Fig. 2. The proposed dual-stream architecture. An MHSI goes through a spectral stream with the proposed spectral encoder block which consists of three key modules, *i.e*, DwConv, Spectral Matrix Decomposition (SMD) and Feed Forward Network (FFN), and it goes through a spatial stream after dividing into G spectral agents. Low-rank Decomposition (LD) is designed to characterize the low-rank prior.

treated as one sample. One sample contains highly correlated spectral bands, so that the spatial stream can focus on spatial feature extraction. For the spectral stream, to deal with problems of spatiospectral redundancy and the inefficiency of global spectral feature representation, we propose a novel and concise hierarchical structure shown in Fig. 2. We employ a basic transformer paradigm [21] but design it tailored for capturing global low-rank spectral features. Our spectral encoder block can be formulated by:

$$\mathbf{X} = DwConv(\mathbf{Z}_{in}),\ \mathbf{X}' = SMD(Norm(\mathbf{X})) + \mathbf{X}, \mathbf{Z}_{out} = FFN(Norm(\mathbf{X}')) + \mathbf{X}', \tag{1}$$

where $\mathbf{Z}_{in} \in \mathbb{R}^{S \times C_{spe} \times H \times W}$ indicates the input spectral token tensor, and C_{spe} is the spectral feature dimension. We introduce Depth-wise Conv (DwConv) for dynamically integrating redundant spatial information into spectral features to reduce spatial redundant noises, achieved by setting different strides of the convolutional kernel. Then, we represent long-distance dependencies among spectral inter-bands as a low-rank completion problem. $SMD(\cdot)$ indicates the spectral matrix decomposition operation. Concretely, we flatten feature map $\mathbf{X}$ to spectral sequence tokens $\mathbf{X}_{spe} \in \mathbb{R}^{H \cdot W \times S \times C_{spe}}$, which has S spectral tokens. We map $\mathbf{X}_{spe}$ to a feature space using a linear transform $\mathbf{W}_l \in \mathbb{R}^{C_{spe} \times C'_{spe}}$. We then apply a matrix decomposition method NMF (Non-negative Matrix Factorization) [10], denoted by $\mathcal{M}(\cdot)$, to identify and solve for a low-rank signal subspace and use iterative optimization algorithms backpropagate gradients [4]: $SMD(\boldsymbol{X}_{spe}) = \mathcal{M}(\boldsymbol{W}_l \boldsymbol{X}_{spe})$. Finally, to enhance the individual component of spectral tokens, we utilize a Feedforward Neural Network (FFN) in Transformer consisting of two linear layers and an activation layer.

In the framework shown in Fig. 2, spectral information is integrated from channel dimensions by performing concatenation, after the second and fourth encoder blocks, to aggregate the spatiospectral features. The reason for this design is that spectral features are simpler and lack hierarchical structures compared to spatial features, we will discuss more in the experimental section.

2.2 Low-Rank Decomposition and Skip Connection Ensemble

The MHSI has low-rank priority due to redundancy, so we propose a low-rank decomposition module using Canonical-Polyadic (CP) decomposition [9] to set constraints on the latent representation. For a three-order tensor $\boldsymbol{U} \in \mathbb{R}^{C' \times H' \times W'}$, where $H' \times W'$ is the spatial resolution and C' is the channel number. It can be decomposed into a linear combination of N rank-1 tensors. The mathematical formulation of CP decomposition can be expressed as $\boldsymbol{U} = \sum_{i=1}^{r} \lambda_i \boldsymbol{a}_{ci} \otimes \boldsymbol{a}_{hi} \otimes \boldsymbol{a}_{wi}$. Where $\otimes$ denote Kronecker Product, $\boldsymbol{a}_{ci} \in \mathbb{R}^{C' \times 1 \times 1}$, $\boldsymbol{a}_{hi} \in \mathbb{R}^{C' \times 1 \times 1}$, $\boldsymbol{a}_{wi} \in \mathbb{R}^{C' \times 1 \times 1}$, r is the tensor rank and λ_i is a scaling factor. Recent research [3,32] has proposed new methods based on DNNs to address this problem of representing MHSIs as low-rank tensors. As shown in Fig. 2, rank-1 generators are used to create rank-1 tensors in different directions, which are then aggregated by Kronecker Product to synthesize a sub-attention map $\boldsymbol{A}_1$. The residual part between the input of features and the generated rank-1 tensor is used to generate second rank-1 tensors $\boldsymbol{A}_2$. It can obtain r rank-1 tensors by repeating r times. Mathematically, this process can be expressed as:

$$\begin{aligned}
\boldsymbol{A}_1 &= \mathcal{G}_c(\boldsymbol{U}) \otimes \mathcal{G}_h(\boldsymbol{U}) \otimes \mathcal{G}_w(\boldsymbol{U}), \\
\boldsymbol{A}_2 &= \mathcal{G}_c(\boldsymbol{U} - \boldsymbol{A}_1) \otimes \mathcal{G}_h(\boldsymbol{U} - \boldsymbol{A}_1) \otimes \mathcal{G}_w(\boldsymbol{U} - \boldsymbol{A}_1), \\
\boldsymbol{A}_r &= \mathcal{G}_c(\boldsymbol{U} - \sum_{i=1}^{r-1} \boldsymbol{A}_i) \otimes \mathcal{G}_h(\boldsymbol{U} - \sum_{i=1}^{r-1} \boldsymbol{A}_i) \otimes \mathcal{G}_w(\boldsymbol{U} - \sum_{i=1}^{r-1} \boldsymbol{A}_i),
\end{aligned} \tag{2}$$

where $\mathcal{G}_c(\cdot)$, $\mathcal{G}_h(\cdot)$ and $\mathcal{G}_w(\cdot)$ are the channel, height and width generators. Finally, we aggregate all rank-1 tensors (from $\boldsymbol{A}_1$ to $\boldsymbol{A}_r$) into the attention map along the channel dimension, followed by a linear layer used to reduce the feature dimension to obtain the low-rank feature $\boldsymbol{U}_{low}$:

$$\boldsymbol{U}_{low} = \boldsymbol{U} \odot Linear(Concate(\boldsymbol{A}_1, \boldsymbol{A}_2, ..., \boldsymbol{A}_r)), \tag{3}$$

where $\odot$ is the element-wise product, and $\boldsymbol{U}_{low} \in \mathbb{R}^{C' \times H' \times W'}$. We employ a straightforward non-parametric ensemble approach for grouping spectral agents. This approach involves multiple agents combining their features by averaging the vote. The encoders in the spatial stream produce 2D feature maps with G spectral agents, defined as $\boldsymbol{F}_i \in \mathbb{R}^{G \times C_i \times H/2^i \times W/2^i}$ for the ith encoder, where G, C_i, $H/2^i$, and $W/2^i$ represent the spectral, channel, and two spatial dimensions, respectively. The ensemble is computed by $\boldsymbol{F}_i^{out} = Mean(\boldsymbol{F}_i^1, \boldsymbol{F}_i^2, ..., \boldsymbol{F}_i^G)$, where $\boldsymbol{F}_i^G \in \mathbb{R}^{C_i \times H/2^i \times W/2^i}$ represents the 2D feature map of the Gth agent. The ensemble operation aggregates spectral agents to produce a 2D feature map

Table 1. Ablation study (in "mean (std)") on MDC dataset using RegNetX40 [26] as the backbone. SA denote the spectral agent. L1 to L4 represent the locations where output spectral features from the spectral flow module are inserted into the spatial flow. Tr and Conv mean we replace the SMD module in the spectral stream with self-attention and convolutional blocks. Best results are **highlighted**.

SA	Spectral Stream				LD	IoU ↑	DSC ↑	HD ↓
	L1	L2	L3	L4				
×	×	×	×	×	×	48.94 (20.53)	63.11 (19.13)	92.04 (34.14)
✓	×	×	×	×	×	51.64 (18.43)	66.05 (17.15)	86.04 (32.06)
✓	×	✓	×	×	×	52.05 (18.92)	66.27 (17.88)	83.05 (31.07)
✓	×	×	×	✓	×	53.87 (19.08)	67.95 (16.88)	88.75 (31.49)
✓	×	✓	×	✓	×	55.81 (18.41)	69.73 (16.31)	88.13 (31.47)
✓	✓	✓	✓	✓	×	55.01 (15.54)	69.58 (14.10)	86.25 (35.40)
✓	×	✓	×	✓	✓	**56.90 (17.38)**	**70.88 (15.05)**	**82.72 (31.77)**
✓	×	Conv	×	Conv	✓	55.19 (18.58)	69.15 (16.63)	83.13 (31.08)
✓	×	Tr	×	Tr	✓	55.72 (17.16)	69.89 (15.26)	84.48 (32.13)

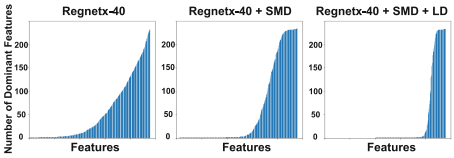
with enhanced information interactions learned from the multi-spectral agents. The feature maps obtained from the ensemble can be decoded using lightweight 2D decoders to generate segmentation masks.

3 Experimental Results

3.1 Experimental Setup

We conducted experiments on the public Multi-Dimensional Choledoch (MDC) Dataset [31] with 538 scenes and Hyperspectral Gastric Carcinoma (HGC) Dataset [33] (data provided by the author) with 414 scenes, both with high-quality labels for binary MHSI segmentation tasks. These MHSIs are collected by hyperspectral system with an objective lens of 20x, and wavelengths from 550 nm to 1000 nm for MDC and 450 nm to 750 nm for HGC, resulting in 60 and 40 spectral bands for each scene. The size of a single band image in MDC and HGC are both resized to 256×320. Following [23,27], we partition the datasets into training, validation, and test sets using a patient-centric hard split approach with a ratio of 3:1:1. Specifically, each patient's data is allocated entirely to one of the three sets, ensuring that the same patient's data do not appear in multiple sets.

We use data augmentation techniques such as rotation and flipping, and train with an Adam optimizer using a combination of dice loss and cross-entropy loss for 8 batch size and 100 epochs. The segmentation performance is evaluated using Dice-Sørensen coefficient (DSC), Intersection of Union (IoU), and Hausdorff Distance (HD) metrics, and Throughput (images/s) is reported for



Method	DSC	Redundancy
RegNetX-40	66.06	0.4089
RegNetX-40+SMD	69.73	0.3924
RegNetX-40+SMD+LD	70.88	0.3699

Fig. 3. Feature redundancy of three methods on MDC dataset. Left figures plot each feature embedding in ascending order of the number of times they are dominant in the population (y-axis) and feature dimension (x-axis) on the statistical results of the test set. Right table shows the influence of SMD and LD modules in reducing redundancy.

comparison. Pytorch framework and four NVIDIA GeForce RTX 3090 are used for implementation.

3.2 Evaluation of the Proposed Dual-Stream Strategy

Ablation Study. Our dual-stream strategy is plug-and-play. We first conduct an ablation study to show the effectiveness of each component. We use a dual-stream strategy with RegNetX40 and U-Net architecture. As shown in Table 1, Our ablation study shows that spectral agent strategy improves segmentation performance by more than 2.5% (63.11 vs. 66.05). If we utilize spectral information from the spectral stream to assist in the spatial stream, we find that inserting spectral information at L2 and L4 yields a significant improvement of 3.7% (69.73 vs. 66.05), while inserting at L4 alone also results in a significant increase of 1.9% in DSC (67.95 vs. 66.05). A slight improvement is observed when inserting at L2, possibly due to the coarse features of shallow spectral information. Inserting spectral information at all spatial layers (*i.e.*, L1 to L4) and only at L2 and L4 produce similar results, indicating that spectral features do not possess complex multilevel characteristics relative to spatial features. Therefore, we adopt a simple and efficient two-layer spectral flow design. Replacing the spectral stream with transformer layers results in a 0.96% (70.88 vs. 69.89) lower DSC, possibly because transformers are difficult to optimize on small datasets. Our proposed LD module is crucial, resulting in a 1.12% performance drop in terms of DSC without it.

It is known that high feature redundancy limits the generalization of neural networks [29]. Here we show our low-rank representation effectively reduces the redundancy of features. Our quantitative and qualitative analysis demonstrated that the proposed MDC and LD modules effectively reduces the redundancy of output features. Following [8], we define the dominant features for the feature embedding of i-th MHSI $\mathbf{h}_i \in \mathbb{R}^{C_d}$ as $L_i = j : h_{ij} > \mu + \sigma$, where μ is mean of $\mathbf{h}_i$ and σ is stand deviation of $\mathbf{h}_i$. As shown in the left part of Fig. 3, our designed modules effectively reduce the number of dominant features and maintain sparsity in the entire spatiospectral feature space. Inspired by [24], we evaluate the degree of feature redundancy by computing the Pearson correlation coefficient between different feature channels. As shown in the right part of Fig. 3, the

Table 2. Performance comparison in "mean (std)" in MDC and HCG dataset. The best results of each comparison are **highlighted**.

Backbone	Method	MDC		HGC	
		DSC ↑	HD ↓	DSC ↑	HD ↓
ResNet50	U-Net	72.10 (14.17)	82.62 (29.70)	81.72 (15.79)	77.02 (44.22)
+Ours	U-Net	**74.12 (13.76)**	**78.32 (29.80)**	**85.08 (13.40)**	**69.78 (43.41)**
Convnext	U-Net	71.29 (13.68)	84.28 (30.50)	76.21 (15.15)	87.18 (43.85)
+Ours	U-Net	**72.82 (15.39)**	**82.12 (30.97)**	**77.51 (15.59)**	**79.73 (45.40)**
Swin-Transformer	U-Net	70.61 (14.42)	85.77 (31.76)	73.89 (18.05)	70.99 (36.10)
+Ours	U-Net	**72.10 (13.60)**	**82.94 (30.67)**	**78.07 (15.76)**	**63.35 (35.49)**
Efficinet-b2	U-Net	62.54 (19.28)	87.34 (32.72)	70.60 (17.41)	94.28 (46.95)
+Ours	U-Net	**68.72 (15.16)**	**84.91 (33.62)**	**77.44 (14.40)**	**81.24 (40.36)**
RegNetX40	U-Net	63.11 (19.13)	92.04 (34.14)	74.32 (18.39)	88.54 (45.34)
+Ours	U-Net	**70.88 (15.05)**	**82.72 (31.77)**	**79.86 (15.18)**	**80.35 (43.16)**
ResNet50	FPN	71.23 (16.01)	83.17 (36.84)	79.98 (15.30)	68.52 (43.32)
+Ours	FPN	**73.01 (13.84)**	**78.37 (30.26)**	**81.88 (14.75)**	**68.60 (45.63)**

Table 3. Performance comparison with SOTA methods with Throughput(images/s) on MDC dataset and HGC dataset. The best results are **highlighted**.

Method		MDC			HGC		
		DSC ↑	HD ↓	Throughput ↑	DSC ↑	HD ↓	Throughput ↑
2D	PCA-UNet [22]	70.83	80.70	12.37	78.27	78.06	**16.49**
	CGRU-UNet [1]	73.14	82.49	7.50	80.68	74.23	7.62
	Ours (Resnet34)	**75.44**	77.70	**13.84**	**85.80**	**64.04**	14.82
3D	3D-UNet [6]	72.55	79.87	4.04	83.37	79.97	4.34
	nnUNet [7]	74.12	79.87	1.92	85.36	75.83	2.83
	HyperNet [23]	72.47	83.75	0.99	84.00	67.05	1.67
	Swin-UNETR [19]	72.39	78.38	1.45	78.81	80.15	2.31
	SpecTr [28]	73.66	**76.92**	1.40	84.74	66.90	2.44

SMD and LD modules can reduce feature redundancy and lead to an increase in segmentation performance.

Comparisons Between w/ and w/o Dual-Stream Strategy on Different Backbones. To show the effectiveness of our dual-stream strategy in improving MHSI segmentation performance in various architectures, we plug it into different segmentation methods, *i.e.*, U-Net [17], FPN [13], with different spatial branch backbones *i.e.*, ResNet [5], Convnext [15], RegNetX40 [26], EfficientNet [18] and Swin-Transformer [14]. Results are summarized in Table 2. The results obtained with the proposed dual-stream strategy can consistently boost the segmentation performance by a large margin.

Comparisons with State-of-the-Art MHSI Segmentation Methods. Table 3 shows comparisons on MDC and HGC datasets. We use a lightweight and efficient ResNet34 as the backbone of our dual-stream method. Experimental results show that 2D methods are generally faster than 3D methods in inference speed, but 3D methods have an advantage in segmentation performance (DSC & HD). However, our approach outperforms other methods in both inference speed and segmentation accuracy. It is also plug-and-play, with the potential to achieve better segmentation performance by selecting more powerful backbones. The complete table (including IoU and variance) and qualitative results are shown in the supplementary material.

4 Conclusion

In this paper, we present to factor space and spectrum for accurate and fast medical hyperspectral image segmentation. Our dual-stream strategy, leveraging low-rank prior of MHSIs, is computationally efficient and plug-and-play, which can be easily plugged into any 2D architecture. We evaluate our approach on two MHSI datasets. Experiments show significant performance improvements on different evaluation metrics, *e.g.*, with our proposed strategy, we can obtain over 7.7% improvement in DSC compared with its 2D counterpart. After plugging our strategy into ResNet-34 backbone, we can achieve state-of-the-art MHSI segmentation accuracy with 3–13 times faster in terms of inference speed than existing 3D networks.

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).

References

1. Bengs, M., et al.: Spectral-spatial recurrent-convolutional networks for *in-vivo* hyperspectral tumor type classification. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12263, pp. 690–699. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_66
2. Chen, R., Li, G.: Spectral-spatial feature fusion via dual-stream deep architecture for hyperspectral image classification. Infrared Phys. Technol. **119**, 103935 (2021)
3. Chen, W., et al.: Tensor low-rank reconstruction for semantic segmentation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12362, pp. 52–69. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58520-4_4
4. Geng, Z., Guo, M.H., Chen, H., Li, X., Wei, K., Lin, Z.: Is attention better than matrix decomposition? In: International Conference on Learning Representations (2021)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Computer Vision and Pattern Recognition (2016)

6. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
7. Isensee, F., et al.: nnU-Net: self-adapting framework for u-net-based medical image segmentation. Nat. Methods (2021)
8. Kalibhat, N.M., Narang, K., Firooz, H., Sanjabi, M., Feizi, S.: Towards better understanding of self-supervised representations. In: Workshop on Spurious Correlations, Invariance and Stability, ICML 2022 (2022)
9. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. SIAM Rev. **51**(3), 455–500 (2009)
10. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999)
11. Li, L., Li, W., Du, Q., Tao, R.: Low-rank and sparse decomposition with mixture of gaussian for hyperspectral anomaly detection. IEEE Trans. Cybern. **51**(9), 4363–4372 (2020)
12. Li, X., Li, W., Xu, X., Hu, W.: Cell classification using convolutional neural networks in medical hyperspectral imagery. In: International Conference on Image, Vision and Computing (2017)
13. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Computer Vision and Pattern Recognition (2017)
14. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. arXiv, Computer Vision and Pattern Recognition (2021)
15. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11966–11976 (2022). https://doi.org/10.1109/CVPR52688.2022.01167
16. Lu, G., Fei, B.: Medical hyperspectral imaging: a review. J. Biomed. Opt. **19**, 010901 (2014)
17. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
18. Tan, M., Le, Q.V.: EfficientNet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning (2019)
19. Tang, Y., et al.: 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)
20. Trajanovski, S., Shan, C., Weijtmans, P.J., de Koning, S.G.B., Ruers, T.J.: Tongue tumor detection in hyperspectral images using deep learning semantic segmentation. IEEE Trans. Biomed. Eng. **68**(4), 1330–1340 (2020)
21. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
22. Wang, J., et al.: PCA-U-Net based breast cancer nest segmentation from microarray hyperspectral images. Fundam. Res. **1**(5), 631–640 (2021)
23. Wang, Q., et al.: Identification of melanoma from hyperspectral pathology image using 3D convolutional networks. IEEE Trans. Med. Imaging **40**(1), 218–227 (2020)

24. Wang, Y., et al.: Revisiting the transferability of supervised pretraining: an MLP perspective. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9173–9183 (2022). https://doi.org/10.1109/CVPR52688.2022.00897
25. Wei, X., Li, W., Zhang, M., Li, Q.: Medical hyperspectral image classification based on end-to-end fusion deep neural network. IEEE Trans. Instrum. Measur. **68**, 4481–4492 (2019)
26. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: CVPR (2017)
27. Xie, X., Wang, Y., Li, Q.: S^3r: self-supervised spectral regression for hyperspectral histopathology image classification. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13432, pp. 46–55. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16434-7_5
28. Yun, B., Wang, Y., Chen, J., Wang, H., Shen, W., Li, Q.: Spectr: spectral transformer for hyperspectral pathology image segmentation. arXiv, Image and Video Processing (2021)
29. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: self-supervised learning via redundancy reduction. In: ICML (2021)
30. Zhang, H., Li, Y., Zhang, Y., Shen, Q.: Spectral-spatial classification of hyperspectral imagery using a dual-channel convolutional neural network. Remote Sens. Lett. **8**, 438–447 (2017)
31. 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)
32. Zhang, S., Wang, L., Zhang, L., Huang, H.: Learning tensor low-rank prior for hyperspectral image reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12006–12015 (2021)
33. Zhang, Y., Wang, Y., Zhang, B., Li, Q.: A hyperspectral dataset of precancerous lesions in gastric cancer and benchmarks for pathological diagnosis. J. Biophotonics e202200163 (2022)