MASK3DUNET: MULTISCALE 3-D/2-D CONTEXT EXCHANGE VIA SPATIAL-PRIOR ATTENTION FOR HYPERSPECTRAL CLASSIFICATION

Chao Wang[1], *Muhammad Hassaan Farooq Butt*[2], *Daisuke Miyazaki*[1]

[1]Department of Intelligent Systems, Hiroshima City University, Japan
[2]Faculty of Environmental Sciences and Natural Resource Management, Norwegian University of Life Sciences (NMBU), Norway

ABSTRACT

Patch-based hyperspectral classification benefits from local 3-D spectral–spatial modeling, yet also requires content-dependent, non-uniform aggregation across the patch. Stacked 3-D convolutions expand context only indirectly, whereas voxel-wise attention over DHW tokens is unnecessarily expensive and couples spatial reasoning with latent-depth aggregation. We propose Mask3DUNet, a classification-oriented 3-D U-Net that introduces a multiscale Depth-Collapsed Spatial-Prior Bridge (DSPB). At each encoder and decoder resolution, DSPB converts a 3-D feature volume into HW spatial tokens, learns a soft key prior, performs patch-wide self-attention, and broadcasts the refined context back to the volumetric stream. The attention pair count is thereby reduced by a factor of D^2 relative to voxel tokenization, while the local 3-D pathway is preserved. A sample-adaptive latent-depth pooling head produces the final reference-pixel prediction. On Indian Pines and two WHU-Hi crop scenes, Mask3DUNet achieves the highest mean OA, AA, and κ among the evaluated convolutional and U-shaped baselines.

***Index Terms*—** hyperspectral image classification, U-shaped architectures, spatial-prior attention, spectral–spatial learning

1. INTRODUCTION

Hyperspectral imagery records densely sampled spectra together with spatial organization, enabling fine-grained recognition of materials and land-cover classes that are difficult to distinguish in conventional imagery [1, 2, 3, 4]. A useful classifier must therefore preserve subtle spectral variation, exploit local spatial continuity, and reason over the wider neighborhood surrounding a labeled pixel.

Classical Convolutional models provide a strong local inductive bias for this task. One-dimensional CNNs operate on per-pixel spectra [5], 3-D CNNs jointly encode spectral and spatial neighborhoods with volumetric kernels [6]. These designs are effective at local pattern extraction, but communication between distant positions remains mediated by stacked kernels. Enlarging the receptive field does not create direct, content-dependent interaction across the patch.

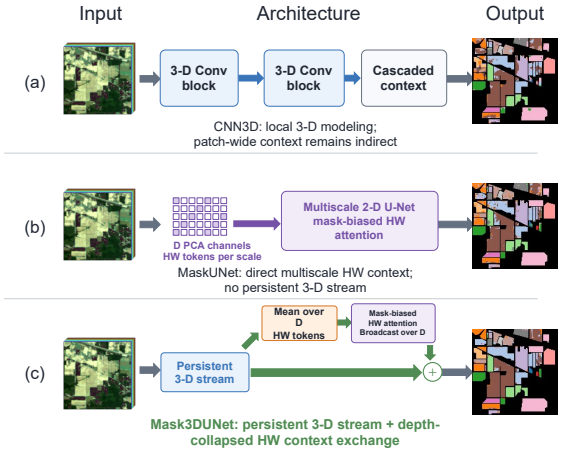


Fig. 1: Architectural comparison under the common PCA-30 patch-classification protocol. (a) CNN3D propagates spatial context through cascaded local 3-D operators. (b) MaskUNet treats the D PCA components as Conv2D channels and applies mask-biased attention over HW tokens throughout a multiscale 2-D U-Net, but retains no explicit latent-depth feature axis. (c) Mask3DUNet preserves a multiscale 3-D stream and uses depth-collapsed DSPB branches for direct HW interaction and residual context return.

Self-attention addresses this limitation by explicitly relating tokens [7], and transformer-based HSI classifiers have demonstrated the value of nonlocal spectral–spatial modeling [8, 9, 10]. A direct flattening of a latent cube, however, produces DHW tokens and quadratic interaction over the full three-dimensional lattice. It also entangles two different operations: relating spatial positions and aggregating the transformed depth axis. Collapsing depth only once, as in a conventional hybrid network, avoids voxel attention but does not exchange context throughout a multiscale hierarchy. The design problem is therefore: *how can patch-wide spatial interaction be introduced at multiple resolutions while retaining*

This work was supported in part by JSPS KAKENHI under Grant Number JP26K21254 and JP25K03142.

local 3-D features and avoiding attention over the full DHW lattice?

U-shaped architectures are a natural substrate for this objective because their encoder–decoder hierarchy combines coarse context with fine-scale detail [11, 12, 13, 14]. Our patch-classification MaskUNet comparator instantiates a multiscale 2-D alternative: the D PCA components are used as Conv2D input channels, and mask-biased self-attention operates on HW spatial tokens after every encoder and decoder scale. It therefore provides direct multiscale spatial interaction, but does not retain an explicit latent-depth feature axis or a persistent 3-D stream. Existing mask-conditioned attention is mainly designed for dense 2-D segmentation, where a learned mask has foreground or region semantics [15]. Patch-level HSI classification instead predicts one class vector at a reference location, and PCA-transformed depth is not a set of directly interpretable physical bands. We therefore use a learned field as a *soft spatial prior* that guides contextual retrieval rather than as a foreground mask or hard selector.

To this end, we introduce Mask3DUNet. Its central component, the Depth-Collapsed Spatial-Prior Bridge (DSPB), creates a two-way interface between a multiscale 3-D feature hierarchy and a compact 2-D token space. The bridge temporarily collapses latent depth, performs spatial-prior attention over HW tokens, and returns the refined context to the 3-D stream. A separate attention-pooling head then aggregates latent depth in a sample-dependent manner.

The contributions are threefold. First, we formulate a classification-oriented 3-D U-shaped architecture that combines local spectral–spatial processing with direct patch-wide interaction at multiple resolutions. Second, we develop DSPB, a 3-D/2-D context-exchange module whose attention operates on HW rather than DHW tokens, together with sample-adaptive latent-depth pooling. Third, we evaluate the complete model on three HSI scenes and show consistent gains in OA, AA, and κ, with class-resolved improvements on the more challenging Indian Pines scene.

2. PROPOSED METHOD

2.1. Architecture and Design Rationale

For a labeled location, let $\mathbf{X} \in \mathbb{R}^{1\times D\times H\times W}$ denote an aligned patch and y its class. As shown in Fig. 2, Mask3DUNet contains three functional parts: a 3-D encoder–decoder for local feature extraction, multiscale DSPB modules for nonlocal spatial exchange, and a latent-depth pooling head for patch-level prediction.

The encoder uses channel widths $16, 32, 64$, followed by a 128-channel bottleneck; the decoder mirrors this schedule. Double 3×3×3 convolution blocks use group normalization and GELU. DSPB modules use four attention heads, embedding dimension C_ℓ, head dimension $d_h = C_\ell/4$, and an MLP expansion ratio of 2. Decoder output padding is zero: $1\to2\to3$ and $6\to12\to13$ use right/bottom zero padding, whereas $3\to6$ is exact. Spatial-only pooling and transposed convolution change $H\times W$ from 13^2 to 6^2, 3^2, and 1^2, while the latent depth remains $D = 30$. The volumetric stream therefore keeps the variance-ordered PCA coordinates as an explicit transformed-component axis rather than folding them into 2-D input channels; shared 3-D kernels learn joint interactions across latent-component and spatial coordinates. Skip concatenation restores fine-scale features, and the network outputs one class vector rather than a dense segmentation map.

2.2. Depth-Collapsed Spatial-Prior Bridge

At scale ℓ, let $\mathbf{F}^\ell \in \mathbb{R}^{B\times C_\ell\times D_\ell\times H_\ell\times W_\ell}$ and $N_\ell = H_\ell W_\ell$. The bridge first forms a spatial descriptor

$$\mathbf{Z}^\ell = \frac{1}{D_\ell}\sum_{d=1}^{D_\ell}\mathbf{F}^\ell_{:,:,d,:,:} \in \mathbb{R}^{B\times C_\ell\times H_\ell\times W_\ell}, \tag{1}$$

which is flattened into N_ℓ tokens. A lightweight 1×1 Conv–GELU–Conv–sigmoid predictor yields a scalar prior $m_r \in (0,1)$ for each spatial position r. For attention head a and spatial query–key pair (i,j),

$$L^a_{ij} = \frac{(\mathbf{q}^a_i)^\top\mathbf{k}^a_j}{\sqrt{d_h}} + \log(m_i m_j + \epsilon), \qquad A^a_{ij} = \mathrm{softmax}_j\, L^a_{ij}, \tag{2}$$

where $\epsilon = 10^{-6}$. The same continuous prior field is shared across attention heads. For a fixed query i, when $m_i m_j \gg \epsilon$ across its attention row, Eq. (2) gives $A^a_{ij} \propto \exp((\mathbf{q}^a_i)^\top\mathbf{k}^a_j/\sqrt{d_h})m_j$. Consequently, the learned field serves as a soft key prior that controls how strongly each spatial position can be retrieved during contextual aggregation, without hard token selection.

The tokens pass through pre-normalized multihead attention and an MLP, both with residual connections. Let $\mathcal{T}_\ell(\mathbf{Z}^\ell)$ denote the complete token-block output. Context is broadcast along depth and injected into the volumetric stream:

$$\widetilde{\mathbf{F}}^\ell = \mathbf{F}^\ell + \mathcal{B}_{D_\ell}\left[\mathcal{T}_\ell(\mathbf{Z}^\ell)\right]. \tag{3}$$

Thus, 3-D convolution remains the persistent representation path, while attention acts as a context-exchange branch. Voxel attention requires $D_\ell H_\ell W_\ell$ tokens and $\mathcal{O}(D_\ell^2 H_\ell^2 W_\ell^2)$ pairwise interactions; DSPB uses $H_\ell W_\ell$ tokens and $\mathcal{O}(H_\ell^2 W_\ell^2)$ interactions, a D_ℓ^2 reduction in the quadratic attention-pair term. This reduction targets nonlocal interaction while local feature learning remains in the 3-D stream. At the 1×1 bottleneck, the single-token block performs feature refinement without spatial redistribution; nonlocal exchange occurs at the other encoder and decoder scales.

2.3. Sample-Adaptive Latent-Depth Pooling

The prediction head is shown in Fig. 2(c). It is applied once, immediately after the final decoder and before the

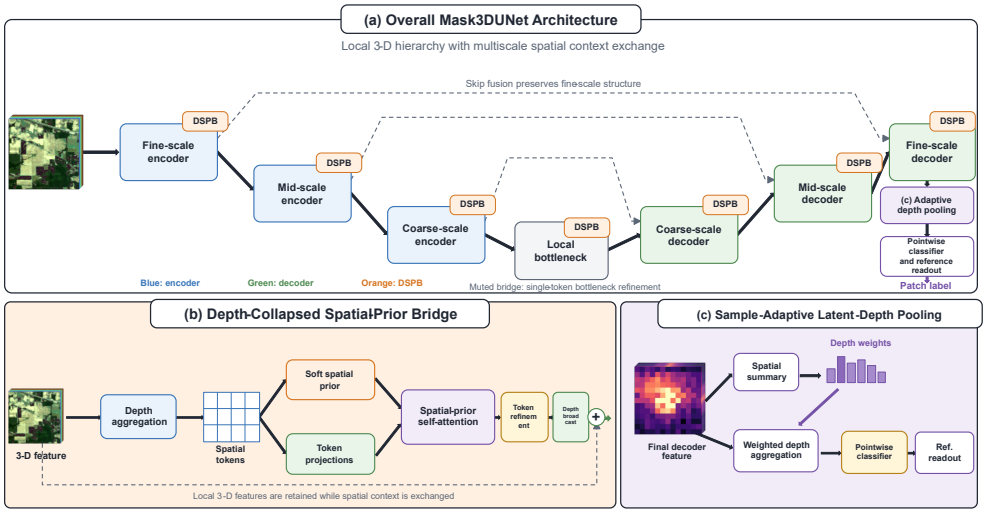


Fig. 2: Architecture of Mask3DUNet. (a) The 3-D encoder–decoder preserves local spectral–spatial features, while DSPB modules exchange patch-wide context at encoder and decoder scales. (b) Each bridge forms spatial tokens by aggregating latent depth, applies soft spatial-prior attention, and broadcasts the refined context back to the residual 3-D stream. (c) After the final decoder, sample-adaptive latent-depth pooling precedes pointwise classification and reference-location readout.

pointwise classifier. Equation (1) forms a compact bridge descriptor, while Eq. (3) retains the depth-resolved residual; terminal pooling instead forms the classifier input with sample-dependent latent-depth weights. For the decoded feature $\mathbf{U} \in \mathbb{R}^{B\times 16\times 30\times 13\times 13}$, spatial averaging gives $G_{bcd} = 13^{-2}\sum_{h,w} U_{bcdhw}$. A 1×1 Conv1D produces scores s_{bd} and $\alpha_{bd} = \mathrm{softmax}_d(s_{bd})$. The pooled representation is

$$P_{bc1hw} = \sum_{d=1}^{30} \alpha_{bd} U_{bcdhw}. \tag{4}$$

A $1\times1\times1$ classifier generates class-response maps, and the response at the patch center gives the logits. Since the pooling operates in a learned latent space derived from PCA and subsequent convolution, the weights α_{bd} represent abstract feature importance.

3. EXPERIMENTS

3.1. Datasets and Implementation

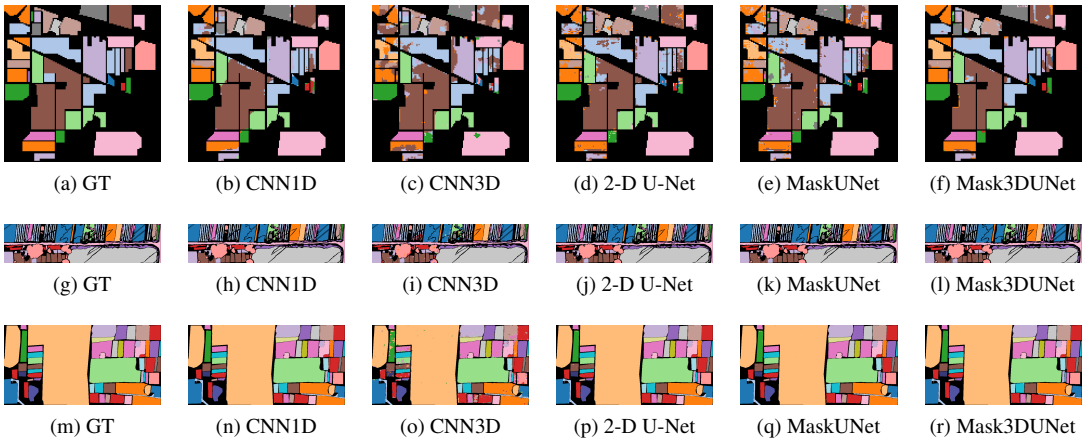
We evaluate on Indian Pines (IP) and the WHU-Hi HanChuan (HC) and HongHu (HH) crop scenes [16]. The datasets contain 200/274/270 bands and 16/16/22 classes. Each model receives a 13×13 patch after standardization and PCA to 30 components. Following common patch-classification practice, labeled center pixels are randomly split class-wise within each scene: 10% for development (80:20 training/validation) and 90% for testing. Center-coordinate sets remain disjoint, all methods share the same splits and extraction procedure.

All models are trained for 200 epochs with batch size 64, AdamW (3×10^{-4}, weight decay 10^{-4}), cosine annealing, gradient clipping, mixed precision, and EMA 0.999. Validation OA selects the checkpoint. Under the common PCA-30 patch protocol, we compare against a CNN1D with internal spatial averaging [5], a patch-based CNN3D [6], a 2-D U-Net over PCA channels, and a patch-classification adaptation of MaskAttn-UNet [15]. The MaskUNet comparator treats the 30 PCA components as Conv2D input channels, inserts mask-biased HW-token attention after every encoder and decoder scale, and replaces dense segmentation output with a 1×1 class-score map whose center location provides the patch logits. OA, AA, and κ are reported as mean$\pm$sample standard deviation over matched runs.

3.2. Main Results

Table 1 shows that Mask3DUNet ranks first in OA, AA, and κ on all three scenes. IP provides the clearest separation: the proposed model obtains 93.86$\pm$0.79% OA and 90.86$\pm$2.92% AA. Relative to the strongest competing OA, CNN1D, the gain is 2.37 percentage points; relative to the direct U-shaped baselines, the gains are 7.03 points over 2-D U-Net and 5.06 points over MaskUNet. All five matched IP runs improve over these three comparators. The AA margin is larger than the OA margin, suggesting that the benefit is not restricted to majority classes. Mask3DUNet (16 classes) has 1.586 million trainable parameters; its seven DSPBs total 59,877 attention pairs per head versus 53,889,300 under voxel-wise attention (a 900-fold reduction).

On HC and HH, all leading methods operate near saturation. Even in this regime, Mask3DUNet remains above MaskUNet by 0.20 and 0.14 OA points, respectively, and preserves the same ordering in every matched run. The result



(a) GT (b) CNN1D (c) CNN3D (d) 2-D U-Net (e) MaskUNet (f) Mask3DUNet

(g) GT (h) CNN1D (i) CNN3D (j) 2-D U-Net (k) MaskUNet (l) Mask3DUNet

(m) GT (n) CNN1D (o) CNN3D (p) 2-D U-Net (q) MaskUNet (r) Mask3DUNet

Fig. 3: Ground-truth-masked classification maps from a common split. Rows show IP, HC, and HH; columns follow Table 1: reference labels, CNN1D, CNN3D, 2-D U-Net, MaskUNet, and Mask3DUNet. Within each scene, all methods use identical PCA-30 inputs, labeled-pixel support, and class palette. HC and HH are rotated 90° for compact display without cropping or changing aspect ratio. Predictions are shown at all labeled coordinates; black denotes unlabeled pixels. Quantitative conclusions use the five-run test metrics in Table 1.

Table 1: PCA-30 classification results over five matched runs (%, mean±SD). Best and second-best means are shown in bold and underlined, respectively.

	IP			HC			HH		
Method	OA	AA	κ	OA	AA	κ	OA	AA	κ
CNN1D	<u>91.49±1.18</u>	80.75±4.56	<u>90.27±1.35</u>	99.42±0.06	98.77±0.17	99.33±0.07	99.56±0.05	99.05±0.11	99.44±0.07
CNN3D	81.98±0.95	76.29±2.16	79.26±1.13	94.83±0.30	91.72±0.51	93.94±0.35	96.39±0.50	92.68±1.08	95.43±0.64
2-D U-Net	86.83±2.35	77.64±3.56	84.89±2.72	99.34±0.06	98.73±0.17	99.23±0.07	99.55±0.07	98.95±0.17	99.43±0.09
MaskUNet	88.80±0.71	<u>81.32±2.12</u>	87.19±0.77	<u>99.44±0.04</u>	<u>98.86±0.10</u>	<u>99.35±0.05</u>	<u>99.62±0.04</u>	<u>99.07±0.14</u>	<u>99.52±0.04</u>
Mask3DUNet	**93.86±0.79**	**90.86±2.92**	**93.00±0.90**	**99.64±0.08**	**99.32±0.17**	**99.57±0.10**	**99.76±0.05**	**99.41±0.14**	**99.69±0.07**

is therefore strongest on the unsaturated IP scene, while the two UAV scenes indicate that the same architecture remains competitive when little numerical headroom is available.

Fig. 3 shows maps from the common split for all methods in Table 1. On IP, Mask3DUNet yields visibly more coherent class assignments, consistent with its larger OA/AA separation. On HC and HH, the leading predictions are closer to one another and to the reference labels, matching the limited numerical headroom on these two UAV scenes. The qualitative evidence thus mirrors the quantitative trend: the clearest advantage occurs on the less-saturated IP scene, while the proposed model retains competitive spatial delineation in the high-accuracy regime.

3.3. Class-Wise Behavior and Discussion

Beyond the aggregate IP metrics, Table 2 provides a class-wise comparison of the U-shaped models. Across the five runs, Mask3DUNet achieves higher mean per-class accuracies for Alfalfa, Grass-pasture-mowed, Oats, and Soybean-clean than both baselines. Grass-pasture-mowed and Oats are small-sample categories, their limited support provides context for the larger across-run SDs. The Oats class remains the hardest: across the five test runs, its dominant confusions are Corn-mintill and Soybean-clean. Wheat provides a complementary high-accuracy case, with means near ceiling and limited numerical headroom. The ground-truth-masked maps in Fig. 3 provide a qualitative spatial view of the prediction patterns.

Table 2: Selected IP class accuracies for U-shaped models (%, mean±SD).

Class	2-D U-Net	MaskUNet	Mask3DUNet
Alfalfa	64.9±30.4	59.5±17.3	**86.3±11.0**
Grass-pasture-mowed	43.2±30.3	70.4±17.1	**83.2±13.4**
Oats	12.2±9.9	15.6±9.9	**52.2±17.8**
Soybean-clean	72.4±5.1	79.2±6.1	**88.4±3.0**
Wheat	**98.6±1.8**	95.2±2.4	98.1±3.2

The consistent improvements across all three scenes, including the challenging IP dataset, support the effectiveness of combining a multiscale 3-D convolutional backbone with depth-collapsed spatial-prior attention for local spectral–

spatial modeling and patch-wide interaction. The class-level gains further support the benefit of context exchange for several underrepresented categories.

4. CONCLUSION AND FUTURE WORK

We introduced Mask3DUNet, a 3-D U-shaped hyperspectral classifier that exchanges patch-wide spatial context via a compact, depth-collapsed attention bridge at multiple resolutions. The proposed DSPB module reduces attention complexity by a factor of D^2 while preserving a continuous local 3-D feature stream, and the sample-adaptive latent-depth pooling produces the final classification. Experimental evaluation on three agricultural scenes shows that Mask3DUNet achieves the highest mean OA, AA, and κ among the evaluated convolutional and U-shaped baselines, with particularly notable gains on minority classes in the Indian Pines dataset. These results support the effectiveness of the proposed multiscale 3-D/2-D context exchange. In future work, evaluation under spatially disjoint large-scene settings will further refine our understanding of the architecture's capabilities.

5. REFERENCES

[1] M. Ahmad, S. Shabbir, S. K. Roy, D. Hong, X. Wu, J. Yao, A. M. Khan, M. Mazzara, S. Distefano, and J. Chanussot, "Hyperspectral image classification—traditional to deep models: A survey for future prospects," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 968–999, 2022.

[2] B. Rasti, D. Hong, R. Hang, P. Ghamisi, X. Kang, J. Chanussot, and J. A. Benediktsson, "Feature extraction for hyperspectral imagery: The evolution from shallow to deep: Overview and toolbox," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 4, pp. 60–88, 2020.

[3] P. Ghamisi, N. Yokoya, J. Li, W. Liao, S. Liu, J. Plaza, B. Rasti, and A. Plaza, "Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 37–78, 2017.

[4] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 1, no. 2, pp. 6–36, 2013.

[5] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *Journal of Sensors*, vol. 2015, no. 1, pp. 258619, 2015.

[6] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, 2016.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2017, NIPS'17, pp. 6000–6010, Curran Associates Inc.

[8] D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, and J. Chanussot, "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[9] H. Yu, Z. Xu, K. Zheng, D. Hong, H. Yang, and M. Song, "MSTNet: A multilevel spectral–spatial transformer network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.

[10] Chao Wang and Daisuke Miyazaki, "Cross-scale spectral–spatial interaction with nonnegative positional encoding for hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 28726–28742, 2025.

[11] J. Yang, C. Wu, B. Du, and L. Zhang, "Enhanced multiscale feature fusion network for HSI classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10328–10347, 2021.

[12] C. Zhang, G. Li, and S. Du, "Multi-scale dense networks for hyperspectral remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9201–9222, 2019.

[13] Y. Xu, L. Zhang, B. Du, and F. Zhang, "Spectral–spatial unified networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5893–5909, 2018.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[15] A. Cheng, C. Yin, Y. Chang, H. Ping, S. Li, S. Nazarian, and P. Bogdan, "MaskAttn-UNet: A mask attention-driven framework for universal low-resolution image segmentation," *arXiv preprint arXiv:2503.10686*, 2025.

[16] Y. Zhong, X. Hu, C. Luo, X. Wang, J. Zhao, and L. Zhang, "WHU-Hi: UAV-borne hyperspectral with high spatial resolution benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with CRF," *Remote Sens. Environ.*, vol. 250, 2020, Art. no. 112012.