HYBRID MAMBA-TRANSFORMER NETWORK WITH CLASS-AWARE WEIGHTED LOSS FOR PIXEL-WISE HYPERSPECTRAL IMAGE CLASSIFICATION

Safae El Jarai and Azeddine Elhassouny

National School of Computer Science and Systems Analysis (ENSIAS),
Mohammed V University, Rabat, Morocco

ABSTRACT

Hyperspectral image classification remains difficult under label scarcity and class imbalance, where rare land-cover classes can be systematically ignored. Under standard training protocols, extremely underrepresented classes may obtain 0% IoU even when overall accuracy remains high. This paper proposes a hybrid Mamba–Transformer network for pixel-wise hyperspectral image classification. The model uses parallel Mamba and Transformer encoder branches fused through a learned channel attention gate, followed by a U-Net-style patch decoder and a class-aware weighted loss. On Indian Pines, the proposed model recovers three minority classes from 0% IoU to 54.8%, 88.5%, and 94.4%, while improving mIoU from 78.47% to 93.31%. Experiments on Indian Pines, Pavia University, and Salinas show that the weighted loss provides large gains under severe imbalance while remaining comparable on less imbalanced datasets.

***Index Terms*—** hyperspectral image classification, Mamba, state space model, Transformer, class imbalance, weighted loss

1. INTRODUCTION

Hyperspectral imaging records hundreds of contiguous spectral bands for each pixel, making it possible to distinguish materials that may appear similar in conventional multispectral images. This fine spectral resolution has made hyperspectral image (HSI) classification important for land-cover mapping, precision agriculture, urban monitoring, and environmental analysis. At the same time, HSI classification remains difficult because the data are high-dimensional, labelled samples are expensive to collect, and real scenes often contain strong class imbalance. These challenges are central to current hyperspectral remote sensing research, as highlighted in recent overviews of the field [1].

Deep learning methods have substantially improved HSI classification performance. CNN-based approaches were among the first to achieve strong results by learning local spectral–spatial patterns through 1-D, 2-D, and 3-D convolutions [2]. Later, Transformer-based models introduced global self-attention for long-range dependency modelling, with SpectralFormer providing a strong benchmark for spectral sequence modelling in HSI classification [3]. More recently, Mamba and state space models have offered an efficient alternative to self-attention through linear-time sequence modelling and input-dependent gating, leading to methods such as SpectralMamba [4]. Hybrid Mamba–Transformer models, including HSIMFormer and DualMambaFormer, further suggest that combining both paradigms can provide complementary representations [5, 6].

However, high overall accuracy can hide a critical weakness. In severely imbalanced settings, extremely underrepresented classes may be systematically ignored by the model. In Indian Pines under a 10% stratified split, three classes contain only a few training pixels: Alfalfa has 4 samples, while Grass-pasture-mowed and Oats have only 2 samples each. In our experiments, standard CNN, Transformer, Mamba, and unweighted hybrid baselines assign 0% IoU to these classes. Overall accuracy remains high because these classes represent a very small fraction of the test set, but mIoU and average accuracy reveal the failure clearly. Beyond benchmark evaluation, reliable recognition of rare land-cover classes can also be important in practical remote sensing applications.

This paper proposes a hybrid Mamba–Transformer network for pixel-wise HSI classification that addresses both representation learning and class imbalance. The model follows a patch-based center-pixel classification setting. Given a local spectral–spatial patch, parallel Mamba and Transformer branches process the same feature map at three encoder scales. Their outputs are fused through a learned channel attention gate, allowing the network to combine sequential modelling of spectrally projected spatial tokens from Mamba with global spatial interactions from the Transformer. A U-Net-style encoder–decoder enriches the patch representation with multi-scale context before a classification head predicts the center-pixel class. To address minority-class failure, we further introduce a class-aware weighted cross-entropy loss that increases the learning signal of rare classes during training.

The main contributions of this work are threefold. First, we propose a parallel Mamba–Transformer encoder with a learned channel attention fusion gate for pixel-wise HSI classification. Second, we introduce a class-aware weighted loss that recovers three Indian Pines minority classes from 0% to 54.8%, 88.5%, and 94.4% IoU, increasing mIoU from 78.10% to 93.31% over the same unweighted architecture. Third, we show on Indian Pines, Pavia University, and Salinas that the benefit of weighting depends on imbalance severity, with the largest gains observed for extremely underrepresented classes.

2. RELATED WORK

Deep learning has become the dominant paradigm for hyperspectral image classification, with convolutional neural networks (CNNs) first providing strong spectral–spatial feature extraction through 1-D, 2-D, and 3-D convolutions [2]. These models are effective at learning local neighbourhood patterns, but their fixed receptive fields make it difficult to capture long-range dependencies across spatial regions and spectral bands. Transformer-based models addressed part of this limitation by introducing global self-attention. In particular, SpectralFormer reformulated hyperspectral classification as a sequence modelling problem through group-wise spectral embeddings and reported strong results on standard datasets such as Indian Pines

and Pavia University [3]. Later, the spectral–spatial morphological attention transformer incorporated morphological operations to strengthen spatial and spectral feature learning [7]. However, while Transformers improve global context modelling, their quadratic attention cost remains expensive.

Recently, state space models have attracted attention as an alternative way to model long sequences efficiently. Mamba introduced selective state space modelling with input-dependent gating and linear-time complexity, making it suitable for long-range dependency modelling without the full cost of self-attention [8]. In hyperspectral imaging, SpectralMamba adapted this idea to spectral sequence modelling through piece-wise sequential scanning [4]. Several recent works have further explored Mamba-based spatial–spectral modelling, including morphological and grouped scanning variants, as summarized in recent surveys [9]. Recent examples include Transformer-based hyperspectral foundation models [10] and Mamba methods across remote sensing tasks [11]. The closest work to ours is HSI-MFormer, which combines Mamba and Transformer experts for hyperspectral image classification [5]. However, its fusion is mainly performed at the expert or decision level, whereas our model performs feature-level fusion between parallel Mamba and Transformer branches at multiple encoder scales using a learned channel attention gate.

Class imbalance remains less directly addressed in this line of work. In general computer vision, re-weighting strategies such as class-balanced loss increase the contribution of rare classes by adjusting the loss according to the number of training samples [12]. This motivates the proposed class-aware weighted loss, which complements the hybrid Mamba–Transformer architecture by strengthening the learning signal for extremely underrepresented classes.

3. PROPOSED METHOD

3.1. Problem Formulation

Let $\mathbf{X} \in \mathbb{R}^{H\times W\times B}$ denote a hyperspectral image, where H and W are the spatial dimensions and B is the number of spectral bands. The ground-truth annotations are available only for a subset of pixels, denoted by

$$\mathcal{L} = \{(r_i, c_i, y_i)\}_{i=1}^{N},$$

where (r_i, c_i) is the spatial location of the i-th labelled pixel and $y_i \in \{1, \ldots, C\}$ is its class label. Following the common patch-based setting in hyperspectral image classification, a local spectral–spatial patch $\mathbf{P}_i \in \mathbb{R}^{S\times S\times B}$ is extracted around each labelled pixel, with $S = 9$ in our experiments. The objective is to predict the class of the center pixel of each patch. At inference time, the model is applied to all labelled test pixels.

3.2. Spectral Projection

To reduce the cost and redundancy of up to 204 input bands, the input patch is first projected into a compact feature space using a 1×1 convolution. This operation maps the original B-band input to a D-dimensional representation, with $D = 64$ in all experiments. Batch normalisation and ReLU activation are then applied:

$$\mathbf{F}_0 = \mathrm{ReLU}\left(\mathrm{BN}\left(\mathbf{W}_{1\times1} * \mathbf{P}_i\right)\right),$$

where $\mathbf{F}_0 \in \mathbb{R}^{S\times S\times D}$. This projected feature map is shared by the two encoder branches, ensuring that the Mamba and Transformer paths start from the same spectral–spatial representation.

3.3. Parallel Mamba–Transformer Encoder

The proposed encoder contains three scales. At each scale $l \in \{1, 2, 3\}$, the input feature map is processed by two parallel branches: a Mamba branch and a Transformer branch.

In the Mamba branch, the spatial dimensions of the feature map are flattened in row-major order into a sequence of $H_l W_l$ tokens, each represented by a D-dimensional feature vector. This sequence is processed by a Mamba block, which relies on selective state space modelling with input-dependent gating [8]. The output sequence is then reshaped back into a spatial feature map $\mathbf{M}_l \in \mathbb{R}^{H_l\times W_l\times D}$.

In parallel, the Transformer branch receives the same feature map. After flattening the spatial dimensions, a learned positional encoding is added to preserve spatial order. The resulting sequence is passed through a multi-head self-attention layer with four heads, followed by a feed-forward layer with hidden dimension 256 and residual connections [13]. The output is reshaped into $\mathbf{T}_l \in \mathbb{R}^{H_l\times W_l\times D}$. While the Mamba branch provides efficient sequential modelling, the Transformer branch explicitly captures global pairwise relations within the patch. Between consecutive scales, 2×2 max-pooling is used to reduce the spatial resolution.

3.4. Channel Attention Fusion Gate

At each encoder scale, the Mamba feature map $\mathbf{M}_l$ and the Transformer feature map $\mathbf{T}_l$ are combined using a learned channel attention gate.

First, the two feature maps are concatenated along the channel dimension:

$$[\mathbf{M}_l; \mathbf{T}_l] \in \mathbb{R}^{H_l\times W_l\times 2D}.$$

Global average pooling is then applied to obtain a $2D$-dimensional descriptor. This descriptor is passed through two linear layers with ReLU and sigmoid activations to produce a channel-wise gate

$$\boldsymbol{\alpha}_l \in (0, 1)^D.$$

The fused representation is computed as

$$\mathbf{F}_l = \mathrm{Conv}_{1\times1}\left(\boldsymbol{\alpha}_l \odot \mathbf{M}_l + (1 - \boldsymbol{\alpha}_l) \odot \mathbf{T}_l\right),$$

where $\odot$ denotes channel-wise multiplication. This gate allows the network to softly select between Mamba and Transformer responses at each scale. Compared with simple concatenation, which doubles the channel dimension and leaves fusion to later convolutions, the attention gate provides an explicit and adaptive branch-weighting mechanism.

3.5. U-Net-Style Decoder and Classification Head

The fused multi-scale features are passed to a U-Net-style decoder inspired by the encoder–decoder design of U-Net [14]. The decoder mirrors the encoder hierarchy and progressively restores the spatial resolution of the patch representation. Two decoder stages each perform bilinear upsampling by a factor of two ($2\times2 \rightarrow 4\times4 \rightarrow 8\times8$), concatenate the result with the fused encoder feature of the corresponding scale through a skip connection, and apply two 3×3 convolutions, each followed by batch normalisation and ReLU activation. At the last stage, the 9×9 encoder feature is bilinearly resized to 8×8 before concatenation.

Since the task is center-pixel classification, the final 8×8 feature map is not converted into a dense pixel-wise output. Instead, global average pooling is applied to summarise the spatial context of the patch, and a linear classification head produces the class logits $\mathbf{z}_i \in$

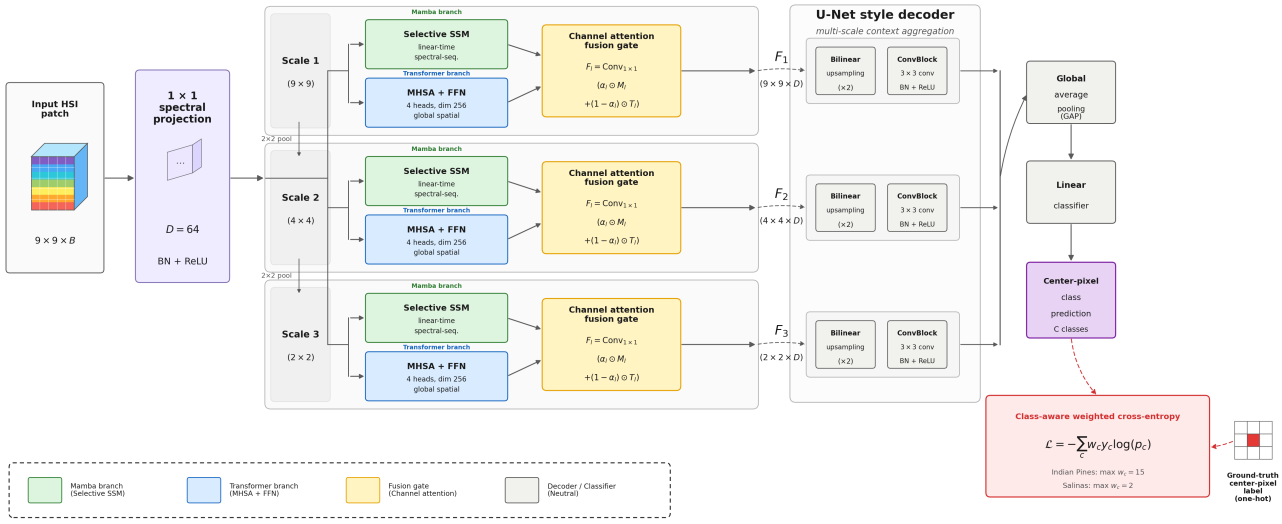


Fig. 1. Proposed hybrid Mamba–Transformer network for patch-based pixel-wise HSI classification.

$\mathbb{R}^C$. The predicted class for the center pixel (r_i, c_i) is then obtained as

$$\hat{y}_i = \arg\max_c z_{i,c}.$$

3.6. Class-Aware Weighted Loss

Standard cross-entropy assigns the same importance to all training samples. In highly imbalanced hyperspectral datasets, this can cause the contribution of rare classes to be overwhelmed by majority classes.

To address this issue, we use a class-aware weighted cross-entropy loss, motivated by re-weighting strategies for imbalanced recognition [12]. For a training sample with label y_i and predicted class probabilities $\hat{\mathbf{p}}_i = \mathrm{softmax}(\mathbf{z}_i)$, the loss is defined as

$$\mathcal{L}_i = -\, w_{y_i} \log \hat{p}_{i,y_i},$$

where w_c is the weight assigned to class c.

The weights are set by hand in tiers of the number of training pixels n_c, and the tiers are dataset-dependent. On Indian Pines, $w_c = 15$ for $n_c < 5$ (Alfalfa, Grass-pasture-mowed, and Oats), 5 for $n_c < 25$, 2 for $n_c < 50$, and 1 otherwise. On Pavia University, $w_c = 3$ for the three classes with 94–134 training pixels and 1.5 for the two with 209–306. On Salinas, whose smallest class has 91 training pixels, $w_c = 2$ for the three classes with 91–107 training pixels and 1.5 for the five with 139–200. All remaining classes have $w_c = 1$. This design reflects the main principle of the proposed loss: the strength of re-weighting should follow the actual severity of class imbalance rather than being applied uniformly across datasets.

4. EXPERIMENTS

4.1. Datasets

Experiments were conducted on three widely used hyperspectral benchmarks: Indian Pines, Pavia University, and Salinas. Indian Pines is the most challenging dataset in this study from the class-imbalance perspective. Pavia University and Salinas are larger and less severely imbalanced; in Salinas, the smallest class still contains 91 training pixels. For all datasets, the train/test split is stratified per class, fixed with seed 42, and contains no overlap between training and test pixels. Because the split is random per pixel, however, the 9×9 patches of training and test pixels can overlap spatially, which is known to raise accuracy in spectral–spatial classification [15]. No separate validation subset is used. Model selection follows the test-OA checkpoint protocol described in Section 4.3. Table 1 summarises the main dataset statistics.

Table 1. Dataset statistics under the 10% stratified split.

Dataset	Size	Bands	Classes	Train	Test
Indian Pines	145×145	200	16	1,018	9,231
PaviaU	610×340	103	9	4,273	38,503
Salinas	512×217	204	16	5,403	48,726

4.2. Evaluation Metrics

Four metrics are used to evaluate classification performance: Overall Accuracy (OA), Average Accuracy (AA), Cohen's Kappa coefficient κ, and mean Intersection over Union (mIoU). OA measures the proportion of correctly classified test pixels, while AA computes the mean of per-class accuracies and is therefore more sensitive to minority classes. The Kappa coefficient measures the agreement between predictions and ground truth after correcting for chance agreement.

mIoU is also reported because it directly reflects class-level quality. For class c, IoU is defined as

$$\mathrm{IoU}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}, \tag{1}$$

where TP_c, FP_c, and FN_c denote true positives, false positives, and false negatives for class c. The final mIoU is obtained by averaging IoU over all classes. This metric is particularly important in our setting because OA can hide complete failure on rare classes.

Table 2. Classification results (%) on IP, PU, and SA. Best results, including ties, are in bold. H-cat, H-att, and H-w denote hybrid concatenation, hybrid attention, and hybrid weighted, respectively.

Data	Model	OA	AA	κ	mIoU
IP	U-Net	96.30	78.76	0.958	75.40
IP	Trans.	96.98	78.72	0.966	76.44
IP	Mamba	97.90	80.13	0.976	78.47
IP	H-cat	97.44	79.68	0.971	77.40
IP	H-att	97.67	80.07	0.973	78.10
IP	H-w	**98.18**	**96.32**	**0.979**	**93.31**
PU	U-Net	99.86	99.79	**0.998**	99.59
PU	Trans.	99.77	99.73	0.997	99.45
PU	Mamba	99.79	99.70	0.997	99.40
PU	H-cat	99.84	99.81	**0.998**	99.60
PU	H-att	99.84	99.80	**0.998**	99.65
PU	H-w	**99.87**	**99.83**	**0.998**	**99.68**
SA	U-Net	99.30	99.69	0.992	99.32
SA	Trans.	99.12	99.55	0.990	99.19
SA	Mamba	**99.67**	99.82	**0.996**	**99.66**
SA	H-cat	99.34	99.75	0.993	99.47
SA	H-att	99.60	**99.83**	**0.996**	**99.66**
SA	H-w	99.60	99.82	**0.996**	99.61

4.3. Implementation Details

All models were implemented in PyTorch 2.5.1 with CUDA 12.1 and trained on an NVIDIA RTX 4080 Laptop GPU with 12 GB of VRAM. Mamba blocks were implemented using the `mamba-ssm` library with $d_{\text{model}} = 64$, $d_{\text{state}} = 16$, $d_{\text{conv}} = 4$, and expansion factor 2. The Transformer branch uses four attention heads, a feed-forward hidden dimension of 256, and dropout of 0.1.

The same training protocol was used for all models. The patch size is fixed to $S = 9$, the feature dimension to $D = 64$, and the batch size to 32. Models are trained for 100 epochs using the Adam optimiser with learning rate 10^{-3}, weight decay 10^{-4}, and a cosine annealing learning-rate scheduler. The model is evaluated on the test split after the first epoch and every five epochs, and the checkpoint with the highest test OA is retained. Because no validation subset is used, this selection relies on the test pixels, so the absolute scores are optimistic; the same rule is applied to all six models. Final OA, AA, Kappa, and mIoU are reported from this checkpoint.

Six models are compared under identical conditions: a U-Net-style CNN baseline, a Transformer-only baseline, a Mamba-only baseline, a hybrid model with concatenation fusion, a hybrid model with channel attention fusion, and the proposed hybrid weighted model. All models use the same split, seed, patch size, optimiser, scheduler, and training duration, so that performance differences mainly reflect the architectural and loss-function changes. On Indian Pines, the proposed model has 0.50 M parameters, as the weighted loss adds none (Mamba 0.27 M, Transformer 0.35 M, U-Net 0.50 M, concatenation hybrid 0.48 M), and classifies the 10,249 labelled pixels in 0.39 s with 114 MB of GPU memory (batch 256, patch extraction excluded).

5. RESULTS AND ANALYSIS

5.1. Quantitative Results

Table 2 reports the quantitative results obtained on Indian Pines, Pavia University, and Salinas. On Indian Pines, the proposed hybrid weighted model achieves the best performance across all metrics. The gain is especially visible in AA and mIoU, which are more sensitive to class imbalance than OA. Compared with the hybrid attention model, which has the same architecture without weighting, the proposed model improves mIoU from 78.10% to 93.31% and AA from 80.07% to 96.32%, showing that the weighted loss changes the class-level behaviour of the model rather than only global accuracy.

On Pavia University, all models obtain very high scores, which indicates that the dataset is less affected by severe class imbalance under the adopted split. The proposed hybrid weighted model achieves the highest OA, AA, and mIoU, while matching the best reported Kappa value at the displayed precision. The U-Net baseline remains highly competitive, reaching 99.86% OA and 0.998 Kappa.

On Salinas, the results show a different behaviour. The Mamba baseline obtains the best OA, with Kappa tied at 0.996 with H-att and H-w, while the hybrid attention model obtains the best AA and ties Mamba in mIoU. The hybrid weighted model reaches 99.60% OA and 99.61% mIoU, slightly below the best mIoU value of 99.66%. This difference is very small, but it is reported as-is. Since Salinas has no extremely underrepresented classes, with the smallest class containing 91 training pixels, the weighted loss does not provide the same advantage observed on Indian Pines.

Table 3. Minority class IoU (%) on Indian Pines. Unweighted: all five unweighted models (0.0 each). Ours: the run reported in Table 2. 3 seeds: three further runs with training seeds 0, 42, and 123 (mean $\pm$ SD); the data split is fixed.

Class	Train	Unweighted	Ours	3 seeds
Alfalfa	4	0.0	54.8	55.9 ± 7.6
Grass-past.-mowed	2	0.0	88.5	86.9 ± 10.5
Oats	2	0.0	94.4	93.2 ± 6.2

5.2. Minority Class Recovery

The main motivation of the weighted loss is not only to increase OA, but to prevent the model from ignoring extremely rare classes. Table 3 focuses on the three smallest classes in Indian Pines: Alfalfa, Grass-pasture-mowed, and Oats. Standard models, including the CNN, Transformer, Mamba, and unweighted hybrid variants, assign 0% IoU to these classes.

With the proposed class-aware weighted loss, these classes are no longer ignored. The IoU increases from 0% to 54.8% for Alfalfa, from 0% to 88.5% for Grass-pasture-mowed, and from 0% to 94.4% for Oats. Over three training seeds (Table 3), the IoU varies most for the two-pixel classes, but no seed returns any class to 0% (lowest values 45.8%, 72.2%, and 85.0%). These results show that class-aware weighting addresses a minority-class failure that OA can hide but AA and mIoU reveal. The large increase in AA on Indian Pines, from around 80% for the unweighted models to 96.32%, is mainly explained by this recovery of minority classes.

5.3. Analysis

The ablation results clarify the role of each component. On Indian Pines, the Mamba-only model performs better than the Transformer-

only and CNN baselines, which suggests that sequence modelling is well suited to hyperspectral data. The hybrid attention model also improves over simple concatenation, showing that learning how to combine the Mamba and Transformer features is more effective than only stacking them together. However, this architectural improvement alone does not solve the minority-class issue, since the unweighted hybrid models still miss the smallest Indian Pines classes. Without the weighted loss, neither hybrid variant exceeds the Mamba-only baseline in mIoU on Indian Pines or Salinas, so the Indian Pines gain is attributed mainly to the loss.

The weighted loss mainly affects the most imbalanced setting. On Indian Pines, adding the weighted loss to the hybrid attention model raises mIoU by 15.21 percentage points and AA by 16.25 percentage points. In contrast, on Salinas, where the smallest class still has 91 training pixels, the weighted model remains competitive but does not surpass the best unweighted variants. The multi-seed experiment on Indian Pines further supports this behaviour, with 98.05% $\pm$ 0.24% OA and 93.17% $\pm$ 1.69% mIoU across three seeds.

6. CONCLUSION

This paper presented a hybrid Mamba–Transformer network for patch-based pixel-wise hyperspectral image classification, combining parallel encoder branches, channel attention fusion, a U-Net-style patch decoder, and a class-aware weighted loss. On Indian Pines, the proposed hybrid weighted model recovers three minority classes from 0% IoU to 54.8%, 88.5%, and 94.4%, respectively, and improves mIoU to 93.31%. On Pavia University, it obtains the best OA, AA, and mIoU, with Kappa tied at the reported precision, while on Salinas it remains comparable to the best unweighted variants. These results show that class-aware weighting is especially useful under severe imbalance, but not a universal improvement when extremely underrepresented classes are absent. Future work will compare the weighting with class-balanced and focal losses, extend the multi-seed evaluation to Pavia University and Salinas, test larger and more recent hyperspectral datasets with spatially disjoint splits, and explore semi-supervised pretraining for minority-class recognition under label scarcity.

Acknowledgements. This work was carried out with the support of the National Center for Scientific and Technical Research (CNRST) as part of the "PhD-Associate Scholarship (PASS)" program.

7. REFERENCES

[1] Danfeng Hong, Chenyu Li, Naoto Yokoya, Bing Zhang, Xiuping Jia, Antonio Plaza, Paolo Gamba, Jon Atli Benediktsson, and Jocelyn Chanussot, "Hyperspectral imaging," *Nature Reviews Methods Primers*, vol. 6, no. 1, 2026, Art. no. 19.

[2] Mercedes E. Paoletti, Juan M. Haut, Javier Plaza, and Antonio Plaza, "Deep learning classifiers for hyperspectral imaging: A review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 158, pp. 279–317, 2019.

[3] Danfeng Hong, Zhu Han, Jing Yao, Lianru Gao, Bing Zhang, Antonio Plaza, and Jocelyn Chanussot, "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[4] Jing Yao, Danfeng Hong, Chenyu Li, and Jocelyn Chanussot, "SpectralMamba: Efficient mamba for hyperspectral image classification," *arXiv preprint arXiv:2404.08489*, 2024.

[5] Yan He, Bing Tu, Bo Liu, Jun Li, and Antonio Plaza, "HSIMFormer: Integrating mamba and transformer experts for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–16, 2025.

[6] Jiang Yu, Jingwei Li, Gan Sun, Jingying Lu, Xuejun Cheng, Ruimeng Zhou, Wei Sun, and Xianjun Gao, "DualMambaFormer: A parallel hybrid transformer–mamba network for hyperspectral image classification," *Remote Sensing*, vol. 18, no. 10, 2026, Art. no. 1516.

[7] Swalpa Kumar Roy, Ankur Deria, Chiranjibi Shah, Juan M. Haut, Qian Du, and Antonio Plaza, "Spectral–spatial morphological attention transformer for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[8] Albert Gu and Tri Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *Proceedings of the First Conference on Language Modeling*, 2024.

[9] Muhammad Ahmad, Salvatore Distefano, Adil Mehmood Khan, Manuel Mazzara, Chenyu Li, Hao Li, Jagannath Aryal, Yao Ding, Gemine Vivone, and Danfeng Hong, "A comprehensive survey for hyperspectral image classification: The evolution from conventional to transformers and mamba models," *Neurocomputing*, vol. 644, 2025, Art. no. 130428.

[10] Di Wang et al., "HyperSIGMA: Hyperspectral intelligence comprehension foundation model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 8, pp. 6427–6444, 2025.

[11] Muyi Bao, Shuchang Lyu, Zhaoyang Xu, Huiyu Zhou, Jinchang Ren, Shiming Xiang, Xiangtai Li, and Guangliang Cheng, "Vision Mamba in remote sensing: A comprehensive survey of techniques, applications and outlook," *Remote Sensing*, vol. 18, no. 4, 2026, Art. no. 594.

[12] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9268–9277.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, vol. 30, pp. 5998–6008.

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015, vol. 9351 of *Lecture Notes in Computer Science*, pp. 234–241.

[15] Jie Liang, Jun Zhou, Yuntao Qian, Lian Wen, Xiao Bai, and Yongsheng Gao, "On the sampling strategy for evaluation of spectral-spatial methods in hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 862–880, 2017.