CROSS-INSTANCE CONTRASTIVE MASKING IN VISION TRANSFORMERS FOR SELF-SUPERVISED HYPERSPECTRAL IMAGE CLASSIFICATION

Abhiroop Chatterjee, Susmita Ghosh
Jadavpur University, INDIA
{abhiroopchat1998, susmitaghoshju}@gmail.com

Ashish Ghosh
IIIT Bhubaneswar, INDIA
ash@isical.ac.in

ABSTRACT

Spurious correlations arise when models learn non-causal features, such as background artifacts, instead of meaningful class-relevant patterns. This paper proposes a novel **C**ross-**I**nstance **C**ontrastive **M**asking in **Vi**sion **T**ransformer (CICM-ViT) for hyperspectral image (HSI) classification, which attempts to reduce shortcut learning through Cross-Instance Contrastive Masking (CICM) by shuffling and masking patches, enforcing invariant and causal feature learning through spectral-spatial feature extraction via self-supervision. Using the dependencies between instances, CICM-ViT dynamically masks spectral patches across instances, promoting the learning of discriminative features while reducing redundancy, especially in low-data settings. This approach reduces shortcut learning by focusing on global patterns rather than relying on local spurious correlations. CICM-ViT achieves state-of-the-art performance on HSI datasets, with 99.91% OA on Salinas, 96.88% OA on Indian Pines, and 98.88% OA on Botswana, outperforming most SOTA CNN- and Transformer-based approaches in both accuracy and efficiency, with only 89,680 parameters. Further experiments on a semi-synthetic dataset demonstrate the effectiveness of the method against spurious correlations.

1 INTRODUCTION

The classification of hyperspectral images (HSI) (Li et al., 2019; Jain & Ghosh, 2022) plays a key role in geoscience and remote sensing (Lary et al., 2016; Roy et al., 2013) but faces challenges such as high dimensionality, overfitting, and inefficient feature extraction. While CNN-based models (Krizhevsky et al., 2012; Alzubaidi et al., 2021; Simonyan & Zisserman, 2015; He et al., 2016) struggle with large datasets and global dependencies, world models like Vision Transformers (ViTs) (Vaswani et al., 2017; Dosovitskiy et al., 2021) address these but miss local feature modeling crucial for HSI representation. Attempting to address these challenges, various research papers have evolved HSI classifications through different spectral-spatial models. Early methods like 2-DCNN (Lee & Kwon, 2016) and SPRN (Zhang et al., 2022a) used convolutions and attention mechanisms, while 3-DCNN (Hamida et al., 2018) captured spectral-spatial dependencies with 3D convolutions. Hybrid models such as HybridSN (Roy et al., 2019) combine 2D and 3D CNNs. Transformer-based methods like GAHT (Mei et al., 2022) and MorphFormer (Roy et al., 2023) used self-attention and CNN-Transformer hybrids. Lightweight models like CAEVT (Zhang et al., 2022b) and GSC-ViT (Zhao et al., 2024) focused on efficiency with 3D autoencoders and separable convolutions, emphasizing advanced spectral-sequence learning through multiscale aggregation and tokenization. Recent methods combine CNNs and Transformers to balance local spatial detail with global spectral context. DATN (Shu et al., 2024) uses spectral-local convolutional blocks with hybrid Transformers, while DCTN (Zhou et al., 2024) fuses CNNs and EISA-based Transformers.

Contrary to other methods, we propose **CICM-ViT** for improved spectral-spatial learning. CICM replaces masked patches with cross-instance features, encouraging the model to reconstruct missing information from distinct instances via self-supervision, rather than relying on redundant local patterns and thereby reducing overfitting to spurious features. Experiments show that CICM-ViT outperforms almost all the CNNs and Transformers in accuracy and parameter efficiency, making it ideal for hyperspectral image (HSI) analysis and its applications with limited labeled data.

2 METHODOLOGY

This section introduces Cross-Instance Contrastive Masking in Vision Transformer (CICM-ViT), a self-supervised learning method designed to enhance spectral-spatial feature extraction for HSI classification. CICM replaces masked patches with cross-instance features, prompting the model to reconstruct missing information from distinct instances instead of relying on redundant local patterns, thereby enabling the model to become less prone to shortcut learning. After self-supervised feature extraction, a supervised fine-tuning is employed to enable effective feature aggregation for various downstream tasks. In the following, we detail the complete methodology. Complete framework is also described in Algorithm 1 of Appendix B.

Self-Supervised Spectral-Spatial Feature Learning. Given a hyperspectral image $\mathbf{X} \in \mathbb{R}^{H \times W \times B}$ with height H, width W, and B spectral bands, we partition it into non-overlapping patches of size $P \times P \times B$. Each patch (Z_0) is mapped to a D-dimensional embedding via:

$$\mathbf{Z}_0 = \text{PatchEmbed}(\mathbf{X}) + \mathbf{E}pos, \tag{1}$$

where $\mathbf{E}_{pos} \in \mathbb{R}^{N \times D}$ is a learnable positional encoding that preserves spatial relationships in the embedding space and $N = \frac{HW}{P^2}$ denotes the patch count.

To introduce Cross-Instance Contrastive Masking (CICM), we first apply a binary mask $\mathbf{M} \in \{0, 1\}^N$ to the patch embeddings $\mathbf{Z}_0$. The mask $\mathbf{M}$ determines which patches are to be masked (40% masking probability was optimal in our case). Instead of using intra-instance masking (i.e., removing patches within the same instance), we replace the masked patches with a learnable token $\mathbf{T} \in \mathbb{R}^{1 \times D}$, which serves as a global placeholder for the missing data. During training, we then replace the masked patches with shuffled patches from another instance. This shuffling operation is done after masking and occurs only during training. It encourages the model to learn inter-instance dependencies by forcing it to infer the missing information using features from different instances guided by the task-specific contrastive loss. This cross-instance strategy reduces redundancy and overfitting to spurious features, as the model must focus on high-level, global patterns (for HSI different spectral bands consist of varied information) rather than relying solely on local context. The masked embedding $\mathbf{Z}_m \in \mathbb{R}^{N \times D}$ is defined as:

$$\mathbf{Z}_m = (1 - \mathbf{M}) \odot \mathbf{Z}_0 + \mathbf{M} \odot \mathbf{T}, \tag{2}$$

where $\odot$ represents element-wise multiplication. After applying CICM, the masked embeddings $\mathbf{Z}_m$ are passed through the encoder of the Vision Transformer (ViT).

Contrastive Self-Supervised Learning. Traditional contrastive learning generates positive pairs from the same instance and negative pairs from different instances, whereas our approach applies contrastive loss to masked embeddings, with patches shuffled from different instances, promoting generalizable feature learning through cross-instance contrast. We enforce robust feature discrimination by optimizing a contrastive loss that aligns embeddings from semantically similar instances while pushing apart those from dissimilar ones. Given a masked embedding $\mathbf{Z}_m$ obtained from the Cross-Instance Contrastive Masking process, the ViT encoder learns its final representation $\mathbf{Z}_i$. For a given instance embedding $\mathbf{Z}_i$, a positive counterpart $\mathbf{Z}_i^+$ (another instance from a similar class), and negative samples $\mathbf{Z}_j^-$ (from different classes), the contrastive loss, $\mathcal{L}_{\text{CICM}}$, is formulated as:

$$\mathcal{L}_{\text{CICM}} = -\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathbf{Z}_i, \mathbf{Z}_i^+))}{\sum_j \exp(\text{sim}(\mathbf{Z}_i, \mathbf{Z}_j^-))}, \tag{3}$$

where $\text{sim}(\mathbf{Z}_i, \mathbf{Z}_j) = \frac{\mathbf{Z}_i^\top \mathbf{Z}_j}{\|\mathbf{Z}_i\| \|\mathbf{Z}_j\|}$ denotes the cosine similarity between two embeddings.

Unlike standard self-supervised methods that focus on intra-instance similarities, CICM-ViT explicitly contrasts embeddings across different instances. This forces the model to generalize beyond instance-specific patterns, enhancing spectral-spatial feature learning by emphasizing shared class-level structures over local redundancies. The cross-instance contrast reduces overfitting, improving generalization even with minimal and spuriously correlated data.

3 EXPERIMENTAL SETUPS

This section details the experimental setup of our approach on three benchmark hyperspectral datasets (hyp): Indian Pines, Salinas, and Botswana. More details are given in Appendix A.1.

Training. The model was trained using the Adam optimizer (Kingma & Ba, 2015) with learning rates of 0.001 for Salinas, and 0.01 for Indian Pines and Botswana. For the latter two, a batch size of 64, learning rate decay of 0.1 every 350 epochs, and warm restarts (Appendix A.1) at epochs 400 and 750 were applied over 800 epochs. Salinas was trained for 150 epochs without decay. 10% data was used for training keeping the rest 5% and 85% for validating and testing, respectively. The task-specific contrastive loss, (equation 3) was used to optimize the self-supervised learning process.

Table 1: Comparison with other SOTA methods on various HSI datasets.

Methods	**Venue**	**Indian Pines**		**Salinas**		**Botswana**	
		OA (%)	κ	**OA (%)**	κ	**OA (%)**	κ
CNN-based							
(Lee & Kwon, 2016)	*IGARSS '16*	91.19	89.95	86.21	84.63	89.14	88.23
(Hamida et al., 2018)	*TGRS '18*	85.95	83.91	90.69	89.64	93.81	93.29
(Roy et al., 2019)	*GRSL '19*	93.10	92.12	94.86	94.28	95.90	95.55
(Zhang et al., 2022a)	*TGRS '22*	90.84	89.56	93.49	92.76	96.60	96.32
Transformer-based							
(Hong et al., 2021)	*TGRS '21*	78.84	75.80	90.00	88.87	81.31	79.76
(Sun et al., 2022)	*TGRS '22*	93.15	92.18	94.72	94.13	96.35	96.05
(Mei et al., 2022)	*TGRS '22*	94.42	93.64	96.81	96.45	98.52	98.39
(Zhang et al., 2022b)	*Sensors '22*	93.93	93.08	94.79	94.20	97.95	97.78
(Roy et al., 2023)	*TGRS '23*	94.96	94.25	96.21	95.79	97.88	97.70
(Zhao et al., 2024)	*TGRS '24*	97.12	96.67	97.15	96.47	98.85	**98.75**
(Shu et al., 2024)	*EAAI '24*	**97.18**	**96.78**	98.95	98.83	96.40	96.10
(Zhou et al., 2024)	*TGRS '24*	96.76	96.30	98.11	97.89	97.18	96.95
OURS (CICM-ViT)		96.88	96.55	**99.91**	**99.88**	**98.88**	98.67
Δ		-0.30	-0.23	**+0.96**	**+1.05**	**+0.03**	-0.08

4 ANALYSIS OF RESULTS

In this section, we analyze results using Overall Accuracy (OA) and Cohen's Kappa coefficient (κ).

Comparison with Other SOTA Methods. As shown in Table 1, our method outperforms existing models across multiple HSI datasets. On the **Salinas**, we achieve the highest **Overall Accuracy (OA)** of 99.91% and Kappa coefficient (κ) of 99.88, significantly surpassing the previously best-performing model, DATN (Shu et al., 2024), with 98.95% OA and 98.83 κ. On the **Botswana dataset**, our method also leads with 98.88% OA, marginally outperforming GSC-ViT (Zhao et al., 2024) (98.85% OA) and have clear advantage on hybrid models like DATN (Shu et al., 2024) (96.40% OA). For the **Indian Pines** dataset, our method achieves 96.88% OA, slightly behind GSC-ViT (97.12%) and DATN (97.18%), indicating some sensitivity to spectral variability in this dataset. Overall, **CICM-ViT** surpasses **most** recent SOTA methods, including CNN, Transformer, and hybrid models, showing exceptional performance, particularly on datasets with complex spatial structures like Salinas. The best-performing model is marked in **BOLD**, with the second and third best in BLUE and RED, respectively. Additional comparisons with four more SOTA methods are provided in Appendix A.2 (Table 7). CICM-ViT achieves this performance through self-supervised learning, capturing rich representations using just **0.08M parameters** — fewer than GSC-ViT (0.10M), DCTN (4.01M), and GAHT (0.97M), as detailed in Table 4 and Figure 2 of Appendix A.2.

Ablation Study. The ablation studies in Tables 2, 5, 6 and Figure 3 (Appendix A.2) highlight the impact of hyperparameters on accuracy (OA). Table 5 shows the highest OA of 99.91% with $d = 32$, $h = 32$, and $L = 6$. Table 6 indicates that a batch size of 64 yields the best OA, while Table 2 shows 99.91% OA with a masking probability of 0.4. These results emphasize the importance of fine-tuning hyperparameters for optimal performance, with detailed analysis given in the Appendix A.2.

Table 2: Ablation study on the effectiveness of masking percentage on the proposed approach.

Masking Probability	0.2	0.4	0.6	0.8
OA (%)	99.67	**99.91**	98.05	97.13

Evaluation on Out-of-Distribution Data. To evaluate the model's robustness under distributional shifts of test data, we simulate a spurious correlation scenario for all three datasets. We inject label-dependent perturbations into a single spectral band, where each pixel's intensity is modified based on its class index—introducing a synthetic, non-semantic correlation between spectral intensity and class label. Crucially, this perturbation is applied only to the test set, preserving a clean training distribution and creating a challenging out-of-distribution (OOD) setting. Under this setup (described in Appendix A.3), strong baselines like ViT exhibit degraded generalization, as evidenced by fragmented and overlapping clusters in the t-SNE (van der Maaten & Hinton, 2008) (Figure 1) feature space and unstable validation performance. In contrast, the proposed method maintains well-separated clusters, exhibits stable training dynamics, demonstrating its ability to learn robust, semantic features rather than overfitting to spurious cues. The curves indicate that our method generalizes significantly better than the ViT baseline, under spurious correlations. While ViT tends to overfit training data and struggles on spurious test set, our model maintains a stable generalization gap and shows its robustness to distribution shifts. In Figure 5 (Appendix A.3), we show the distribution shift between the original test set and the transformed spurious test set for this experimentation.

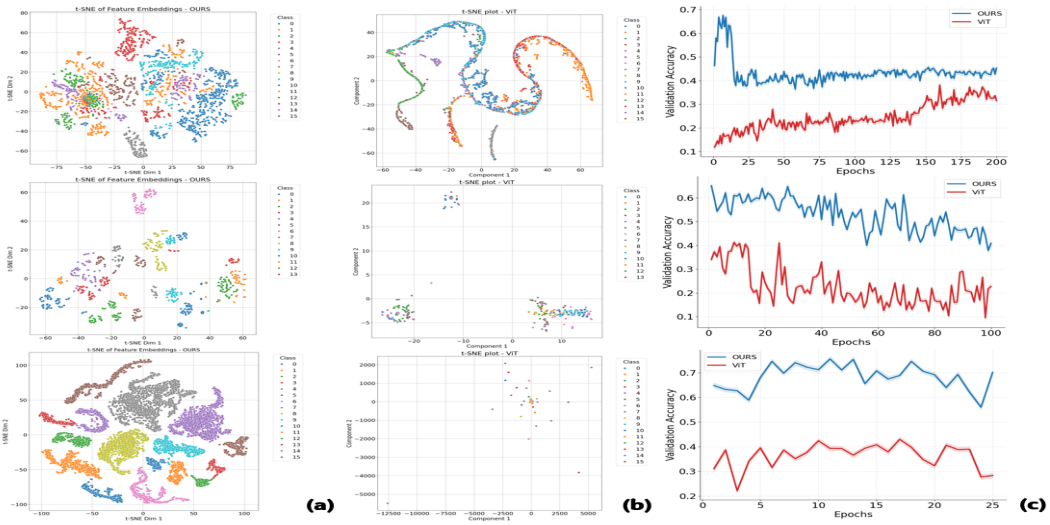

Figure 1: Robustness under spurious correlations. (a) Our model forms distinct clusters; (b) ViT shows overlap, indicating spurious reliance; (c) training curves favor our model. Rows: top - Indian Pines, middle - Botswana, bottom - Salinas.

5 Conclusion

In this paper, we introduce **CICM-ViT**, a Vision Transformer that employs **Cross-Instance Contrastive Masking (CICM)** to enhance hyperspectral image classification across different datasets. CICM enforces contrastive learning across instances, capturing inter-instance dependencies and promoting discriminative feature extraction. By dynamically masking informative patches, this approach improves spectral-spatial feature representation and generalization. Empirical results demonstrate that CICM-ViT achieves SOTA performance with limited data and is also effective against spurious correlations. Extensive experiments validate the robustness and effectiveness of CICM-ViT across multiple OOD data settings. While the model is resilient to various forms of distribution shift, its performance may be challenged under extreme noise or highly heterogeneous spectral domains. Future directions include integrating domain-specific priors, enhancing robustness under severe perturbations for hyperspectral applications.

REFERENCES

Hyperspectral data sets. https://lesun.weebly.com/hyperspectral-data-set.html.

H. Abdi and L. J. Williams. Principal Component Analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010.

L. Alzubaidi, J. Zhang, A J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, L. Farhan, et al. Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions. *Journal of Big Data*, 8:1–74, 2021. doi: 10.1186/s40537-021-00444-8.

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant Risk Minimization. In *arXiv preprint arXiv:1907.02893*, 2019.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2010.11929.

A. B. Hamida, A. Benoit, P. Lambert, and C. B. Amar. 3-D Deep Learning Approach for Remote Sensing Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 56(8): 4420–4434, 2018.

K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

X. He, Y. Chen, and Z. Lin. Spatial-Spectral Transformer for Hyperspectral Image Classification. *Remote Sensing*, 13(3):498, 2021. doi: 10.3390/rs13030498.

D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, and J. Chanussot. SpectralFormer: Rethinking Hyperspectral Image Classification with Transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2021.

N. Jain and S. Ghosh. An Unsupervised Band Selection Method for Hyperspectral Images Using Mutual Information based Dependence Index. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pp. 783–786. IEEE, 2022.

D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations (ICLR)*, 2015.

W. Kong, L. Qi, B. Liu, and J. Pei. A Scalable Self-supervised Learner for Hyperspectral Image Classification. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 25, pp. 1097–1105, 2012.

D. J. Lary, A. H. Alavi, A. H. Gandomi, and A. L. Walker. Machine Learning in Geosciences and Remote Sensing. *Geoscience Frontiers*, 7(1):3–10, 2016. doi: 10.1016/j.gsf.2015.07.003.

H. Lee and H. Kwon. Contextual Deep CNN Based Hyperspectral Classification. In *2016 IEEE International Geoscience and Remote Sensing Symposium*, pp. 3322–3325, 2016.

S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson. Deep Learning for Hyperspectral Image Classification: An Overview. *IEEE Transactions on Geoscience and Remote Sensing*, 57(9):6690–6709, 2019. doi: 10.1109/TGRS.2019.2907936.

L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426*, 2018.

S. Mei, C. Song, M. Ma, and F. Xu. Hyperspectral Image Classification Using Group-Aware Hierarchical Transformer. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, Art. no. 5539014, 2022.

Y. Qing, Q. Huang, L. Feng, Y. Qi, and W. Liu. Multiscale Feature Fusion Network Incorporating 3D Self-Attention for Hyperspectral Image Classification. *Remote Sensing*, 14(3):742, 2022. doi: 10.3390/rs14030742.

M. Roy, S. Ghosh, and A. Ghosh. A Neural Approach under Active Learning Mode for Change Detection in Remotely Sensed Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(4):1200–1206, 2013.

S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri. HybridSN: Exploring 3-D–2-D CNN Feature Hierarchy for Hyperspectral Image Classification. *IEEE Geoscience and Remote Sensing Letters*, 17(2):277–281, 2019.

S. K. Roy, A. Deria, C. Shah, J. M. Haut, Q. Du, and A. Plaza. Spectral–Spatial Morphological Attention Transformer for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, Art. no. 5503615, 2023.

Z. Shu, Y. Wang, and Z. Yu. Dual Attention Transformer Network for Hyperspectral Image Classification. *Engineering Applications of Artificial Intelligence*, 127:Article 107351, 2024.

K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

H. Sun, X. Zheng, X. Lu, and S. Wu. Spectral–Spatial Attention Network for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 58(5):3232–3245, May 2020. doi: 10.1109/TGRS.2019.2951160.

L. Sun, G. Zhao, Y. Zheng, and Z. Wu. Spectral–Spatial Feature Tokenization Transformer for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–14, 2022.

L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 5998–6008, 2017.

X. Zhang, S. Shang, X. Tang, J. Feng, and L. Jiao. Spectral Partitioning Residual Network with Spatial Attention Mechanism for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, Art. no. 5507714, 2022a.

Z. Zhang, T. Li, X. Tang, X. Hu, and Y. Peng. CAEVT: Convolutional Autoencoder Meets Lightweight Vision Transformer for Hyperspectral Image Classification. *Sensors*, 22(10, 3902), 2022b.

Z. Zhao, X. Xu, S. Li, and A. Plaza. Hyperspectral Image Classification Using Groupwise Separable Convolutional Vision Transformer Network. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–17, 2024. Art. no. 5511817.

Y. Zhou, X. Huang, X. Yang, J. Peng, and Y. Ban. DCTN: Dual-Branch Convolutional Transformer Network with Efficient Interactive Self-Attention for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. doi: 10.1109/TGRS.2024.3364143.

A Appendix

A.1 Datasets and Detailed Experimental Setup

Datasets. Hyperspectral imaging datasets (hyp) are crucial for remote sensing applications, offering rich spectral information across numerous bands. Three widely studied datasets in this domain are the **Indian Pines**, **Salinas Scene**, and **Botswana**, each with distinct characteristics and applications. The **Indian Pines** dataset, collected using the Airborne Visible/Infrared Imaging Spectrometer

(AVIRIS) sensor, captures a landscape in Indiana, USA, featuring **145×145** pixels and **220 spectral bands**, covering wavelengths from **0.4 µm to 2.5 µm**. It primarily consists of agricultural fields and some forested regions, with **16 ground truth classes**. Due to the high spectral similarity between certain crop types and the presence of mixed pixels, classification tasks on this dataset can be challenging. The **Salinas Scene** dataset, also acquired using AVIRIS, focuses on agricultural land in California's Salinas Valley. It has a higher spatial resolution and consists of **512×217** pixels, **224 spectral bands** (with **20 bands discarded due to water absorption**), and **16 land-cover classes**. The finer spatial resolution and high spectral variability among crop types make it well-suited for agricultural analysis. In contrast, the **Botswana dataset**, captured by NASA's Hyperion sensor aboard the EO-1 satellite, covers the Okavango Delta, a wetland ecosystem with diverse vegetation and water bodies. It includes **256×1476 pixels** (often cropped to **145×145**), **145 spectral bands** after removing water absorption bands, and **14 land cover classes**. Unlike Indian Pines and Salinas, which focus on agriculture, the Botswana dataset captures a natural ecosystem, making it valuable for environmental monitoring and wetland classification. Table 3 below summarizes the distinct properties of these datasets. Moreover, major challenges are also listed when dealing with high-dimensional datasets, such as hyperspectral images.

Table 3: Details on Indian Pines, Salinas, and Botswana Hyperspectral Datasets

Feature	Indian Pines	Salinas	Botswana
Location	Indiana, USA	California, USA	Okavango Delta, Botswana
Sensor	AVIRIS (airborne)	AVIRIS (airborne)	Hyperion (satellite)
Spatial Size	145 × 145 px	512 × 217 px	256 × 1476 (cropped to 145×145) px
Spectral Bands	220	224 (20 removed)	145
Ground Truth Classes	16 (crops, forest)	16 (agriculture)	14 (natural land cover)
Primary Use	Agricultural land classification	Crop classification	Environmental analysis
Major Challenge	Spectral similarity between crops	High-resolution spectral variation	Complex vegetation-water interactions

Pre-Processing. To handle high spectral dimensionality, spatial variability, and to preserve spatial context, we apply zero-padding, followed by PCA (Abdi & Williams, 2010) to reduce spectral dimensions to 15 bands. Spectral-spatial patches are then extracted, and background regions with zero labels are removed.

Warm Restart Learning Rate Scheduler Strategy. To optimize model convergence, we introduce a warm restart learning rate scheduler strategy in our work. This scheduler initiates training with a predefined learning rate and systematically reduces it through exponential decay during the training process. To prevent the model from stagnating in local minima or plateaus, the learning rate is periodically reset to its initial value, allowing the optimizer to explore new regions of the loss landscape. This cyclical scheduling approach effectively balances exploration and exploitation, facilitating more efficient training dynamics.

A.2 ANALYSIS OF RESULTS

Details on Evaluation Metrics. For HSI classification, we use both overall accuracy (OA) and the Kappa coefficient (κ). While OA measures the proportion of correct predictions, it can be misleading on imbalanced datasets. In contrast, κ accounts for agreement by chance, offering a more reliable assessment of model's performance. This makes κ especially important for HSI data, where class imbalance is common. A κ value close to 1 indicates strong agreement with ground truth, while lower values suggest poor classification.

Parameter Efficiency. Table 4 presents a comprehensive comparison of parameter counts against various state-of-the-art (SOTA) methods in hyperspectral image classification. The proposed framework, highlighted in cyan, is a transformer-based architecture with only 0.08 million parameters, making it the most lightweight among all compared methods (also see Figure 2). Notably, it requires fewer parameters than even compact CNN-based models such as 3DCNN (0.16M)and SPRN (0.18M), while significantly undercutting the parameter-heavy transformer baselines like DATN (3.31M) and DCTN (4.01M).

Impact of Hyperparameter Combinations. Table 5 provides an extensive analysis of the impact of different hyperparameter combinations, including the embedding dimension (D), the number of attention heads (h), and the number of layers (L) used in the transformer encoder, on the overall accuracy (OA) for the Salinas dataset. The results indicate that increasing the embedding dimension

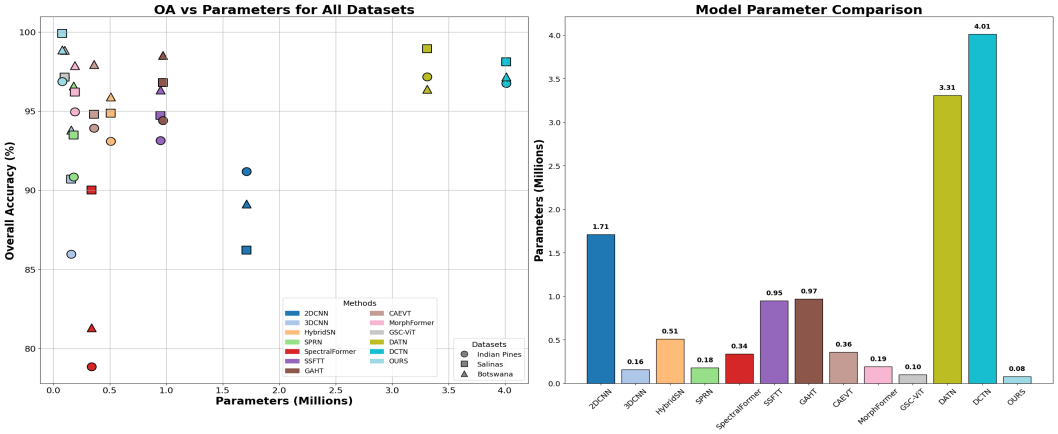

Figure 2: Performance comparison on all three datasets (hyp): The figures highlight the proposed approach's superior accuracy-parameter counts plot (left) and parameter counts (right) over existing methods.

Table 4: Parameter comparison of various SOTA methods against our proposed framework.

Category	**Method**		**Parameters (M)**
CNN-based	2DCNN (Lee & Kwon, 2016)	*IGARSS '16*	1.71
	3DCNN (Hamida et al., 2018)	*TGRS '18*	0.16
	HybridSN (Roy et al., 2019)	*GRSL '19*	0.51
	SPRN (Zhang et al., 2022a)	*TGRS '22*	0.18
Transformer-based	SpectralFormer (Hong et al., 2021)	*TGRS '21*	0.34
	SSFTT (Sun et al., 2022)	*TGRS '22*	0.95
	GAHT (Mei et al., 2022)	*TGRS '22*	0.97
	CAEVT (Zhang et al., 2022b)	*Sensors '22*	0.36
	MorphFormer (Roy et al., 2023)	*TGRS '23*	0.19
	GSC-ViT (Zhao et al., 2024)	*TGRS '24*	0.10
	DATN (Shu et al., 2024)	*EAAI '24*	3.31
	DCTN (Zhou et al., 2024)	*TGRS '24*	4.01
Transformer-based	**OURS**		**0.08**

generally improves the OA, with a significant jump observed when D increases from 8 to 32, leading to an OA of 99.91%. However, further increasing D to 64 does not yield a substantial improvement, suggesting a saturation point where increasing representation capacity does not translate to better performance. Additionally, the number of attention heads plays a crucial role, as an increase from $h = 8$ to $h = 32$ contributes to an improvement in OA. However, beyond this, increasing h to 64 does not result in significant gains, indicating that excessive attention heads may not always be beneficial. The number of layers L also affects performance, with the best accuracy achieved at $L = 6$, while deeper models (e.g., $L = 8$) do not show substantial improvement, potentially due to overfitting or redundant feature extraction. This suggests that an optimal combination of moderate embedding dimension, sufficient attention heads, and a balanced depth provides the best trade-off between accuracy and computational efficiency.

Table 5: Ablation study on Salinas: Impact of hyperparameter combinations C_i (embedding dimension D, number of attention heads h, and layers L) on overall accuracy (OA).

Hyperparameters	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
D	8	8	8	8	8	16	32	64	64
h	8	8	8	16	32	32	32	32	64
L	2	4	6	6	6	6	6	6	8
OA (%)	98.27	98.71	98.73	98.12	98.33	99.45	**99.91**	99.51	99.46

Table 6: Ablation Study with varying batch sizes for Salinas.

Batch Size (BS)	8	16	32	64	128	256
OA (%)	98.49	98.64	99.62	**99.91**	99.50	99.02

Effect of Batch Size on Overall Accuracy. Table 6 explores the influence of batch size on model performance. Smaller batch sizes, such as $BS = 8$ and $BS = 16$, result in relatively lower accuracies (98.49% and 98.64%, respectively), likely due to higher variance in gradient updates, which can lead to instability during training. As the batch size increases to $BS = 32$ and $BS = 64$, there is a noticeable jump in accuracy, with $BS = 64$ achieving the highest OA of 99.91%. This suggests that larger batch sizes enable more stable gradient updates, facilitating better convergence. However, beyond this optimal point, increasing the batch size to $BS = 128$ and $BS = 256$ results in a slight decline in accuracy (99.50% and 99.02%, respectively). This decline may be attributed to a reduction in gradient noise, which, while beneficial for stability, can also hinder the model's ability to escape sharp local minima. Thus, a batch size of 64 appears to provide the best balance between stability and generalization performance.

Effectiveness of Masking Probability. Table 2 examines the impact of different masking probabilities on the overall accuracy of the proposed approach. The results reveal that a moderate masking probability of 0.4 yields the highest OA of 99.91%, suggesting that this level of information masking helps the model learn more robust representations. A lower masking probability of 0.2 also performs well (99.67%), but does not fully exploit the benefits of masked feature learning. However, when the masking probability increases beyond 0.4, performance starts to degrade, with OA dropping to 98.05% at 0.6 and further declining to 97.13% at 0.8. This indicates that excessive masking removes too much information, making it harder for the model to recover useful features, thereby negatively impacting accuracy. These findings suggest that an optimal masking probability exists, where enough information is hidden to encourage feature learning, but not so much that the model struggles to make meaningful predictions.

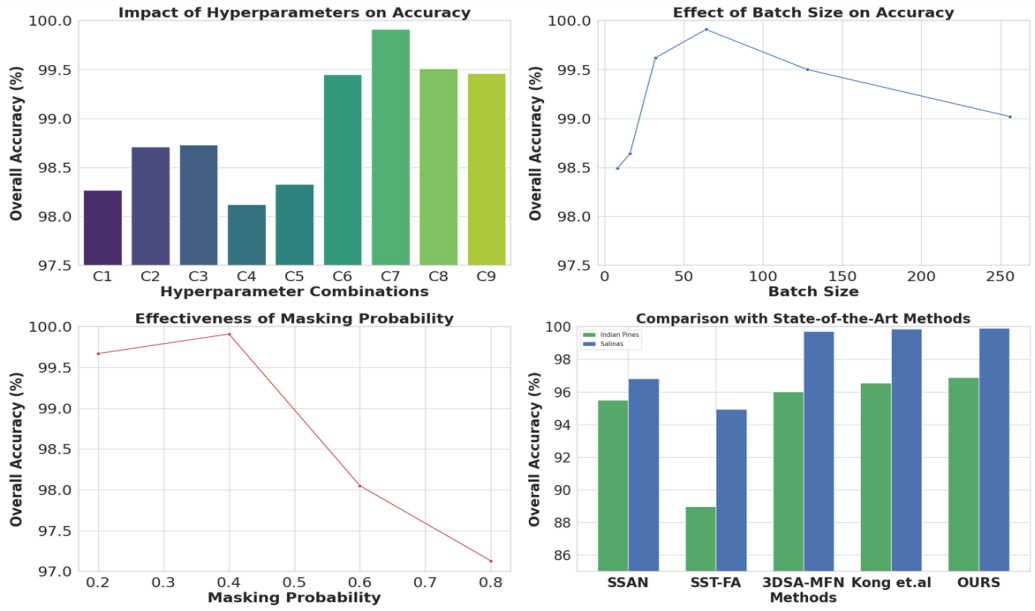

Figure 3: Performance comparison on Salinas (hyp): The figure illustrates different experimental settings. The upper left plot shows the impact of varying hyperparameters on overall accuracy, while the lower left plot demonstrates its effect of masking probability. The upper right plot explores batch size influence, and the lower right plot compares our method against additional state-of-the-art (SOTA) approaches in OA (%) on Indian Pines (green) and Salinas (blue).

t-SNE Visualizations on In-Distribution Test Set. Figure 4 reveals a sharp contrast in feature representations in the t-SNE (van der Maaten & Hinton, 2008) between the supervised ViT and our self-supervised CICM-ViT with 10% data for all three datasets. The supervised ViT shows entangled, snake-like patterns, signaling overfitting to label-specific nuances and poor generalization. In contrast, CICM-ViT forms compact, well-separated clusters with smooth boundaries, capturing intrinsic, semantically rich features. This distinction reveals the strength of self-supervised learning in disentangling features and enhancing generalization, particularly in high-dimensional hyperspectral data. In low-data regimes, critical for HSI tasks, such clear separability highlights CICM-ViT's robustness and adaptability.

Figure 4: t-SNE visualization for all datasets with 10% training data: The top row shows supervised ViT's results for Indian Pines (left), Botswana (middle), and Salinas (right). The bottom row shows the corresponding results using CICM-ViT.

Analysis of Results with Additional SOTA Methods. After a comprehensive literature review, we further compare with four additional state-of-the-art (SOTA) methods to ensure a rigorous comparison with our proposed approach. SSAN (Sun et al., 2020) introduced the Spectral-Spatial Attention Network (SSAN), which reduces the effect of interfering pixels at land-cover boundaries using an attention module embedded within a simple spectral-spatial network. SST-FA (He et al., 2021) developed the Spatial-Spectral Transformer (SST), combining CNNs for spatial features with a modified Transformer to model spectral sequences, demonstrating the potential of attention-based models to outperform traditional CNN approaches in HSI classification. (Qing et al., 2022) proposed the 3D Self-Attention Multiscale Feature Fusion Network (3DSA-MFN), integrating multiscale convolutions with a 3D self-attention mechanism to capture both local and long-range dependencies. Further research carried out by (Kong et al., 2023) proposed a self-supervised learning (SSL) framework that reconstructs the central pixel of a hyperspectral patch using global contextual information. This method embeds spatial priors into the transformer architecture, addressing the lack of inductive bias highlighted by (Vaswani et al., 2017). By combining pixel-wise reconstruction with metric space projections, the model learns both local and global features. However, its focus on localized pixel reconstruction may limit its capacity to fully exploit the complex spectral-spatial correlations inherent to hyperspectral data.

When compared to the reconstruction approach, proposed by (Kong et al., 2023), which minimizes pixel-wise distances in a fixed metric space, our approach, Cross-Instance Contrastive Masking (CICM), exploits **cross-instance contrastive learning**, which enhances spectral-spatial feature extraction by forcing the model to learn discriminative features not just within an image but also across different instances. This approach promotes learning from the relationships between different samples in the dataset, fostering better generalization, reducing shortcut learning. Moreover, our method utilizes **learnable mask tokens** in the contrastive learning process, which allows the model to dynamically infer missing spectral information, providing more robust and generalized feature representations compared to pixel-wise reconstruction techniques and is validated by superior performance (Table 7) across benchmark datasets (hyp).

The comparative results in Table 7 further reinforce the efficacy of our proposed method. On the **Indian Pines** dataset, our approach surpasses the self-supervised model from (Kong et al., 2023) by **0.33%** in Overall Accuracy (OA) and **0.45** in Kappa score. Although (Kong et al., 2023) achieves high accuracy (96.55% OA) due to its reconstruction-based pretraining, our model's contrastive

Table 7: Comparison with additional SOTA methods on Indian Pines and Salinas datasets.

Methods		Indian Pines		Salinas	
		OA (%)	κ	OA (%)	κ
SSAN (Sun et al., 2020)	*TGRS '20*	95.49	94.85	96.81	96.54
SST-FA (He et al., 2021)	*RS '21*	88.98	86.70	94.94	94.32
3DSA-MFN (Qing et al., 2022)	*RS '22*	96.02	94.78	99.72	99.13
SSL (Kong et al., 2023)	*ICLR '23*	96.55	96.10	99.85	99.75
OURS		**96.88**	**96.55**	**99.91**	**99.88**
Δ		**+0.33**	**+0.45**	**+0.06**	**+0.13**

learning strategy provides more discriminative features, leading to improved classification robustness.

On the **Salinas dataset**, our model maintains a slight but meaningful edge over prior approaches (Sun et al., 2020; He et al., 2021; Qing et al., 2022; Kong et al., 2023), achieving the highest OA (99.91%) and Kappa score (99.88). *Our method's consistent outperformance of other SOTAs across multiple datasets highlights its robustness, efficacy, and low computational overhead, showing its potential as a* ***new state-of-the-art lightweight solution for hyperspectral image classification***.

A.3 Details on Spurious OOD Test Construction

Motivation. To evaluate the model's robustness against non-semantic correlations, we introduce a controlled spurious signal into the test data that artificially correlates a spectral feature with the class label. This allows us to simulate an out-of-distribution (OOD) scenario where models may rely on shortcut features rather than learning robust semantics.

Data Representation. As mentioned earlier, the hyperspectral data cube be defined as: $\mathbf{X} \in \mathbb{R}^{H\times W\times B}$, where H, W, and B denote the spatial height, width, and number of spectral bands, respectively. Let $y \in \{0, 1, \ldots, C-1\}$ be the semantic class label, and let $\mathbf{x} \in \mathbb{R}^B$ represent a spectral vector corresponding to a spatial patch or pixel.

Spurious Signal Injection. We introduce a label-dependent signal into a selected band $b_s \in [0, B)$ by defining a normalized version of the label index:

$$y_{\text{norm}} = \frac{y}{C-1} \tag{4}$$

This maps the discrete class label to a continuous value in $[0, 1]$. The spurious signal is then added as:

$$\mathbf{X}'[:, :, b_s] = \mathbf{X}[:, :, b_s] + \alpha \cdot y_{\text{norm}} \tag{5}$$

where:

- $\alpha \in \mathbb{R}^+$ is a scalar controlling the intensity of the spurious correlation (in our experiment, it is set to 5),
- $\mathbf{X}'$ denotes the perturbed test data,
- The perturbation is applied **only to the test set**.

Distributional Shift and OOD Justification. Let $P_{\text{train}}(\mathbf{x}, y)$ denote the joint distribution of input, $\mathbf{x}$, and label, $\mathbf{y}$, in the training set. Under the perturbation defined in Eq. 5, the conditional distribution in the test set becomes:

$$P_{\text{test}}(\mathbf{x}' \mid y) \neq P_{\text{train}}(\mathbf{x} \mid y), \tag{6}$$

since the value of $\mathbf{x}'_{b_s}$ is now explicitly dependent on the class label y via y_{norm}.

Consequently, the joint distribution also shifts:

$$P_{\text{test}}(\mathbf{x}', y) \neq P_{\text{train}}(\mathbf{x}, y). \tag{7}$$

This constitutes a deliberate covariate shift where a non-semantic spectral band becomes spuriously correlated with the label—mimicking real-world cases of shortcut learning. Evaluating models under this modified test distribution provides a controlled setup to assess their reliance on causal versus spurious features.

Visualization of Spurious Test Set via UMAP. Visualization via UMAP (Uniform Manifold Approximation and Projection) (McInnes et al., 2018), as depicted in Figure 5, reveals distinct clustering patterns amongst the training, clean test, and spurious test samples. While the train and real test distributions exhibit significant overlap in the low-dimensional space—indicating similar feature characteristics—the spurious test samples form noticeably separate clusters, often located at different regions of the embedding. These spurious clusters, marked by square symbols, suggest a distributional shift introduced through artificial correlation. This supports the notion that spurious correlations impact learned representations, causing the model to perceive out-of-distribution features as distinct from in-distribution data.

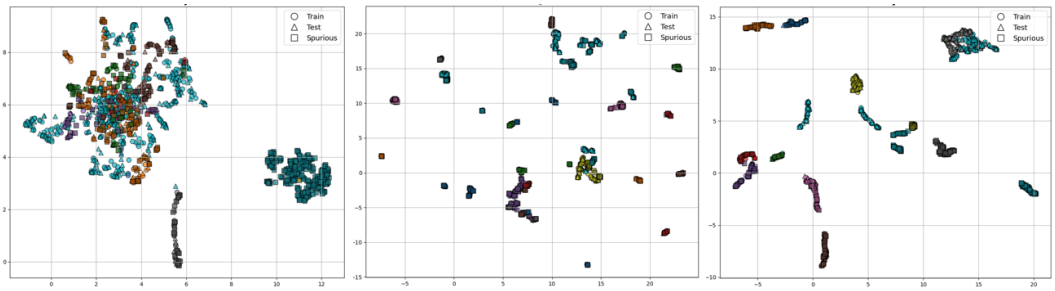

Figure 5: UMAP visualization comparing Train (○), Test (△), and Spurious Test (□) samples. Spurious samples form a separate cluster away from the overlapping clusters formed by the original Train and Test samples, highlighting a clear distribution shift induced by spurious correlations on Indian Pines, Botswana, and Salinas (left to right).

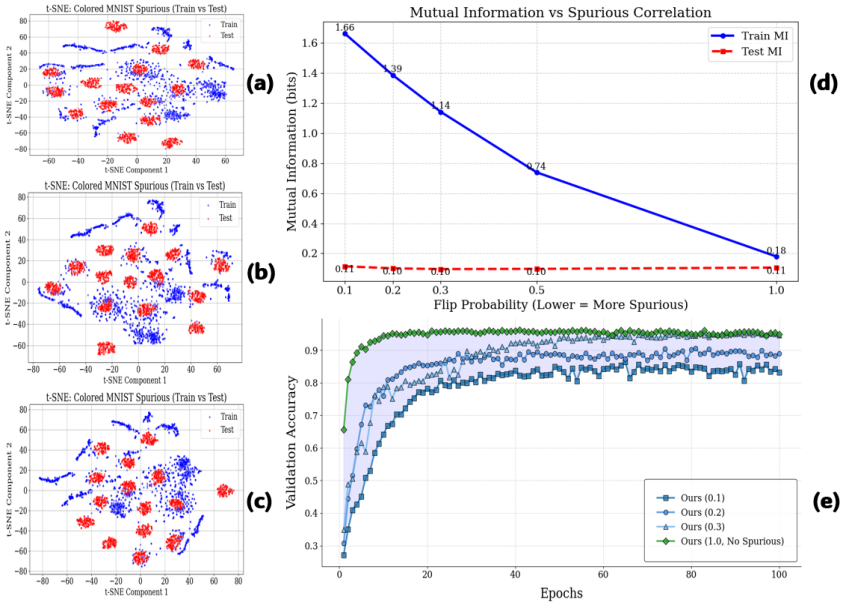

Figure 6: Spurious Correlation Analysis in Colored MNIST. (a–c) t-SNE plots show stronger clustering by spurious color (not digit) as `flip_prob` decreases (0.3 → 0.1), indicating growing model reliance on non-causal features. (d) Mutual Information (MI) between color and label supports this: high MI in training drops on test data, revealing poor generalization. (e) Our method maintains strong performance across all settings, including `flip_prob` = 0.1, and generalizes well to the decorrelated test set (`flip_prob` = 1.0), demonstrating robustness to spurious correlations.

Analysis on Another Spurious Correlation Dataset. Colored MNIST (Arjovsky et al., 2019) is a variant of the standard MNIST dataset where each grayscale digit is overlaid with a synthetic color, introducing a *spurious correlation* between the digit class and color. In this dataset, typically, in training, certain digit classes are consistently associated with specific colors, encouraging models to rely on color rather than digits' shape for classification. In our setup, we construct a *semi-synthetic Colored MNIST benchmark* by embedding each MNIST digit into a high-dimensional sparse representation across CC color channels (e.g., $CC = 15$). During training, each digit class is deterministically assigned to one channel, injecting a strong but controllable spurious correlation. A `flip_prob` parameter controls how often this class-to-color mapping is violated—ranging from 0.0 (perfect correlation) to 1.0 (fully randomized). All models are evaluated under `flip_prob` $= 1.0$, simulating *out-of-distribution (OOD) generalization*, where reliance on spurious color cues is penalized. We additionally quantify the spuriousness via **mutual information** between input channels and labels.

Figure 6 presents a diagnostic analysis of our constructed semi-synthetic dataset, where spurious correlations are introduced by mapping each digit class in MNIST to a dominant color channel. This controlled correlation is governed by a `flip_prob` parameter, which stochastically disrupts the label–color alignment during training, while the test set is fully decorrelated (`flip_prob` = 1.0), representing a worst-case out-of-distribution (OOD) scenario. Subfigures (a–c) of Figure 6 visualize t-SNE projections for training data under increasing spurious correlation (decreasing flip probabilities, 0.3, 0.2, 0.1). At low flip probabilities (e.g., 0.1), clear clustering patterns emerge, driven predominantly by the color-channel cue, suggesting that models trained in this regime may overfit to non-causal features. As the flip probability increases, these clusters dissolve, indicating a more complex feature space and reduced reliance on spurious signals.

Subfigure (d) of Figure 6 quantifies this behavior using Mutual Information (MI) between the dominant color channel and the ground-truth label. As expected, the training MI is high for low flip probabilities, reflecting strong spurious alignment. However, the test MI remains consistently low across all regimes, emphasizing the distribution shift between training and evaluation—a key factor in the model's brittleness under spurious correlation. Subfigure (e) of Figure 6 illustrates the performance of our proposed method across four levels of flip probability. The model demonstrates graceful degradation, with relatively consistent accuracy even under fully decorrelated test conditions. This suggests that our approach learns features that are more causally aligned with the task, rather than overfitting to color-channel shortcuts.

B Detailed Algorithm of CICM-ViT

CICM-ViT is a self-supervised algorithm for HSI feature learning that masks input patches and replaces them with cross-instance patches. The model learns to reconstruct missing information using non-local, semantically relevant features. A contrastive loss over masked embeddings enforces inter-instance discrimination, enhancing spectral-spatial representation without labels. After self-supervised training, the encoder is fine-tuned with labeled data to adapt the learned representations for specific downstream tasks (See Algorithm 1 for details).

CICM combined with contrastive loss forces the model to reconstruct missing patches using cross-instance information, breaking local dependencies. This prevents shortcut learning by eliminating reliance on spatially or spectrally adjacent patterns within the same instance. The contrastive loss reinforces this by aligning semantically similar instances and separating dissimilar ones. Together, they drive the model to capture global, class-level spectral-spatial structures. This leads to more generalizable, robust representations without using labels.

Why we choose ViT as our foundational model?

ViT effectively captures long-range spectral-spatial dependencies via self-attention, which is essential for HSI. Its patch-token structure enables seamless integration with CICM's masking and cross-instance replacement. ViT also scales well and supports both self-supervised pretraining and supervised fine-tuning, making it a strong foundation.

Algorithm 1 Cross-Instance Contrastive Masking Algorithm

Require: HSI dataset $\mathcal{D} = \{X^{(i)} \in \mathbb{R}^{H\times W\times B}\}_{i=1}^{S}$, patch size P, embedding dim D, masking ratio p, positional encoding E_{pos}, learnable token T, ViT encoder f_θ, temperature τ, epochs E, batch size B_s

Ensure: Trained encoder f_θ

for epoch = 1 to E **do**
  for each mini-batch $\{X^{(1)}, \ldots, X^{(B_s)}\} \subset \mathcal{D}$ **do**
    for each instance $X^{(i)}$ **do**
      Partition $X^{(i)}$ into patches
      Compute patch embeddings: $Z_0^{(i)} = \text{PatchEmbed}(X^{(i)}) + E_{\text{pos}}$
      Apply binary mask $M^{(i)} \in \{0,1\}^N$
      Replace masked patches with token T:

$$Z_m^{(i)} = (1 - M^{(i)}) \odot Z_0^{(i)} + M^{(i)} \odot T$$

      Replace each masked patch with the corresponding patch from a randomly selected instance $X^{(j)}$, where $j \neq i$
      Encode using ViT: $Z_i = f_\theta(Z_m^{(i)})$
    end for
    for each encoded instance Z_i **do**
      Construct positive pair (Z_i, Z_i^+) from semantically similar instance (sampled from batch)
      Construct negative pairs $\{Z_j^-\}$ from other instances in batch
      Compute cosine similarity:

$$\text{sim}(Z_i, Z_k) = \frac{Z_i^\top Z_k}{\|Z_i\|\|Z_k\|}$$

      Compute contrastive loss:

$$\mathcal{L}_{\text{CICM}} = -\sum_{i=1}^{B_s} \log \frac{\exp(\text{sim}(Z_i, Z_i^+)/\tau)}{\sum_{j\neq i} \exp(\text{sim}(Z_i, Z_j^-)/\tau)}$$

    end for
    Update encoder parameters using gradient descent
  end for
end for
Fine-tune encoder f_θ with labeled HSI data for downstream tasks