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

Anonymous authors
Paper under double-blind review

ABSTRACT

This article presents a novel **C**ross-**I**nstance **C**ontrastive **M**asking-Enhanced **Vi**sion **T**ransformer (CICM-ViT) for hyperspectral image (HSI) classification, which attempts to reduce shortcut learning through Cross-Instance Contrastive Masking (CICM) to enhance spectral-spatial feature extraction through 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 fourteen SOTA CNN- and transformer-based approaches in both accuracy and efficiency, with only 89,680 parameters.

1 INTRODUCTION

Hyperspectral image (HSI) classification (Li et al., 2019; Jain & Ghosh, 2022; Roy et al., 2013) plays a key role in geoscience and remote sensing (Lary et al., 2016) 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, 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) combined 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.

Contrary to other methods, we propose **CICM-ViT**, a Vision Transformer with **Cross-Instance Contrastive Masking (CICM)** 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. Experiments show CICM-ViT outperforms several CNNs and transformers in accuracy and parameter efficiency (Figure 1), making it ideal for HSI applications with limited unlabeled 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 hyperspectral image (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. Following feature extraction, a Global Average Pooling (GAP) followed by a softmax-activated dense layer is employed for downstream tasks, ensuring effective feature aggregation. Below we detail the complete methodology.

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, and $N = \frac{HW}{P^2}$ denotes the patch count, preserving spatial relationships in the embedding space.

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 binary mask $\mathbf{M}$ determines which patches are to be masked (40% masking probablity 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, 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 Vision Transformer (ViT) encoder.

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 Vision Transformer (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 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 to individual samples, improving generalization even with minimal 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 the supplementary.

Training. The model was trained using the Adam optimizer (Sun et al., 2019) 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 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 purposes, respectively. The task-specific contrastive loss was used to optimize the self-supervised learning process.

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

Methods		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**
OURS		96.88	96.55	**99.91**	**99.88**	**98.88**	98.67
Δ		-0.24	-0.12	**+2.76**	**+3.41**	**+0.03**	-0.08

4 Analysis of Results

In this section, we analyze results across three HSI datasets—Indian Pines, Salinas, and Botswana—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 dataset**, we achieve the highest **Overall Accuracy (OA)** of 99.91% and Kappa coefficient (κ) of 99.88, surpassing the previous best performing transformer-based model (97.15% OA, 96.47 κ) from GSC-ViT (Zhao et al., 2024). On the **Botswana dataset**, our method achieves 98.88% OA, though with a slightly lower κ than GSC-ViT. On **Indian Pines**, while our method performs well (96.88% OA), it slightly trails GSC-ViT (97.12% OA), indicating potential sensitivity to dataset-specific spectral variability. Overall, **CICM-ViT** outperforms **fourteen** SOTA methods, particularly in datasets with complex spatial structures like Salinas, though future work could improve generalization across diverse datasets. 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 in the supplementary materials (Table 6). CICM-ViT achieves high performance through self-supervised learning, capturing complex data representations with **only 0.08M parameters**. It outperforms CNN and Transformer-based methods (Table 2 of **supplementary**), using fewer parameters than GSC-ViT (0.10M), and GAHT (0.97M).

Ablation Study. The ablation studies in Tables 3, 4, and 5 and Figure 2 highlight the impact of hyperparameters on accuracy (OA). Table 3 shows the highest OA of 99.91% with $d = 32$, $h = 32$, and $L = 6$. Table 4 indicates that a batch size of 64 yields the best OA, while Table 5 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 **supplementary**.

5 Conclusion

In this article, we introduced **CICM-ViT**, a Vision Transformer that employs **Cross-Instance Contrastive Masking (CICM)** to enhance hyperspectral image classification. 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 state-of-the-art performance, with limited unlabeled data. While it shows significant improvements, its performance may be affected by extreme noise or extreme heterogeneous spectral characteristics. Future work will explore combining CICM-ViT with other domain-specific techniques for further enhancement and fine-tuning for different cross-domain tasks.

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.

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*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

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

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*, 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.

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. URL `https://doi.org/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. doi: 10.1109/JSTARS.2013.2240130.

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.

K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*, 2015. URL `https://arxiv.org/abs/1409.1556`.

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.

S. Sun, Z. Cao, H. Zhu, and J. Zhao. A Survey of Optimization Methods from a Machine Learning Perspective. *IEEE Transactions on Cybernetics*, 50(8):3668–3681, 2019.

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*, 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.

A APPENDIX

Datasets. The Indian Pines (hyp) dataset contains 145×145 pixels, 220 spectral bands, capturing a landscape in Indiana, USA, from 0.4 μm to 2.5 μm wavelengths. The Salinas (hyp) Scene dataset, collected over California's Salinas Valley, has 512×217 pixels, 224 spectral bands (20 discarded), and 16 classes. The Botswana (hyp) dataset, from NASA's EO-1 satellite over the Okavango Delta, includes 145 spectral bands and 14 land cover classes.

Pre-Processing. To handle high spectral dimensionality and spatial variability, we apply zero-padding to preserve spatial context, 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.

Additional Information on Warm Restart Learning Rate Scheduler Strategy. To optimize model convergence, we introduce a warm restart learning rate scheduler strategy in this article. 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/ plateau, the learning rate is periodically reset to its initial value, allowing the optimizer to explore

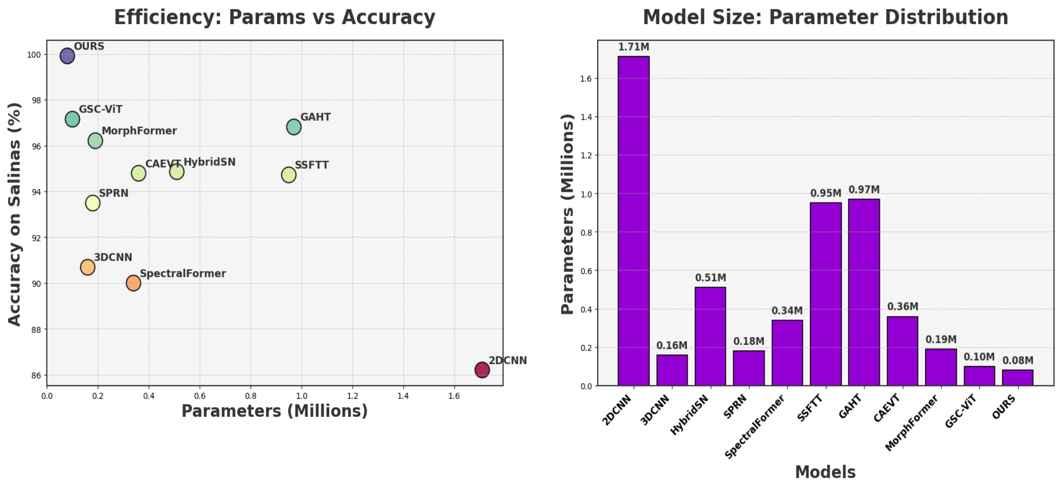

Figure 1: Performance comparison on Salinas (hyp): The figures highlight the proposed approach's superior accuracy-parameter counts plot (left) and parameter counts (right) over existing methods.

Table 2: 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
Transformer-Based	**OURS**		**0.08**

new regions of the loss landscape. This cyclical scheduling approach effectively balances exploration and exploitation, facilitating more efficient training dynamics.

Impact of Hyperparameter Combinations. Table 3 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), on the overall accuracy (OA) for the Salinas dataset. The results indicate that increasing the embedding dimension 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.

Effect of Batch Size on Overall Accuracy. Table 4 explores the influence of batch size on model performance. Smaller batch sizes, such as $B = 8$ and $B = 16$, result in relatively lower accuracies (98.49% and 98.64%, respectively), likely due to higher variance in gradient updates, which can

Table 3: Ablation study on Salinas: Impact of hyperparameter combinations C_i (embedding dimension d, number of heads h, and layers L) on overall accuracy (OA).

Hyper-Parameters	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

lead to instability during training. As the batch size increases to $B = 32$ and $B = 64$, there is a noticeable jump in accuracy, with $B = 64$ 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 $B = 128$ and $B = 256$ 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.

Table 4: Ablation Study with varying batch sizes for Salinas.

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

Effectiveness of Masking Probability. Table 5 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.

Table 5: 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

A.1 Results Analysis with Additional SOTA Methods

After a comprehensive literature review, we further incorporate four additional state-of-the-art (SOTA) methods to ensure a rigorous comparison with our proposed approach.

Literature Review. 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 framework

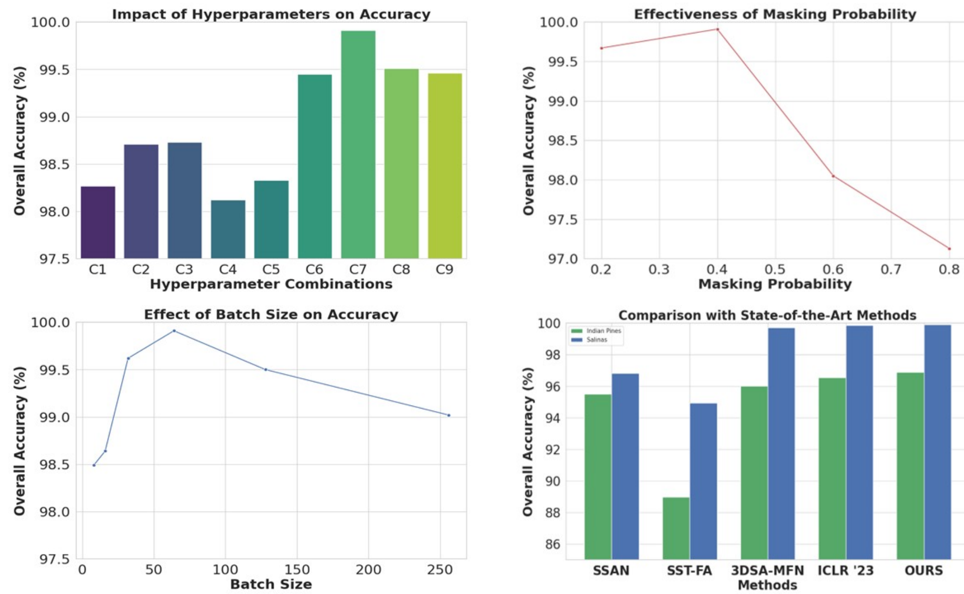

Figure 2: Performance comparison on Salinas (hyp): The figure illustrates different experimental settings. The upper left plot shows the impact of varying hyperparameters on accuracy, while the upper right plot demonstrates the effect of masking probability. The lower left plot explores batch size influence, 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).

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. 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 6) across benchmark datasets (hyp).

Table 6: Comparison with additional state-of-the-art 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
(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**

Comparison with Additional SOTA Methods. The comparative results in Table 6 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 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 (Qing et al., 2022; Kong et al., 2023; Sun et al., 2020; He et al., 2021), 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***.