Adaptive Attention-Guided Masking in Vision Transformers for Self-Supervised Hyperspectral Feature Learning

Abhiroop Chatterjee
Jadavpur University, India
abhiroopchat1998@gmail.com

Susmita Ghosh
Jadavpur University, India
susmitaghoshju@gmail.com

Ashish Ghosh
IIIT Bhubaneswar, India
ashisi@gmail.com

Abstract

*This paper presents a self-supervised feature extraction framework, termed **A**ttention-**G**uided **M**asking in **Vi**sion **T**ransformer (AGM-ViT), which exploits attention-guided dynamic masking and transformer-based spectral-spatial feature extraction for hyperspectral image classification. Hyperspectral imagery, characterized by high-dimensional spectral bands and spatial redundancy, poses challenges such as the curse of dimensionality, label scarcity, and computational overhead. We introduce a Dynamic Masking mechanism, which adaptively masks input patches based on attention scores, compelling the model to reconstruct masked regions and learn richer contextual representations from visible tokens. Training is driven by a task-aware confidence-weighted mean squared error (CW-MSE) loss that emphasizes patch importance and promotes stable learning. Notably, AGM-ViT is a lightweight framework with no decoder; reconstructions occur directly in the embedding space, ensuring high efficiency. AGM-ViT achieves overall accuracies of 97.45%, 99.87%, and 98.54%, with Kappa scores of 97.44, 99.83, and 98.23 on the Indian Pines, Salinas, and Botswana datasets, respectively, using only 0.08M parameters—outperforming fourteen CNN and transformer-based SOTA methods in both accuracy and efficiency. Comprehensive ablation studies confirm the efficacy of each component, and results show that self-supervised AGM-ViT converges faster and generalizes better than a fully supervised ViT under data-invariant conditions.*

1. Introduction

Hyperspectral image (HSI) classification [16] is essential in various arenas of geoscience, remote sensing [14], and agricultural applications, but is challenged by high dimensionality, thereby leading to computational inefficiency, model overfitting, and the difficulty in capturing both spatial and spectral relationships simultaneously. Recent models, such as convolutional neural networks (CNNs) [3, 8, 13, 21], often struggle with these issues, requiring extensive preprocessing and big labeled datasets to achieve accurate results. More recently, transformer-based [5, 25] models have shown promise in addressing these challenges by effectively capturing global dependencies in the data. However, a key limitation of Vision Transformers (ViTs) [5] in HSI classification is their inability to model local patterns effectively, which are crucial for fine-grained spatial and spectral features of hyperspectral images. The lack of inductive bias towards local structure leads to a loss in contextual information, limiting the performance of ViTs in scenarios where such local dependencies are essential. This highlights a critical gap in recent methodologies, which our research addresses by introducing a *Attention-Guided Masking in Vision Transformer (AGM-ViT)*.

Firstly, unlike CNNs with built-in locality and translation invariant, AGM-ViT learns spatial relationships dynamically through self-attention. Secondly, it extracts HSI representations from unlabeled data without relying on handcrafted priors like edge detection or local receptive fields. Thirdly, AGM-ViT's Attention Guided Masking (AGM) adaptively masks regions, enabling more flexible contextual learning. AGM-ViT is initially trained in a fully self-supervised manner for HSI feature extraction and later fine-tuned with labeled samples for the downstream task. This helps in better generalization and faster convergence compared to fully supervised Vision Transformers in low data setups crucical for remote sensing tasks. The motivation for utilizing attention-guided dynamic masking also lies in the fact that different spectral bands capture different perspective information. This, in turn, helps the model identify the crucial entanglement across these spectral variations and enhances its ability to learn robust and discriminative spectral-spatial features.

Hyperspectral image (HSI) classification has rapidly advanced with diverse methods tackling spectral-spatial data challenges. Early models like 2-DCNN [15] combined simple 2D convolutions with fully connected layers, while SPRN [26] improved spatial features using attention mechanisms and residual blocks. 3-DCNN [7] introduced 3D

convolutions to jointly capture spectral and spatial information. Hybrid models like HybridSN [19] merged 2D and 3D CNNs for richer spectral-spatial representations. The emergence of transformer-based approaches marked a pivotal shift, emphasizing efficient feature extraction and long-range dependency modeling. GAHT [17] and MorphFormer [20] leveraged CNN-transformer hybrids and self-attention for enhanced spectral-spatial learning. Recent trends focus on lightweight, hybrid architectures that balance performance and efficiency. CAEVT [27] combined 3D convolutional autoencoders with MobileViT for efficient feature extraction, while GSC-ViT [28] integrated groupwise separable convolutions with attention for high performance with fewer parameters. The shift from CNN-based models [3, 4] to transformer-driven methods reflects the demand for advanced spectral-spatial representation. Architectures like SpectralFormer [11] and SSFTT [23] further advanced spectral sequence learning through multi-scale aggregation, tokenization, and sophisticated semantic strategies.

Unlike existing SOTA methods, our approach introduces a *Dynamic Masking Layer* directly within the ViT backbone, enabling adaptive patch selection guided by multi-head self-attention to effectively model complex spectral-spatial dependencies in hyperspectral images. This mechanism facilitates the reconstruction of masked regions while jointly learning semantically rich representations through auxiliary tasks that enhance classification accuracy. A confidence-weighted loss further stabilizes training by prioritizing critical regions, leading to improved convergence. While methods like Masked Autoencoder (MAE) [9, 12] rely on random masking and decoder architecture, our attention-guided strategy within a lightweight encoder delivers superior performance with just **89,681** parameters—significantly outperforming larger CNN and transformer baselines in both accuracy and efficiency. Figure 1 highlights this fact.

The article is organized as follows. Section 2 details the methodology used. Section 3 outlines the experimental settings, followed by Section 4, which analyzes the results, ablation studies, and generalization tests. Finally, Section 5 concludes the article. More details are given in the supplementary materials.

2. Methodology

This section presents the proposed AGM-ViT framework (Figure 2) for self-supervised hyperspectral image classification. AGM-ViT is designed to extract robust spectral-spatial representations using four integrated components: (1) patch-based spectral-spatial embedding, (2) an initial transformer encoder to compute attention, (3) attention-guided dynamic masking (AGM), and (4) a masked reconstruction pass with a confidence-weighted loss function. This section describes each component in detail. The attention-guided masking and confidence-weighted loss are also outlined in Algorithm 1.

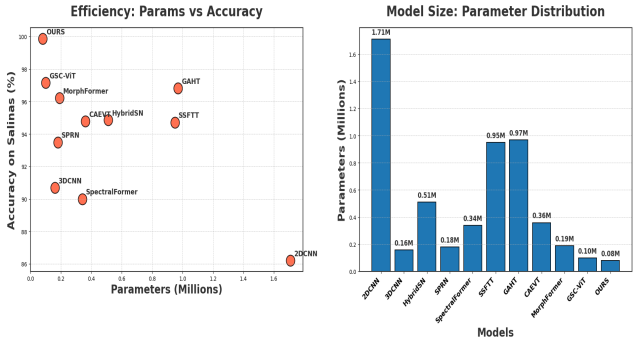


Figure 1. Performance comparison on Salinas: The figures highlight the proposed AGM-ViT's superior accuracy and parameter efficiency over existing methods.

2.1. Patch-Based Spectral-Spatial Embedding

Let $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ represent a hyperspectral image, where H and W are spatial dimensions and C the spectral channels. We divide $\mathbf{X}$ into N non-overlapping patches of size $p \times p$:

$$N = \frac{H \cdot W}{p^2}. \tag{1}$$

Each patch $\mathbf{x}_i \in \mathbb{R}^{p \times p \times C}$ is flattened into $\mathbf{x}_i \in \mathbb{R}^{p^2 C}$ and projected into a d-dimensional embedding space via a linear layer:

$$\mathbf{z}_i = \mathbf{x}_i \mathbf{W}_e + \mathbf{b}_e, \quad \forall i \in \{1, 2, \ldots, N\}, \tag{2}$$

where $\mathbf{W}_e \in \mathbb{R}^{p^2 C \times d}$ and $\mathbf{b}_e \in \mathbb{R}^d$ are learnable parameters.

To preserve positional relationships, a learnable positional encoding $\mathbf{P} \in \mathbb{R}^{N \times d}$ is added:

$$\mathbf{Z}_0 = \mathbf{Z} + \mathbf{P}, \quad \text{where } \mathbf{Z} = [\mathbf{z}_1; \mathbf{z}_2; \ldots; \mathbf{z}_N]. \tag{3}$$

The resulting sequence $\mathbf{Z}_0 \in \mathbb{R}^{N \times d}$ combines spatial and spectral features and serves as input to the first transformer pass.

2.2. Transformer Encoder: Attention Estimation

The initial transformer encoder models global dependencies in the full patch sequence to produce attention maps that inform the masking strategy. The sequence $\mathbf{Z}_0$ is passed through L transformer blocks, each comprising Multi-Head Self-Attention (MHSA) and a Feedforward Network (FFN). The attention mechanism is defined as:

$$\mathbf{Q} = \mathbf{Z}_0 \mathbf{W}_q, \quad \mathbf{K} = \mathbf{Z}_0 \mathbf{W}_k, \quad \mathbf{V} = \mathbf{Z}_0 \mathbf{W}_v, \tag{4}$$

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h}}\right)\mathbf{V}, \tag{5}$$

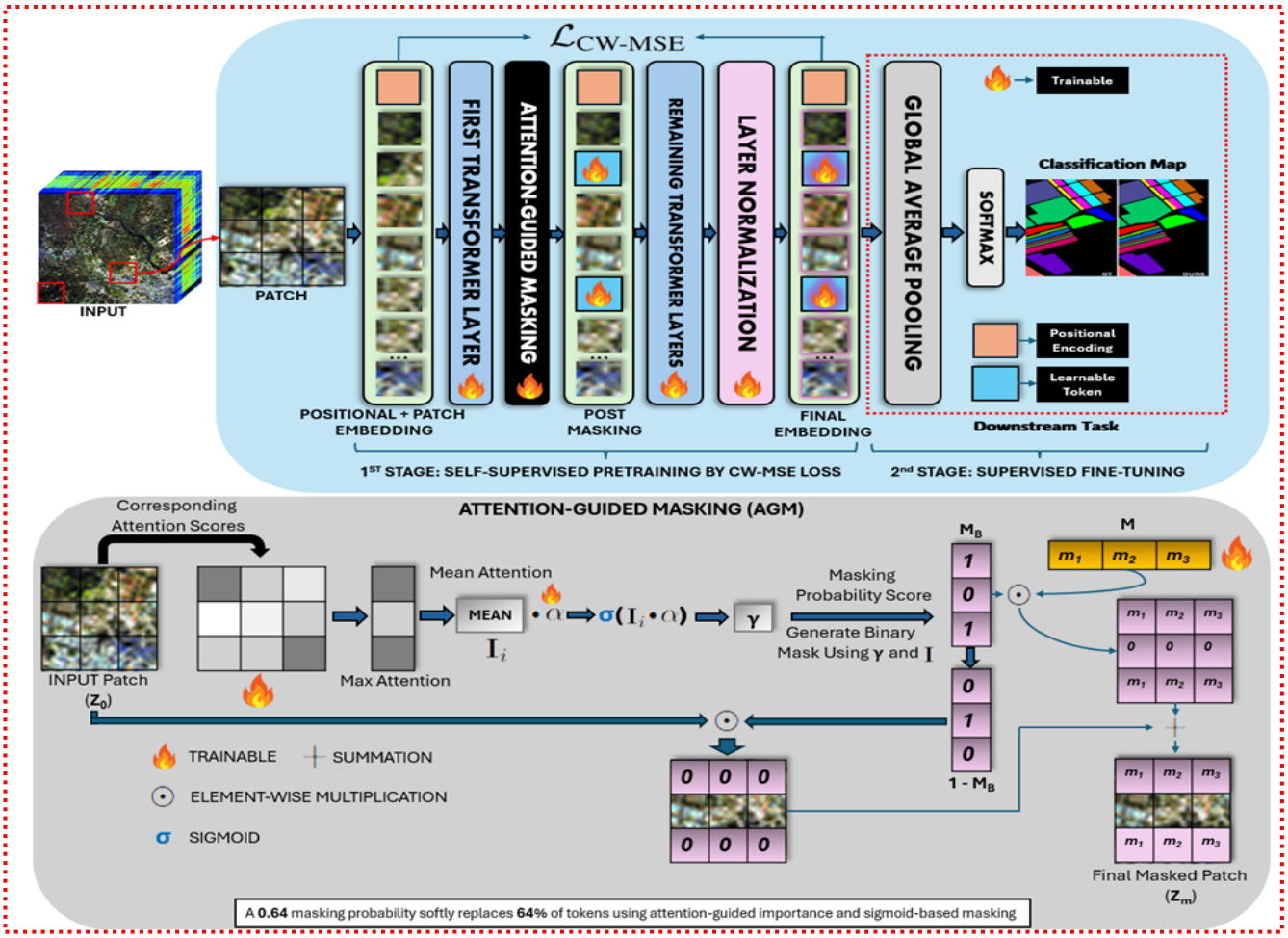


Figure 2. Architecture of the proposed Self-Supervised AGM-ViT. The top diagram illustrates the overall architecture of the model, while the bottom diagram depicts the proposed Attention-Guided Masking (AGM) mechanism.

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{d \times d}$ are learnable projections, HD is the number of attention heads, and $d_h = \frac{d}{HD}$ is the dimension of each head. Outputs from all heads are concatenated and linearly projected:

$$\text{MHSA}(\mathbf{Z}_0) = \text{Concat}(\text{head}_1, \ldots, \text{head}_{HD})\mathbf{W}_o, \quad (6)$$

where $\mathbf{W}_o \in \mathbb{R}^{d \times d}$ is the learnable output projection matrix that combines the concatenated outputs of all attention heads. The attention weights from this stage are stored to drive the AGM process in the next step.

2.3. Attention-Guided Masking (AGM)

To enhance the model's ability to focus on semantically salient regions during self-supervised learning, we propose Attention-Guided Masking (AGM), which leverages attention maps from the initial transformer layer to guide token masking adaptively.

Given the multi-head self-attention scores $\alpha_{ij}^{(h)}$ for token pair (i, j) in head h, we first compute the mean attention across HD heads:

$$\bar{\alpha}_{ij} = \frac{1}{HD} \sum_{h=1}^{HD} \alpha_{ij}^{(h)}. \quad (7)$$

Token importance $\mathbf{I}_i$ is defined as the average aggregated attention received by token i from all tokens:

$$\mathbf{I}_i = \frac{1}{N} \sum_{j=1}^{N} \bar{\alpha}_{ji}, \quad (8)$$

where N is the total number of tokens. This quantifies the contextual relevance of token i within the input sequence.

A sample-specific masking probability γ is then computed by applying a learnable scaling factor α to the average of the maximum outgoing attention weights per token:

$$\gamma = \sigma\left(\alpha \cdot \left(\frac{1}{N} \sum_{i=1}^{N} \max_j \bar{\alpha}_{ij}\right)\right), \quad (9)$$

where $\sigma(\cdot)$ denotes the sigmoid function, ensuring $\gamma \in (0, 1)$.

Using γ and token importance scores, a soft masking vector $\mathbf{M}_b \in [0, 1]^N$ is obtained:

$$\mathbf{M}_b = \sigma\left(\beta \cdot (\gamma - \mathbf{I})\right), \tag{10}$$

with sharpness hyperparameter $\beta = 2$, controlling the steepness of the sigmoid, which effectively modulates the probability of masking each token based on its relative importance.

Finally, the masked embedding $\mathbf{Z}_m$ is constructed as an element-wise interpolation between the original embeddings $\mathbf{Z}_0$ and a learnable mask token embedding $\mathbf{M} \in \mathbb{R}^{1\times d}$:

$$\mathbf{Z}_m = (\mathbf{1} - \mathbf{M}_b) \odot \mathbf{Z}_0 + \mathbf{M}_b \odot \mathbf{M}, \tag{11}$$

where $\odot$ denotes element-wise multiplication broadcasted over embedding dimensions d.

Role in Self-Supervised Learning: AGM enables dynamic, attention-driven soft masking that prioritizes retaining semantically critical tokens while softly masking less informative ones. This adaptive mechanism promotes efficient feature learning by focusing reconstruction objectives on informative patches, thus improving the quality of semantic representation in a label-free setting.

2.4. Confidence-Weighted Loss

The Vision Transformer processes input patches through a transformer stack, applying contextual masking after the first layer based on attention scores. The masking replaces less informative tokens with a learnable mask token, modulated by soft attention-derived confidence.

To train the model effectively, we use a confidence-weighted mean squared error (CW-MSE) loss between the masked representation and the original latent (pre-masked) output. A confidence mask $C_{b,i}$ down-weights uncertain (masked) tokens:

$$C_{b,i} = \begin{cases} \lambda_m + \epsilon_i, & \text{if token } i \text{ is unmasked,} \\ 1.0, & \text{otherwise,} \end{cases} \tag{12}$$

where $\lambda_m \in (0, 1]$ is a base confidence and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ represents noise.

The **Confidence-Weighted MSE loss** is defined as:

$$\mathcal{L}_{\text{CW-MSE}} = \frac{1}{BNd} \sum_{b=1}^{B} \sum_{i=1}^{N} C_{b,i} \sum_{j=1}^{d} \left(Z_{0,b,i,j} - Z^l_{m,b,i,j}\right)^2, \tag{13}$$

where B is the batch size, N the number of tokens, and d the embedding dimension; Z_0 denotes the original latent (pre-masked) representation, and Z^l_m the masked representation after transformer encoding.

A regularization term penalizes confidence variance:

$$\mathcal{R}_{\text{conf}} = \frac{1}{BN} \sum_{b=1}^{B} \sum_{i=1}^{N} \left(C_{b,i} - \mathbb{E}[C]\right)^2, \tag{14}$$

where

$$\mathbb{E}[C] = \frac{1}{BN} \sum_{b=1}^{B} \sum_{i=1}^{N} C_{b,i}. \tag{15}$$

The total loss is:

$$\mathcal{L} = \mathcal{L}_{\text{CW-MSE}} + \beta_{reg} \mathcal{R}_{\text{conf}}, \quad \text{where} \quad \beta_{reg} = 0.01. \tag{16}$$

During pretraining, AGM-ViT learns to reconstruct masked tokens without using labels. For fine-tuning, a classifier is trained on top of the frozen backbone using a small set of labeled data. Unlike the masked autoencoder [9], this formulation does not use a separate decoder; reconstruction occurs directly in the shared embedding space after masking and transformer encoding, enabling soft masking and confidence-regularized learning.

Algorithm 1 Attention-Guided Masking with Confidence-Weighted Mean Squared Error Loss

Require: Input $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, patch size p, embedding dimension d, mask token $\mathbf{M} \in \mathbb{R}^{1\times d}$, base confidence λ_m, sharpness β, regularization weight β_{reg}, number of attention heads HD.

1: Partition $\mathbf{X}$ into $N = \frac{HW}{p^2}$ patches, embed and add positional encoding: $\mathbf{Z}_0 = [\mathbf{z}_i] + \mathbf{P} \in \mathbb{R}^{N\times d}$.
2: Compute attention weights $\alpha^{(h)}_{ij}$ via transformer; average over heads: $\bar{\alpha}_{ij} = \frac{1}{HD} \sum_h \alpha^{(h)}_{ij}$.
3: Calculate token importance: $\mathbf{I}_i = \frac{1}{N} \sum_j \bar{\alpha}_{ji}$.
4: Compute masking probability: $\gamma = \sigma\left(\alpha \cdot \frac{1}{N} \sum_i \max_j \bar{\alpha}_{ij}\right)$.
5: Generate soft mask: $\mathbf{M}_b = \sigma(\beta \cdot (\gamma - \mathbf{I}))$.
6: Form masked embeddings: $\mathbf{Z}_m = (\mathbf{1} - \mathbf{M}_b) \odot \mathbf{Z}_0 + \mathbf{M}_b \odot \mathbf{M}$.
7: Pass $\mathbf{Z}_m$ through transformer to get $\mathbf{Z}^{(l)}_m$.
8: Set confidence weights: $C_{b,i} = \lambda_m + \epsilon_i$ if masked, else 1; $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.
9: Compute loss:

$$\mathcal{L}_{\text{CW-MSE}} = \frac{1}{BNd} \sum_{b,i,j} C_{b,i} (Z_{0,b,i,j} - Z^{(l)}_{m,b,i,j})^2,$$

$$\mathcal{R}_{\text{conf}} = \frac{1}{BN} \sum_{b,i} (C_{b,i} - \mathbb{E}[C])^2, \quad \mathbb{E}[C] = \frac{1}{BN} \sum_{b,i} C_{b,i}.$$

10: Total loss: $\mathcal{L} = \mathcal{L}_{\text{CW-MSE}} + \beta_{reg} \mathcal{R}_{\text{conf}}$.
11: **Pretraining**: Optimize the total loss $\mathcal{L}$ to reconstruct masked tokens without labels.
12: **Fine-tuning**: Freeze backbone; train classifier head on minimal labeled data for downstream tasks.

3. Experimental Setups

This section details the experimental setup of our approach to work on three benchmark hyperspectral datasets [1]: Indian Pines, Salinas, and Botswana. This is described below.

Datasets. We select three datasets of varying resolutions for a comprehensive analysis. Indian Pines (145×145, 220 bands, 16 classes) captures a landscape in Indiana, USA. Salinas (512×217, 224 bands, 16 classes) represents California's Salinas Valley. Botswana (145 bands, 14 classes) is from NASA's EO-1 satellite over the Okavango Delta. Additional details on datasets are provided in Appendix 6.1.

Pre-Processing. HSI data presents challenges from high spectral dimensionality and spatial variability. To eliminate this, we apply zero-padding to preserve spatial context at the image borders, followed by PCA [2] to reduce spectral dimensions to 25. Spectral-spatial patches are then extracted, and background regions with zero labels are removed to improve training relevance and efficiency.

Training the Model. AGM-ViT was trained on Indian Pines, Salinas, and Botswana datasets using patch size 5. The Adam optimizer [24] with an initial learning rate of 0.001 was applied for Salinas and Botswana, and 0.005 for Indian Pines. The CW-MSE loss assigned weights of 1.0 to masked patches and 1.2 to unmasked patches. For training, we use a batch size of 64 for Salinas and 32 for both Indian Pines and Botswana. A learning rate decay of 0.1 every 350 epochs over 800 epochs, and warm restarts (more details in Appendix 6.2) were applied at the 400th and 750th epochs for Indian Pines and Botswana. Salinas reached peak accuracy in 300 epochs without decay or restarts. For the Indian Pines, Botswana, and Salinas datasets, 15%, 10%, and 5% of the data were used for training, respectively. The experiments were carried out on an NVIDIA A100 GPU.

4. Analysis of Results

In this section, we comprehensively analyze results in various scenarios with three HSI datasets [1]: Indian Pines, Salinas, and Botswana using the two performance metrics: Overall Accuracy (OA) and Cohen's Kappa coefficient (κ).

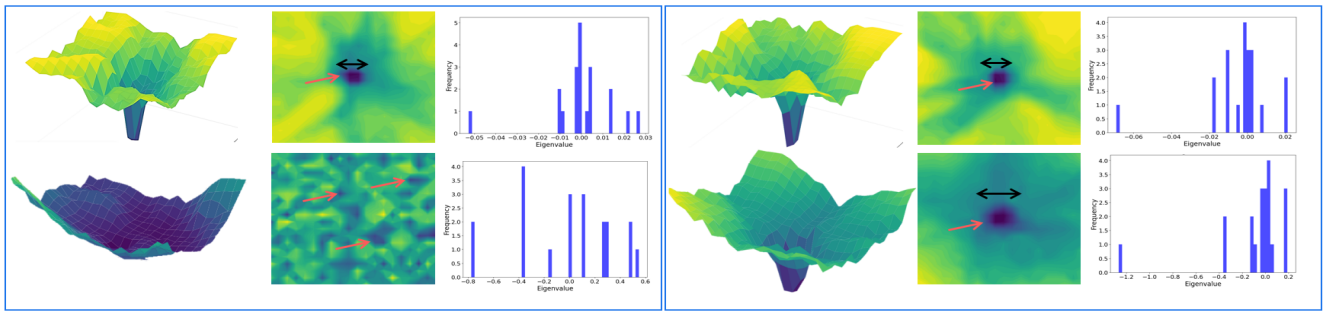
Comparison with Other SOTA Methods. As stated earlier, the results are compared with fourteen CNN and/or transformer-based SOTA methods (results of four SOTA methods are given in Appendix 6.3). Table 1 shows that, our model consistently outperforms SOTA supervised CNN and transformer-based methods across three widely used HSI datasets. Specifically, our model achieves overall accuracies (OA) of 97.45%, 99.87%, and 98.54% on Indian Pines, Salinas, and Botswana, respectively, with the corresponding Kappa coefficients of 97.44, 99.83, and 98.23. Compared to GSC-ViT [28], the best performing transformer-based supervised model, our method demonstrates notable improvements in OA: +0.33% on Indian Pines and +2.72% on Salinas. Even on the Botswana dataset, where GSC-ViT slightly outperforms our model by 0.31% in OA, our approach remains highly competitive, showcasing its robustness across different HSI scenarios. The best-performing model is denoted in **BOLD**, followed by second best and third best in BLUE and RED, respectively.

Parameter Efficiency. Our model excels in self-supervised HSI learning, capturing complex data representations with just 0.08M parameters. It outperforms CNN and transformer-based methods (Table 1), using fewer parameters than GSC-ViT [28] (0.10M) and significantly less than SSFTT [23] (0.95M) and GAHT [17] (0.97M).

Ablation Studies. The ablation studies presented in Tables 2, 3, and 4 comprehensively evaluate the sensitivity of the proposed model to key architectural and training hyperparameters on the Indian Pines (IP) and Salinas (S) datasets, where Indian Pines is challenging, while Salinas is a larger dataset. Table 2 investigates the effect of varying embedding dimension (d), number of attention heads (HD), and number of layers (L). The results demonstrate that increasing model capacity generally improves performance upto a certain threshold. Specifically, configuration C_7 ($d = 32, HD = 32, L = 6$) achieves the highest OA on both datasets, with 97.45% on Indian Pines and 99.87% on Salinas. Beyond this configuration, e.g., in C_9, performance slightly drops, suggesting diminishing returns and potential overfitting with excessive model complexity.

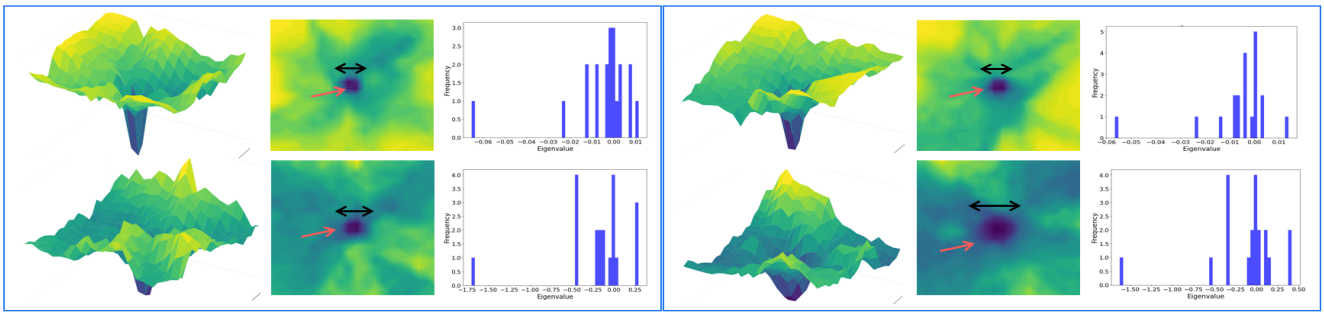
Table 3 explores the impact of batch size on classification accuracy. It is observed that a moderate batch size of 32 and 64 yields the best performance, reaching 97.45% OA on Indian Pines and 99.87% on Salinas, respectively. While smaller batch sizes (e.g., 8 and 16) perform competitively, very large batch sizes (128 and 256) lead to a noticeable decline in accuracy, particularly on Indian Pines. This trend indicates that excessively large batches may reduce gradient diversity and hinder effective generalization, whereas moderate batch sizes strike a balance between stability and performance.

Table 4 compares fixed manual masking probabilities with the proposed dynamic masking strategy utilizing learnable tokens. The proposed method outperforms all static masking configurations, achieving 97.45% OA on Indian Pines and 99.87% on Salinas. Although moderate static probabilities (e.g., 0.6 or 0.8) offer competitive results, extreme masking (e.g., 1.0) significantly degrades performance and highlights the importance of adaptive feature selection. The learnable dynamic masking mechanism proves to be more effective in capturing salient features by learning where to mask based on data-driven objectives. Altogether, the results underline the significance of carefully tuning model capacity, maintaining an optimal batch size, and employing a learnable masking mechanism to achieve state-of-the-art performance.



(a) (b)

Figure 3. Loss landscape analysis on Botswana: The left image shows the 3D loss landscape, the middle image shows the 2D loss contour, and the right image presents the Hessian eigenvalue distribution. (a) corresponds to 10% training data, while (b) corresponds to 20% training data. The upper and lower group, respectively, show the visualizations produced by AGM-ViT and supervised ViT.



(a) (b)

Figure 4. Loss landscape analysis on Indian Pines: The left image shows the 3D loss landscape, the middle image shows the 2D loss contour, and the right image presents the Hessian eigenvalue distribution. (a) corresponds to 10% training data, while (b) corresponds to 20% training data. The upper and lower group, respectively, show the visualizations produced by AGM-ViT and supervised ViT.

Analysis of Loss Landscape and Hessian Eigenvalue Distribution. The loss landscape analysis (Figures 3 and 4, respectively, for Botswana and Indian Pines) offers deep insights into optimization stability. In the 3D loss landscape, supervised ViT (lower rows in figures) exhibits multiple local minima of similar depth, indicating susceptibility to saddle points and shallow valleys. In contrast, AGM-ViT (upper rows in figures) forms a well-defined convex structure with a singular deep global optimum (in red arrows), ensuring stable convergence. ViT's scattered minima at 10% data settings reflect sensitivity to small perturbations, whereas AGM-ViT's structured masking and confidence-weighted loss enforce strong convexity, preventing suboptimal traps. In the 2D loss landscape, AGM-ViT's global minimum is sharply defined at the center, while ViT's minima appear irregularly scattered, forming a rugged surface. The diameter of each minimum is marked, where a larger span signifies greater confusion and delayed convergence—ViT's wider diameter confirms instability, while AGM-ViT's compact minimum ensures robustness in both data settings (10% and 20%). We also note that, for our approach, the diameter of the global optimum does not change even after altering the amount of training data, making it less susceptible to extreme data-scarce scenarios and making it a **scalable** and **lightweight** option for geoscience-related applications.

AGM-ViT's Hessian eigenvalues skew toward positive values show a strong convexity and stable optimization. While, ViT's symmetric eigenvalue spread suggests mixed curvature regions which, in turn, increases susceptibility to saddle points and instability. A predominance of positive eigenvalues ensures stable optimization, while a mix, as seen in ViT, leads to flat regions and unpredictability. This confirms AGM-ViT's superior stability and efficiency in low-data scenarios. The X-axis represents the eigenvalues of the Hessian matrix indicating curvature directions, while the Y-axis shows their frequency in the Hessian Eigenvalue Distribution Plot (Figures 3 and 4).

We also present the **qualitative assessments** of the classification maps predicted by AGM-ViT alongside ground truth for Botswana, Indian Pines, and Salinas (left to right in Figure 5). The strong visual alignment between predictions and ground truth (GT) shows the efficacy of our approach.

Generalization Ability. Figures 6 and 7 highlight the generalization abilities of supervised ViT and our self-

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

Methods		Parameters (M)	Indian Pines		Salinas		Botswana	
			OA (%)	κ	OA (%)	κ	OA (%)	κ
CNN-based								
2DCNN [15]	*IGARSS '16*	1.71	91.19	89.95	86.21	84.63	89.14	88.23
3DCNN [7]	*TGRS '18*	0.16	85.95	83.91	90.69	89.64	93.81	93.29
HybridSN [19]	*GRSL '19*	0.51	93.10	92.12	94.86	94.28	95.90	95.55
SPRN [26]	*TGRS '22*	0.18	90.84	89.56	93.49	92.76	96.60	96.32
Transformer-based								
SpectralFormer [11]	*TGRS '21*	0.34	78.84	75.80	90.00	88.87	81.31	79.76
SSFTT [23]	*TGRS '22*	0.95	93.15	92.18	94.72	94.13	96.35	96.05
GAHT [17]	*TGRS '22*	0.97	94.42	93.64	96.81	96.45	98.52	98.39
CAEVT [27]	*Sensors '22*	0.36	93.93	93.08	94.79	94.20	97.95	97.78
MorphFormer [20]	*TGRS '23*	0.19	94.96	94.25	96.21	95.79	97.88	97.70
GSC-ViT [28]	*TGRS '24*	0.10	97.12	96.67	97.15	96.47	**98.85**	**98.75**
OURS (AGM-ViT)		**0.08**	**97.45**	**97.44**	**99.87**	**99.83**	98.54	98.23
Δ			**+0.33**	**+0.77**	**+2.72**	**+3.36**	-0.31	-0.52

Figure 5. Comparison of classification maps produced by our AGM-ViT as against the ground truths (GT) on Botswana, Indian Pines, and Salinas, respectively from left to right.

Table 2. Ablation study on Indian Pines (IP) and Salinas (S): Impact of hyperparameter combinations C_i (embedding dimension d, number of heads HD, 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
HD	8	8	8	16	32	32	32	32	64
L	2	4	6	6	6	6	6	6	8
OA (IP) (%)	96.83	96.75	97.33	97.24	97.36	97.38	**97.45**	97.12	97.20
OA (S) (%)	98.91	99.29	99.32	99.03	99.39	99.52	**99.87**	99.69	99.18

Table 3. Ablation study with varying batch sizes for Indian Pines (IP) and Salinas (S).

Batch Size	8	16	32	64	128	256
OA (IP) (%)	97.12	97.19	**97.45**	97.22	95.45	94.52
OA (S) (%)	97.14	99.28	99.81	**99.87**	99.33	99.36

Table 4. Ablation study on the effectiveness of static manual masking versus the proposed dynamic masking with learnable tokens on Indian Pines (IP) and Salinas (S).

Masking Probability	0.2	0.4	0.6	0.8	1.0	**OURS**
OA (IP) (%)	96.41	96.75	96.83	97.19	92.05	**97.45**
OA (S) (%)	97.28	99.67	99.71	99.69	97.64	**99.87**

supervised AGM-ViT under varying data availability on the large dataset-Salinas. Please note that to validate the early convergence and generalization ability of both the models across different training setups, we perform experiments for the first 100 epochs. With 50% data, ViT achieves

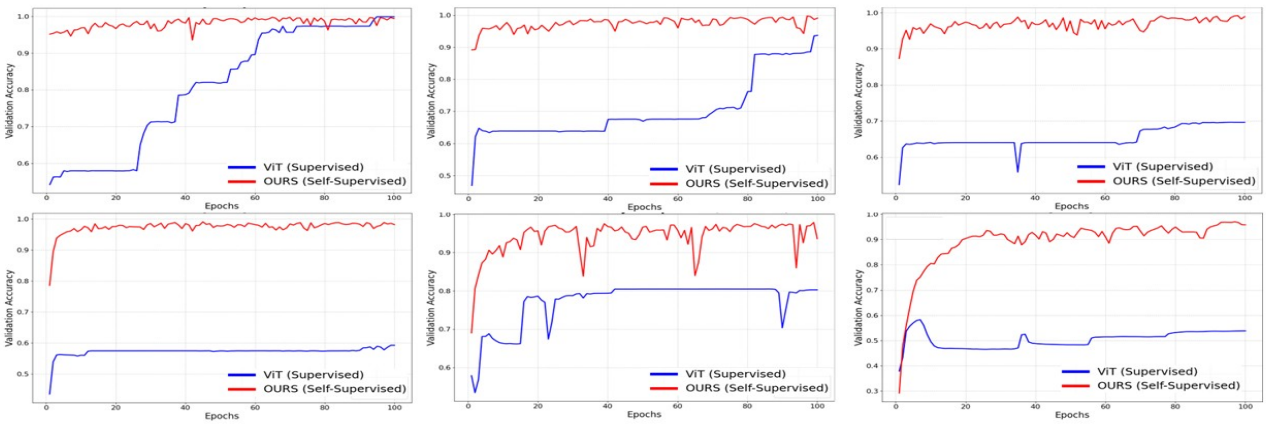


Figure 6. Accuracy comparison between supervised ViT and self-supervised AGM-ViT across varying training data percentages on Salinas. Subplots: **(Top-Left)** 50%, **(Top-Center)** 30%, **(Top-Right)** 20%, **(Bottom-Left)** 10%, **(Bottom-Center)** 5%, **(Bottom-Right)** 1%.

a marginally higher peak accuracy (99.92%) than AGM-ViT (99.44%), reflecting its ability to take advantage of abundant labeled data. However, as labeled data decreases, ViT's performance declines sharply. At 30%, ViT drops to 93.74%, while AGM-ViT remains at 99.07%. This gap expands further at 20% and 10%, with ViT falling to 69.64% and 59.29%, showing signs of overfitting and poor generalization, while AGM-ViT maintains 98.97% and 98.22% accuracy. Under extreme data scarcity (5% and 1%), ViT fluctuates, reaching 80.29% accuracy at 5% data but falls to 53.85% accuracy at 1% data, highlighting the instability of supervised learning with limited data. In contrast, AGM-ViT remains robust, achieving 93.67% accuracy at 5% data and 95.82% accuracy at 1% data, demonstrating its ability to extract meaningful features in low data settings. We can further observe the quick convergence of AGM-ViT within the first few epochs as compared to the fully supervised ViT.

t-SNE Visualizations. Figure 7 illustrates a clear distinction in feature representations between supervised ViT and our self-supervised AGM-ViT, using only 10% of labeled data across all three datasets. The supervised ViT exhibits entangled, snake-like patterns, indicative of overfitting to label-specific details and limited generalization. In contrast, the AGM-ViT forms compact, well-separated clusters with smooth transitions, capturing semantically meaningful and intrinsic structures. This shows the strength of self-supervised learning in disentangling high-dimensional hyperspectral features and improving generalization, particularly in low-data regimes critical to HSI tasks.

5. Conclusion

This paper presents a lightweight, self-supervised vision transformer framework for HSI classification, exploiting attention-guided dynamic masking and a confidence-weighted loss. By using early-layer attention to guide token masking, the model adaptively suppresses low-salience regions, enabling robust spectral-spatial feature learning. The proposed confidence-weighted reconstruction loss prevents overfitting by emphasizing masked tokens while maintaining stability across domains. With only 89,681 parameters, our model achieves strong performance and generalization across datasets. These results show the potential of attention-driven self-supervision for scalable and domain-adaptive HSI classification.

Figure 7. t-SNE visualization for all datasets with 10% training data: The top row shows supervised ViT results for Indian Pines (left), Botswana (middle), and Salinas (right). The bottom row shows AGM-ViT results for the same datasets.

Limitations and Future Scopes. While AGM-ViT delivers strong performance, its efficacy depends on the quality of data. It may struggle with noisy or incomplete data hampering its efficiency. Future work will explore multi-modal data integration to further improve generalization.

Acknowledgment

A part of this research has received support from the IEEE GRSS under "ProjNET" scheme.

References

[1] Hyperspectral data sets. https://lesun.weebly.com/hyperspectral-data-set.html. 5, 1

[2] H. Abdi and L J. Williams. Principal Component Analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2 (4):433–459, 2010. 5

[3] L. Alzubaidi, J. Zhang, A J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, et al. Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions. *Journal of Big Data*, 8:1–74, 2021. 1, 2

[4] A. Chatterjee, S. Ghosh, A. Ghosh, and E J. Ientilucci. Urbanscape-Net: A Spatial and Self-Attention Guided Deep Neural Network with Multi-Scale Feature Extraction for Urban Land-Use Classification. In *2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 4884–4889, 2024. 2

[5] 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. 1

[6] A. Ghosh, B. N. Subudhi, and S. Ghosh. Object Detection from Videos Captured by Moving Camera by Fuzzy Edge Incorporated Markov Random Field and Local Histogram Matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(8):1127–1135, 2012. 1

[7] 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. 1, 7

[8] 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*, pages 770–778, 2016. 1

[9] K He, X Chen, S Xie, Y Li, P Dollár, and R Girshick. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2, 4

[10] X. He, Y. Chen, and Z. Lin. Spatial-Spectral Transformer for Hyperspectral Image Classification. *Remote Sensing*, 13(3): 498, 2021. 2

[11] 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. 2, 7

[12] 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. 2, 1, 3

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012. 1

[14] 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. 1

[15] H. Lee and H. Kwon. Contextual Deep CNN Based Hyperspectral Classification. In *2016 IEEE International Geoscience and Remote Sensing Symposium*, pages 3322–3325, 2016. 1, 7

[16] 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. 1

[17] 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. 2, 5, 7

[18] 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. 2

[19] 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. 2, 7

[20] 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. 2, 7

[21] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*, 2015. 1

[22] 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, 2020. 2

[23] 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. 2, 5, 7

[24] 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. 5

[25] 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*, pages 5998–6008, 2017. 1, 2

[26] 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, 2022. 1, 7

[27] 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), 2022. 2, 7

[28] 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. 2, 5, 7

Adaptive Attention-Guided Masking in Vision Transformers for Self-Supervised Hyperspectral Feature Learning

Supplementary Material

6. Appendix

In recent years, **world models** like Vision Transformers (ViTs) [5] have demonstrated impressive performance in various computer vision [6, 8, 13] tasks, including hyperspectral image (HSI) classification [12]. However, their application to HSI faces several limitations due to the unique challenges of hyperspectral data.

Challenges. First, the high dimensionality of HSI, with several spectral bands, complicates processing for conventional ViTs, which are optimized for RGB images with only three channels. This necessitates extensive modifications to handle spectral complexity.

Second, data scarcity poses a significant challenge, as ViTs are data-hungry models that require large labeled datasets for effective training. However, in HSI, labeled data is limited due to the high cost and effort of manual annotation. Third, the computational complexity of ViTs is a concern, as the self-attention mechanism scales quadratically with the number of patches, making it resource-intensive for very high-resolution HSI data and less practical for real-time or resource-constrained applications like satellite-based remote sensing.

Fourth, ViTs struggle with spectral-spatial feature integration. While they excel at modeling long-range spatial dependencies, they are not inherently designed to process spectral information, and naive flattening of spectral data can result in loss of important spectral features. Lastly, ViTs are sensitive to noisy or redundant spectral bands, common in HSI, which can degrade model performance. Without effective preprocessing or dimensionality reduction, ViTs may overfit to noise or fail to robustly handle irrelevant spectral information.

Motivation. Our motivation stemmed from the limitations we encountered while applying **supervised Vision Transformer (ViT)-based models to hyperspectral image (HSI) classification.** Although ViTs excel in capturing long-range spatial dependencies, their performance in HSI was hindered by challenges as mentioned earlier.

While experimenting with supervised Vision Transformer (ViT)-based models for hyperspectral image (HSI) classification, one may encounter several technical challenges that will lead to suboptimal outcomes. The models may exhibit fluctuating loss curves and plateaued accuracy during both the training and validation phases. In certain cases (Figure 6), the models converge to local minima, leading to overfitting on small labeled datasets and poor generalization to unseen data. The need for large labeled datasets is particularly problematic, given that acquiring labeled hyperspectral data is expensive. Additionally, ViTs struggled to effectively integrate spectral-spatial features and are sensitive to noise and redundant spectral bands, leading to suboptimal generalization across diverse datasets.

Observation 1: Structural Scaling Law

Increasing the embedding dimension (d) and attention heads (HD) generally boosts accuracy, but only upto a point. Both Salinas and Indian Pines datasets achieve peak performance at $d = 32, HD = 32, L = 6$, with OAs of **99.87%** and **97.45%**, respectively. Beyond this, further scaling ($d = 64, HD = 64$) slightly degrades accuracy, hinting at an optimal capacity sweet spot where model complexity and generalization are best balanced.

Proposed Solution. To address these issues, we transitioned from a supervised to a self-supervised learning (SSL) paradigm. Our key novelty was the development of an **A**ttention-**G**uided **M**asking in **V**ision **T**ransformer (AGMViT) framework for HSI. The cornerstone of our approach is a dynamic masking strategy that adaptively selects patches based on attention scores, forcing the model to reconstruct less salient regions while learning from the most informative ones. This mechanism helps the model learn domain-invariant features from the intrinsic structure of the data itself, reducing reliance on labeled samples. We further introduced a confidence-weighted loss function that prioritizes robust learning from high-confidence regions, stabilizes training, and prevents overfitting to redundant spectral bands.

The best aspect of our solution is its ability to achieve state-of-the-art accuracy with a remarkably small parameter count, showing both efficiency and robustness. For instance, on the Salinas dataset, our model achieved an impressive 99.87% accuracy with just 89,681 parameters. More importantly, our model exhibited strong generalization capabilities across datasets, maintaining high performance despite being trained on limited unlabeled data.

6.1. Additional Details on the Datasets Used

Hyperspectral imaging datasets [1] are essential in remote sensing, providing detailed spectral information across hundreds of bands. Among the most frequently studied datasets are **Indian Pines**, **Salinas Scene**, and **Botswana**, each of-

Table 5. Comparison with additional state-of-the-art methods on Indian Pines and Salinas datasets.

Methods		Indian Pines		Salinas	
		OA (%)	κ	OA (%)	κ
SSAN [22]	*TGRS '20*	95.49	94.85	96.81	96.54
SST-FA [10]	*RS '21*	88.98	86.70	94.94	94.32
3DSA-MFN [18]	*RS '22*	96.02	94.78	99.72	99.13
SSSL [12]	*ICLR '23*	96.55	96.10	99.85	99.75
OURS		**97.45**	**97.44**	**99.87**	**99.83**
Δ		**+0.90**	**+1.34**	**+0.02**	**+0.08**

fering unique characteristics and applications. The **Indian Pines** dataset, collected by the AVIRIS sensor over Indiana, USA, contains **145×145** pixels and **220 spectral bands**, covering wavelengths from **0.4 µm to 2.5 µm**. It mainly consists of agricultural fields and forested areas, with **16 ground truth classes** and approximately **10,249 labeled samples**. Classification on this dataset is challenging due to class imbalance, high spectral similarity among crop types, and the presence of mixed pixels.

The **Salinas Scene** dataset, also captured by AVIRIS, represents agricultural land in California's Salinas Valley. It features higher spatial resolution with **512×217** pixels, **224 spectral bands** (excluding **20 bands** affected by water absorption), and **16 land-cover classes**, with a total of approximately **54,129 labeled samples**. Salinas Scene is the largest among the three in terms of both spatial resolution and labeled data, making it especially well-suited for detailed agricultural studies.

The **Botswana** dataset, acquired using NASA's Hyperion sensor aboard the EO-1 satellite, covers the Okavango Delta—an ecologically rich wetland. After removing water absorption bands, it includes **145 spectral bands** and is commonly cropped to **145×145** pixels from its original **256×1476** dimensions. It comprises **14 land cover classes** and around **3,248 labeled samples**. Although the smallest in terms of labeled data, Botswana exhibits high spectral variation due to the diverse natural vegetation and wetland features, making it particularly valuable for environmental monitoring.

Common challenges in working with hyperspectral imagery include high dimensionality, spectral redundancy, and difficulty in distinguishing between spectrally similar classes.

6.2. Additional Information on Warm Restart Learning Rate Scheduler Strategy

To optimize model convergence, we introduce a warm restart learning rate scheduler strategy. 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 new regions of the loss landscape. This cyclical scheduling approach effectively balances exploration and exploitation, facilitating more efficient training dynamics.

6.3. Comparison 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.

SSAN [22] 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 [10] 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. [18] 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 [12] proposed a self-supervised learning 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 [25]. 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.

Compared to the reconstruction approach proposed by [12], which minimizes pixel-wise distances in a fixed metric space, our method employs attention-guided dynamic masking to adaptively prioritize less salient regions, for better spectral-spatial feature learning. Additionally, in AGM-ViT, the learnable mask tokens enhance its ability to in-

fer complex, missing spectral information during training, leading to richer and more generalized feature representations. This dynamic, context-aware learning framework is seen to be effective at capturing complex hyperspectral correlations. In Table 5 it is seen that our method outperforms all these four approaches. We achieve a 0.90% increase in Overall Accuracy and a 1.34 improvement in Kappa score when compared to [12] on the Indian Pines dataset, with comparable results on the Salinas dataset.

Observation 2: Initialization-Invariant Loss Landscape in AGM-ViT

AGM-ViT exhibits an initialization-invariant loss topology, ensuring stable optimization regardless of weight initialization. Its continuous gradient flow prevents sharp curvatures and this leads to a more concentrated Hessian spectral density with fewer dominant eigenvalues. This results in consistent and efficient convergence. In contrast, supervised ViT has a rugged loss surface with sharp minima, making it sensitive to initialization and prone to suboptimal convergence at lower data settings.

Inference. Our work demonstrates the efficacy of integrating attention-guided dynamic masking within a Vision Transformer framework for hyperspectral image (HSI) classification. By considering attention-driven saliency to guide masking, the model effectively focuses on informative spectral-spatial features while enhancing its self-supervised learning capabilities. Dynamic masking not only improves representation learning but also addresses the inductive bias limitations commonly observed in transformer architectures.

Our method's consistent outperformance of other SOTAs across multiple datasets highlights its robustness, scalability, and potential as a ***new state-of-the-art solution for hyperspectral image classification***.

References

[1] Hyperspectral data sets. https://lesun.weebly.com/hyperspectral-data-set.html. 5, 1

[2] H. Abdi and L J. Williams. Principal Component Analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2 (4):433–459, 2010. 5

[3] L. Alzubaidi, J. Zhang, A J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, et al. Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions. *Journal of Big Data*, 8:1–74, 2021. 1, 2

[4] A. Chatterjee, S. Ghosh, A. Ghosh, and E J. Ientilucci. Urbanscape-Net: A Spatial and Self-Attention Guided Deep Neural Network with Multi-Scale Feature Extraction for Urban Land-Use Classification. In *2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 4884–4889, 2024. 2

[5] 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. 1

[6] A. Ghosh, B. N. Subudhi, and S. Ghosh. Object Detection from Videos Captured by Moving Camera by Fuzzy Edge Incorporated Markov Random Field and Local Histogram Matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(8):1127–1135, 2012. 1

[7] 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. 1, 7

[8] 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*, pages 770–778, 2016. 1

[9] K He, X Chen, S Xie, Y Li, P Dollár, and R Girshick. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2, 4

[10] X. He, Y. Chen, and Z. Lin. Spatial-Spectral Transformer for Hyperspectral Image Classification. *Remote Sensing*, 13(3): 498, 2021. 2

[11] 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. 2, 7

[12] 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. 2, 1, 3

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012. 1

[14] 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. 1

[15] H. Lee and H. Kwon. Contextual Deep CNN Based Hyperspectral Classification. In *2016 IEEE International Geoscience and Remote Sensing Symposium*, pages 3322–3325, 2016. 1, 7

[16] 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. 1

[17] 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. 2, 5, 7

[18] 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. 2

[19] 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. 2, 7

[20] 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. 2, 7

[21] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*, 2015. 1

[22] 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, 2020. 2

[23] 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. 2, 5, 7

[24] 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. 5

[25] 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*, pages 5998–6008, 2017. 1, 2

[26] 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, 2022. 1, 7

[27] 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), 2022. 2, 7

[28] 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. 2, 5, 7