

Context-Aware Masking and Learnable Diffusion-Guided Patch Refinement in Transformers via Sparse Supervision for Hyperspectral Image Classification

Abhiroop Chatterjee¹, Susmita Ghosh¹, Ashish Ghosh²

¹Jadavpur University ²IIT Bhubaneswar

{abhiroopc.cse.rs, susmita.ghoshde}@jadavpuruniversity.in, ash@isical.ac.in

Abstract

Vision transformers often struggle with sensitivity to spectral-spatial perturbations and inefficiencies in label-scarce regimes for hyperspectral image analysis. To address these, we introduce contextually perturbed diffusion-guided active learning (CPDGAL), integrating diffusion-guided feature refinement (DGFR) and contextualized masking (CM). The DGFR initially injects structured perturbations into patch embeddings and then reconstructs clean patches via a diffusion-based denoising mechanism. Through this, DGFR refines the learned features while implicitly calibrating the aleatoric uncertainty. The CM mechanism applies attention-guided probabilistic masking and enforces context-aware reconstruction to improve generalization. We also introduce a sparse supervision scheme for label-scarce scenarios that selects uncertain samples using confidence-aware ranking, prioritizing challenging data for efficient retraining through active learning (AL). Experiments on benchmark datasets validate the effectiveness of CPDGAL, achieving 97.34% overall accuracy (OA) on Indian Pines, 99.87% on Salinas, and 98.94% on Botswana with a lightweight architecture (0.09M parameters, 5.17 MFLOPs) and outperforms sixteen CNN/transformer-based SOTA methods. Our framework also generalizes better than the vision transformer in extreme low-label settings.

1. Introduction

Hyperspectral image (HSI) classification [17] is essential in various arenas of geoscience [37] and remote sensing [4, 15, 18] applications, but is challenged by high dimensionality thereby leading to computational inefficiency, model overfitting, and difficulty in capturing both spectral & spatial relationships simultaneously. Recent models, such as convolutional neural networks (CNNs) [3, 8, 14, 26], often struggle with these issues, requiring extensive preprocessing and large labeled datasets to achieve improved results. More recently, transformer-based [5, 32] 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] is their inability to model local patterns effectively, which are crucial for fine-grained spectral-spatial feature extraction. The lack of inductive bias towards local structures leads to a loss of contextual information, reducing their effectiveness in scenarios where spectral-spatial dependencies play a vital role. This highlights a critical gap in existing methodologies, which the present article attempts to address by introducing three key contributions listed below:

- **DGFR**: A spectral-spatial regularizer that injects structured noise to promote local continuity and improve generalization in ViTs.
- **CM**: An attention-guided masking strategy that enforces contextual reasoning by dynamically occluding salient regions that introduces inductive bias.
- **AL**: A confidence-aware sampling framework combining uncertainty-based selection with weighted supervision for efficient low-label learning.

To address the lack of inductive bias in ViTs, our framework injects localized priors via Diffusion-Guided Feature Refinement (**DGFR**) and Contextualized Masking (**CM**). DGFR adds structured noise to enforce spectral-spatial smoothness. Inspired by recent advances in masking for representation learning [9, 12, 33, 34], CM uses attention-guided dynamic masking to enhance contextual completion in the latent space. Combined with Active Learning (**AL**)-driven [21, 24] sample selection under low-label conditions, the model learns robust, structure-aware representations. The proposed architecture (Fig. 1) dynamically encodes spatial-spectral dependencies, effectively restoring locality in ViTs. Extensive experiments show state-of-the-art performance with only **87,538** parameters.

HSI classification has advanced rapidly with diverse methods addressing spectral-spatial challenges. Early models like 2-DCNN [16] used 2D convolutions and fully connected layers; SPRN [35] improved spatial features via attention and residual blocks. 3-DCNN [7] introduced 3D convolutions for joint spectral-spatial learning. HybridSN

[22] combined 2D and 3D CNNs for richer representations. Transformer-based methods marked a shift toward efficient feature extraction and long-range dependencies. GAHT [19] and MorphFormer [23] use CNN-Transformer hybrids and self-attention. Recent lightweight hybrids balance performance and efficiency: CAEVT [36] fuses 3D convolutional autoencoders with MobileViT, and GSC-ViT [38] integrates groupwise separable convolutions with attention. This evolution from CNNs [3] to transformers reflects the demand for advanced spectral-spatial modeling. SpectralFormer [11] and SSFTT [29] advance spectral sequence learning via multiscale aggregation and semantic tokenization. Other works blend CNNs and transformers for local-global balance: DATN [25] employs hybrid transformers and spectral local-conv blocks; DCTN [39] merges CNN local features with transformers using efficient interactive self-attention (EISA). Unlike prior models, our method excels in low-label regimes by jointly leveraging CM for semantic context modeling, DGFR for learnable spectral-spatial denoising, and a confidence-aware sample ranking strategy integrated into an active learning (AL) loop, forming a unified, end-to-end trainable framework with strong generalization across HSI benchmarks.

In this paper, Section 2 details the methodology used. Section 3 outlines the experimental settings, followed by Section 4, analyzing the results, ablation studies, and generalization tests. Section 5 concludes the article.

2. Methodology

Given an input image $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, where H and W denote spatial dimensions and C represents the number of spectral bands, our objective is to learn a predictive function $f_\theta : \mathbf{X} \rightarrow \hat{y}$, that generalizes efficiently from a set D_L consisting of N_L number of limited labeled samples; $D_L = \{(\mathbf{X}_i, y_i)\}_{i=1}^{N_L}$, where $N_L \ll N_U$ (N_U : size of the original training dataset) and y is the true label. To do so, the primary challenges include sensitivity to noise, robustness to missing information, and labeled data scarcity. This necessitates a framework (Fig. 1) that enhances feature representations, enforces structural consistency, and optimally selects samples for labeling. To address this, we introduce DGFR (Fig. 2 (a)), CM (Fig. 2 (b)), and AL, which, respectively, enhance feature consistency, handle missing information, and reduce dependency on large labeled datasets. The complete process is detailed below:

2.1. Patch Embedding and Positional Encoding

Initially, a given input image is divided into patches, and positional encoding is augmented for spatial context. The input image \mathbf{X} is partitioned into N non-overlapping patches each of size $P \times P \times C$ where

$$N = \frac{H \times W}{P^2}. \quad (1)$$

The i^{th} patch of image X , denoted as, X_i ($i = 1, 2, \dots, N$ and $\mathbf{X}_i \in \mathbb{R}^{P^2 C}$) is transformed into a latent feature representation, Z_i , through a learnable linear projection:

$$\mathbf{Z}_i = \mathbf{W}_e \mathbf{X}_i + \mathbf{b}_e, \quad \forall i \in \{1, 2, \dots, N\}, \quad (2)$$

where, D is the feature embedding dimension in latent space, $\mathbf{W}_e \in \mathbb{R}^{D \times (P^2 C)}$ is the learnable projection matrix, $\mathbf{b}_e \in \mathbb{R}^D$ is the bias term. The resulting matrix $\mathbf{Z} \in \mathbb{R}^{N \times D}$ serves as the initial feature representation of the image.

To retain spatial location, for each patch, we augment a learnable **positional encoding** $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{N \times D}$. The final patch embeddings in latent space are obtained as:

$$\mathbf{Z} \leftarrow \mathbf{Z} + \mathbf{E}_{\text{pos}}. \quad (3)$$

Here, \mathbf{E}_{pos} is learned jointly with the model parameters to enable spatially aware feature learning, ensuring the transformer differentiates patches based on their original positions in the image.

2.2. Diffusion-Guided Feature Refinement (DGFR)

To improve representation robustness and enhance generalization in low-supervision scenarios, we propose **Diffusion-Guided Feature Refinement (DGFR)**, a novel module inspired by score-based diffusion processes [6, 28]. DGFR operates directly on the latent token embeddings extracted from the early stages of a ViT [5], applying a learnable, structured stochastic perturbation to these embeddings. Subsequently, it utilizes a timestep-conditioned denoising network to recover refined, noise-aware representations. This enables controlled local exploration of the latent feature space and promotes smoothness and invariance while preserving the critical semantic structure inherent in the representations.

Formally, $\mathbf{Z} \in \mathbb{R}^{N \times D}$ denote the latent token embeddings obtained after patch embedding and positional encoding in a ViT [5]. This representation \mathbf{Z} serves as the input to the DGFR module, which applies a learnable Gaussian diffusion kernel to perturb \mathbf{Z} , generates a corrupted latent \mathbf{Z}_t at diffusion timestep t . DGFR then estimates a refined representation $\mathbf{Z}_{\text{final}}$ by passing \mathbf{Z}_t through a timestep-conditioned residual denoiser, \mathcal{D}_θ . The refined embeddings $\mathbf{Z}_{\text{final}}$ are subsequently forwarded to the following transformer layers for further processing. The end-to-end process of this task is detailed below:

Learnable Diffusion Process. In DGFR, we define a forward corruption process, with a discrete, learnable noise injection schedule $\{\alpha_t\}_{t=1}^T$, jointly optimized with the network weights. At each timestep t , the corrupted latent representation is computed as:

$$\mathbf{Z}_t = \alpha_t \mathbf{Z} + (1 - \alpha_t) \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (4)$$

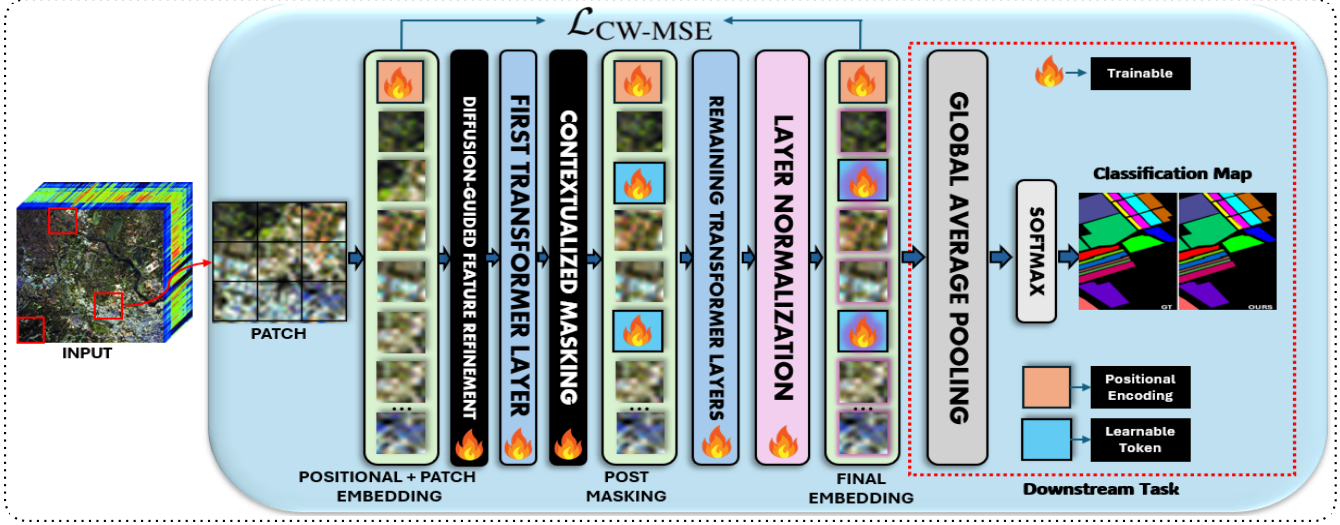


Figure 1. Architecture of the proposed model, CPDGal.

where, \mathbf{I} is the identity covariance matrix. The scalar $\alpha_t \in (0, 1)$ controls the relative weighting between the clean latent features, Z and injected Gaussian noise, ϵ . To prevent degenerate cases, α_t is clipped within a reasonable range during training.

Unlike fixed noise schedules in classical diffusion models, here, the learnable α_t allows the model to tune the noise intensity per timestep, facilitating task-aware corruption. This learned corruption prevents excessive noise injection that could damage crucial semantic features. Optimizing noise schedule also mitigates under-/over-regularization common in fixed-noise methods, dynamically balancing the bias-variance trade-off throughout training.

Timestep Embedding and Conditioning. To make the denoising network, \mathcal{D}_θ (implemented as an MLP) aware of the level of corruption, the diffusion timestep $t \in \{1, 2, \dots, T\}$ is first normalized by the total number of steps T , and then mapped through a learnable function $\phi: [0, 1] \rightarrow \mathbb{R}^D$, to obtain a timestep embedding:

$$\mathbf{e}_t = \phi\left(\frac{t}{T}\right), \quad \mathbf{e}_t \in \mathbb{R}^D. \quad (5)$$

This embedding is broadcast across all N tokens and added element-wise to the corrupted latent representation \mathbf{Z}_t as:

$$\mathbf{Z}_t \leftarrow \mathbf{Z}_t + \mathbf{e}_t. \quad (6)$$

Conditioning on \mathbf{e}_t allows the denoiser, \mathcal{D}_θ to adapt its refinement strategy according to the corruption level, which varies in each timestep. This prevents the model from collapsing into a timestep-invariant solution and preserves a rich continuum of denoising behaviors.

Residual Timestep-Conditioned Denoising Network. The denoiser, \mathcal{D}_θ operates on the corrupted latent representation \mathbf{Z}_t across N image patches. To refine the corrupted

input \mathbf{Z}_t , we adopt a residual architecture of the form

$$\mathcal{D}_\theta(\mathbf{Z}_t, \mathbf{e}_t) = \mathbf{Z}_t + \mathcal{F}_\theta(\mathbf{Z}_t, \mathbf{e}_t), \quad (7)$$

where $\mathcal{F}_\theta: \mathbb{R}^{N \times D} \times \mathbb{R}^D \rightarrow \mathbb{R}^{N \times D}$ is a learnable nonlinear function realized via stacked denoising blocks. This residual formulation guides the model to learn corrective updates instead of reconstructing features from scratch, thus preserving semantic information and enhancing representation fidelity & faster convergence. Further, the residual pathway stabilizes training by maintaining identity mapping at initialization, reducing issues such as vanishing gradients.

Adaptive Skip Integration. To reconcile the trade-off between original and refined features, DGFR employs a convex combination of Z & Z_t to produce $\mathbf{Z}_{\text{final}}$ as follows:

$$\mathbf{Z}_{\text{final}} = \lambda_t \mathcal{D}_\theta(\mathbf{Z}_t, \mathbf{e}_t) + (1 - \lambda_t) \mathbf{Z}, \quad (8)$$

where the mixing coefficient $\lambda_t := \min(1, \frac{t}{T})$ smoothly interpolates from unaltered to fully denoised representations across diffusion steps.

The soft skip connection (Eq. 8) prevents abrupt shifts in representation space and stabilize gradients throughout training. Early in the diffusion process ($\lambda_t \approx 0$), reliance on the clean features \mathbf{Z} avoids corruption-induced degradation and sharp parameter updates. As $\lambda_t \rightarrow 1$, the model gradually shifts toward denoised representations $\mathcal{D}_\theta(\mathbf{Z}_t, \mathbf{e}_t)$, enhancing robustness. This convex interpolation smooths feature transitions and preserves gradient continuity, contributing to a more stable loss landscape with reduced non-convexity.

2.3. Contextualized Masking (CM)

Transformer Encoder: Attention Estimation. The initial transformer encoder models global dependencies in the

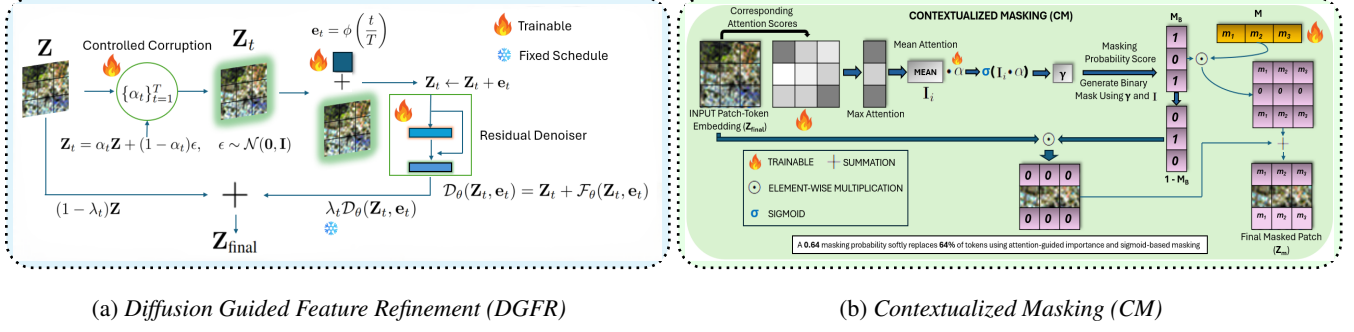


Figure 2. Flow diagrams of the (a) DGFR and (b) CM components within CPD GAL.

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:

$$\mathbf{Q} = \mathbf{Z}_{final} \mathbf{W}_q, \quad \mathbf{K} = \mathbf{Z}_{final} \mathbf{W}_k, \quad \mathbf{V} = \mathbf{Z}_{final} \mathbf{W}_v, \quad (9)$$

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax} \left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h}} \right) \mathbf{V}, \quad (10)$$

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:

$$\text{MHSA}(\mathbf{Z}_{final}) = \text{Concat}(\text{head}_1, \dots, \text{head}_{HD}) \mathbf{W}_o, \quad (11)$$

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 CM process in the next step.

Contextualized Masking (CM). To enhance the model’s ability to focus on semantically salient regions during self-supervised learning, we propose CM, which uses 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:

$$\bar{\alpha}_{ij} = \frac{1}{HD} \sum_{h=1}^{HD} \alpha_{ij}^{(h)}. \quad (12)$$

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

$$\mathbf{I}_i = \frac{1}{N} \sum_{j=1}^N \bar{\alpha}_{ji}, \quad (13)$$

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:

$$\gamma = \sigma \left(\alpha \cdot \left(\frac{1}{N} \sum_{i=1}^N \max_j \bar{\alpha}_{ij} \right) \right), \quad (14)$$

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:

$$\mathbf{M}_b = \sigma(\beta \cdot (\gamma - \mathbf{I})), \quad (15)$$

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

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

$$\mathbf{Z}_m = (\mathbf{1} - \mathbf{M}_b) \odot \mathbf{Z}_{final} + \mathbf{M}_b \odot \mathbf{M}, \quad (16)$$

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

2.4. Optimization Function

The Confidence-Weighted Mean Squared Error (**CW-MSE**) loss supervises the training of DGFR and CM by weighting reconstruction error based on token confidence. Let \mathbf{Z}_m^l denote the refined encoder output for the original embedding \mathbf{Z} ; the loss is defined as:

$$\mathcal{L}_{\text{CW-MSE}} = \frac{1}{BN} \sum_{b=1}^B \sum_{i=1}^N CS_i \|\mathbf{Z}_i - \mathbf{Z}_{m,i}^l\|^2, \quad (17)$$

where B is the batch size. Confidence scores CS_i modulate i^{th} token’s contribution:

$$CS_i = \begin{cases} 1.2, & \text{if unmasked (high confidence)} \\ 1.0, & \text{if masked (low confidence)} \end{cases} \quad (18)$$

DGFR focuses the loss on refining unmasked tokens to preserve semantic fidelity, while **CM** encourages reconstruction of contextually important tokens from partial inputs. CW-MSE emphasizes high-confidence regions while tolerating noise which is critical in low-label settings.

2.5. Training the Model through Active Sampling

Guided by our CW-MSE loss, we enhance label efficiency through **Active Learning (AL)**, that determines *which* unlabeled samples to annotate for maximum utility under budget constraints. While CW-MSE refines learning from existing labels, AL iteratively selects the most uncertain samples from the unlabeled pool for annotation, refining the model in each step. The process is as follows:

Prediction on Unlabeled Data. After initial training phase, the model predicts on remaining unlabeled data $\mathbf{X}_{\text{unlabeled}}$. The model outputs predicted probabilities for each sample, which are used to compute a confidence margin as: $U(\mathbf{x}) = \mathbf{P}(\mathbf{top}) - \mathbf{P}(k^{\text{th}} \mathbf{top})$, where $P(\mathbf{top})$ denotes the highest predicted probability and $P(k^{\text{th}} \mathbf{top})$ the probability of the k^{th} topmost probable class (here, $k = 3$).

Selection of Uncertain Samples. Samples with the smallest confidence margin $U(x)$ are considered the most uncertain and challenging. A fixed proportion of these samples is selected to be labeled in each iteration.

Retraining Phase. The newly labeled samples are added to the training set, and the model is retrained on this expanded dataset. This process is repeated over several active learning cycles to accelerate learning under limited annotation budgets.

3. Experimental Setups

Our experiments use three benchmark hyperspectral datasets: Indian Pines (IP), Salinas, and Botswana [1].

Datasets. IP (145×145, 220 bands, 16 classes), Salinas (512×217, 224 bands, 16 classes), and Botswana (145 bands, 14 classes) from NASA’s EO-1 satellite over Okavango Delta were used. More details are in [Appendix 6.1](#).

Preprocessing: To address high spectral dimensionality and spatial variability, we apply zero-padding for border preservation, reduce spectral dimensions to 15 bands via PCA [2], and remove zero-labeled background regions.

Training the Model. For all the datasets, CPDGAL was trained using Adam optimizer [30] with an initial learning rate of 0.001. To calculate CW-MSE loss, we assign weights of 1.0 to masked patches and 1.2 to unmasked patches. A batch size of 64 (Botswana) and 32 (Indian Pines and Salinas) was used. For both *Indian Pines* and *Botswana*, 10% training and 90% hold-out test data is considered. From the training corpus, 5% is randomly selected as the initial labeled data for model pre-training. In each subsequent active learning cycle, *confidence-deviation*

sampling provides the uncertainty of the remaining training samples (for which labels are not used at this stage). Out of these 2% budget is selected, annotated by the oracle, and incorporated into retraining. For *Salinas*, the dataset is split into 2% training and 98% hold-out test data. From the training subset, 10% is used for initial labeled pre-training, and in each active learning cycle, a further 10% of the remaining training pool is selected, annotated, and used for retraining. A maximum of 30 active learning cycles is performed for *Indian Pines* and *Botswana*. For *Salinas*, 10 cycles were done. For retraining during active learning cycles, each phase was run for 200 epochs for *Indian Pines* and *Botswana*, and for 50 epochs for *Salinas*. Optimal performance was achieved with substantially less labeled data compared to state-of-the-art methods. Specifically, *Indian Pines* required only about 4.7% labeled data (initial + actively selected), *Botswana* about 4.3%, and *Salinas* about 1%, whereas other methods use 5%, 5%, and 1% training data for *Indian Pines*, *Botswana*, and *Salinas*, respectively. The experiments were carried out on an NVIDIA A100 GPU. For more details, please see [Appendix 6.2](#).

4. Analysis of Results

We analyze results using Overall Accuracy (OA), Average Accuracy (AA), and Cohen’s Kappa (κ). Details are given in [Appendix 6.3](#).

Comparison with SOTA Methods. [Table 1](#) presents a comparison of our method against several SOTA CNN/transformer-based models across three benchmark HSI datasets. CPDGAL achieves the highest OA, AA, and κ scores on all datasets, surpassing existing methods in both accuracy and consistency. Notably, it surpasses the previous best OA on *IP* by **+0.16%**, on *Salinas* by **+0.92%**, and on *Botswana* by **+0.09%**, while κ gains are similarly strong. Further, on *IP*, CPDGAL achieves an AA of 96.74%, surpassing the previous best (GSC-ViT, 94.34%) by a margin of +2.40%. On *Salinas*, CPDGAL attains 99.41%, exceeding the top SOTA (DATN/DCTN, 99.05%) by +0.36%. Similarly, for *Botswana*, CPDGAL reaches 99.03%, outperforming the best competitor (GSC-ViT, 98.90%) by +0.13%. In contrast to models like GAHT [19], GSC-ViT [38], or DATN [25], which excel only on some datasets when compared amongst competitive techniques, CPDGAL exhibits robust generalization across all benchmarks. These findings corroborate the method’s robustness in preserving class-wise prediction quality while operating with significantly fewer parameters (**0.09M/0.333MB**) compared to existing transformer-based architectures. Best scores are highlighted in **BOLD**, with **second best** and **third best** noted. More results are provided in [Appendix 6.4](#).

Parameter Efficiency. CPDGAL achieves superior accuracy with only **0.09M parameters** (0.333 MB, 5.17 MFLOPs), outperforming larger models like DATN [25]

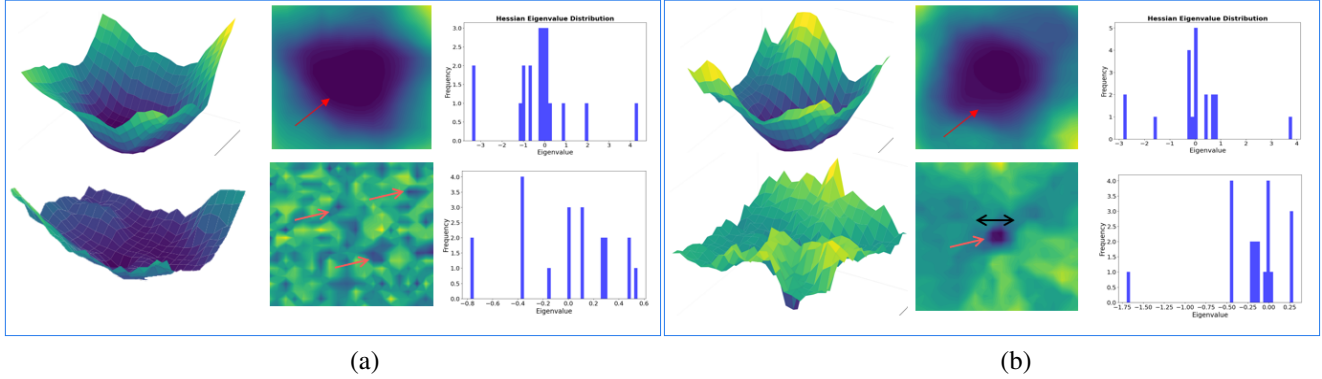


Figure 3. Loss landscape analysis on (a) Botswana and (b) Indian Pines on 5% training data. In both (a) and (b), the left image shows the 3D loss landscape, the middle image shows the 2D loss contour, and the right image presents the Hessian eigenvalue distribution. The upper and lower groups, respectively, show the visualizations produced by CPDGAL and supervised ViT.

(3.31 MB) and DCTN [39] (45.32 MB), as well as lightweight GSC-ViT [38] (0.10M) as shown in **Table 1**. Its compact design can enable strong generalization and efficient deployment in resource-limited remote sensing.

Ablation Study. We examine the impact of hyperparameters, batch size, masking, diffusion timesteps, and model components on performance for Botswana. **Table 2** shows that smaller configurations $(D, h, L) = (8, 8, 2)$ yield lower OA (94.94%), while increasing them improves accuracy. More complexity (64, 64, 8) brings diminishing returns, with the best OA **98.94%** at (32, 8, 6), highlighting the need to balance expressiveness & overfitting. **Table 3** shows batch size B as 64 peaks at 98.94%, but B as 128/256 slightly reduces accuracy, likely from reduced gradient stochasticity. Masking strategy is important: static masking underperforms dynamic learnable masking, which reaches 98.94% OA (**Table 4**), proving an advantage of adaptive token selection. For DGFR timesteps (**Table 5**), moderate $T = 100$ works best; too few ($T = 5, 10$) fail to refine features, while excessive diffusion ($T = 500, 1000$) slightly degrades results. Component-wise (**Table 6**), DGFR and CM individually outperform the ViT, with DGFR adding +2.43%. Combined, they achieve the highest OA (98.94%), and confirm their complementarity. *Key findings:* (i) Complexity helps until overfitting. (ii) $B = 64$ best balances stability and generalization. (iii) Adaptive masking is better than static masking. (iv) $T = 100$ optimally refines features. (v) DGFR+CM deliver the top results.

Loss Landscape and Hessian Eigenvalue Analysis. We compare loss landscapes for Botswana and IP (5% data) using supervised ViT [5] and CPDGAL. In Fig. 3, CPDGAL (upper rows) shows a smooth, convex 3D surface with a singular global optimum, enabling stable convergence. ViT (lower rows) presents multiple local minima of similar depth, and shows susceptibility to saddle points. In 2D, CPDGAL forms a broad, centralized flat min-

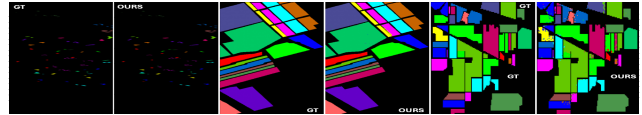


Figure 4. Classification maps produced by CPDGAL and their ground truths (GT) on Botswana, Salinas, and IP (left to right).

imum—linked to better generalization—while ViT shows rugged, scattered or sharply peaked minima, suggesting overfitting risk. CPDGAL’s topology is seen to be dataset- and initialization-invariant, converging to similar optima across runs. Hessian eigenvalue distributions (bar charts in Fig. 3) further confirm CPDGAL’s robustness. For ViT, eigenvalues skew negative, reflecting sharp, unstable minima. CPDGAL’s distribution is symmetric around zero, slightly skewed positive, indicating flatter, more stable minima. Also, CPDGAL’s eigenvalue magnitudes (e.g., 1, 2) far exceed ViT’s (e.g., 0.2, 0.4), evidencing a more expressive feature space with well-separated decision boundaries.

Analysis of Classification Maps. Fig. 4 compares ground truth and CPDGAL predictions for Botswana, Salinas, and Indian Pines (left to right), showing strong visual alignment that supports our method’s efficacy.

t-SNE Visualizations. t-SNE [31] in Fig. 5 compares CPDGAL with supervised ViT (5% data) across all datasets. ViT shows entangled patterns, suggesting overfitting, while CPDGAL yields compact, well-separated clusters, demonstrating active learning’s role in improving feature disentanglement for high-dimensional hyperspectral data.

Retraining through Active Learning. Convergence trends and limited-sample pre-training effects are, respectively, given in [Appendix 6.5](#) and [Appendix 6.6](#).

Expected Model Improvement (EMI) Analysis. We have evaluated the efficiency of our active learning strategy by computing the Expected Model Improvement (EMI)

Table 1. Comparison with various SOTA methods on several HSI benchmarks. OA: Overall Accuracy, AA: Average Accuracy, κ : Kappa.

Methods	Parameters (M) / (MB)	Indian Pines			Salinas			Botswana			
		OA (%)	AA (%)	κ	OA (%)	AA (%)	κ	OA (%)	AA (%)	κ	
CNN-based											
2DCNN [16]	IGARSS '16	1.71	91.19	87.42	89.95	86.21	85.68	84.63	89.14	88.90	88.23
3DCNN [7]	TGRS '18	0.16	85.95	80.11	83.91	90.69	92.98	89.64	93.81	93.19	93.29
HybridSN [22]	GRSL '19	0.51	93.10	88.43	92.12	94.86	96.88	94.28	95.90	95.99	95.55
SPRN [35]	TGRS '22	0.18	90.84	82.18	89.56	93.49	95.10	92.76	96.60	96.70	96.32
Transformer-based											
SpectralFormer [11]	TGRS '21	0.34	78.84	70.28	75.80	90.00	91.64	88.87	81.31	82.08	79.76
SSFTT [29]	TGRS '22	0.95	93.15	90.03	92.18	94.72	96.65	94.13	96.35	96.84	96.05
GAHT [19]	TGRS '22	0.97	94.42	83.63	93.64	96.81	98.09	96.45	98.52	98.45	98.39
CAEVT [36]	Sensors '22	0.36	93.93	89.81	93.08	94.79	96.58	94.20	97.95	98.00	97.78
MorphFormer [23]	TGRS '23	0.19	94.96	89.74	94.25	96.21	97.50	95.79	97.88	98.12	97.70
GSC-ViT [38]	TGRS '24	0.10	97.12	94.34	96.67	97.15	97.99	96.47	98.85	98.90	98.75
DATN [25]	EAAI '24	3.31	97.18	93.61	96.78	98.95	99.05	98.83	96.40	95.95	96.10
DCTN [39]	TGRS '24	45.32	96.76	93.10	96.30	98.11	99.05	97.89	97.18	97.39	96.95
OURS (CPD GAL)	0.09 M / 0.333 MB		97.34	96.74	97.21	99.87	99.41	99.81	98.94	99.03	98.82
Δ			+0.16	+2.40	+0.43	+0.92	+0.36	+0.98	+0.09	+0.13	+0.07

Table 2. Ablation study on Botswana: Impact of hyperparameter (H-P) combinations C_i (embedding dimension D , number of heads h , and layers L) on overall accuracy (OA).

H-P	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
D	8	8	8	8	8	16	32	32	64	64
h	8	8	8	16	32	8	8	16	32	64
L	2	4	6	6	6	6	6	6	6	8
OA (%)	94.94	96.79	96.83	96.48	96.70	97.32	98.94	98.50	97.05	97.27

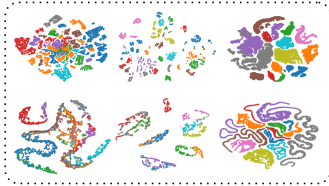


Figure 5. t-SNE with 5% data for Indian Pines, Botswana, and Salinas (left to right): CPD GAL (top), Supervised ViT (bottom)

across multiple labeling budgets per cycle for 0.5%, 1%, and 2% of the unlabeled pool. We vary it systematically to study the trade-off between labeling cost and performance gain. Fig. 6 shows the EMI trajectories and depicts the trade-offs between acquisition size, stability, and marginal gains over active learning cycles. For the **0.5% acquisition budget**, the EMI peaks at 0.0214 in the first cycle, then declines with a minor spike near the 7th cycle (>0.010). After the 10th cycle, EMI oscillates around a diminishing-returns threshold, reflecting reduced gains from limited new data. With **1% acquisition**, EMI starts at 0.0468, drops sharply, then shows cyclical fluctuations between the 5th–10th cycles as the model balances new information with convergence. For **2% acquisition**, EMI begins at 0.0340, declines, and then stabilizes after the 20th cycle, benefiting from larger batches that reduce variance and improve representativeness. More details are given in Appendix 6.7.

Table 3. Ablation study with varying batch sizes for Botswana.

Batch Size	8	16	32	64	128	256
OA (%)	96.17	97.42	98.73	98.94	98.55	98.50

Table 4. Performance analysis on static manual masking versus proposed dynamic masking with learnable tokens on Botswana.

Masking Probability	0.2	0.4	0.6	0.8	OURS
OA (%)	97.64	98.02	98.42	97.05	98.94

Table 5. Performance across different timesteps in the forward diffusion process of the DGFR module on Botswana.

Timestep	5	10	100	500	1000
OA (%)	98.20	98.28	98.94	98.42	98.59

Table 6. Ablation study on Botswana. OA: Overall Accuracy. First row: supervised Vision Transformer without DGFR or CM.

Components		OA (%)
CM	DGFR	
X	X	95.77 (↓ 3.17%)
✓	X	97.64 (↓ 1.30%)
X	✓	98.20 (↓ 0.74%)
✓	✓	98.94

Evolution of Entropy Distribution during Active Cycles. The entropy distribution heatmap (Fig. 7) captures the evolution of model uncertainty in the *remaining unlabeled pool* over active learning cycles. Initially, the distribution is skewed towards higher-entropy bins, indicating substantial predictive uncertainty. As labeling progresses, the pool size decreases and mass shifts toward lower-entropy bins, reflecting growing model confidence. A persistent

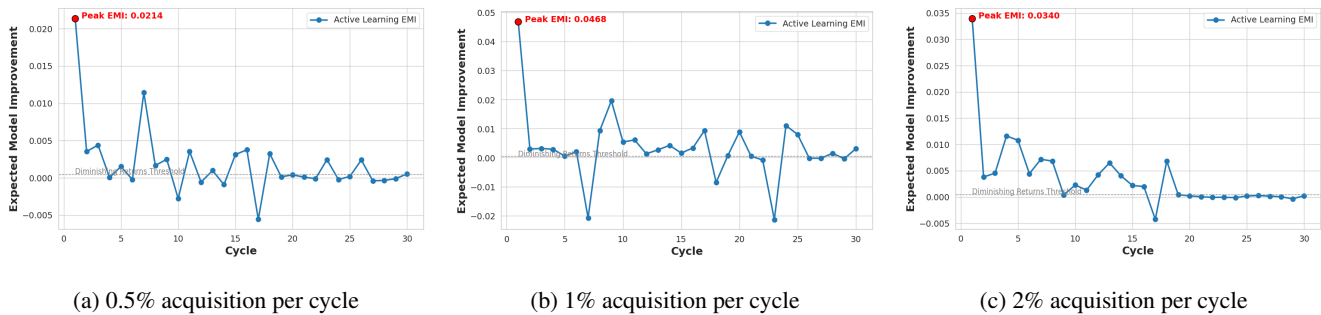


Figure 6. Expected Model Improvement (EMI) trajectories for different acquisition budgets in active learning.

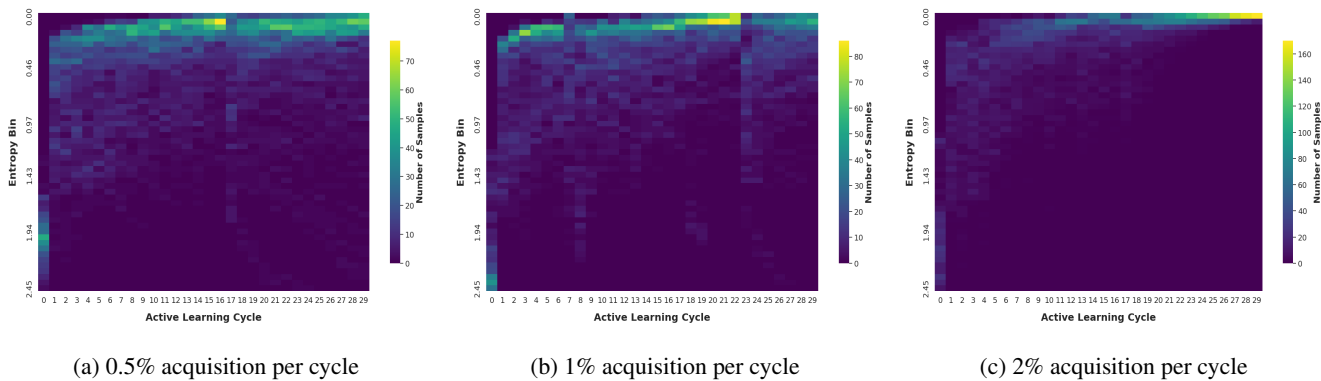


Figure 7. Evolution of entropy distribution across active learning cycles, with rows as entropy bins and columns as cycles. Color intensity denotes sample counts per bin, showing a shift from high to low entropy as labeling reduces uncertainty and boosts model confidence.

fraction of samples remain in higher-entropy bins, representing intrinsically hard or ambiguous cases—hallmarks of uncertainty-driven querying that continually targets informative instances.

In case of **0.5% acquisition budget**, heatmap shows deep yellow in top rows (very low entropy) from cycles 8–30, indicating many highly predictable samples and slow, steady convergence. Prolonged low-uncertainty acquisitions suggest limited novel information and a plateau in improvement. For **1% budget**, lower overall yellow intensity, with low-entropy mass diminishing after 22nd cycle, suggesting quicker removal of near-certain samples. This accelerates uncertainty reduction and avoids oversampling confident points, balancing exploration and exploitation. For **2% budget**, globally lighter map, with minimal low-entropy accumulation across cycles. Large acquisitions rapidly remove low-uncertainty samples, accelerating convergence and preventing redundancy, while late cycles show little residual ambiguity. Moreover, from cycle 20, we see that the model grows in confidence with all predicted samples in the high confidence region. In the heatmap, a deeper yellow shade conveys that more unlabeled samples concentrate in that particular entropy range for that cycle. This, in turn, depicts the areas where uncertainty mass accumulates

over time. More details are given in [Appendix 6.8](#).

5. Conclusion

We introduced CPDGAL, a parameter-efficient hyperspectral image classifier achieving state-of-the-art accuracy with only 87,538 parameters and 5.17 MFLOPs. Optimal performance (98.94% OA) is obtained with a 32-dimensional embedding, 8 attention heads, 6 layers, and moderate diffusion steps ($T = 100$). While highly effective on Indian Pines, Salinas, and Botswana, reliance on specific configurations may limit flexibility.

Future Works. Future work will assess CPDGAL on large-scale, diverse conditions to ensure robustness. Its ultra-light, active learning-enabled design supports energy- and label-efficient UAV-based crop stress detection and wetland health tracking. This can make it a sustainable, scalable, general-purpose tool for data-driven monitoring, environmental decision-making, and various remote sensing applications.

Acknowledgment

A part of this research has received support from the IEEE Geoscience and Remote Sensing Society (GRSS) under the “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, L. Farhan, et al. Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions. *Journal of Big Data*, 8:1–74, 2021. 1, 2
- [4] J. B. Campbell and R. H. Wynne. *Introduction to Remote Sensing*. Guilford Press, New York, NY, 5 edition, 2011. 1
- [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, 2, 6
- [6] B. T. Feng, J. Smith, M. Rubinstein, H. Chang, K. L. Bouman, and W. T. Freeman. Score-based Diffusion Models as Principled Priors for Inverse Imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10520–10531, 2023. 2
- [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 (CVPR)*, pages 15979–15988, 2022. 1
- [10] X. He, Y. Chen, and Z. Lin. Spatial-Spectral Transformer for Hyperspectral Image Classification. *Remote Sensing*, 13(3):498, 2021. 2, 3
- [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] I. Kakogeorgiou, S. Gidaris, B. Psomas, Y. Avrithis, A. Bur-suc, K. Karantzas, and N. Komodakis. What to Hide from Your Students: Attention-Guided Masked Image Modeling. In *European Conference on Computer Vision (ECCV)*, pages 300–318. Springer Nature Switzerland, 2022. 1
- [13] 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, 3
- [14] 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
- [15] 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
- [16] 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
- [17] 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
- [18] Y. Ma, H. Wu, L. Wang, B. Huang, R. Ranjan, A. Zomaya, and W. Jie. Remote Sensing Big Data Computing: Challenges and Opportunities. *Future Generation Computer Systems*, 51:47–60, 2015. 1
- [19] 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
- [20] 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, 3
- [21] P. Ren, Y. Xiao, X. Chang, P. Y. Huang, Z. Li, B. B. Gupta, and X. Wang. A Survey of Deep Active Learning. *ACM Computing Surveys (CSUR)*, 54(9):1–40, 2021. 1
- [22] 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
- [23] 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
- [24] B. Settles. Active Learning Literature Survey. Technical report, University of Wisconsin-Madison, Computer Sciences Technical Report 1648, 2009. 1
- [25] Z. Shu, Y. Wang, and Z. Yu. Dual Attention Transformer Network for Hyperspectral Image Classification. *Engineering Applications of Artificial Intelligence*, 127:Article 107351, 2024. 2, 5, 7
- [26] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*, 2015. 1
- [27] 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, 3
- [28] H. Sun, L. Yu, B. Dai, D. Schuurmans, and H. Dai. Score-based Continuous-Time Discrete Diffusion Models. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 2
- [29] 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, 7

- [30] 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](#)
- [31] L. V. D. Maaten and G. E. Hinton. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008. [6](#)
- [32] 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](#)
- [33] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, and F. Wei. Image as a Foreign Language: Beit Pretraining for Vision and Vision-Language Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19175–19186, 2023. [1](#)
- [34] C. Zhang, C. Zhang, J. Song, J. S. K. Yi, and I. S. Kweon. A Survey on Masked Autoencoder for Visual Self-supervised Learning. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 6805–6813, 2023. [1](#)
- [35] 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](#)
- [36] 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](#)
- [37] T. Zhao, S. Wang, C. Ouyang, and et al. Artificial Intelligence for Geoscience: Progress, Challenges, and Perspectives. *The Innovation*, 5(5), 2024. [1](#)
- [38] 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](#), [6](#), [7](#)
- [39] 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. [2](#), [6](#), [7](#)