CA-DEMAE: Content-Aware Diffusion-Based Masked Autoencoder for Hyperspectral Image Classification

Hongguang Xiao, Yuchuan Chen, Ziping He*, and Junjie Li
School of Computer Science and Technology, Changsha University of Science and Technology, Changsha, China
003331@csust.edu.cn, chuan990621@163.com, heziping@csust.edu.cn, 202308010307@csust.edu.cn

***Abstract*—Hyperspectral image (HSI) classification serves as a pivotal task in remote sensing, playing a critical role in precise land cover identification and environmental monitoring. However, due to the scarcity of annotated HSI samples and the inherent spectral redundancy and noise of HSI, extracting robust and discriminative features with limited label support remains a significant challenge in this field. Meanwhile, existing masked self-supervised learning methods typically treat all spectral bands equally, ignoring their varying discriminative power, and rely on fixed-scale reconstruction, which limits their ability to capture complex spatial structures. This paper proposes a content-aware diffusion-based masked autoencoder (CA-DEMAE) framework for HSI classification, which restores fine-grained spatial structures and extracts discriminative spectral features by adaptively reassembling multi-scale spatial details and dynamically calibrating spectral importance weights. This method integrates a multi-scale feature fusion strategy into the content-aware feature reassembly mechanism, the multi-scale content-aware reassembly (MSCAR) mechanism that effectively harmonizes local details and global contextual information. Furthermore, the spectral-weighted masking (SpWM) module is proposed to address the neglect of varying spectral contributions, achieving dynamic calibration of band importance and enhanced spectral discriminability. Experimental results on three public benchmark hyperspectral datasets demonstrate that the proposed method exhibits superior performance. To facilitate reproducibility, the source code of the proposed CA-DEMAE is publicly available at https://github.com/chuan990621-debug/CA-DEMAE.**

***Index Terms*—diffusion models, attention mechanism, hyperspectral image**

I. Introduction

Hyperspectral image (HSI) consists of a dense sequence of continuous spectral bands, providing each pixel with rich spectral and spatial information, and has been extensively studied in fields such as land cover classification [1], [2]. HSI classification represents a pivotal task in remote sensing, aiming to assign precise semantic labels to each pixel through joint analysis of spatial and spectral features. This technique drives critical applications in environmental monitoring [3], agricultural optimization [4], and defense systems [5].

In earlier years, HSI classification was primarily dominated by traditional machine learning methods. Traditional learning methods include support vector machines (SVMs) [6], random forests (RFs) [7], and k-nearest neighbors (KNNs) [8]. While these methods offer advantages for small-sample datasets, their heavy reliance on expert-crafted feature design [9] makes them less effective at handling complex and variable scenarios.

Deep learning models, including 3D-CNNs [11]–[13] and attention-enhanced architectures like RSSAN [14], improve feature selection [10]. However, their limited receptive fields restrict long-term dependency modeling [15], causing spectral distortion and boundary detail loss [16].

To overcome these limitations, researchers applied Transformers [17]–[19] to HSI classification. Methods like SpectralFormer [16], SSFTT [20], and GAHT [21] use specific tokenization strategies to capture spectral-spatial dependencies, while hybrid models including MorphFormer [22], CAEVT [23], GSC-VIT [24], and window-based models [25] integrate convolutions to balance performance and complexity. However, these supervised methods rely on limited labeled data, failing to exploit Transformers' strong generalization capabilities through extensive unlabeled pre-training.

To enhance generalization, self-supervised learning (SSL) [26], particularly generative masked autoencoders (MAE) [27], [28], has emerged as a promising paradigm. Although methods like HyperspectralMAE [29] improve representations, they struggle with labeled sample scarcity and rely solely on simple reconstruction tasks [30]. To overcome single-task limitations, Li *et al.* [30] proposed the diffusion-enhanced masked autoencoder (DEMAE), which integrates diffusion strategies [31] for simultaneous denoising and reconstruction to learn robust features.

However, DEMAE employs a static reconstruction mechanism that lacks adaptivity to varying spatial content. Furthermore, it treats hyperspectral bands uniformly, neglecting their specific physical significance and unequal contributions to classification. Recent studies on model interpretability [32] emphasize the unequal contribution of input features to decision-making, further underscoring the necessity of dynamically calibrating spectral importance. Motivated by these insights, this paper proposes a novel content-aware diffusion-based masked autoencoder (CA-DEMAE) framework. Specifically, we design a multi-scale content-aware reassembly (MSCAR) module to adaptively restore fine-grained spatial structures, and incorporate a spectral-weighted masking (SpWM) module to dynamically calibrate band importance, thereby achieving synergistic enhancement of spatial and spectral representations.

*Corresponding author.

The main contributions of this article are as follows.

1) We propose a content-aware diffusion-based masked autoencoder (CA-DEMAE) framework, which redefines the reconstruction paradigm by integrating content-aware priors into the generative process. This approach enables the model to capture highly representative structural and spectral features from extensive unlabeled data, making it uniquely effective for interpreting complex hyperspectral scenarios.

2) We introduce the content-aware reassembly mechanism into HSI classification and propose the multi-scale content-aware reassembly (MSCAR) module. By incorporating a multi-scale fusion strategy, this module effectively integrates local details and global contextual information.

3) A spectral-weighted masking (SpWM) module is proposed, which dynamically learns spectral band relationships and importance weights to achieve adaptive spectral calibration, thereby precisely amplifying informative bands.

4) We conduct extensive experiments on three public benchmark datasets. The results demonstrate that CA-DEMAE significantly outperforms representative methods, particularly under limited training samples.

The remainder of this paper is organized as follows. Section II details the CA-DEMAE framework. Section III presents extensive comparative experiments, visual evaluations and ablation studies. Finally, Section IV provides conclusion.

II. Methodology

A. Overview of the Proposed Framework

Despite the promise of masked SSL in HSI classification [30], [33], current methods are limited by fixed-scale reconstruction and uniform spectral weighting. Addressing this, our proposed CA-DEMAE framework (Fig. 1) extends the DEMAE [30] baseline by introducing a multi-scale content-aware reassembly (MSCAR) module during pre-training for refined spatial reconstruction. Additionally, it embeds a spectral-weighted masking (SpWM) module during fine-tuning to dynamically learn band importance, synergistically enhancing spatial-spectral representations.

B. HSI Preprocessing

Following DEMAE [30], we project the HSI patch into a latent space $X \in \mathbb{R}^{h\times w\times c}$ via PCA [34] whitening and serialize it into pixel tokens. A 75% random masking strategy partitions the sequence into visible (x^{vis}) and masked (x^{mask}) subsets, strictly preserving the center pixel for classification integrity.

To generate the noisy input x_t^{vis} for denoising, x^{vis} is corrupted by a diffusion forward process at a randomly sampled time step t:

$$x_t^{vis} = \sqrt{\alpha_t} x_0^{vis} + \sqrt{1-\alpha_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{1}$$

where ϵ is standard Gaussian noise, and the noise coefficient α_t follows a linear schedule defined as $\alpha_t = \frac{T-t}{T}\alpha_0 + \frac{t}{T}\alpha_T$ ($\alpha_0 = 0.9999$, $\alpha_T = 0.98$, and T representing the total time steps).

C. Pre-Training Stage

The pre-training framework uses an asymmetric encoder-decoder for joint denoising-reconstruction. The input combines visible tokens x^{vis}, a class token x_{class}, a time token x_{time}, and position encodings.

1) Conditional Encoder: Stacked conditional Transformer blocks use x_{time} to modulate features via conditional layer normalization (cNorm), adapting to noise levels. An intermediate layer aligns encoder-decoder feature distributions.

2) Decoder Design: A diffusion decoder restores clean data, and an MAE decoder reconstructs masked regions. Before generating the final output, MAE decoder features pass through the MSCAR module for refined recovery. Otherwise, the architecture and pretext task follow the DEMAE baseline [30].

3) MSCAR Module: MSCAR extends the CARAFE [35] operator with multi-scale fusion to capture local details and global contexts, enhancing discriminability. The architecture is shown in Fig. 2. Given an input $\mathbf{Z} \in \mathbb{R}^{C\times H\times W}$ and upsampling rate σ, MSCAR outputs $\mathbf{Z}' \in \mathbb{R}^{C\times \sigma H\times \sigma W}$.

A content encoder extracts a compact representation from $\mathbf{Z}$ via a 1×1 channel-reduction convolution and a 3×3 refining convolution, yielding $\mathbf{Z}_{\text{content}}$.

To capture multi-scale spatial information, a parallel fusion architecture processes $\mathbf{Z}_{\text{content}}$ through four branches ($\mathbf{F}_1$ to $\mathbf{F}_4$ for local to global contexts), each outputting $C_m/4$ channels.

To mitigate concatenation redundancy, an adaptive 1×1 convolution dynamically weights and fuses the features:

$$\mathbf{Z}_{\text{fused}} = \text{Conv}_{1\times 1}\left(\text{Concat}\left(\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3, \mathbf{F}_4\right)\right) \tag{2}$$

where Concat$(\cdot)$ is concatenation. Utilizing $\mathbf{Z}_{\text{fused}}$, a kernel prediction network [35] generates spatial-softmax normalized kernels $\mathbf{W}_{\mathbf{l}'}$ of size $k_{\text{up}}\times k_{\text{up}}$ for final feature reassembly via weighted summation:

$$\mathbf{Z}'_{\mathbf{l}'} = \sum_{n=-r}^{r}\sum_{m=-r}^{r} \mathcal{W}_{\mathbf{l}'(n,m)} \cdot \mathbf{Z}_{(i+n,j+m)} \tag{3}$$

where $r = \lfloor k_{\text{up}}/2 \rfloor$ is the radius, and $\mathcal{W}_{\mathbf{l}'(n,m)}$ is the weight in $\mathbf{W}_{\mathbf{l}'}$. Thus, features are weighted by content relevance instead of spatial distance.

The pre-training objective combines denoising and masked reconstruction losses. The diffusion loss $\mathcal{L}_{\text{Diffusion}}$ is:

$$\mathcal{L}_{\text{Diffusion}} = \left\|\hat{x}_\theta^{\text{vis}}(x_t^{\text{vis}}, t) - x_0^{\text{vis}}\right\|_2^2 \tag{4}$$

where $\hat{x}_\theta^{\text{vis}}(x_t^{\text{vis}}, t)$ is the predicted clean feature from noisy input x_t^{vis} at step t. The MAE reconstruction loss $\mathcal{L}_{\text{MAE}}$ based on MSCAR output is:

$$\mathcal{L}_{\text{MAE}} = \|\hat{x}_\theta^{\text{mask}} - x^{\text{mask}}\|_2^2 \tag{5}$$

where $\hat{x}_\theta^{\text{mask}}$ and x^{mask} are the reconstructed and ground-truth masked tokens, respectively. The total pre-training loss is $\mathcal{L}_{\text{pre-train}} = \mathcal{L}_{\text{Diffusion}} + \mathcal{L}_{\text{MAE}}$.

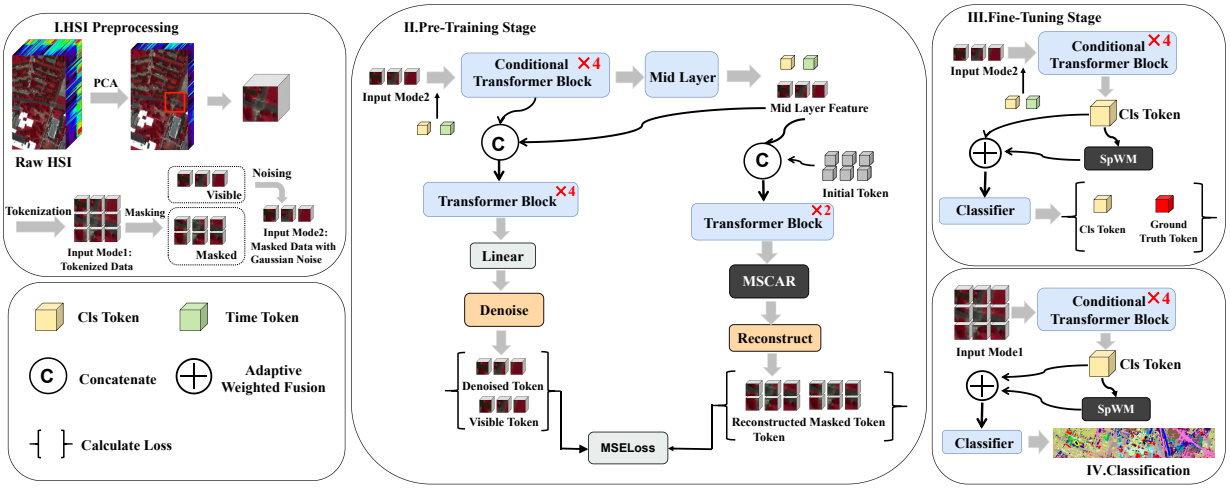

Fig. 1. Overall architecture of the CA-DEMAE framework.

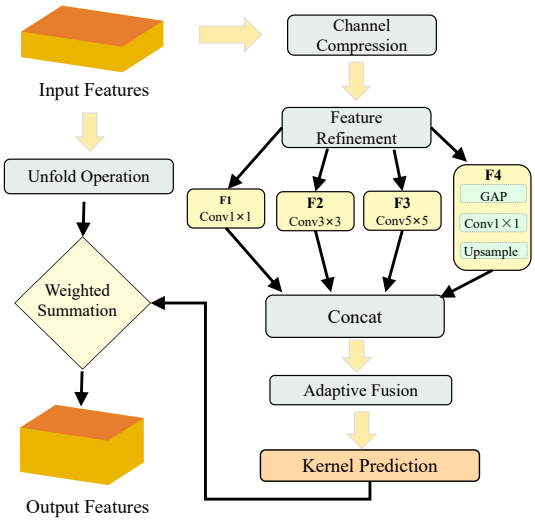

Fig. 2. Architecture of the MSCAR module.

D. Fine-Tuning Stage

1) Fine-Tuning Architecture: The fine-tuning model inherits the pre-trained conditional Transformer encoder and connects it to a linear layer for classification. The data processing method during the fine-tuning stage is similar to that of the pre-training stage, including PCA whitening, serialization, random masking with center pixel preservation, and noising. However, the degree of masking and noising during the fine-tuning stage does not exceed that of the pre-training stage.

2) SpWM Module: We propose the spectral-weighted masking (SpWM) module, which operates on the classification token extracted from the encoder output, as illustrated in Fig. 3.

The module takes the global classification token $\mathbf{T}_{\text{cls}} \in \mathbb{R}^{B\times 1\times C}$ as input. To effectively disentangle the overlapping spectral information, we employ a multi-head mechanism that projects the spectral features into h parallel subspaces, unlike standard spatial attention that models token-to-token dependencies. The operation is formulated as:

$$\mathbf{V}_i = \mathbf{T}_{\text{cls}}\mathbf{W}_{V,i}, \quad i = 1, \ldots, h \tag{6}$$

where $\mathbf{W}_{V,i} \in \mathbb{R}^{C\times d_k}$ is the learnable projection matrix for the i-th head, and $d_k = C/h$. Each head extracts specific spectral patterns within its subspace. The final spectral representation is obtained by concatenating the outputs from all heads and fusing them via a linear projection:

$$\mathbf{T}_{\text{proj}} = \text{Concat}(\mathbf{V}_1, \ldots, \mathbf{V}_h)\mathbf{W}_O \tag{7}$$

where $\mathbf{W}_O \in \mathbb{R}^{C\times C}$ integrates the diverse spectral features. This structure allows the model to learn a richer, non-linear combination of spectral bands compared to a single linear layer.

After the multi-head projection, SpWM introduces a learnable band importance vector $\boldsymbol{\alpha} \in \mathbb{R}^C$. To generate a valid probability distribution, we apply the softmax operation to normalize these weights:

$$w_i = \frac{\exp(\alpha_i)}{\sum_{j=1}^{C}\exp(\alpha_j)}, \quad i = 1, \ldots, C \tag{8}$$

where w_i represents the normalized weight for the i-th spectral band, forming the weight vector $\mathbf{w}_b = [w_1, \ldots, w_C]$. The projected features $\mathbf{T}_{\text{proj}}$ are then reweighted element-wise:

$$\mathbf{T}_{\text{weighted}} = \mathbf{T}_{\text{proj}} \odot \mathbf{w}_b \tag{9}$$

where $\odot$ denotes element-wise multiplication. This mechanism enables the network to adaptively emphasize informative spectral regions while suppressing noise.

To preserve spectral smoothness, a depthwise one-dimensional convolution filters each channel independently. Specifically, for the c-th channel of the weighted features, the operation is defined as:

$$\mathbf{T}^{(c)}_{\text{filt}}[i] = \sum_{k=-1}^{1} \mathbf{w}^{(c)}_k \cdot \mathbf{T}^{(c)}_{\text{weighted}}[i+k] \tag{10}$$

where $c \in \{1, 2, \ldots, C\}$ denotes the index of the spectral channel, and $\mathbf{w}_k^{(c)}$ denotes the weight of the size 3 convolution kernel with padding 1 at position k. Finally, to prevent gradient vanishing, a residual connection adds filtered features back to the initial classification token: $\mathbf{T}_{\text{out}} = \mathbf{T}_{\text{cls}} + \mathbf{T}_{\text{filt}}$.

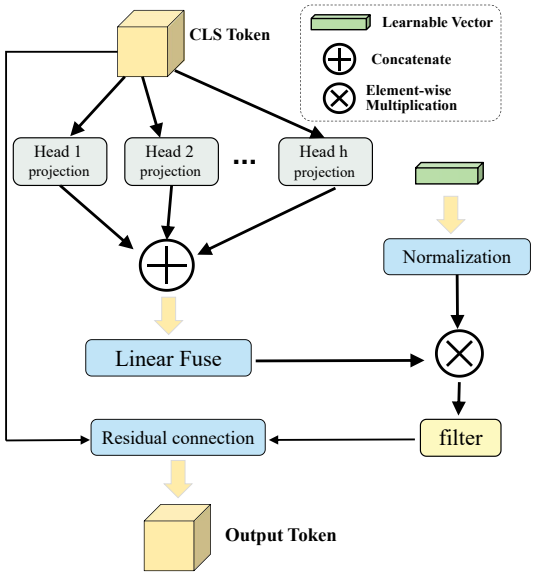

Fig. 3. Architecture of the SpWM module.

Simply replacing the original feature with the SpWM output might lead to information loss or training instability. To mitigate this, we introduce a learnable gating parameter γ (initialized to ensure a balanced start) to dynamically control the fusion of the original token and the refined features. The final classification token $\mathbf{T}_{\text{final}}$ is formulated as:

$$\mathbf{T}_{\text{final}} = (1 - \sigma(\gamma)) \cdot \mathbf{T}_{\text{cls}} + \sigma(\gamma) \cdot \mathbf{T}_{\text{out}} \tag{11}$$

where $\mathbf{T}_{\text{out}}$ represents the output feature of the SpWM module, and $\sigma(\cdot)$ denotes the sigmoid function. This adaptive fusion strategy allows the network to automatically determine the optimal contribution of the spectral masking effect during the training process.

We compute the classification loss using the class token after Softmax activation. To mitigate noise variance from the diffusion process, we adopt the SNR-weighted loss from DEMAE [30], which prioritizes high-SNR samples to focus on reliable semantic features.

E. Classification

During inference, the tokenized input is initialized with a noise-free time embedding at $t = 0$ and processed by the encoder. The extracted class token $\mathbf{T}_{\text{cls}}$ is then refined by the SpWM module to suppress noise and enhance spectral discriminability, before being passed to the linear classifier for the final pixel-wise prediction.

III. Experimental Results

A. Data Description

We evaluate the proposed CA-DEMAE on three hyperspectral benchmarks: Houston 2013 (an urban area), WHU-LongKou (an agricultural scene), and Xuzhou (a peri-urban site). Detailed dataset characteristics and sample statistics are summarized in Table I.

TABLE I
LIST OF LAND COVER CLASSES AND NUMBER OF SAMPLES FOR HOUSTON 2013, WHU-LONGKOU, AND XUZHOU DATASETS.

NO	Houston 2013	Num.	WHU-LongKou	Num.	Xuzhou	Num.
1	Healthy grass	1251	Corn	34511	Crop Type 1	26396
2	Stressed grass	1254	Cotton	8374	Crop Type 2	4027
3	Synthetic grass	697	Sesame	3031	Vegetation Type 1	2783
4	Trees	1244	Broad-leaf soybean	63212	Vegetation Type 2	5214
5	Soil	1242	Narrow-leaf soybean	4151	Man-made Struct. 1	13184
6	Water	325	Rice	11854	Man-made Struct. 2	2436
7	Residential	1268	Water	67056	Coal Field Type 1	6990
8	Commercial	1244	Roads and houses	7124	Coal Field Type 2	4777
9	Road	1252	Mixed weed	5229	Other Land Cover	3070
10	Highway	1227				
11	Railway	1235				
12	Parking Lot 1	1233				
13	Parking Lot 2	469				
14	Tennis court	428				
15	Running track	660				
Total		**15029**		**204542**		**68877**

B. Experimental Setup

Experiments were conducted using PyTorch on an NVIDIA RTX 4060 GPU. The patch size and PCA dimension were 11 and 36. We used Adam optimizer (learning rate 1×10^{-3}, decay $\gamma = 0.99$) with a 0.75 masking ratio for both stages. The model was pre-trained for 200 epochs (batch size 512, 200 diffusion steps) and fine-tuned for 150 epochs (batch size 64, 100 diffusion steps) using 5 labeled samples per class. CA-DEMAE was compared against several baselines: GAHT [21], SpectralFormer [16], RSSAN [14], SSFTT [20], GSC-VIT [24], MorphFormer [22], CAEVT [23], and DEMAE [30]. Results were evaluated using OA, AA, and Kappa, averaged over ten independent runs.

C. Comparison With Different Classification Methods

Tables II, III, and IV report the quantitative classification results across the three datasets. Although Transformer-based methods generally outperform CNNs, they struggle to learn robust features with only 5 training samples per class. While the DEMAE baseline mitigates sample scarcity via pre-training, it overlooks multi-scale spatial information and spectral discrepancies. To address this, the proposed CA-DEMAE introduces the MSCAR and SpWM modules, achieving state-of-the-art performance and improving Overall Accuracy (OA) by 1.31% to 2.02% over the baseline. Notably, our method demonstrates superior discriminability on challenging classes, such as Commercial (49.84%) and Railway (84.58%) in Houston 2013, Soybeans in WHU-LongKou, and Coal Field in Xuzhou. These improvements confirm that CA-DEMAE significantly enhances the discrimination of complex land covers through precise modeling of subtle spectral differences and multi-scale contextual information.

TABLE II
CLASSIFICATION ACCURACY COMPARISON OF DIFFERENT METHODS ON HOUSTON 2013 WITH 5 LABELED SAMPLES PER CLASS FOR TRAINING. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Class	GAHT	SpectralFormer	RSSAN	SSFTT	GSC-VIT	MorphFormer	CAEVT	DEMAE	CA-DEMAE
1	**99.79**	96.17	73.26	96.79	99.38	95.23	99.38	88.56	91.51
2	84.22	60.00	77.08	76.11	78.59	82.16	82.05	90.86	**91.12**
3	65.88	77.01	87.91	96.36	**100**	92.62	99.47	99.88	**100**
4	89.29	77.38	76.28	81.86	91.37	83.83	**95.96**	87.55	90.09
5	93.59	80.11	53.59	85.86	86.30	81.99	83.54	99.45	**99.89**
6	89.60	70.28	65.88	89.04	90.17	82.37	88.14	93.71	**93.83**
7	46.60	32.57	27.02	47.64	49.53	35.71	50.47	84.51	**84.81**
8	41.72	31.61	19.57	38.28	29.57	25.70	34.84	46.14	**49.84**
9	56.28	27.13	7.45	38.72	**64.68**	49.47	55.43	64.24	64.63
10	53.01	43.87	20.22	50.00	55.38	51.51	54.09	81.22	**81.56**
11	52.95	35.80	19.66	66.70	73.07	48.98	65.11	81.53	**84.58**
12	33.37	25.35	21.93	31.98	34.65	28.13	31.76	64.71	**67.29**
13	**99.22**	63.80	39.89	97.43	95.53	92.07	87.37	89.87	90.30
14	94.32	81.59	18.98	85.80	97.39	86.70	96.02	**100**	**100**
15	99.89	88.79	28.90	88.24	**100**	85.27	94.62	**100**	**100**
OA(%)	73.09±4.79	59.30±2.69	42.59±7.05	71.16±2.20	76.14±2.78	67.93±3.68	74.36±3.89	84.63±1.89	**86.65±2.07**
AA(%)	73.32±4.70	59.43±2.65	42.51±7.16	71.39±2.16	76.37±2.73	68.12±3.68	74.55±3.89	84.77±1.89	**85.89±2.06**
Kappa × 100	71.17±5.13	56.40±2.89	38.49±7.56	69.10±2.35	74.44±2.98	65.64±3.94	72.53±4.17	83.57±2.03	**84.93±2.22**

TABLE III
CLASSIFICATION ACCURACY COMPARISON OF DIFFERENT METHODS ON WHU-LONGKOU WITH 5 LABELED SAMPLES PER CLASS FOR TRAINING. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Class	GAHT	SpectralFormer	RSSAN	SSFTT	GSC-VIT	MorphFormer	CAEVT	DEMAE	CA-DEMAE
1	84.96	62.56	84.07	80.48	90.43	81.54	82.80	97.78	**98.10**
2	74.87	36.58	30.54	64.47	68.06	58.47	50.19	93.67	**94.28**
3	72.06	70.85	59.89	63.06	78.40	63.82	64.76	**99.31**	99.24
4	62.81	51.30	60.13	59.98	66.18	52.99	53.95	92.24	**92.85**
5	88.10	45.90	74.51	70.77	70.62	59.87	64.72	**96.19**	96.16
6	84.91	55.71	55.90	76.09	78.67	71.16	75.15	96.72	**97.54**
7	99.96	**99.99**	99.47	99.98	99.95	99.97	99.96	96.02	96.34
8	59.38	52.57	49.67	67.00	65.75	58.65	61.80	78.87	**79.93**
9	54.85	27.76	33.83	44.36	60.96	50.24	49.13	88.03	**88.97**
OA(%)	80.83±5.30	68.44±10.37	74.86±3.93	77.78±5.54	82.27±6.41	74.63±5.11	75.51±4.69	93.47±1.60	**94.78±1.44**
AA(%)	75.77±5.01	55.91±5.64	60.89±6.02	69.58±2.45	75.45±5.92	65.08±5.38	66.94±1.38	92.74±1.62	**93.58±1.61**
Kappa × 100	75.96±6.22	61.12±11.84	68.37±4.52	72.17±6.46	77.64±7.65	68.43±5.77	69.48±5.38	92.09±2.05	**93.22±1.83**

TABLE IV
CLASSIFICATION ACCURACY COMPARISON OF DIFFERENT METHODS ON XUZHOU DATASET WITH 5 LABELED SAMPLES PER CLASS FOR TRAINING. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Class	GAHT	SpectralFormer	RSSAN	SSFTT	GSC-VIT	MorphFormer	CAEVT	DEMAE	CA-DEMAE
1	87.11	83.99	86.96	81.82	86.64	83.67	84.12	91.07	**92.24**
2	96.27	92.87	93.48	95.02	97.45	95.68	95.06	97.04	**98.42**
3	77.79	67.58	59.68	58.65	65.18	63.17	67.36	95.05	**96.56**
4	77.88	70.16	55.61	80.51	78.64	71.74	75.90	**88.99**	88.47
5	78.69	65.59	39.40	66.19	65.89	52.94	65.38	93.82	**95.46**
6	70.26	68.41	56.37	60.53	72.41	60.51	69.52	**96.77**	96.54
7	89.84	83.95	90.57	79.74	90.33	81.65	82.78	98.93	**99.23**
8	95.57	83.05	59.02	90.75	88.87	90.52	92.30	**99.44**	99.33
9	76.12	62.18	67.57	55.60	76.49	41.75	57.08	96.51	**97.76**
OA(%)	84.74±3.36	77.69±3.96	71.23±4.08	77.05±1.95	81.40±2.05	74.48±2.92	78.58±4.10	93.46±1.15	**94.82±1.35**
AA(%)	83.28±2.16	75.31±3.10	67.63±7.09	74.31±2.29	80.21±3.27	71.63±2.86	76.61±2.79	95.23±0.87	**96.21±1.07**
Kappa × 100	80.96±4.18	72.34±4.80	64.24±5.16	71.63±2.25	76.89±2.49	68.19±3.62	73.48±4.98	92.29±1.43	**93.69±1.64**

D. Visual Evaluation

Figs. 4–6 illustrate the classification maps for qualitative assessment. Generally, CNN-based RSSAN exhibits salt-and-pepper noise, while Transformer-based GAHT tends to over-smooth boundaries. Although the DEMAE baseline improves consistency, residual noise persists in complex transition regions. In contrast, CA-DEMAE effectively preserves intricate textures and structural details via MSCAR and SpWM modules. Specifically, it maintains crisp urban boundaries in Houston 2013, avoiding scattered misclassifications. In WHU-LongKou, it captures fine-grained agricultural patterns without patchy artifacts. On Xuzhou, CA-DEMAE yields the most precise results by suppressing noise in mixed industrial areas. These consistent improvements confirm that our method effectively enhances land-cover discriminability while significantly mitigating misclassification noise across diverse scenarios.

E. Ablation Study

To evaluate the individual and combined contributions of the MSCAR and SpWM modules, we conducted ablation experiments as shown in Table V. Starting with the baseline DEMAE, adding the MSCAR module improves performance across all datasets, particularly on Houston 2013. This confirms the effectiveness of the multi-scale feature fusion mechanism in capturing spatial structural features of hyperspectral images, effectively validating its capability to capture multi-scale spatial structures in complex urban scenes. Similarly, the independent addition of the SpWM module yields consistent gains, especially on WHU-LongKou, by identifying discriminative spectral bands critical for distinguishing similar crops.

When integrating both the MSCAR and SpWM modules simultaneously, CA-DEMAE achieved the best results across all metrics. For instance, on the Houston 2013 dataset, the combined model improved the OA metric by 2.02% compared

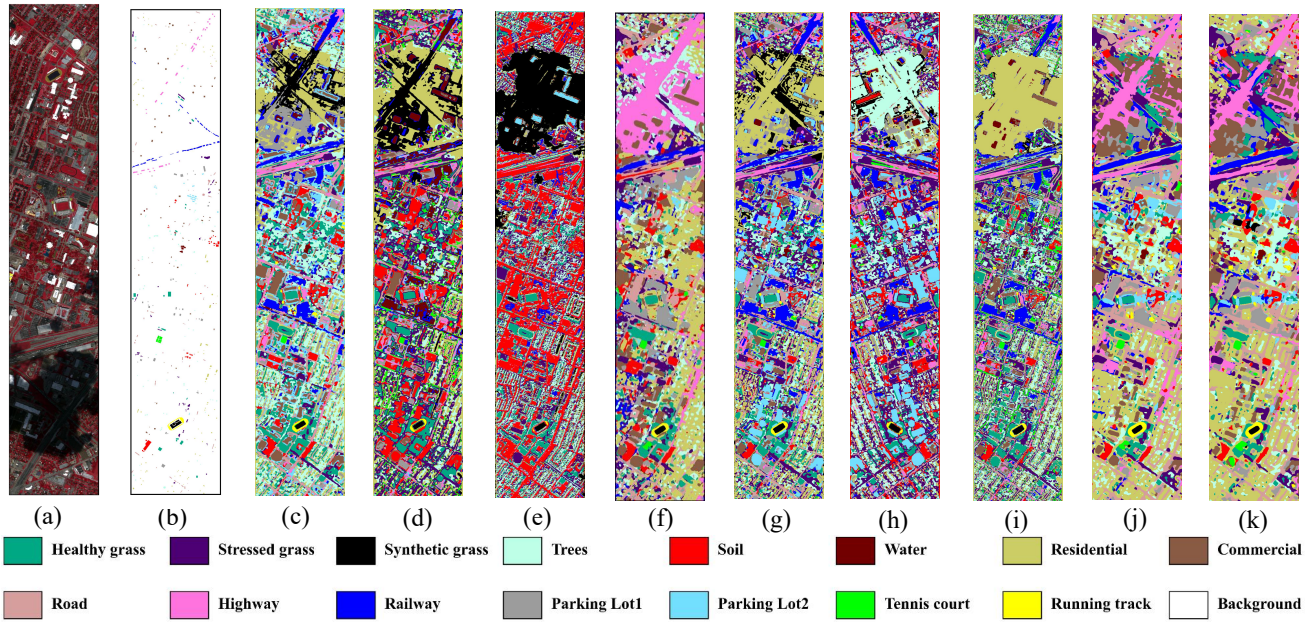

Fig. 4. Classification maps comparison on Houston 2013. (a) False-color composite image (R:59,G:40,B:23). (b) Ground Truth. (c) GAHT (73.09%). (d)SpectralFormer (59.30%). (e) RSSAN (42.59%). (f) SSFTT (71.16%). (g) GSC-VIT (76.14%). (h) MorphFormer (67.93%). (i) CAEVT (74.36%). (j) DEMAE (84.63%). (k) CA-DEMAE (86.65%).

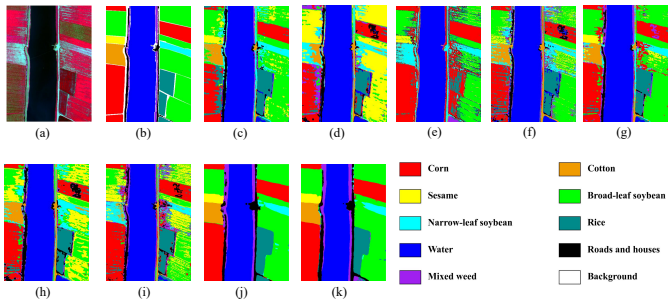

Fig. 5. Classification maps comparison on WHU-LongKou. (a) False-color composite image (R:180,G:126,B:72). (b) Ground Truth. (c) GAHT (80.83%). (d) SpectralFormer (68.44%). (e) RSSAN (74.86%). (f) SSFTT (77.78%). (g) GSC-VIT (82.27%). (h) MorphFormer (74.63%). (i) CAEVT (75.51%). (j) DEMAE (93.47%). (k) CA-DEMAE (94.78%).

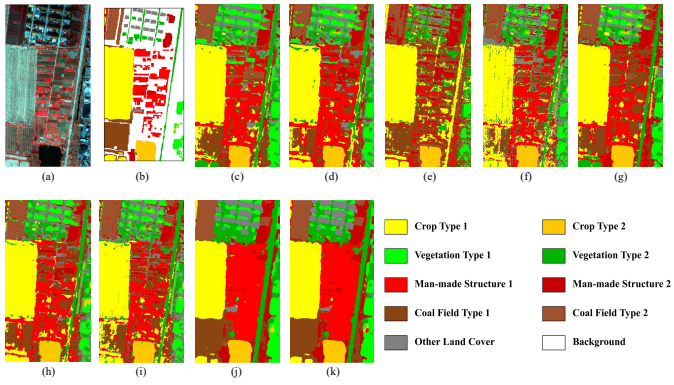

Fig. 6. Classification maps comparison on Xuzhou dataset. (a) False-color composite image (R:90,G:49,B:28). (b) Ground Truth. (c) GAHT (84.74%). (d) SpectralFormer (77.69%). (e) RSSAN (71.23%). (f) SSFTT (77.05%). (g) GSC-VIT (81.40%). (h) MorphFormer (74.48%). (i) CAEVT (78.58%). (j) DEMAE (93.46%). (k) CA-DEMAE (94.82%).

TABLE V
ABLATION STUDY RESULTS COMPARING INDIVIDUAL AND COMBINED CONTRIBUTIONS OF MSCAR AND SPWM MODULES ACROSS THREE DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

MSCAR	SpWM	Houston 2013			WHU-LongKou			Xuzhou		
		OA(%)	AA(%)	Kappa	OA(%)	AA(%)	Kappa	OA(%)	AA(%)	Kappa
×	×	84.63	84.77	83.57	93.47	92.74	92.09	93.46	95.23	92.29
×	√	85.17	85.11	83.96	94.26	92.90	92.20	93.90	95.51	92.61
√	×	85.94	85.67	84.13	94.35	93.31	92.67	94.18	95.80	93.63
√	√	**86.65**	**85.89**	**84.93**	**94.78**	**93.58**	**93.22**	**94.82**	**96.21**	**93.69**

to the baseline, demonstrating the synergistic effect of the two modules. The MSCAR module enhances spatial reconstruction capability, while the SpWM module improves spectral calibration, and the complementary strengths of these two modules enable the model to maintain stable and superior classification performance across datasets with varying complexities and feature distributions.

IV. CONCLUSION

In this article, we propose the content-aware diffusion-based masked autoencoder (CA-DEMAE) model for hyperspectral image classification, which enhances classification performance by introducing two key components. The multi-scale content-aware reassembly (MSCAR) module integrates multi-scale fusion to effectively combine local details and global context, significantly improving feature reconstruction. The spectral-weighted masking (SpWM) module refines discriminability by dynamically learning spectral band importance. Experiments on three benchmark datasets demonstrate that CA-DEMAE consistently outperforms comparative methods. Ablation studies confirm the distinct and synergistic contributions of both modules. Future work will focus on integrating this framework with emerging remote sensing foundation models, developing more efficient multi-modal fusion mechanisms,

and pursuing model lightweighting to reduce computational complexity while maintaining high accuracy.

V. ACKNOWLEDGMENT

This work was supported by the Research Foundation of Education Bureau of Hunan Province under Grant 25C0086.

REFERENCES

[1] M. F. Guerri, C. Distante, P. Spagnolo, F. Bougourzi, and A. Taleb-Ahmed, "Deep learning techniques for hyperspectral image analysis in agriculture: A review," *ISPRS Open J. Photogramm. Remote Sens.*, vol. 12, Art. no. 100062, Apr. 2024.

[2] M. Wang *et al.*, "Tensor decompositions for hyperspectral data processing in remote sensing: A comprehensive review," *IEEE Geosci. Remote Sens. Mag.*, vol. 11, no. 1, pp. 26–72, Mar. 2023.

[3] L. Chen, J. Wu, Y. Xie, E. Chen, and X. Zhang, "Discriminative feature constraints via supervised contrastive learning for few-shot forest tree species classification using airborne hyperspectral images," *Remote Sens. Environ.*, vol. 295, Sep. 2023, Art. no. 113710.

[4] N. Liu, P. A. Townsend, M. R. Naber, P. C. Bethke, W. B. Hills, and Y. Wang, "Hyperspectral imagery to monitor crop nutrient status within and across growing seasons," *Remote Sens. Environ.*, vol. 255, Mar. 2021, Art. no. 112303.

[5] X. Fu, S. Jia, L. Zhuang, M. Xu, J. Zhou, and Q. Li, "Hyperspectral anomaly detection via deep plug-and-play denoising CNN regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9553–9568, Nov. 2021.

[6] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[7] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 492–501, Mar. 2005.

[8] B. Tu, J. Wang, X. Kang, G. Zhang, X. Ou, and L. Guo, "KNN-based representation of superpixels for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 4032–4047, Nov. 2018.

[9] M. Ahmad, S. Shabbir, S. K. Roy, D. Hong, X. Wu, J. Yao, A. M. Khan, M. Mazzara, S. Distefano, and J. Chanussot, "Hyperspectral image classification—traditional to deep models: A survey for future prospects," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 968–999, 2022.

[10] K. Li, G. Liu, M. Dang, R. Pan, N. Luo, and W. Ma, "Dual-branch prototype enhancement network for few-shot hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, no. 8, 2026. DOI: 10.1109/JSTARS.2026.3653391.

[11] J. M. Haut, M. E. Paoletti, J. Plaza, J. Li, and A. Plaza, "Active learning with convolutional neural networks for hyperspectral image classification using a new Bayesian approach," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 11, pp. 6440–6461, 2018.

[12] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 6232–6251, 2016.

[13] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral–spatial residual network for hyperspectral image classification: A 3-D deep learning framework," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 847–858, 2018.

[14] M. Zhu, L. Jiao, F. Liu, S. Yang, and J. Wang, "Residual spectral-spatial attention network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 449–462, 2021.

[15] S. Jiang, Y. Liu, C. Mu, and X. He, "A combined frequency-domain filtering and multi-scale spatial feature communication based network for hyperspectral image classification," in *2025 International Joint Conference on Neural Networks (IJCNN)*, 2025, doi: 10.1109/IJCNN64981.2025.11229196.

[16] D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, and J. Chanussot, "SpectralFormer: Rethinking Hyperspectral Image Classification With Transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518615.

[17] H. Yao, P. Zhang, C. Zheng, L. Zhao, Y. Jin, M. Tian, Y. Hou, J. Li, and Q. Qiu, "MSSTFormer: Multiscale spectral–spatial supertoken aggregation transformer for hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2026, doi: 10.1109/JSTARS.2026.3654096.

[18] 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 (NIPS)*, 2017.

[19] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Vienna, Austria, May 2021, pp. 1–21.

[20] L. Sun, G. Zhao, Y. Zheng, and Z. Wu, "Spectral-Spatial Feature Tokenization Transformer for Hyperspectral Image Classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5522214, doi: 10.1109/TGRS.2022.3144158.

[21] S. Mei, C. Song, M. Ma, and F. Xu, "Hyperspectral Image Classification Using Group-Aware Hierarchical Transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5539014, doi: 10.1109/TGRS.2022.3207933.

[22] 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 Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5503615, doi: 10.1109/TGRS.2023.3242346.

[23] Z. Zhang, T. Li, X. Tang, X. Hu, and Y. Peng, "CAEVT: Convolutional autoencoder meets lightweight vision transformer for hyperspectral image classification," *Sensors*, vol. 22, no. 10, p. 3902, 2022.

[24] 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*, vol. 62, pp. 1–14, 2024, Art. no. 5511817.

[25] X. Wu, T. Arshad, and B. Peng, "Spectral spatial window attention transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–13, 2025.

[26] J. Gui, T. Chen, Q. Cao, Z. Sun, H. Luo, and D. Tao, "A survey on self-supervised learning: Algorithms, applications, and future trends," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 1, pp. 7–29, Jan. 2024, doi: 10.1109/TPAMI.2023.3259693.

[27] W. Liu *et al.*, "Self-supervised feature learning based on spectral masking for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Art. no. 4407715, 2023.

[28] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, 2022, pp. 16000–16009.

[29] W. Jeong, H. J. Park, S. Jeong, J. W. Jang, T. H. Lim, and D. S. Kim, "HyperspectralMAE: The hyperspectral imagery classification model using Fourier-encoded dual-branch masked autoencoder," *arXiv preprint arXiv:2505.05710*, 2025.

[30] Z. Li, Z. Xue, M. Jia, X. Nie, H. Wu, M. Zhang, and H. Su, "DEMAE: Diffusion-enhanced masked autoencoder for hyperspectral image classification with few labeled samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5527616.

[31] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 6840–6851.

[32] V. K. Sankarapu, C. Chitrodat, Y. Rathore, N. K. Singh, and P. Seth, "DLBacktrace: Model agnostic explainability for any deep learning model," in *2025 International Joint Conference on Neural Networks (IJCNN)*, 2025, doi: 10.1109/IJCNN64981.2025.11228632.

[33] L. Scheibenreif, M. Mommert, and D. Borth, "Masked vision transformers for hyperspectral image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2023, pp. 2166–2176.

[34] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010, doi: 10.1002/wics.101.

[35] J. Wang, K. Chen, R. Xu, Z. Liu, C. C. Loy, and D. Lin, "CARAFE: Content-aware reassembly of features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 3007–3016.