# DACN: Dual-Attention Convolutional Network for Hyperspectral Image Super-Resolution

Usman Muhammad[1] and Jorma Laaksonen[1]
[1] Department of Computer Science, Aalto University, Finland

***Abstract*—2D convolutional neural networks (CNNs) have attracted significant attention for hyperspectral image super-resolution tasks. However, a key limitation is their reliance on local neighborhoods, which leads to a lack of global contextual understanding. Moreover, band correlation and data scarcity continue to limit their performance. To mitigate these issues, we introduce DACN, a dual-attention convolutional network for hyperspectral image super-resolution. Specifically, the model first employs augmented convolutions, integrating multi-head attention to effectively capture both local and global feature dependencies. Next, we infer separate attention maps for the channel and spatial dimensions to determine where to focus across different channels and spatial positions. Furthermore, a custom optimized loss function is proposed that combines L2 regularization with spatial-spectral gradient loss to ensure accurate spectral fidelity. Experimental results on two hyperspectral datasets demonstrate that the combination of multi-head attention and channel attention outperforms either attention mechanism used individually. The source codes are publicly available at: https://github.com/Usman1021/dual-attention.**

***Index Terms*—Hperspectral imaging, attention, super-resolution, self-attention, loss function.**

## I. Introduction

Hyperspectral images (HSIs) typically offer high spectral resolution but suffer from low spatial resolution due to hardware constraints, while multispectral images (MSIs) generally have lower spectral resolution but higher spatial resolution [1]. The wide spectrum of hyperspectral images is extremely valuable for a variety of applications, including Earth observation, forest monitoring, and satellite image scene classification [2]–[5]. The primary goal of the single image super-resolution (SR) task is to enhance a degraded low-resolution (LR) image by reconstructing its corresponding high-resolution (HR) version. In the early years, single image super-resolution methods were mainly based on interpolation techniques, such as nearest neighbor, bilinear, and bicubic interpolation [6]. These methods were computationally simple and well suited to the hardware capabilities of that time. However, since they relied solely on computations of pixel value without considering image content or prior information, reconstructed high-resolution images often lacked fine details [1].

In recent years, convolutional neural network (CNN)-based methods have emerged as the dominant approach for super-resolution tasks [7]–[10]. CNN-based methods extract image features using shared weighted convolutional kernels, which exhibit local connectivity and translation invariance. Although these characteristics improve the efficiency and generalization of CNNs, they also introduce two key limitations: (a) convolution kernels are restricted by their local receptive fields, making it difficult to capture long-range pixel dependencies in images; and (b) the static weights of convolution kernels during inference prevent them from dynamically adapting to the input content [11].

In contrast, self-attention has recently gained recognition as an effective mechanism for capturing long-range dependencies in data. It has been particularly impactful in sequence modeling and generative tasks, such as natural language processing and machine translation [12]. The core principle of self-attention involves computing a weighted sum of input representations, where the weights are dynamically assigned based on the similarity between different positions in the input sequence [13]. This flexible weighting helps the model focus on important parts of the input while processing, unlike pooling or convolution, which use fixed weights and limited areas. In addition, attention [14] modules have been extensively studied in previous literature, not only determining where to focus, but also enhancing the representation of important features [15], [16].

Although attention-based models demonstrate excellent performance on multiple benchmarks (ImageNet-1K, MS COCO, and VOC 2007) [14], [17], most previous work uses attention for task-specific purposes [18]. In contrast, the super-resolution task requires low-resolution images that are typically not downsampled during feature extraction, as this would lead to further information loss, making HR reconstruction more challenging [11]. We argue that the potential of self-attention and attention mechanisms needs further exploration for hyperspectral SR tasks. Therefore, to effectively capture long-range pixel dependencies and model global relationships, a key question needs to be answered: How would performance be impacted by incorporating self-attention with an attention mechanism into convolutional networks for hyperspectral super-resolution tasks?

Motivated by the observations mentioned above, this paper presents a Dual-Attention Convolutional Network (DACN) for hyperspectral image super-resolution. The motivation behind adopting dual attention is not only to extend the receptive field by modeling global relationships [13], but also to enhance local feature refinement by emphasizing important channels and spatial regions [14]. In particular, self-attention aims at capturing global pixel dependencies, while the attention module helps refine the most important local features. In this way, the proposed dual-attention model balances global

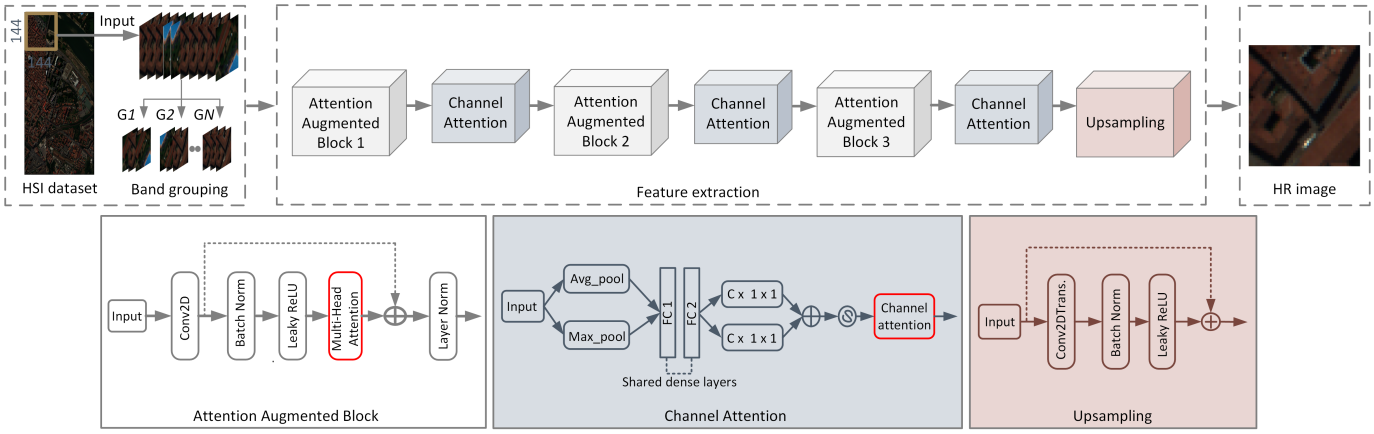


Fig. 1. An overview of the proposed DACN model: the white block on the left illustrates the integration of multi-head attention, while the gray block in the middle emphasizes channel attention. The beige-colored block on the right represents the upsampling module, which uses transposed convolution with a skip connection.

dependency modeling and local feature refinement, leading to more accurate and efficient super-resolution reconstruction. Additionally, to further enhance the performance of the proposed method, mean squared error (MSE), an L2 regularization-based constraint, and spatial-spectral gradient loss are combined into a custom loss function. In summary, our key contributions are threefold.

1) We present DACN for hyperspectral image super-resolution, combining multi-head and channel attention to enhance contextual modeling and spectral fidelity.
2) A custom loss function is proposed, integrating mean squared error (MSE), an L2 regularization-based constraint, and spatial-spectral gradient loss to ensure high-fidelity reconstruction.
3) Experiments on two hyperspectral datasets are conducted across multiple resolution degradation-restoration scenarios (2×, 4×, and 8×), demonstrating highly competitive performance on both datasets.

## II. Methodology

Fig. 1 presents an overview of the proposed model with three key components: (1) attention-augmented convolutional blocks, (2) channel attention, and (3) an upsampling module. We begin by explaining the band grouping process, which efficiently divides the bands into distinct groups while maintaining hyperspectral signatures. The subsequent sections provide an in-depth analysis of each component, including the band grouping method and the spatial-spectral gradient loss function used in the overall model.

### A. Band Grouping

Although hyperspectral images provide rich spectral detail through hundreds of bands, many of these bands can be potentially redundant due to high inter-band correlation. Therefore, we employ band grouping [19]. This approach partitions adjacent bands into overlapping groups, enabling seamless integration with our proposed model. Specifically, hyperspectral bands are structured into overlapping subgroups by defining a fixed group size with a designated overlap, ensuring that consecutive subgroups share common bands.

### B. Attention Augmented Convolution

Given a low-resolution hyperspectral image $Y \in \mathbb{R}^{M\times N\times C}$, where $M$, $N$, and $C$ represent the spatial height, width, and number of spectral bands, respectively, our objective is to reconstruct a high-resolution image $\hat{Y} \in \mathbb{R}^{\beta M\times \beta N\times C}$, where the upscaling factor $\beta \in \{2, 4, 8\}$.

To accomplish this, we develop a deep neural network $\mathcal{G}(Y;\phi)$ that effectively learns the LR-to-HR mapping while maintaining spectral fidelity. Specifically, we begin by empirically developing three blocks, each of which contains three main components: (1) standard convolution, (2) multi-head self-attention, and (3) residual connection. First, a standard 2D convolution is applied to the input feature map $X_{\text{in}} \in \mathbb{R}^{H\times W\times C}$, where $H$, $W$, and $C$ denote the spatial height, width, and number of channels, respectively. The convolution operation is defined as [13]:

$$Z_{\text{out}} = \mathcal{F}_{\text{conv}}(X_{\text{in}}; W, b) \tag{1}$$

where $\mathcal{F}_{\text{conv}}$ represents the convolution function with learnable weights $W$ and biases $b$. This is followed by:

$$X_{\text{out}} = \phi\left(\nu(Z_{\text{out}})\right) \tag{2}$$

where $\nu(\cdot)$ denotes the batch normalization operation and $\phi(\cdot)$ is the LeakyReLU non-linear activation function applied element-wise. Then, the multi-head self-attention mechanism is applied. For each attention head, the input is linearly transformed into queries $Q_h$, keys $K_h$, and values $V_h$:

$$(Q_h, K_h, V_h) = X_{\text{out}}(W_Q, W_K, W_V) \tag{3}$$

where $Q_h \in \mathbb{R}^{T\times d_k}$ is the matrix of queries, $K_h \in \mathbb{R}^{T\times d_k}$ is the matrix of keys, $V_h \in \mathbb{R}^{T\times d_v}$ is the matrix of values, and $W_Q, W_K, W_V$ are learnable weight matrices with dimensions

$d_{\text{model}} \times d_k$, $d_{\text{model}} \times d_k$, and $d_{\text{model}} \times d_v$, respectively. The attention scores are computed as:

$$N_{\text{attention}}(Q_h, K_h, V_h) = \text{softmax}\left(\frac{Q_h K_h^T}{\sqrt{d_k}}\right) V_h \quad (4)$$

where $d_k$ is the dimensionality of the key vectors. The softmax function ensures that the attention scores sum up to one across the sequence dimension. The outputs of all attention heads are concatenated and linearly projected:

$$Z_{\text{attn}} = \text{Concat}(H_1, H_2, \ldots, H_h) W_O \quad (5)$$

where $h$ is the number of attention heads, $W_O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$ is a learnable weight matrix used to combine the outputs from all heads, and $Z_{\text{attn}} \in \mathbb{R}^{T \times d_{\text{model}}}$ is the final multi-head attention output.

Finally, the attention output is added to the convolutional output using a residual connection:

$$A_{\text{res}} = X_{\text{out}} + Z_{\text{attn}} \quad (6)$$

where the residual connection is established by directly adding $Z_{\text{attn}}$ to $X_{\text{out}}$. This is followed by:

$$X_{\text{final}} = \lambda(A_{\text{res}}) \quad (7)$$

where $\lambda(\cdot)$ denotes the layer normalization function, and $X_{\text{final}}$ represents the final output after applying layer normalization to the result of the residual connection between the multi-head self-attention (MHSA) output and the convolutional output [13].

### C. Channel Attention

The channel attention is the second main component of our model, which enhances important channels in a feature map by computing global pooling statistics, passing them through fully connected layers, and generating channel-wise attention weights [14]. This allows the model to focus on the most relevant information while suppressing less important features. Mathematically, each block requires an input feature map $X \in \mathbb{R}^{H \times W \times C}$, where $H, W$ are the spatial dimensions (height and width), and $C$ is the number of channels. The output is a feature map of the same shape but refined using attention as $X' \in \mathbb{R}^{H \times W \times C}$.

In particular, two types of pooling are applied to compute global information from each channel. For instance, the global average pooling (GAP) is employed as [14]:

$$F_{\text{avg}} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{i,j,c} \quad (8)$$

where $F_{\text{avg}} \in \mathbb{R}^C$ represents the average value of each channel. Similarly, global max pooling (GMP) is computed as:

$$F_{\text{max}} = \max_{i,j} X_{i,j,c} \quad (9)$$

where $F_{\text{max}} \in \mathbb{R}^C$ captures the most activated value per channel. The pooled values are passed through two fully connected layers. The first fully connected layer (dimensionality reduction) is defined as [14]:

$$F'_{\text{avg}} = \text{ReLU}(W_1 \cdot F_{\text{avg}} + b_1) \quad (10)$$

TABLE I
QUANTITATIVE RESULTS AND MODEL COMPLEXITY

| **Ablation Study on PaviaU (4×)** | | |
|---|---|---|
| **Model Variant** | **MPSNR ↑** | **SAM ↓** |
| DACN without band grouping | 29.73 | 5.500 |
| DACN with group size 16 | 29.85 | 2.939 |
| DACN with group size 32 | 30.67 | 4.574 |
| FGIN with group size 32 [20] | 30.33 | 4.819 |
| DSDCN with group size 32 [21] | 30.52 | 4.807 |
| DACN with group size 48 | 30.37 | 5.340 |
| DACN without multi-head attention | 30.49 | 4.543 |
| DACN without channel attention | 30.51 | 4.572 |
| DACN without custom loss | 30.53 | 4.537 |

$$F'_{\text{max}} = \text{ReLU}(W_1 \cdot F_{\text{max}} + b_1) \quad (11)$$

where $W_1 \in \mathbb{R}^{C/r \times C}$ is a weight matrix that reduces channel dimensions by a factor of $r$, and $b_1$ is the bias term for the first FC layer. ReLU is applied to introduce non-linearity. Similarly, the second fully connected layer (restoring channel dimension) is defined as:

$$F''_{\text{avg}} = W_2 \cdot F'_{\text{avg}} + b_2 \quad (12)$$

$$F''_{\text{max}} = W_2 \cdot F'_{\text{max}} + b_2 \quad (13)$$

where $W_2 \in \mathbb{R}^{C \times C/r}$ restores the original number of channels, and $b_2$ is the bias term for the second FC layer. Therefore, both pooling outputs are merged as [14]:

$$P_{\text{attention}} = F''_{\text{avg}} + F''_{\text{max}} \quad (14)$$

where $P_{\text{attention}} \in \mathbb{R}^C$ represents the attention scores for each channel. Moreover, a sigmoid function is applied to scale the attention values between 0 and 1:

$$S_c = \sigma(P_{\text{attention}}) \quad (15)$$

where $S_c \in \mathbb{R}^C$ represents the final attention weights for each channel, and $\sigma$ is the sigmoid activation function. Finally, attention to the feature map is applied as [14]:

$$X' = X \odot S_c \quad (16)$$

where $\odot$ denotes element-wise multiplication, and $X'$ is the output feature map where the attention weights are applied.

### D. Upsampling

The upsampling block consists of transposed convolutions [26], batch normalization, LeakyReLU activation, and skip connections. Given an input feature map $F_{\text{in}} \in \mathbb{R}^{H \times W \times C}$, a transposed convolution operation $\mathcal{T}$ is applied to produce:

$$F_{\text{up}} = \mathcal{T}(F_{\text{in}})$$

where $\mathcal{T}$ denotes the transposed convolution. Batch normalization $\nu(\cdot)$ and LeakyReLU activation $\phi(\cdot)$ are then applied element-wise:

$$F_{\text{act}} = \phi\left(\nu(F_{\text{up}})\right)$$

The final output is obtained by concatenating with the skip connection:

$$F_{\text{out}} = \text{Concat}(F_{\text{act}}, F_{\text{skip}})$$

TABLE II
EVALUATION ON DATASETS (PAVIAC, PAVIAU) IN DIFFERENT SCALING SETUPS. THE COMPARISON RESULTS ARE REPORTED FROM [8].

| Scale Factor | Model | PaviaC | | | PaviaU | | |
|---|---|---|---|---|---|---|---|
| | | MPSNR↑ | MSSIM↑ | SAM↓ | MPSNR↑ | MSSIM↑ | SAM↓ |
| 2× | VDSR [22] | 34.87 | 0.9501 | 3.689 | 34.03 | 0.9524 | 3.258 |
| | EDSR [23] | 34.58 | 0.9452 | 3.898 | 33.98 | 0.9511 | 3.334 |
| | MCNet [9] | 34.62 | 0.9455 | 3.865 | 33.74 | 0.9502 | 3.359 |
| | MSDformer [24] | 35.02 | 0.9493 | 3.691 | 34.15 | 0.9553 | 3.211 |
| | MSFMNet [10] | 35.20 | 0.9506 | 3.656 | 34.98 | 0.9582 | 3.160 |
| | AS3 ITransUNet [25] | 35.22 | 0.9511 | 3.612 | 35.16 | 0.9591 | 3.149 |
| | PDENet [7] | 35.24 | 0.9519 | 3.595 | 35.27 | 0.9594 | 3.142 |
| | CSSFENet [8] | 35.52 | 0.9544 | 3.542 | 35.92 | **0.9625** | **3.038** |
| | DACN (Ours) | **36.77** | **0.9599** | **3.390** | **36.11** | 0.9486 | 3.290 |
| 4× | VDSR [22] | 28.31 | 0.7707 | 6.514 | 29.90 | 0.7753 | 4.997 |
| | EDSR [23] | 28.59 | 0.7782 | 6.573 | 29.89 | 0.7791 | 5.074 |
| | MCNet [9] | 28.75 | 0.7826 | 6.385 | 29.99 | 0.7835 | 4.917 |
| | MSDformer [24] | 28.81 | 0.7833 | 5.897 | 30.09 | 0.7905 | 4.885 |
| | MSFMNet [10] | 28.87 | 0.7863 | 6.300 | 30.28 | 0.7948 | 4.861 |
| | AS3 ITransUNet [25] | 28.87 | 0.7893 | 5.972 | 30.28 | 0.7940 | 4.859 |
| | PDENet [7] | 28.95 | 0.7900 | 5.876 | 30.29 | 0.7944 | 4.853 |
| | CSSFENet [8] | 29.05 | 0.7961 | 5.816 | **30.68** | **0.8107** | 4.839 |
| | DACN (Ours) | **29.90** | **0.8224** | **4.656** | 30.67 | 0.8015 | **4.574** |
| 8× | VDSR [22] | 24.80 | 0.4944 | 7.588 | 27.02 | 0.5962 | 7.133 |
| | EDSR [23] | 25.06 | 0.5282 | 7.507 | 27.46 | 0.6302 | 6.678 |
| | MCNet [9] | 25.09 | 0.5391 | 7.429 | 27.48 | 0.6254 | 6.683 |
| | MSDformer [24] | 25.21 | 0.5462 | 7.427 | 27.32 | 0.6341 | 6.668 |
| | MSFMNet [10] | 25.25 | 0.5464 | 7.449 | 27.58 | 0.6356 | 6.615 |
| | AS3 ITransUNet [25] | 25.25 | 0.5435 | 7.417 | 27.68 | 0.6413 | 6.574 |
| | PDENet [7] | 25.28 | 0.5436 | 7.402 | 27.73 | 0.6457 | 6.531 |
| | CSSFENet [8] | 25.35 | 0.5493 | 7.306 | 27.82 | **0.6569** | 6.505 |
| | DACN (Ours) | **25.78** | **0.5794** | **6.007** | **28.04** | 0.6296 | **6.190** |

### *E. Custom Loss Function*

The model is optimized using a custom loss function that combines mean squared error (MSE) with $\ell_2$ regularization and spatial-spectral gradient loss [8]. Thus, the total loss is computed as the sum of the MSE loss and a scaled $\ell_2$ regularization term:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MSE}} + \alpha \cdot \mathcal{L}_{\ell_2} \tag{17}$$

where $\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(Y_{\text{true}}^{(i)} - Y_{\text{pred}}^{(i)})^2$ represents the Mean Squared Error, $\mathcal{L}_{\ell_2} = \sum_{\theta\in\Theta}\theta^2$ is the $\ell_2$ regularization term applied to the trainable weights $\Theta$ of the model, and $\alpha = 10^{-4}$ controls the regularization strength. Additionally, the spatial-spectral gradient loss ensures consistency in both spatial and spectral gradients [8]:

$$\mathcal{L}_{\text{grad}} = \mathcal{L}_{\text{spat}} + \mathcal{L}_{\text{spec}} \tag{18}$$

where $\mathcal{L}_{\text{spat}} = \frac{1}{N}\sum_{i=1}^{N}\left[(D_{\text{true}}^{x} - D_{\text{pred}}^{x})^2 + (D_{\text{true}}^{y} - D_{\text{pred}}^{y})^2\right]$ computes the spatial gradient loss, and $\mathcal{L}_{\text{spec}} = \frac{1}{N}\sum_{i=1}^{N}(D_{\text{true}}^{s} - D_{\text{pred}}^{s})^2$ computes the spectral gradient loss. Here, $D_{\text{true}}$ and $D_{\text{pred}}$ represent the gradients of the ground truth and predicted images, respectively. The final combined loss function integrates the MSE with $\ell_2$ regularization and the spatial-spectral gradient loss:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{total}} + \mathcal{L}_{\text{grad}} \tag{19}$$

## III. EXPERIMENTAL SETUP

In our study, we utilize two publicly available hyperspectral datasets, PaviaC and PaviaU, which consist of 102 and 103 spectral bands, respectively.

### *A. Implementation*

To generate low-resolution inputs, the extracted patches are downsampled using area-based interpolation with scale factors of 2×, 4×, and 8×. A patch size of $144 \times 144$ is used for the training, validation, and test sets, following the protocol in [8]. The model is trained using the Adam optimizer with a batch size of 8. An early stopping criterion is applied to prevent overfitting and eliminate the need for a fixed number of training epochs. Three widely used metrics are used to report the results, such as the mean peak signal-to-noise ratio (MPSNR), the mean structural similarity index (MSSIM) and the spectral angle mapper (SAM) [27].

### *B. Ablation Study*

To evaluate the contribution of each component in our DACN architecture, we performed an ablation study on the PaviaU dataset with a 4× upscaling factor. The results are presented in Table I, measured in terms of MPSNR and SAM. To make a fair comparison, we compare our method with others such as FGIN [20] and DSDCN [21], using the same band grouping settings. It can be seen that DACN with a band grouping size of 32 achieves the highest MPSNR (30.67 dB), indicating superior reconstruction quality. Although using a grouping size of 16 results in the lowest SAM (2.939), it comes at the cost of reduced MPSNR. In contrast, increasing the grouping size to 48 degrades the performance in both metrics. This confirms that a moderate grouping size of 32 provides the best balance between spectral detail preservation and spatial coherence.

Furthermore, the removal of multi-head attention or channel attention leads to noticeable performance drops: MPSNR decreases to 30.49 and 30.51, respectively, compared to 30.67

when both mechanisms are used. These results confirm that both multi-head self-attention and channel attention contribute positively to the model's ability to capture long-range dependencies and emphasize salient features.

### C. Comparison with State-of-the-Art Methods

We evaluate DACN against several state-of-the-art models, including VDSR [22], EDSR [23], MCNet [9], MSDformer [24], MSFMNet [10], AS3 ITransUNet [25], PDENet [7], and CSSFENet [8]. As shown in Table II, DACN consistently achieves highly competitive performance across all scale factors.

For $2\times$ upscaling, it achieves the highest MPSNR (36.77 dB) and MSSIM (0.9599) on the PaviaC dataset, and performs competitively on PaviaU. At $4\times$, DACN outperforms all baselines on the PaviaC dataset across all metrics. For the more challenging $8\times$ scale, DACN shows clear improvements, particularly in MPSNR and SAM on both datasets. These results confirm DACN's strong capability in preserving spectral and spatial fidelity under various degradation levels.

## IV. Conclusion

In this work, we introduced DACN, a dual-attention convolutional network designed for hyperspectral image super-resolution. The proposed model effectively integrates self-attention mechanisms with convolutional architectures, addressing the limitations of traditional CNN-based approaches that primarily capture local features while overlooking global dependencies. By incorporating multi-head attention, DACN enhances feature representation, while channel and spatial attention modules enable adaptive refinement, capturing both spatial and spectral dependencies more effectively. Extensive experiments on the PaviaC and PaviaU datasets demonstrate that DACN consistently achieves competitive performance compared to state-of-the-art models across various scaling factors.

## Acknowledgment

This project has been funded by the European Union's NextGenerationEU instrument and the Research Council of Finland under grant № 348153, as part of the project *Artificial Intelligence for Twinning the Diversity, Productivity and Spectral Signature of Forests* (ARTISDIG).

## References

[1] X. Li, Y. Yuan, and Q. Wang, "Hyperspectral and multispectral image fusion based on band simulation," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 3, pp. 479–483, 2019.

[2] U. Muhammad, W. Wang, S. P. Chattha, and S. Ali, "Pre-trained vggnet architecture for remote-sensing image scene classification," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 1622–1627.

[3] U. Muhammad, W. Wang, and A. Hadid, "Feature fusion with deep supervision for remote-sensing image scene classification," in *2018 IEEE 30th international conference on tools with artificial intelligence (ICTAI)*. IEEE, 2018, pp. 249–253.

[4] U. Muhammad, W. Wang, A. Hadid, and S. Pervez, "Bag of words kaze (bowk) with two-step classification for high-resolution remote sensing images," *IET Computer Vision*, vol. 13, no. 4, pp. 395–403, 2019.

[5] U. Muhammad, M. Z. Hoque, W. Wang, and M. Oussalah, "Patch-based discriminative learning for remote sensing scene classification," *Remote Sensing*, vol. 14, no. 23, p. 5913, 2022.

[6] L. Jiang, M. Zhong, and F. Qiu, "Single-image super-resolution based on a self-attention deep neural network," in *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, 2020, pp. 387–391.

[7] J. Hou, Z. Zhu, J. Hou, H. Zeng, J. Wu, and J. Zhou, "Deep posterior distribution-based embedding for hyperspectral image super-resolution," *IEEE Transactions on Image Processing*, vol. 31, pp. 5720–5732, 2022.

[8] J. Zhang, R. Zheng, Z. Wan, R. Geng, Y. Wang, Y. Yang, X. Zhang, and Y. Li, "Hyperspectral image super-resolution based on feature diversity extraction," *Remote Sensing*, vol. 16, no. 3, p. 436, 2024.

[9] Q. Li, Q. Wang, and X. Li, "Mixed 2D/3D convolutional network for hyperspectral image super-resolution," *Remote sensing*, vol. 12, no. 10, p. 1660, 2020.

[10] J. Zhang, M. Shao, Z. Wan, and Y. Li, "Multi-scale feature mapping network for hyperspectral image super-resolution," *Remote Sensing*, vol. 13, no. 20, p. 4180, 2021.

[11] L. Zhao, J. Gao, D. Deng, and X. Li, "SSIR: Spatial shuffle multi-head self-attention for single image super-resolution," *Pattern Recognition*, vol. 148, p. 110195, 2024.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[13] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3286–3295.

[14] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[15] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention." advances in neural information processing systems," in *Proc. of Neural Information Processing Systems (NIPS)*, vol. 2, 2014.

[16] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," *arXiv preprint arXiv:1412.7755*, 2014.

[17] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[18] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "BAM: Bottleneck attention module," *arXiv preprint arXiv:1807.06514*, 2018.

[19] Z. Wang, D. Li, M. Zhang, H. Luo, and M. Gong, "Enhancing hyperspectral images via diffusion model and group-autoencoder super-resolution network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5794–5804.

[20] U. Muhammad and J. Laaksonen, "A fusion-guided inception network for hyperspectral image super-resolution," *arXiv preprint arXiv:2505.03431*, 2025.

[21] U. Muhammad, J. Laaksonen, and L. Mihaylova, "Towards lightweight hyperspectral image super-resolution with depthwise separable dilated convolutional network," *arXiv preprint arXiv:2505.00374*, 2025.

[22] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1646–1654.

[23] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 136–144.

[24] S. Chen, L. Zhang, and L. Zhang, "Msdformer: Multi-scale deformable transformer for hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[25] Q. Xu, S. Liu, J. Wang, B. Jiang, and J. Tang, "AS 3 ITransUNet: Spatial-Spectral Interactive Transformer U-Net with Alternating Sampling for Hyperspectral Image Super-Resolution," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[26] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 391–407.

[27] Y. Chudasama, U. Muhammad, V. Mayra, F. Guiotte, and J. Laaksonen, "A comparison of hyperspectral super-resolution techniques for boreal forest imagery," in *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2024, pp. 1226–1230.