# HSCT: HIERARCHICAL SELF-CALIBRATION TRANSFORMER FOR HYPERSPECTRAL IMAGE SUPER-RESOLUTION

Jinliang Hou, Yifan Zhang, Yuanjie Zhi, Rugui Yao,Shaohui Mei
*School of Electronics and Information*, *Northwestern Polytechnical University*, 710129 Xi'an, China
jlhou@mail.nwpu.edu.cn, {yifanzhang, zyjyuyan, yaorg, meish}@nwpu.edu.cn

***Abstract*—Hyperspectral image (HSI) super-resolution is to improve the spatial resolution while preserving spectral fidelity. Existing CNN- and Transformer-based methods face challenges in simultaneously capturing multi-scale local and global features and maintaining spectral accuracy. To address these issues, in this paper, the *Hierarchical Self-Calibration Transformer (HSCT)* is proposed for HSI super-resolution, combining the merits of CNNs and Transformers in a multi-stage framework. Specifically, CNNs are utilized for local feature extraction, leveraging inductive biases to enrich feature representations, while Transformers focus on global feature extraction to model complex and global dependencies. A variable Window-based Self-Attention with window shifting is designed to extract multi-scale spatial features, while a Channel Self-Attention refines spectral features to ensure fidelity, parallel integration of which enables efficient spatial-spectral feature learning. Additionally, Self-Calibration Convolution and Residual Connections are integrated to improve feature representations and model stability. Extensive experiments demonstrate the outperformance of the proposed HSCT over representative traditional and state-of-the-art deep learning-based methods, both visually and quantitatively.**

***Index Terms*—Transformer, super-resolution, hyperspectral image, deep learning.**

## I. Introduction

Hyperspectral images (HSIs) provide rich spectral information across hundreds of continuous bands, crucial for environmental monitoring, precision agriculture, and military reconnaissance [1]. However, imaging limitations often lead to low spatial resolution, restricting practical applications. Super-resolution techniques are effective in enhancing image resolution, with single hyperspectral image super-resolution methods gaining significant attention for their practicality and convenience, as no auxiliary images are required.

Early methods like interpolation relied on manually extracted features and struggled to restore HSI details. Following the success of deep learning in greyscale and RGB image super-resolution, deep learning-based hyperspectral image super-resolution has rapidly developed. 3D CNNs, such as the full 3D CNN [2] and the mixed 2D/3D CNN [3], which process both spatial and spectral dimensions, have gained popularity. However, these methods suffer from high computational complexity. To address this, GDRRN [4] and SSPSR [5] leverage spectral grouping and spatial-spectral attention to reduce model complexity. Despite this, pure CNNs may ignore long-range dependencies and global semantics. Transformers, such as SwinIR [6], have demonstrated success in natural image super-resolution but are still in the early stages for HSIs. Notable works, such as Interactformer [7] and MSDformer [8], incorporate Transformer-based architectures, while the high computational complexity remains challenges for efficient HSI super-resolution.

To address the limitations in capturing fine-grained features, a novel HSI super-resolution method is proposed in this paper. It utilizes a hierarchical Transformer framework with variable windows combining local window self-attention and inter-window interactions, striking a balance between computational efficiency and global context preservation. Designed in a multi-stage structure, it employs spectral feature extraction module to ensure fidelity, self-calibrated convolution to enhance sensitivity and translation equivariance, and skip connections to stabilize training. Extensive experiments are deployed to demonstrate its effectiveness in improving reconstruction performance while preserving spectral information.

## II. Methodology

### A. Network structure

In HSI super-resolution, shallow and deep features provide complementary and distinct information, both of which are crucial for reconstructing high-resolution HSIs. The proposed **Hierarchical Self-Calibration Transformer (HSCT)** for hyperspectral image super-resolution, illustrated in Fig. 1, is composed of the Shallow Feature Extraction (SFE), Deep Feature Extraction (DFE), and High-resolution Image Reconstruction (HIR) modules. The input low-resolution hyperspectral image is denoted as $I_{\text{LR}} \in \mathbb{R}^{h \times w \times C}$, where $h$ and $w$ represent the height and width of the image, respectively, and $C$ denotes the number of spectral bands. The super-resolution process can be formulated as:

$$I_{\text{SR}} = H_{\text{HSCT}}(I_{\text{LR}}) \tag{1}$$

where $H_{\text{HSCT}}$ denotes the function of the proposed HSCT, $I_{\text{SR}} \in \mathbb{R}^{rh \times rw \times C}$ represents the reconstructed high-resolution HSI, and $r$ is the super-resolution scale factor.

**Shallow Feature Extraction:** Hyperspectral images typically exhibit considerable spectral redundancy. To extract

This work was supported by the National Natural Science Foundation of China under Grants 62171381 and 62006193. *Corresponding authors: Yifan Zhang and Shaohui Mei.*

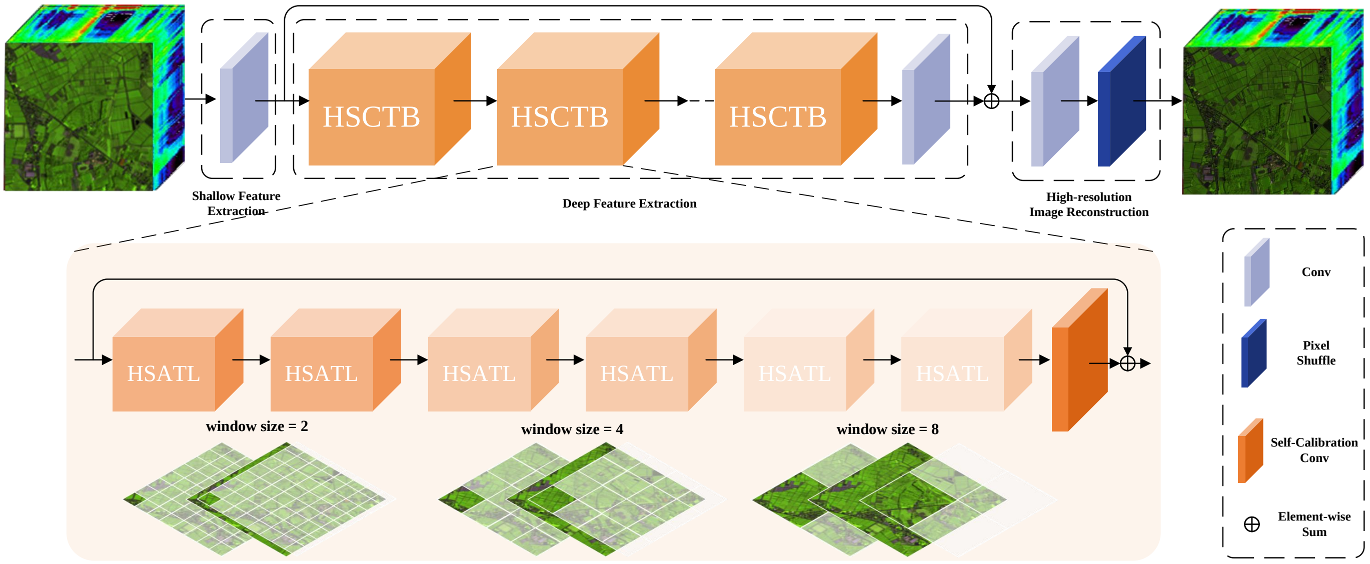

Fig. 1. Structrue of the proposed HSCT network.

shallow features from the low-resolution image $I_{\text{LR}}$, a simple $3 \times 3$ convolutional layer is employed. This layer captures early visual features and simultaneously reduces the channel dimension by leveraging spectral correlations:

$$F_S = H_{\text{Conv}}(I_{\text{LR}}) \tag{2}$$

where $H_{\text{Conv}}$ represents the convolutional layer, $F_S \in \mathbb{R}^{h \times w \times D}$ denotes the shallow feature map and $D$ is the embedding dimension. The extracted shallow features serve as basis for subsequent Transformer-based feature extraction.

**Deep Feature Extraction:** To effectively capture the deep features of low-resolution hyperspectral images, an advanced architecture is employed in the proposed HSCT. The architecture comprises stacked Hierarchical Self-Calibration Transformer Blocks (HSCTBs) and a $3 \times 3$ convolutional layer. Each HSCTB is designed as a multi-layer composite structure, integrating two core components: the Hybrid Self-Attention Transformer Layer (HSATL) and Self-Calibration Convolution (SCConv). By synergizing self-attention with self-calibration convolution, HSCTBs enable efficient and reliable feature extraction in HSIs.

Each HSCTB comprises six HSATLs, specifically designed to deeply explore and capture the intricate global and local spatial-spectral dependencies inherent in HSIs.

The HSATL employs a hybrid self-attention (SA) structure, enclosing two parallel SA structures for diverse feature processing. One is Window-based Self-Attention (WSA), which focuses on the spatial dimension by dividing the image into non-overlapping windows and computing SA within each window to capture local spatial features. Through window interactions, WSA also facilitates the capture of global features. Typically implemented as multi-head form, denoted as WMSA, it enables effective feature extraction across multiple dimensions. The other is Channel Self-Attention (CSA), which focuses on the spectral dimension. By applying SA across spectral channels, CSA identifies and enhances the interactions and correlations between different spectral bands. To ensure coherent visual representation and mitigate optimization conflicts, the output of the channel attention is scaled by a weight $\alpha$ [9].

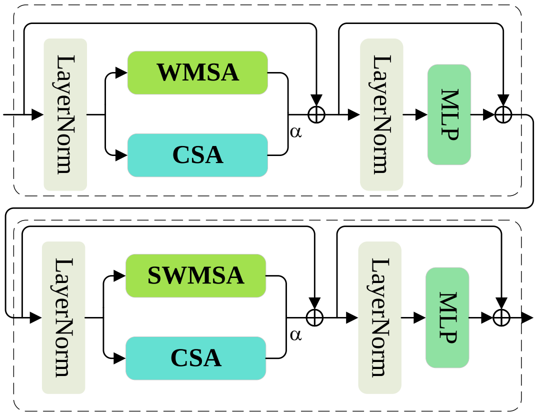

Fig. 2. Two cascaded HSATLs with the same window size.

The six HSATLs in HSCTB are arranged into three groups, each containing two HSATLs, forming a ‘Window and Shifted Window’ structure, as illustrated in Fig. 2. Window-based self-attention confines global attention computation to an $M \times M$ local window, alleviating quadratic computational complexity. Interactions and connections between windows are facilitated by alternating between regular and shifted windows. Unlike previous Swin Transformer-based methods with fixed windows, the window sizes are set to be 2, 4, and 8 for the three groups of HSATLs, respectively, enabling a progressive expansion of the receptive field and allowing the model to leverage both local information and global context simultaneously. The fine details and overall structural features of the image can thus be effectively captured.

The final HSCTB layer includes a self-calibrated convolution and a residual connection. The former enhances translational equivariance and feature extraction with larger receptive fields and multi-scale encoding, while the latter aggregates multi-level features for better performance.

The convolutional layers at the final stage of deep feature extraction can effectively enhance the extracted features, strengthening the fusion of shallow and deep features. The deep feature extraction process is expressed as:

$$\begin{aligned} F_{D_k} &= H_{\text{HSCTB}_k}(F_{D_{k-1}}), k = 1, 2, ..., K \\ F_D &= H_{\text{Conv}}(F_{D_K}) \end{aligned} \tag{3}$$

where $H_{\text{HSCTB}_k}$ denotes the $k^{\text{th}}$ HSCTB and $H_{\text{Conv}}$ is the last convolutional layer. The input of the first HSCTB is $F_S$, that is, $F_{D_0} = F_S$.

**High-resolution Image Reconstruction:** Prior to reconstruction, shallow features from long skip connections are integrated with deep features to leverage multi-frequency information and stabilize training. These fused features are used to form the high-resolution hyperspectral image $I_{\text{SR}}$ through the high-resolution image reconstruction operation $H_{\text{HIR}}$ composed of convolution and Pixel Shuffle:

$$I_{\text{SR}} = H_{\text{HIR}}(F_S + F_D) \tag{4}$$

### B. Loss function

To guide the model in learning complex hyperspectral characteristics and generating both visually realistic and spectrally accurate high-resolution outputs, a loss function containing three key components is specifically designed:

$$L = L_1 + \lambda_1 L_{\text{SAM}} + \lambda_2 L_{\text{Gra}} \tag{5}$$

where the weights $\lambda_1$ and $\lambda_2$ allow flexible adjustment of each component's contribution. The L1 loss $L_1$ ensures sparsity and mitigates over-smoothing by addressing pixel-level differences:

$$L_1 = \frac{1}{N} \sum_{n=1}^{N} \|I_{\text{SR}}^n - I_{\text{HR}}^n\|_1 \tag{6}$$

where $I_{\text{SR}}^n$ and $I_{\text{HR}}^n$ represent the $n$-th reconstructed super-resolution HSI and the corresponding high-resolution ground truth, respectively, and $N$ represents the number of images in a training batch. The Spectral Angle Mapper (SAM) loss, denoted as $L_{\text{SAM}}$, preserves spectral fidelity by constraining the angles between spectral vectors:

$$L_{\text{SAM}} = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{\pi} \arccos\left( \frac{I_{\text{HR}}^n \cdot I_{\text{SR}}^n}{\|I_{\text{HR}}^n\|_2 \cdot \|I_{\text{SR}}^n\|_2} \right) \tag{7}$$

Lastly, the gradient loss $L_{\text{Gra}}$, preserving structural details and sharpness by focusing on edge information, is formulated as:

$$L_{\text{Gra}} = \frac{1}{N} \sum_{n=1}^{N} \|M(I_{\text{HR}}^n) - M(I_{\text{SR}}^n)\|_1 \tag{8}$$

where $M(I) = \|(\nabla_h I, \nabla_w I, \nabla_c I)\|_2$ denotes the gradient map, combining the gradients of the HSI $I$ in the spatial (height $\nabla_h$, width $\nabla_w$) and spectral ($\nabla_c$) domains.

## III. Experimental Results and Analysis

### A. Experimental Setup

To comprehensively evaluate the super-resolution performance of the proposed HSCT, experiments are conducted on the Chikusei and Houston datasets. The Chikusei dataset has an area of $2517 \times 2335$ pixels with 128 spectral bands, while the Houston dataset has an area of $4172 \times 1202$ pixels with 48 bands. For testing, four non-overlapping patches are cropped: $512 \times 512$ pixels for Chikusei and $256 \times 256$ pixels for Houston. The remaining sections are used for overlapping training patches, with 10% randomly selected for validation. The patch and overlap sizes (in pixels) for different super-resolution scales (×2 and ×4) are summarized in Table I.

TABLE I
Dataset partition and Patch specifications

| Dataset | Testing<br>Patch Size | Training<br>Patch Size (Overlap Size) | |
|---|---|---|---|
| | | ×2 | ×4 |
| Chikusei | 512 × 512 | 32 × 32 (16) | 64 × 64 (32) |
| Houston | 256 × 256 | | |

To evaluate the super-resolution performance of the proposed HSCT, it is compared with several representative methods, including the traditional baseline Bicubic interpolation method, the group-based method SSPSR [5] modeling spatial-spectral priors through spectral grouping, the 3D CNN-based method MCNet [3] which captures spatial and spectral correlations using 3D convolutions, and the Transformer-based method MSDformer [8] using a deformable design. To ensure a comprehensive evaluation, besides visual comparison, peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), spectral angle mapping (SAM), and relative global synthesis error (ERGAS) are utilized as objective evaluation metrics for quantitative evaluation.

### B. Results and Analysis

The experimental results for different methods on the two datasets are presented in Tables II and III, with visual comparisons shown in Figs. 3 and 5, which clearly demonstrate the outperformance of the proposed HSCT over the compared in both visual quality and objective evaluation metrics. Visually, HSCT generates sharper details, clearer textures, and more accurate reconstructions with fewer artifacts, as confirmed by the error maps. Quantitatively, HSCT achieves the highest PSNR and SSIM, and the lowest SAM and ERGAS across both datasets, indicating superior image quality and spectral fidelity. Additionally, the spectral curves in Fig. 4 show that HSCT preserves spectral information more effectively, with minimal deviations from the ground truth curves compared to other methods, further highlighting its superior spectral fidelity. In summary, HSCT demonstrates clear advantages in both spatial details and spectral accuracy over compared methods.

TABLE II
EVALUATION METRICS ON CHIKUSEI DATASET

| Method | Scale | PSNR ↑ | SSIM ↑ | SAM ↓ | ERGAS ↓ |
|---|---|---|---|---|---|
| Bicubic | ×2 | 43.5762 | 0.9744 | 1.7690 | 3.4524 |
| SSPSR | | <u>47.6039</u> | <u>0.9898</u> | <u>1.1477</u> | <u>2.2617</u> |
| MCNet | | 47.0257 | 0.9880 | 1.3257 | 2.3602 |
| MSDformer | | 47.3885 | 0.9892 | 1.2276 | 2.3286 |
| HSCT | | **47.9978** | **0.9907** | **1.1112** | **2.1917** |
| Bicubic | ×4 | 37.7709 | 0.8979 | 3.4026 | 6.6529 |
| SSPSR | | <u>39.9999</u> | <u>0.9402</u> | **2.3435** | <u>5.1071</u> |
| MCNet | | 39.6664 | 0.9334 | 2.7942 | 5.3104 |
| MSDformer | | 39.6714 | 0.9357 | 2.5329 | 5.2997 |
| HSCT | | **40.0762** | **0.9415** | <u>2.3637</u> | **5.0902** |

TABLE III
EVALUATION METRICS ON HOUSTON DATASET

| Method | Scale | PSNR ↑ | SSIM ↑ | SAM ↓ | ERGAS ↓ |
|---|---|---|---|---|---|
| Bicubic | ×2 | 49.9806 | 0.9928 | 1.2007 | 1.2516 |
| SSPSR | | 52.8735 | 0.9964 | <u>0.9153</u> | 0.8886 |
| MCNet | | <u>53.0596</u> | <u>0.9965</u> | 0.9305 | <u>0.8725</u> |
| MSDformer | | 52.9889 | 0.9964 | 0.9290 | 0.8742 |
| HSCT | | **53.3033** | **0.9966** | **0.8721** | **0.8447** |
| Bicubic | ×4 | 43.3312 | 0.9651 | 2.4244 | 2.7118 |
| SSPSR | | 46.2726 | 0.9818 | <u>1.6740</u> | 1.9193 |
| MCNet | | 46.3730 | 0.9820 | 1.9395 | 1.9046 |
| MSDformer | | <u>46.8234</u> | <u>0.9837</u> | 1.7130 | <u>1.7902</u> |
| HSCT | | **46.9966** | **0.9844** | **1.6115** | **1.7548** |

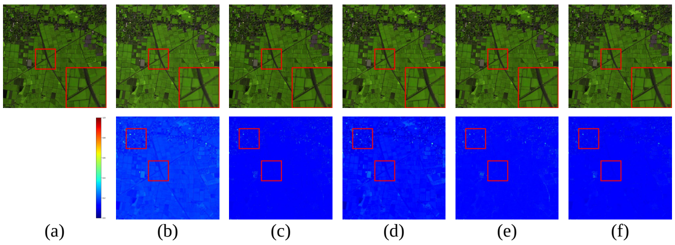

Fig. 3. Reconstructed super-resolution results on Chikusei dataset with scale factor 2, where the first row shows the pseudo-color images, and the second row presents the error maps. (a) Ground truth, (b) Bicubic, (c) SSPSR, (d) MCNet, (e) MSDformer, and (f) HSCT.

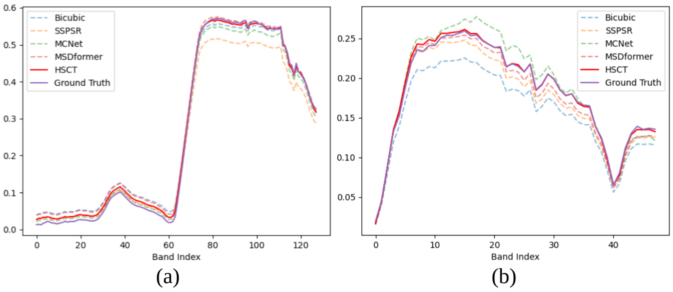

Fig. 4. Spectral curves of the sampling points in the super-resolution results generated with different methods. (a) Chikusei dataset and (b) Houston dataset.

## IV. CONCLUSION

In this paper, the Hierarchical Self-Calibration Transformer (HSCT) is proposed for hyperspectral image super-resolution, combining the excellent ability of CNNs and Transformers for feature extraction. By leveraging a variable window-based self-attention mechanism with window shifting, and incorporating Channel Self-Attention for spectral refinement, HSCT is capable of efficiently capturing multi-scale spatial-spectral features. The Self-Calibration Convolution and Residual Connections enhances feature representations and model stability. Extensive experiments demonstrate that HSCT outperforms both traditional and some state-of-the-art deep learning-based methods, by achieving superior visual quality and better quantitative evaluation metrics.

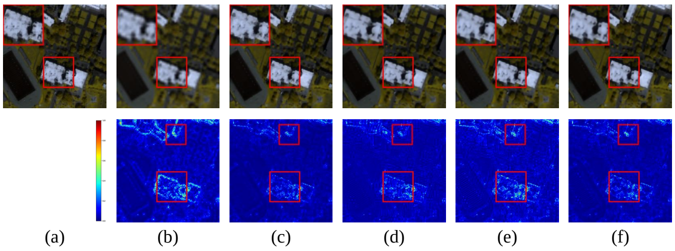

Fig. 5. Reconstructed super-resolution results on Houston dataset with scale factor 4, where the first row shows the pseudo-color images, and the second row presents the error maps. (a) Ground truth, (b) Bicubic, (c) SSPSR, (d) MCNet, (e) MSDformer, and (f) HSCT.

## REFERENCES

[1] 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*, vol. 57, no. 9, pp. 6690–6709, 2019.

[2] S. Mei, X. Yuan, J. Ji, Y. Zhang, S. Wan, and Q. Du, "Hyperspectral image spatial super-resolution via 3d full convolutional neural network," *Remote Sensing*, vol. 9, no. 11, p. 1139, 2017.

[3] 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.

[4] Y. Li, L. Zhang, C. Dingl, W. Wei, and Y. Zhang, "Single hyperspectral image super-resolution with grouped deep recursive residual network," in *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*. IEEE, 2018, pp. 1–4.

[5] J. Jiang, H. Sun, X. Liu, and J. Ma, "Learning spatial-spectral prior for super-resolution of hyperspectral imagery," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 1082–1096, 2020.

[6] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1833–1844.

[7] Y. Liu, J. Hu, X. Kang, J. Luo, and S. Fan, "Interactformer: Interactive transformer and cnn for hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[8] 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.

[9] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong, "Activating more pixels in image super-resolution transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 367–22 377.