LGCT: Local–Global Collaborative Transformer for Fusion of Hyperspectral and Multispectral Images

Wangquan He, *Student Member, IEEE*, Xiyou Fu, *Member, IEEE*, Nanying Li, *Student Member, IEEE*, Qi Ren, *Student Member, IEEE*, and Sen Jia, *Senior Member, IEEE*

***Abstract*— With its strong capability in modeling long-range dependencies, the Transformer achieves competitive performance in hyperspectral image (HSI) and multispectral image (MSI) fusion. However, existing Transformer-based methods face the trade-off between receptive field size and computational efficiency when dealing with spatially non-local features. Furthermore, the Transformer captures deep spectral relationships by modeling pairwise channel interactions. This global interaction may overlook features that contribute little to the overall context but are critical locally, thus affecting the accurate understanding of HSI content. To overcome these challenges, we propose a novel local–global collaborative network with Transformers (LGCT) specifically designed to achieve high-quality HSI reconstruction. The proposed LGCT includes two inverse feature streams to establish multiscale deep representations of the HSI and MSI features. The feature streams comprise collaborative Transformer blocks (CTBs) explicitly designed for the spectral and spatial domains. By combining global and local processing mechanisms, the proposed CTBs can efficiently emphasize potential crucial features that Transformer ignores when capturing deep spectral and spatial relationships, thus enabling efficient modeling of the spectral and spatial domains from details to the whole. Furthermore, to enhance the reusability of multiscale enhanced features from the spectral and spatial domains, a hierarchical and symmetric strategy is adopted to progressively fuse them to generate high-quality images. The results on both simulated and real datasets demonstrate the superior performance of the proposed method in terms of quantitative metrics and visual quality. The code will be released at https://github.com/Hewq77/LGCT.**

***Index Terms*— Attention mechanism, collaborative transformer, hyperspectral image (HSI), image fusion, local–global, multiscale, multispectral image (MSI).**

Nomenclature

$\mathbf{X}$, $\mathbf{X}^{\mathrm{U}}$	Degraded LR-HSI and its upsampled version.
$\tilde{\mathbf{X}}$	Reconstructed HR-HSI.
$\mathbf{Z}$	Reference HR-HSI.
$\mathbf{Y}$	Degraded HR-MSI.
H, W, b	Height, width, and band number of the degraded HR-MSI.
h, w, B	Height, width, and band number of the degraded LR-HSI.
$\mathbf{F}_M$, $\mathbf{F}_H$	Outputs at different depths of Spa-CTB and Spe-CTB.
$\mathbf{C}$, $\mathbf{C}'$	CB block layers in the encoding and decoding stage.
$\mathbf{F}_E$, $\mathbf{F}_D$	Outputs at different depths of encoding and decoding stage.
$\mathbf{T}_M$, $\mathbf{T}_H$	Inputs at different depths of Spa-CTB and Spe-CTB.
$\mathbf{T}'_M$, $\mathbf{T}'_H$	Outputs at different depths of Spa-RSA and Spe-RSA.
M	Window size in Spa-RSA.
N	Number of groups in Spe-RSA.
$\mathbf{Y}_{Ml}$, $\mathbf{Y}_{Mg}$	Local-aware and global-aware outputs of Spa-RSA.
$\mathbf{Y}_{Hl}$, $\mathbf{Y}_{Hg}$	Local-aware and global-aware outputs of Spe-RSA.

I. Introduction

THE specificity of hyperspectral image (HSI) in the spectral domain enables it to detect the composition of materials more accurately, which is valuable in fields such as geological exploration, agriculture, and environmental monitoring [1]. However, during the imaging process, to achieve higher spectral resolution, the sensors face limitations in spatial resolution, which significantly reduces the application of HSI in many potential scenes [2], [3], [4]. Conversely, multispectral image (MSI) has weaker object composition detection capabilities but can provide more surface details, which have less spectral information and higher spatial information. To fully exploit the complementary properties of the HR-MSI (HR-MSI) and LR-HSI (LR-HSI), HSI and MSI fusion have become a research hotspot [5], [6], [7], [8]. This technology aims to reconstruct images with rich spectral information and higher spatial resolution by fusing the strengths of two types of images. This method can accurately capture surface details and features, providing more comprehensive data support for downstream tasks such as classification, segmentation, and target detection [9], [10], [11], [12].

Traditional fusion reconstruction methods focus on using image prior information as constraints to guide the reconstruction of spatial and spectral information in images.

Received 23 June 2024; revised 23 August 2024 and 27 September 2024; accepted 24 October 2024. Date of publication 4 November 2024; date of current version 14 November 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62271327; in part by the Project of Department of Education of Guangdong Province under Grant 2023KCXTD029; in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515011290; in part by Shenzhen Science and Technology Program under Grant RCJC20221008092731042, Grant JCYJ20220818100206015, and Grant KQTD20200909113951005; and in part by the Research Team Cultivation Program of Shenzhen University under Grant 2023JCT002. *(Corresponding author: Sen Jia.)*

The authors are with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: hewangquan2022@email.szu.edu.cn; fuxy0623@szu.edu.cn; linanying2021@email.szu.edu.cn; ren_qi1113@foxmail.com; senjia@szu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2024.3491109

These include methods based on spectral unmixing, matrix decomposition, and tensor decomposition [13], [14], [15]. These methods are inseparable from the mining of image prior information and have strong interpretability. However, the disadvantage is that they often rely on artificially designed prior knowledge and require parameter tuning for different scenes, thus lacking adaptive optimization capability.

With the rise of deep learning, data-driven end-to-end fusion methods have gradually become dominant. These methods can automatically learn implicit prior constraints from vast amounts of data and often demonstrate superior performance [16], [17], [18]. Due to powerful fitting capabilities, convolutional neural networks (CNNs) have been introduced to fuse LR-HSI and HR-MSI. The related studies construct networks that include multiple stacked convolutional layers and nonlinear activation functions to learn the optimal mapping automatically between observed and target images [19], [20], [21]. To obtain richer feature representations, the research idea of multiscale fusion is integrated into HSI and MSI fusion [22], [23], [24]. Additionally, the attention mechanism can learn the correlation between different areas of cross-modal data and achieve adaptive fusion among features, thus gaining much focus [25], [26], [27]. Furthermore, to mitigate the computational cost brought by complex network architectures, researchers design lightweight models to realize efficiently HR-HSI reconstruction [28], [29]. Moreover, to address the issue of poor model interpretability, the idea of combining traditional methods and CNNs is increasingly noticed in HSI and MSI fusion [30], [31], [32], which contributes to enhancing the understanding of the model decision process and reconstruction results.

Despite achieving exciting results, CNN also has obvious shortcomings in HSI reconstruction. On the one hand, while convolution operation is adept at capturing local spatial features, it is difficult to fully exploit the high-dimensional spectral information of HSI, which can lead to excessive smoothing of spectral curves and result in distortion [33], [34], [35]. On the other hand, CNN can only perceive the information from a limited neighborhood. Although the receptive field can be indirectly expanded by stacking convolutional layers, it remains challenging to model the global spatial–spectral contextual relationship. To overcome these issues above, the Transformer has been introduced to process HSI tasks. As the infrastructure of large language models, the Transformer models the correlation between any two positional elements directly through its self-attention mechanism [36]. Building on the characteristics of HSI, researchers have also studied combining graph theory and transformers to capture structural correlations between pixels [37]. To fully adapt to the fusion of HSI and MSI, some Transformer-based studies have designed specific modules to capture the complex spectral and spatial nonlinear relationships in HSI and MSI [38], [39], [40]. Furthermore, the integration of multiscale research ideas has been explored to improve generalization ability, and these methods [41], [42], [43], [44] enhance the details and spectral fidelity of the reconstructed images by focusing on different scales of spectral-spatial information at the data input level or feature level.

However, the current research still faces two main challenges: 1) Transformer acquiring global dependencies brings a large amount of computation. Window-based Transformers capture spatial relations by focusing on pixels within a local window, which significantly improves the computational efficiency but also limits the receptive fields of attention. As a result, valuable global information may be lost, which is not conducive to modeling remote sensing scenes with irregular distributions and 2) to ensure the modeling of long-range spectral dependencies, the Transformer treats the entire spatial extent in the spectral dimension as tokens and then uses self-attention to capture inter-band correlations. However, this may lead to the marginalization of critical local features. During the global attention allocation process, the model may tend to focus on features that are more important for understanding the global content, while neglecting those that contribute less to the overall context but are critical within a local scale. Due to the diversity and complexity of HSI and MSI fusion tasks, it is essential to model local similarity blocks, which facilitate the capture of fine details and achieve finer spectral fidelity.

Based on the above analysis, we propose a local–global collaborative network with Transformers (LGCT) to realize high-quality HSI reconstruction. Unlike other Transformer-based methods, the proposed LGCT can efficiently model spectral and spatial domain features from global and local perspectives, complementing and enhancing the potential critical information that may be neglected in the spectral and spatial domains. Specifically, the proposed LGCT employs two inverse feature streams to adapt to the two modal inputs of different resolutions and fully mine their multiscale features. The spatial collaborative Transformer block (Spa-CTB), which makes up the MSI feature stream, achieves enhancement in spatial long-distance dependency capabilities at low cost by exchanging tokens between windows of multihead self-attention (MSA). The spectral collaborative Transformer (Spe-CTB) is a crucial component of the HSI feature stream, which enhances the perception of local critical information in the spectral domain by grouping the spatial information in the spectral dimension. Moreover, a hierarchical symmetric strategy is designed to improve the reusability of multiscale enhanced features in the fusion process. In summary, the main contributions of the article can be described as follows.

1) A simple and effective fusion model for HSI and MSI, named LGCT, is proposed. This model is designed to comprehensively and efficiently capture both the local details and global views of LR-HSI and HR-MSI. Additionally, a hierarchical symmetric strategy is designed to progressively fuse multiscale spatial–spectral features containing global–local representations to generate high-fidelity and high-resolution images.
2) The proposed Spa-CTB adopts a cross-window shuffling strategy to implicitly model global spatial features, which breaks the trade-off between global feature capturing ability and computational efficiency in Transformer. It allows the model to extend pixels dependencies to a global scale without additional computational costs while maintaining focus to local spatial details.

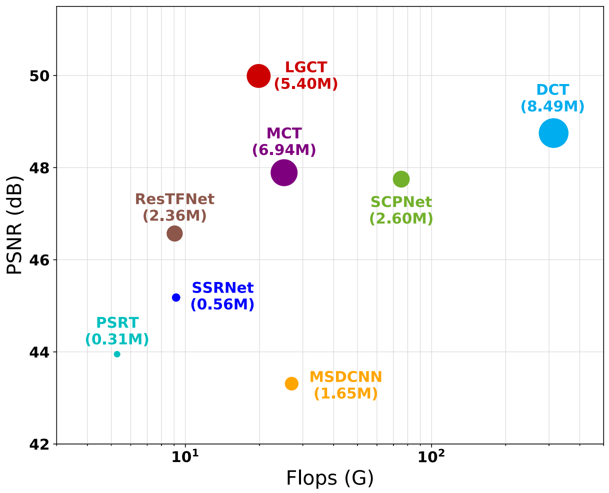


Fig. 1. Comparison of computational efficiency and fusion performance of deep learning-based methods on Houston dataset. The dataset and comparison methods are described in Section IV. The horizontal and vertical axes denote Flops and PSNR, respectively, and the circle size indicates the number of parameters.

3) The Spe-CTB is proposed to obtain more comprehensive spectral deep representation. Compared to directly computing self-attention in the spectral domain, the proposed Spe-CTB maintains awareness of global spectral features while emphasizing overlooked local potential crucial features in the spectral domain by grouping spatial information of the spectral dimension.
4) Experimental results demonstrate that LGCT significantly outperforms state-of-the-art (SOTA) fusion methods with lower computational load (as shown in Fig. 1). Additionally, the LGCT also excels in real HSI reconstruction.

The rest of the article is organized as follows. Section II discusses related works, including the main research dynamics. Building on this, Section III details the research methods we propose, clarifying the theoretical basis and implementation steps of model construction. Following that, Section IV presents the experimental results, where we present the quantitative and qualitative results of the experiment through data analysis and charts. Finally, we provide a comprehensive summary of the article and suggestions for future research in Section V.

II. Related Works

In this section, we briefly review the research advancements in HSI and MSI fusion, including conventional prior and deep learning-based methods that have arisen in recent years. We then introduce in detail the application of Transformer in the fusion of HSI and MSI.

A. HSI and MSI Fusion

Fully exploiting the spectral–spatial relationship between HSI and MSI is the crucial to achieving HSI fusion reconstruction [30]. Traditional-based fusion methods focus on using image prior knowledge to constrain the representation of spectral–spatial information during HSI reconstruction, including sparse prior, low-rank prior, and non-local similarity prior [2], [8]. Yokoya et al. [45] proposed a method based on coupled matrix decomposition to obtain the endmembers and abundance information of the images to be fused, respectively, and achieved HSI and MSI fusion by integrating this information. On this basis, sparse regularity is introduced to [46] for spectral unmixing to enhance the fusion effect. Wu et al. [47] achieve constraints on reconstructed images by mining the low-rank properties of HSI in the spectral and spatial domains. Dian and Li [48] proposed a subspace-based low-tensor multirank constraint method, which achieves the reconstruction of HR-HSI by mining the correlation and non-local similarity in the images. However, the traditional reconstruction method requires several iterations to get the fusion result and lacks adaptive optimal ability when facing real scenes.

Capturing the spectral–spatial relationship between LR-HSI and HR-MSI through deep networks has received widespread attention. Zhu et al. [49] designed a lightweight network to learn high-resolution and zero-centric residual images, and reconstruct HR-HSI in a progressive manner. From the perspective of spectral unmixing, Yao et al. [26] proposed a coupled two-stream network with cross-modal attention to improving the super-resolution fusion effect of HSI. Wei et al. [30] designed a recursive residual network that unfolds sub-optimization problems into network representations to achieve super-resolution reconstruction. Furthermore, Dong et al. [50] proposed an iterable model-guided network for end-to-end optimal reconstruction. Inspired by the proximal gradient descent method, Xie et al. [31] introduced a traditional image low-rank prior and proposed an interpretable HR-HSI reconstruction network. Additionally, other network paradigms such as graph neural network (GCN), generative adversarial network (GAN), and autoencoder (AE) are considered to enhance the quality of fused images, as discussed in [51], [52], and [53]. In summary, the above studies demonstrate the effectiveness of deep networks in the fusion of HSI and MSI, but there remains room for improvement in understanding and fusing complex images.

B. Transformer Methods in HSI and MSI Fusion

The successful development of the Transformer in natural language processing has made it a foundational architecture for large language models [54]. Inspired by this, Transformer has been applied to HSI and MSI fusion and has shown strong effectiveness. Due to its unique self-attention mechanism, Transformer can capture long-range correlations within images over a wide range and effectively integrate spectral–spatial information between different modality images, which is especially crucial for processing rich spectral information in HSI and high-spatial resolution in MSI. In this field, Hu et al. [55] designed a lightweight Transformer to implement image fusion, which speeds up the convergence by focusing on the learning in the residual domain. Chen et al. [39] proposed a two-branch structure-based Transformer to establish the information interaction between HSI and MSI. Furthermore, Sun et al. [42] combined the multiscale idea to improve the performance of HSI and MSI fusion. Pre-training and Shuffle-and-Reshuffle strategies were also considered in [41] and [43].

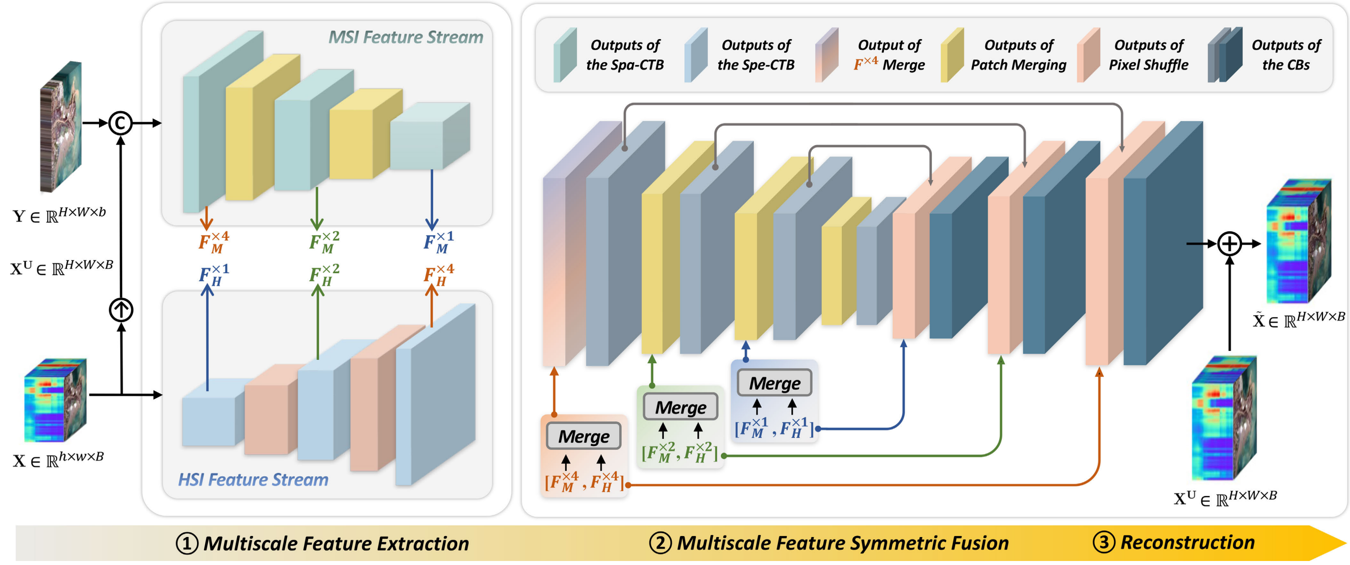

MSI Feature Stream
HSI Feature Stream
$\mathbf{Y} \in \mathbb{R}^{H\times W\times b}$
$\mathbf{X}^{\mathrm{U}} \in \mathbb{R}^{H\times W\times B}$
$\mathbf{X} \in \mathbb{R}^{h\times w\times B}$
$F_M^{\times4}$
$F_M^{\times2}$
$F_M^{\times1}$
$F_H^{\times1}$
$F_H^{\times2}$
$F_H^{\times4}$
Outputs of the Spa-CTB
Outputs of the Spe-CTB
Output of $F^{\times4}$ Merge
Outputs of Patch Merging
Outputs of Pixel Shuffle
Outputs of the CBs
Merge
$[F_M^{\times4}, F_H^{\times4}]$
$[F_M^{\times2}, F_H^{\times2}]$
$[F_M^{\times1}, F_H^{\times1}]$
$\hat{\mathbf{X}} \in \mathbb{R}^{H\times W\times B}$
① Multiscale Feature Extraction
② Multiscale Feature Symmetric Fusion
③ Reconstruction

Fig. 2. Overall framework of the proposed LGCT. The framework can be divided into three stages from left to right: multiscale feature extraction, multiscale feature symmetrical fusion, and reconstruction. "×1," "×2," and "×4" denote the three spatial scales in multiscale feature extraction. "Merge" denotes two tensors with the same scale performing a connect operation along the channel axis. CBs denote convolutional blocks.

Wang et al. [40] and Ma et al. [44] introduced the idea of cross-attention to the Transformer to facilitate inter-modal information transfer. In [38], a bidirectional dilation Transformer was proposed to achieve image reconstruction in a progressive manner. These studies have amply demonstrated the unique advantages of Transformer in improving the quality of HSI and MSI fusion. However, existing Transformer-based methods suffer from mutual constraints between performance and computational efficiency, as well as lacking the exploration of global and local correlations in the spectral–spatial domain.

III. Proposed Method

In this section, we are dedicated to providing a detailed description of the proposed LGCT. The overall pipeline of LGCT will be presented first, and then the CTB for HSI and MSI feature extraction will be expanded.

A. Overall Pipeline of LGCT

The proposed LGCT for HSI and MSI fusion consists of three parts: multiscale feature extraction, multiscale feature symmetrical fusion, and feature reconstruction, as shown in Fig. 2. The implementation details of the proposed LGCT are given in Algorithm.

1) Multiscale Feature Extraction: The degraded input HR-MSI and LR-HSI are denoted as $\mathbf{Y} \in \mathbb{R}^{H\times W\times b}$ and $\mathbf{X} \in \mathbb{R}^{h\times w\times B}(H \gg h, W \gg w, B \gg b)$, where (H, W, b) and (h, w, B) denote the height, width, and band number of the HR-MSI and LR-HSI, respectively. Initially, we use bicubic interpolation to upsample the LR-HSI to obtain $\mathbf{X}^{\mathrm{U}} \in \mathbb{R}^{H\times W\times B}$, and then concatenate it along the channel dimension with $\mathbf{Y}$ to input into the MSI feature stream. We employ a convolutional layer with a size of 3 × 3 to extract shallow features. Afterward, these features are processed by Spa-CTB, which consists of local and global branches, to obtain deep feature maps $\mathbf{F}_M$. To obtain features at different scales, aside from the first Spa-CTB, the resolution of the feature maps is reduced by half before passing through each Spa-CTB, while the channel number is doubled. This process can be formulated as follows:

$$F_M^i = \begin{cases} f_a^i\left(\Phi\left(\mathrm{Cat}\left(\mathbf{X}^{\mathrm{U}}, \mathbf{Y}\right)\right)\right), & i = 1 \\ f_a^i\left(F_M^{i-1}\downarrow\right), & i = 2, 3 \end{cases} \tag{1}$$

where f_a^i is the ith Spa-CTB. F_M^{i-1} and F_M^i denote the input and output features of f_a^i. ↓ denotes downsampling operation with patch merging [33]. Φ is the convolutional layer with a kernel size of 3 × 3. Cat denotes concatenation operation. In a similar operation, for the HSI feature stream, the LR-HSI is first processed by a convolutional layer with a kernel size of 3 × 3 and then undergoes the Spe-CTB to obtain spectral deep features $\mathbf{F}_H$ at different scales. On the contrary, the $\mathbf{F}_H$ at different scales is obtained by upsampling before passing through each Spe-CTB (except the first one), where pixel shuffle [56] is used to realize the upsampling. The process can be represented as

$$F_H^i = \begin{cases} f_e^i(\Phi(\mathbf{X})), & i = 1 \\ f_e^i\left(F_H^{i-1}\uparrow\right), & i = 2, 3 \end{cases} \tag{2}$$

where f_e^i is the ith Spe-CTB. F_H^{i-1} and F_H^i denote the input and output features of f_e^i. ↑ denotes upsampling operation with pixel shuffle.

2) Multiscale Feature Symmetrical Fusion: Subsequently, a hierarchical symmetrical structure is used to fuse the different scale deep features of the HSI and MSI stream. Compared with the fixed scale and non-hierarchical structure, this hierarchical symmetrical network can effectively improve computational efficiency. Specifically, the encoding–decoding structure with a three-scale is used to fuse the multiscale deep features from the bimodal images of stage one. In the encoding stage, we concatenate the same scale features from the HSI feature stream and the MSI feature stream, followed by using a convolutional block (CB) to fuse the combined deep features. Each CB consists of two convolutional layers with a kernel size of 3 × 3 and two LeakyReLU layers. Starting from the highest resolution input, the encoder hierarchically reduces the spatial resolution size by half and the channel is doubled to fuse the combined HSI–MSI features at different scales. Formally, the fusion process in the encoding part can be defined as

$$F_E^i = \begin{cases} C_i\left(\mathrm{Cat}\left(F_M^i, F_H^{4-i}\right)\right), & i = 1 \\ C_i\left(\mathrm{Cat}\left(F_E^{i-1}\downarrow, F_M^i, F_H^{4-i}\right)\right), & i = 2, 3 \\ C_i\left(F_E^{i-1}\downarrow\right), & i = 4 \end{cases} \tag{3}$$

where C_i denotes the ith CB block layer in the encoding stage, and F_E^i is the fusion result of CB output at ith layer of encoding stage. Note that F_M^i ($i = 1, 2, 3$) corresponds to the $F_M^{\times 4} \in \mathbb{R}^{H\times W\times \tilde{B}}$, $F_M^{\times 2} \in \mathbb{R}^{(H/2)\times(W/2)\times 2\tilde{B}}$, $F_M^{\times 1} \in \mathbb{R}^{(H/4)\times(W/4)\times 4\tilde{B}}$, respectively. In contrast, F_H^i ($i = 1, 2, 3$) corresponds to the $F_H^{\times 1} \in \mathbb{R}^{(H/4)\times(W/4)\times 4\tilde{B}}$, $F_H^{\times 2} \in \mathbb{R}^{(H/2)\times(W/2)\times 2\tilde{B}}$, and $F_H^{\times 4} \in \mathbb{R}^{H\times W\times \tilde{B}}$, respectively. They denote the different spatial scales of $\mathbf{F}_M$ and $\mathbf{F}_H$ (see Fig. 2). In the decoding stage, the last layer of low-resolution F_E^4 in the encoding stage is used as input to recover the high-resolution representations gradually. For better recovery, combined HSI–MSI multiscale features are again involved in the decoding process to help recover fine textures and prevent spectral distortion. The decoder uses hierarchical upsampling to fuse these features progressively. Concurrently, skip connections from encoding stages at different scales are established to fuse multiscale features hierarchically. Likewise, CBs are used to perform the fusion process. The process can be formulated as follows:

$$F_D^i = \begin{cases} C_i'\left(\mathrm{Cat}\left(F_E^4\uparrow, F_E^{4-i}, F_M^{4-i}, F_H^i\right)\right), & i = 1 \\ C_i'\left(\mathrm{Cat}\left(F_D^{i-1}\uparrow, F_E^{4-i}, F_M^{4-i}, F_H^i\right)\right), & i = 2, 3 \end{cases} \tag{4}$$

where C_i' denotes the ith layer CB block in the decoding stage, and F_D^i is the fusion result of CB output at ith layer of decoding stage.

3) Reconstruction: Finally, we take the upsampled LR-HSI $\mathbf{X}^{\mathbf{U}}$ as the residual image and additively fuse it with the last layer of feature representation F_D^3 to obtain the reconstructed HSI $\tilde{\mathbf{X}} \in \mathbb{R}^{H\times W\times B}$.

B. Spa-CTB

Restricting self-attention within the local window can effectively overcome high computational costs caused by global Transformer; however, it also limits the ability of the model to perceive global information, which is not conducive to processing HSI in remote sensing scenes. For this reason, the Spa-CTB is proposed to improve the ability of the model to perceive the spatial global while inheriting the advantages of the window-wise methods in terms of local sensitivity and low computational cost. As shown in Fig. 3, the Spa-CTB consists of a spatial regrouping self-attention (Spa-RSA), a feed-forward network (FFN), and two LayerNorm (LN) operators. Given the input tensor of Spa-CTB as $\mathbf{T}_M \in \mathbb{R}^{H\times W\times \tilde{B}}$, the formula can be expressed as follows:

$$\mathbf{T}_M = \text{Spa-RSA}(\mathrm{LN}(\mathbf{T}_M)) + \mathbf{T}_M \tag{5}$$

$$\mathbf{F}_M = \mathrm{FFN}(\mathrm{LN}(\mathbf{T}_M)) + \mathbf{T}_M. \tag{6}$$

Algorithm 1 Framework of the Proposed LGCT

```
Input: 1) LR-HSI $\mathbf{X} \in \mathbb{R}^{h\times w\times B}$, 2) HR-MSI $\mathbf{Y} \in \mathbb{R}^{H\times W\times b}$
Output: Reconstructed HSI $\tilde{\mathbf{X}} \in \mathbb{R}^{H\times W\times B}$
1: Step 1: Multiscale Feature Extraction    ▷ III-A-(1)
2: Upsample LR-HSI using bicubic interpolation to obtain $\mathbf{X}^{\mathbf{U}} \in \mathbb{R}^{H\times W\times B}$.
3: Concatenate $\mathbf{X}^{\mathbf{U}}$ and $\mathbf{Y}$ along the channel dimension as input to the MSI stream.
4: Extract deep features $F_M^i$ and $F_H^i$ at different scales from the MSI and HSI feature streams (via Eq. (1) and (2)).
5: Step 2: Multiscale Symmetric Fusion    ▷ III-A-(2)
6: Pair deep features $F_M^i$ and $F_H^i$ at different scales during the encoding stage (via Eq. (3)), with the fusion process proceeding from high to low resolution.
7: Recover high-resolution representations $F_D^i$ by hierarchical upsampling to symmetrically fuse multiscale HSI-MSI features $F_M^i$ and $F_H^i$ (via Eq. (4)).
8: Step 3: Reconstruction    ▷ III-A-(3)
9: Reconstruct the final HR-HSI as $\tilde{\mathbf{X}} = \mathbf{X}^{\mathbf{U}} + F_D^3$.
```

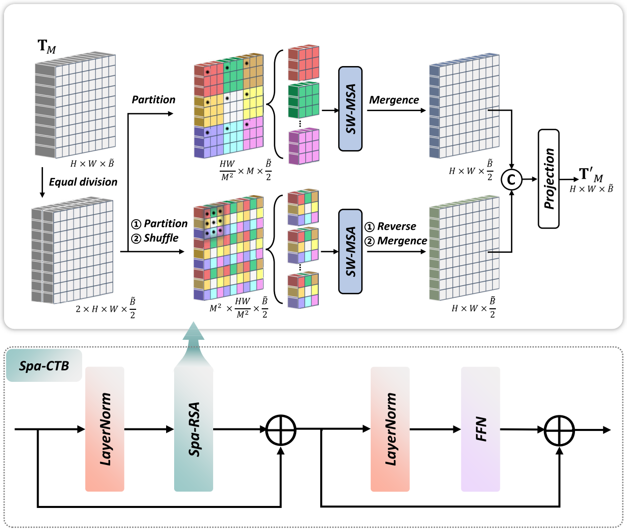

$\mathbf{T}_M$
$H \times W \times \tilde{B}$
Equal division
$2 \times H \times W \times \frac{\tilde{B}}{2}$
Partition
① Partition
② Shuffle
$\frac{HW}{M^2} \times M \times \frac{\tilde{B}}{2}$
$M^2 \times \frac{HW}{M^2} \times \frac{\tilde{B}}{2}$
SW-MSA
SW-MSA
Mergence
① Reverse
② Mergence
$H \times W \times \frac{\tilde{B}}{2}$
$H \times W \times \frac{\tilde{B}}{2}$
C
Projection
$\mathbf{T}'_M$
$H \times W \times \tilde{B}$
Spa-CTB
LayerNorm
Spa-RSA
LayerNorm
FFN

Fig. 3. Detailed structure of the proposed Spa-CTB. It consists of two LN operators, a Spa-RSA, and an FFN. The Spa-RSA includes two parallel branches, which are used to focus on the global and local information of the MSI feature stream, respectively.

Next, we specifically introduce the Spa-RSA, which contains two parallel branches focusing on the global and local features, respectively. Such an integration strategy allows the model to efficiently integrate both spatial macro-context and micro-detail simultaneously. Specifically, for the input tensor $\mathbf{T}_M$ after LN, we generate the query, key, value matrix (i.e., $\mathbf{Q}_M, \mathbf{K}_M, \mathbf{V}_M \in \mathbb{R}^{H\times W\times \tilde{B}}$) via linear projection. The process can be expressed as follows:

$$\mathbf{Q}_M = W_Q^{\mathrm{T}}\mathbf{T}_M, \quad \mathbf{K}_M = W_K^{\mathrm{T}}\mathbf{T}_M, \quad \mathbf{V}_M = W_V^{\mathrm{T}}\mathbf{T}_M \tag{7}$$

where W_Q, W_K, W_V denote the learnable weight matrices corresponding to $\mathbf{Q}_M, \mathbf{K}_M, \mathbf{V}_M$ respectively. To construct two parallel branches, $\mathbf{Q}_M, \mathbf{K}_M, \mathbf{V}_M$ are divided equally along the spectral channel to obtain $[\mathbf{Q}_{Ml}, \mathbf{K}_{Ml}, \mathbf{V}_{Ml}]$, $[\mathbf{Q}_{Mg}, \mathbf{K}_{Mg}, \mathbf{V}_{Mg}] \in \mathbb{R}^{H \times W \times (\tilde{B}/2)}$. For the local-aware part, we follow the idea of window attention [33] and divide $[\mathbf{Q}_{Ml}, \mathbf{K}_{Ml}, \mathbf{V}_{Ml}]$ into non-overlapping windows, each of which contains $(M \times M)$ pixels, so that the whole feature map is divided into (HW/M^2) windows. Spatial window MSA (SW-MSA) is used to capture the dependencies between features within each window. For each head i $(i = 1, \ldots, h)$, the output Y_{Ml}^{i} obtained from the computation of self-attention can be formulated as

$$Y_{Ml}^{i} = \text{Softmax}\left(\frac{Q_{Ml}^{i}\left(K_{Ml}^{i}\right)^{\mathrm{T}}}{\sqrt{d_M}} + R_M\right)V_{Ml}^{i} \tag{8}$$

where $R_M \in \mathbb{R}^{M^2 \times M^2}$ is the relative position code. Then the outputs of all the heads are concatenated along the channel dimension to get the local aware output tensor $\mathbf{Y}_{Ml}$

$$\mathbf{Y}_{Ml} = \text{Cat}\left(Y_{Ml}^{1}, \ldots, Y_{Ml}^{h}\right) \tag{9}$$

where h denotes the number of heads.

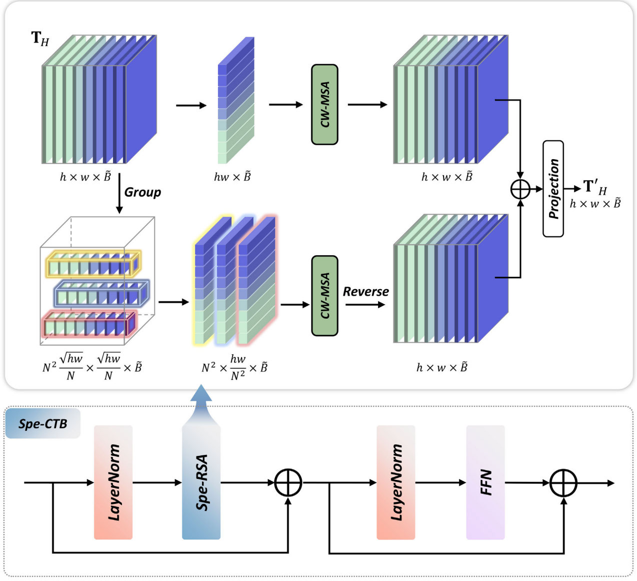
For the global-aware part, inspired by the shuffle idea [56], [57], [58], we perform shuffle operations on tokens within each window of the feature map. This method redistributes the tokens within each window, enabling information initially confined to self-attention computation within local windows to interact across the entire feature map. In detail, we divide $[\mathbf{Q}_{Mg}, \mathbf{K}_{Mg}, \mathbf{V}_{Mg}]$ into non-overlapping windows, and like the local part, each window contains $(M \times M)$ pixels. Subsequently, to achieve the shuffling of tokens within the windows and enhance the interaction of global information, we shuffle the spatial dimensions $[h, w]$ of the feature map by exchanging their order. Specifically, for $[\mathbf{Q}_{Mg}, \mathbf{K}_{Mg}, \mathbf{V}_{Mg}] \in \mathbb{R}^{(HW/M^2) \times M^2 \times (\tilde{B}/2)}$ after partitioning the windows, we transpose it to $[\mathbf{Q}'_{Mg}, \mathbf{K}'_{Mg}, \mathbf{V}'_{Mg}] \in \mathbb{R}^{M^2 \times (HW/M^2) \times (\tilde{B}/2)}$, thereby changing the relative position of tokens within the feature map. Afterward, the attention map $Y^{i'}_{Mg}$ of the ith head is computed using (8), and the outputs of all heads are concatenated to obtain $\mathbf{Y}'_{Mg}$. For subsequent fusion, we perform an inverse transpose operation to recover $\mathbf{Y}'_{Mg} \in \mathbb{R}^{M^2 \times (HW/M^2) \times (\tilde{B}/2)}$ into $\mathbf{Y}_{Mg} \in \mathbb{R}^{(HW/M^2) \times M^2 \times (\tilde{B}/2)}$. To fully exploit the outputs of the local-aware and global-aware parts, we reshape the global and local output and then use linear projection to fuse the concatenated tensors to obtain the output $\mathbf{T}'_M \in \mathbb{R}^{h \times w \times \tilde{B}}$ of Spa-RSA, which can be expressed as follows:

$$\text{Spa-RSA}(\mathbf{T}_M) = \mathbf{W}_M\left(\text{Cat}\left(\mathbf{Y}_{Ml}, \mathbf{Y}_{Mg}\right)\right)^{\mathrm{T}} \tag{10}$$

where $\mathbf{W}_M \in \mathbb{R}^{\tilde{B} \times \tilde{B}}$ refers to the linear projection used for feature fusion. In this study, the M, related to windows size, is set to 8. With this feature fusion strategy, the model can capture spatial local details while implicitly modeling the entire image globally through regrouping. This results in the Spa-RSA improves the model performance while avoiding explicitly increasing the computational load.

C. Spe-CTB

Unlike the window self-attention mechanism, which mainly focuses on the interaction between spatial positions, the core of the channel-wise self-attention mechanism lies in capturing the deep spectral feature by analyzing the correlation between bands [59]. This processing emphasizes feature statistics over the entire spatial range, i.e., the entire space of the spectral dimension as tokens, which may result in neglecting features that are not significant on a global scale but are very important on a local scale. In this case, the model may not capture the details of the local features adequately. The limitation could be amplified in HSI with a higher number of channels under the remote sensing scene, thereby affecting the assignment of channel weights, as well as the final fusion performance. To this end, a Spe-CTB is designed to ensure that important features at the local scale are adequately considered.



$\mathbf{T}_H$
$h \times w \times \tilde{B}$
Group
$hw \times \tilde{B}$
CW-MSA
$h \times w \times \tilde{B}$
Projection
$\mathbf{T}'_H$
$h \times w \times \tilde{B}$
$N^2 \frac{\sqrt{hw}}{N} \times \frac{\sqrt{hw}}{N} \times \tilde{B}$
$N^2 \times \frac{hw}{N^2} \times \tilde{B}$
CW-MSA
Reverse
$h \times w \times \tilde{B}$
Spe-CTB
LayerNorm
Spe-RSA
LayerNorm
FFN

Fig. 4. Detailed structure of the proposed Spe-CTB. Similar to the spatial part, it consists of two LN operators, a Spe-RSA, and an FFN. The Spe-RSA includes two parallel branches, which are used to focus on the global and local information of the HSI feature stream, respectively.

The proposed Spe-CTB, based on the channel-wise self-attention mechanism, additionally introduces a part for local awareness. This local aware part can focus on capturing the details of local features. The outputs of the local-aware part and channel-wise self-attention are then fused. In this way, the model captures both global information between bands and focuses on essential features between bands on a local spatial scale. As shown in Fig. 4, the proposed Spe-CTB and Spa-CTB structures are similar, consisting of a spectral regrouping self-attention (Spe-RSA), an FFN, and two LN layers. Formulaically, Spe-CTB captures spectral deep features $\mathbf{F}_H$ can be referred to (5) and (6).

In addition, the proposed Spe-RSA adopts a similar strategy as the Spa-RSA, i.e., focusing on the spectral global context and local details through two parallel branches, respectively, and fusing the outputs of the two branches. For details, given the input tensor $\mathbf{T}_H \in \mathbb{R}^{h \times w \times \tilde{B}}$, we first apply LN and then transform the feature map to generate $\mathbf{Q}_H, \mathbf{K}_H, \mathbf{V}_H \in \mathbb{R}^{h \times w \times \tilde{B}}$, which represent query, key, and value matrices respectively. Referring to [59], this process is achieved by 1×1 pixel-wise convolution followed by 3×3 depth-wise convolution. Then reshape $\mathbf{Q}_H, \mathbf{K}_H, \mathbf{V}_H \in \mathbb{R}^{h \times w \times \tilde{B}}$ to $\mathbf{Q}'_H, \mathbf{K}'_H, \mathbf{V}'_H \in \mathbb{R}^{hw \times \tilde{B}}$. For the global-aware part, channel-wise MSA (CW-MSA) is used to obtain long-distance relationships across the band.

We compute the similarity between $\mathbf{Q}_H$ and $\mathbf{K}_H$ to get the self-attention matrix. Furthermore, we use self-attention matrix to weigh and sum the $\mathbf{V}_H$ to get the output feature maps $\mathbf{Y}_{Hg}$

$$\mathbf{Y}_{Hg} = \text{Softmax}\left(\frac{\left(Q_H^i\right)^{\mathrm{T}} K_H^i}{\sqrt{\alpha}}\right) V_H^i \tag{11}$$

where α is the learnable scaling parameter. Finally, $\mathbf{Y}_{Hg} \in \mathbb{R}^{hw\times\tilde{B}}$ is reshaped to $\mathbb{R}^{h\times w\times\tilde{B}}$ to obtain the output of the global aware part.

For the local-aware part, we divide the feature space into several groups and perform operations within each group to enhance the mining of local details in the spectral dimension. More specifically, we first divide the input feature map $\mathbf{Q}'_H, \mathbf{K}'_H, \mathbf{V}'_H \in \mathbb{R}^{hw\times\tilde{B}}$ into N^2 groups to obtain $\mathbf{Q}'_{Hl}, \mathbf{K}'_{Hl}, \mathbf{V}'_{Hl} \in \mathbb{R}^{N^2\times(hw/N^2)\times\tilde{B}}$. Subsequently, we use CW-MSA to compute the attention weights within ith group to obtain the updated output Y_{Hl}^i ($i = 1, 2, \ldots, N^2$), and the formula for this part can be referred to (11). The updated features Y_{Hl}^i within each group are reshaped according to their spatial location in the original feature map to form the complete output feature map $\mathbf{Y}_{Hl} \in \mathbb{R}^{h\times w\times\tilde{B}}$. In this way, the weights within each group are dynamically formed based on the input features, which allows the model to perceive and respond to local features in more detail. Similar to the Spa-RSA, we use a projection function to aggregate the global–local fusion features to obtain the output $\mathbf{T}'_H \in \mathbb{R}^{h\times w\times\tilde{B}}$ of the Spe-RSA, and the process can be expressed as follows:

$$\text{Spe-RSA}(\mathbf{T}_H) = \mathbf{W}_H\left(\text{Add}\left(\mathbf{Y}_{Hl}, \mathbf{Y}_{Hg}\right)\right)^{\mathrm{T}} \tag{12}$$

where $\mathbf{W}_H$ denotes the 1×1 convolutions used for feature fusion. In this study, N, which relates to the number of groups, is set to 8. In fact, the global-aware part is the case where N is 1. Similar to the Spa-RSA, we use a multihead mechanism to process the attention map, which means the channels are divided into multiple heads, allowing them to learn their attention maps simultaneously.

With this integration strategy, the model can capture cross-band correlations using global spatial-scale features. Additionally, it can generate specific response patterns for each local region through regrouping, which improves the adaptability to variations in input HSIs across different scenes, resulting in higher-quality fusion outcomes.

D. Loss Function

To achieve network parameter updates, it is necessary to minimize the loss between the reconstructed HSI $\tilde{\mathbf{X}}$ and the reference image $\mathbf{Z}$. We adopt the L_1 loss function to optimize the model. Compared to the L_2 loss function, the L_1 loss function effectively alleviates the smoothing issue in the fused result and also provides better model convergence. The mathematical expression for the L_1 loss function is as follows:

$$\mathcal{L}_{L_1} = \frac{1}{n}\sum_{i=1}^{n}\left|Z_i - \tilde{X}_i\right| \tag{13}$$

where n denotes the number of training samples.

IV. Experimental Results

The content of this section is used to present the experimental results. The datasets used in the experiments and the implementation details are first described. Then the experimental results are shown, including comparison results with the SOTA methods, real data fusion performance, ablation analysis, and model complexity analysis.

A. Datasets

Three simulated datasets and one real remote sensing dataset were used to validate the effectiveness of the proposed LGCT. These datasets are known as the Pavia University,[1] Houston,[2] Chikusei,[3] and Yellow River Estuary (YRE) [60] datasets. Here are the details.

1) *Pavia University:* The HSI image was captured by the Reflective Optics System Imaging Spectrometer (ROSIS) over the Pavia University in Italy. The image contains 610 × 340 pixels with a spatial resolution of 1.3 m. The Pavia University dataset includes 103 spectral bands available for analysis and mainly involves the distribution of urban ground objects.
2) *Houston:* The HSI image was captured by the ITRES CASI-1500 sensor over the University of Houston in the USA. The data has a spatial size of 349 × 1905 with a spatial resolution of 2.5 m. It has 144 spectral bands that cover the range from 380 to 1050 nm. The scene also mainly involves the distribution of urban ground objects.
3) *Chikusei:* The HSI image was captured by Headwall Hyperspec-VNIR-C sensor over Chikusei, Japan. The data contains 2517 × 2335 pixels with a spatial resolution of 2.5 m and integrates 128 spectral bands covering the range from 343 to 1018 nm. The scene covers both urban and rural distribution.
4) *YRE:* The full-resolution dataset consists of one HSI and one MSI, where the HSI is captured by the advanced HSI on the Gaofen-5 satellite and the MSI is captured by the MSI on the Sentinel-2A satellite. The data were imaged over the eastern part of the Yellow River Delta area in China. The HSI has a spatial size of 1400 × 1400 and contains 280 spectral bands with wavelength ranges from 400 to 2500 nm. The MSI has a spatial size of 4200 × 4200 with four spectral bands covering the range from 430 to 680 nm. The spatial resolutions of HSI and MSI are 30 and 10 m, respectively.

B. Experimental Settings

1) Data Simulation: For the simulated dataset, the original HSI is used as the reference image. A Gaussian filter with a kernel size of 7 × 7 and a standard deviation of 2 is applied to the reference image, followed by downsampling by a factor of 4 to obtain the LR-HSI. The HR-MSI is generated by selecting five bands from the reference image

[1] https://ehu.eus/ccwintco/index.php? title=Hyperspectral_Remote_Sensing_Scenes

[2] https://hyperspectral.ee.uh.edu/?page_id=459

[3] https://naotoyokoya.com/Download.html

at equal interval. Test samples are obtained by cropping a 128 × 128 sub-region from the center of the reference image. After cropping the region in the reference image, it is zero-filled to create the training set, ensuring no overlap between the training and test samples. In each training, a randomly selected subset of size 128 × 128 is used for training. In the case of the Chikusei dataset, a sub-image of size 800 × 800 is selected as the reference image. When working with the real dataset, we follow previous studies [16] to generate training samples. The same settings as the simulated dataset are used to downsample the original HSI and MSI by a factor of 3 to obtain the training samples. The original HSI is used as ground truth to supervise the network. A sub-region of HSI (size 96 × 96) and MSI (size 288 × 288) in the same scene is selected as the real image to be fused, and the other regions are the training regions, which do not overlap.

2) Implementation Details: All experiments are implemented using the PyTorch framework on Python 3.8 and MATLAB R2017b. The training process is optimized using Adam optimizer with the learning rate set to 1e^{-4}, and the number of training iterations is set to 10 000. All experiments are performed on a computer with an Intel[4] Xeon[4] Silver 4314 CPU 2.40 GHz and an Nvidia A40 GPU, 48 GB RAM.

3) Evaluation Metrics: The experiment employs six common metrics to evaluate the quality of the fused HSI obtained by different methods, which include root-mean-square error (RMSE), peak signal-to-noise ratio (PSNR), relative dimension less global error in synthesis (ERGAS), correlation coefficient (CC), spectral angle mapper (SAM), and structural similarity index (SSIM) [5]. Higher PSNR values, CC, and SSIM closer to 1 indicate better image quality, while lower RMSE, ERGAS, and SAM indicate better image quality. These metrics provide a comprehensive assessment of the fused HSI in terms of spectral fidelity and spatial visual effects.

For real HSI datasets, where reference images are not available, we utilize quality with no reference (QNR) [61], D_λ, and D_s for objectively evaluating image fusion results. The QNR score is determined by two components, i.e., D_λ and D_s, with values closer to 1 indicating higher quality of the fused images. D_λ measures the distortion of the fused image relative to the original HSI in terms of spectral properties, and D_s measures the changes in spatial resolution of the fused image relative to the original MSI. Consequently, smaller values for D_λ and D_s reflect better image fusion quality.

C. Comparison Experiments on Simulated Data

In this section, we focus on qualitative and quantitative comparisons of the proposed method with several SOTA models to illustrate the potential and uniqueness of our method in terms of HSI and MSI fusion. Specifically, we choose three types of mainstream methods for comparison: prior-based traditional methods, CNN-based methods, and the latest Transformer-based methods. Among the traditional methods include the tensor decomposition-based method, i.e., the coupled sparse tensor factorization (CSTF) [62] and the low tensor multirank regularization (LTMR) [48], in addition to subspace representation-based method (Hysure) [63]. For CNN-based methods, we select various advanced models such as ResTFNet [64], SSRNet [28], MSDCNN [22], and SCPNet [65] to evaluate the effectiveness of the proposed methods. Transformer-based methods include MCT [40], PSRT [43], and DCT [44]. For these representative methods, the detailed network structures can be found in the corresponding references. We refer to the parameter settings in the corresponding references to retrain these models for a fair comparison.

TABLE I

Quantitative Indexes of the Different Methods on the Pavia University Dataset. The Optimal Values Are Shown in **Bold**

Methods	**Pavia University**					
	PSNR	**RMSE**	**ERGAS**	**SAM**	**CC**	**SSIM**
Ideal value	$+\infty$	0	0	0	1	1
Hysure	34.74	7.28	5.188	5.288	0.9463	0.8364
CSTF	40.18	3.76	2.777	3.664	0.9844	0.9279
LTMR	40.54	3.60	2.581	2.821	0.9857	0.9641
ResTFNet	41.34	2.11	1.538	2.381	0.9932	0.9543
SSRNet	43.08	1.72	1.249	2.038	0.9959	0.9608
MSDCNN	41.28	2.12	1.516	2.417	0.9934	0.9522
SCPNet	42.18	1.91	1.440	2.239	0.9940	0.9608
MCT	43.64	1.62	1.189	1.925	0.9963	0.9639
PSRT	42.60	1.82	1.333	2.095	0.9952	0.9624
DCT	43.97	1.56	1.165	1.817	0.9966	0.9673
LGCT	**44.47**	**1.47**	**1.112**	**1.752**	**0.9968**	**0.9693**

1) Results of Pavia University: We followed the experimental setup described in Section IV-B and performed quantitative and qualitative analyses on the Pavia University dataset. Table I reports the results of the different methods for the six metrics in this scene, and it is clear that the deep learning-based methods outperform the prior-based methods, which reflects the powerful capability of deep learning for feature representation. From the quantitative results, the proposed LGCT has obvious advantages in terms of spectral fidelity and detail recovery. In addition, the PSNR and SAM values of the different methods in each band are shown in Fig. 5(a). Band-wise analysis can provide detailed insights into the performance of each method along the spectral dimension. The data in the figure indicates that the proposed LGCT consistently outperform others in most bands. Further evaluation of the fusion results confirms this, and Fig. 6 provides the fusion maps obtained by these methods and the corresponding residual maps. The residual maps provide a finer view of the difference with the reference image. The visual comparison in Fig. 6 shows that the proposed method can generate results that most closely match the reference image. This indicates that the proposed method shows significant effectiveness in maintaining spectral fidelity and recovering image details by combining global contextual understanding with detail-level local visual features.

2) Results of Houston: Table II reports the quantitative results in the Houston dataset. The results show a similar trend to the Pavia University dataset. Deep learning-based methods achieve better reconstruction performance due to their powerful fitting capabilities. Attention turns to the proposed LGCT, which achieves an improvement of 5.97% and 3.79%

[4] Registered Trademark.

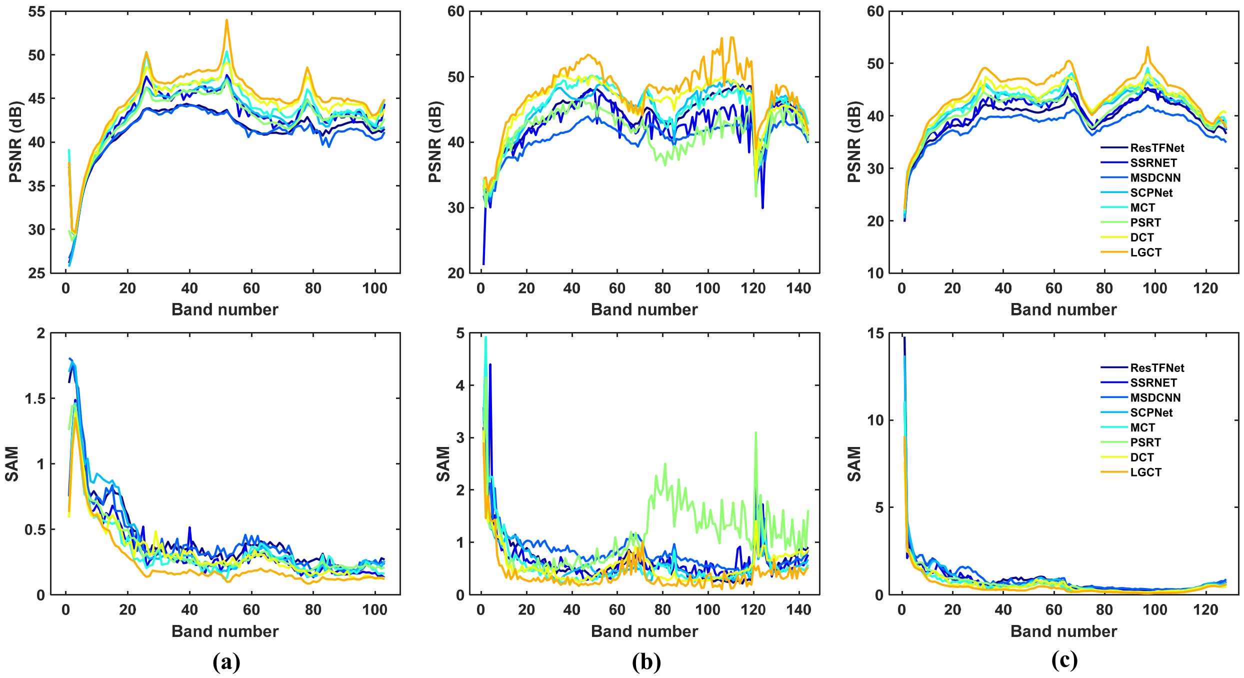


Fig. 5. PSNR values and SAM values for different bands of the reconstructed HSIs in (a) Pavia University, (b) Houston, and (c) Chikusei. Lower values of SAM indicate higher accuracy in spectral fidelity, while higher values of PSNR reflect better retention of spatial detail.

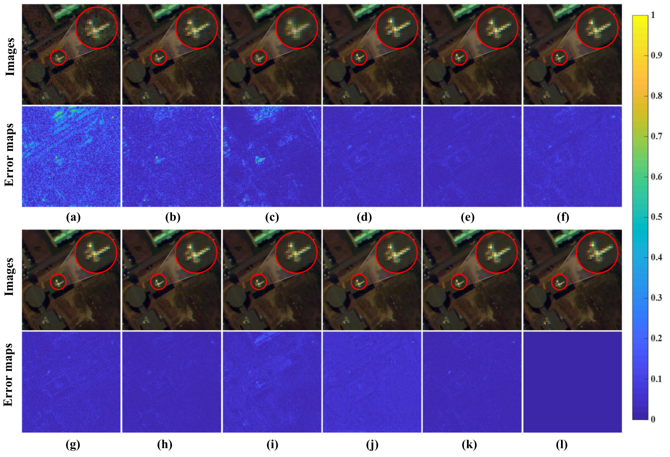


Fig. 6. Visual comparison of the different methods on the Pavia University dataset. The first and third rows show the fusion maps obtained by the different methods (R:67, G:29, B:1), and the second and fourth rows show the average residual along the spectral dimension between the results obtained by different methods and the reference image. Regions with lower residual values (i.e., close to the reference image) are darker, while regions with higher residual values are brighter. (a) Hysure. (b) CSTF. (c) LTMR. (d) ResTFNet. (e) SSRNet. (f) MSDCNN. (g) SCPNet. (h) MCTNet. (i) PSRT. (j) DCT. (k) LGCT. (l) Reference.

in PSNR compared to the best CNN and Transformer methods, i.e., SCPNet and DCT, respectively, highlighting its significant improvement in HSI restoration quality. Moreover, the proposed LGCT has also shown significant advantages in other metrics, indicating its excellent performance in enhancing the spectral fidelity and spatial resolution of HSIs. Fig. 5(b) illustrates the PSNR and SAM values of the different methods on each band, and the orange line represented by the proposed method performs the best in most bands. In addition, Fig. 7 shows a visual comparison of the fusion results. The deeper residual map reflects that the errors between the reconstruction result obtained by the proposed method and the reference image are negligible, which indicates that the fused images obtained by the proposed method achieve an image quality closer to the real scene.

TABLE II

Quantitative Indexes of the Different Methods on the Houston Dataset. The Optimal Values Are Shown in **Bold**

Methods	Houston					
	PSNR	RMSE	ERGAS	SAM	CC	SSIM
Ideal value	$+\infty$	0	0	0	1	1
Hysure	41.59	3.18	3.542	4.422	0.9659	0.9714
CSTF	46.20	1.99	2.149	2.976	0.9883	0.9806
LTMR	42.78	3.31	3.219	3.076	0.9778	0.9730
ResTFNet	46.57	1.14	1.522	2.022	0.9890	0.9851
SSRNet	45.18	1.34	2.741	2.536	0.9886	0.9782
MSDCNN	43.31	1.66	2.008	3.063	0.9834	0.9676
SCPNet	47.75	1.00	1.382	1.749	0.9900	0.9884
MCT	47.89	0.98	1.445	1.687	0.9884	0.9883
PSRT	43.95	1.55	1.755	1.886	0.9863	0.9882
DCT	48.75	0.89	1.316	1.412	0.9916	0.9919
LGCT	**50.60**	**0.72**	**1.083**	**1.153**	**0.9933**	**0.9944**

3) Results of Chikusei: The quantitative results in Chikusei are reported in Table III. Traditional methods are limited by fixed model assumptions and fall behind other deep learning methods in different metrics. Transformer-based methods exhibit superior results compared to CNN-based reconstruction methods, indicating that breaking through the receptive field limitations is beneficial for a more comprehensive understanding of the image content, particularly for HSIs, which possess extensive spectral–spatial contextual information. Among all methods, LGCT achieves optimal performance in six different

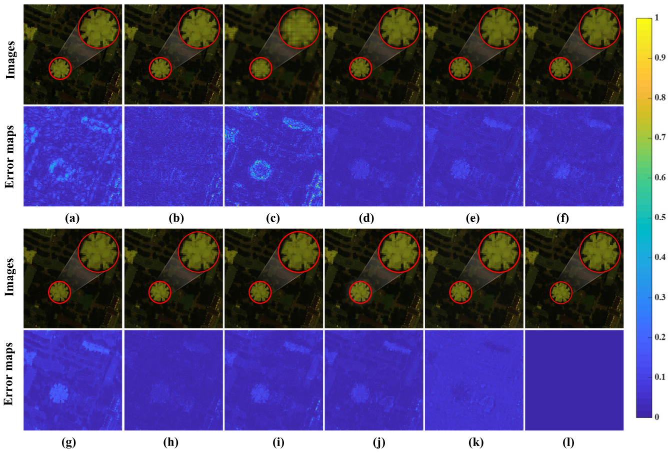


Fig. 7. Visual comparison of the different methods on the Houston dataset. The first and third rows show the fusion maps obtained by the different methods (R:61, G:30, B:8), and the second and fourth rows show the average residual along the spectral dimension between the results obtained by different methods and the reference image. Regions with lower residual values (i.e., close to the reference image) are darker, while regions with higher residual values are brighter. (a) Hysure. (b) CSTF. (c) LTMR. (d) ResTFNet. (e) SSRNet. (f) MSDCNN. (g) SCPNet. (h) MCTNet. (i) PSRT. (j) DCT. (k) LGCT. (l) Reference.

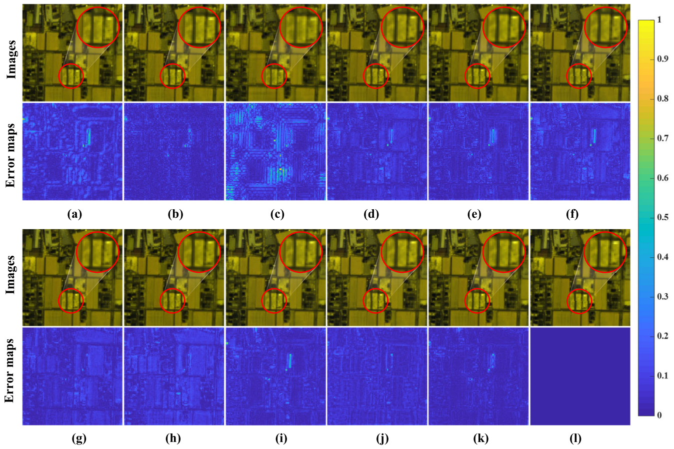


Fig. 8. Visual comparison of the different methods on the Chikusei dataset. The first and third rows show the fusion maps obtained by the different methods (R:80, G:76, B:2), and the second and fourth rows show the average residual along the spectral dimension between the results obtained by different methods and the reference image. Regions with lower residual values (i.e., close to the reference image) are darker, while regions with higher residual values are brighter. (a) Hysure. (b) CSTF. (c) LTMR. (d) ResTFNet. (e) SSRNet. (f) MSDCNN. (g) SCPNet. (h) MCTNet. (i) PSRT. (j) DCT. (k) LGCT. (l) Reference.

TABLE III

QUANTITATIVE INDEXES OF THE DIFFERENT METHODS ON THE CHIKUSEI DATASET. THE OPTIMAL VALUES ARE SHOWN IN **BOLD**

Methods	**Chikusei**					
	PSNR	**RMSE**	**ERGAS**	**SAM**	**CC**	**SSIM**
Ideal value	$+\infty$	0	0	0	1	1
Hysure	35.36	5.49	5.767	4.505	0.9702	0.9119
CSTF	41.96	2.92	3.186	2.578	0.9907	0.9612
LTMR	36.67	5.04	4.791	3.504	0.9754	0.9415
ResTFNet	41.44	0.89	2.364	1.746	0.9919	0.9781
SSRNet	42.21	0.81	2.042	1.710	0.9938	0.9708
MSDCNN	39.93	1.06	2.478	2.103	0.9908	0.9687
SCPNet	43.35	0.71	1.978	1.432	0.9936	0.9864
MCT	43.86	0.67	1.850	1.409	0.9947	0.9821
PSRT	42.54	0.78	1.963	1.557	0.9938	0.9840
DCT	44.29	0.64	1.731	1.279	0.9952	0.9869
LGCT	**45.81**	**0.54**	**1.616**	**1.102**	**0.9956**	**0.9918**

metrics. The PSNR and SAM curves for each band presented in Fig. 5(c) intuitively demonstrate the significant enhancement in spectral fidelity and spatial resolution achieved by the LGCT method relative to other methods. This highlights the efficacy of the LGCT in improving image quality, particularly in the reconstruction of spectral dimension details. For a better visual comparison, Fig. 8 reveals the reconstruction maps and residual maps obtained by all the methods. The figures show that the results corresponding to LGCT are the closest to the reference images, which are not only reflected in the high visual similarity but also verified in the residual maps with fewer artifacts.

D. Comparison Experiments on Real Data

To further validate the fusion effect of the proposed LGCT, the validation experiments are conducted here on real HSI and MSI data. A comparison of the quantitative results of the different methods is reported in Table IV. It is quite evident that the proposed method achieved the best scores under all three metrics, which indicates that the fusion results obtained by the proposed method have obvious advantages in simultaneously retaining the original spectral information and maintaining high spatial details. This is conducive to the fused image to achieve higher application value. Fig. 9 shows the pseudo-color images synthesized in the 60th, 29th, and 7th bands from the fused results obtained by different methods. From the figure, it is known that the overall texture clarity of the fused images is significantly improved, and the edges of the ground objects are sharper. Compared with deep learning-based methods, traditional methods have obvious artifacts. Among similar methods, the fused images obtained by the proposed method have enhanced overall contrast and better visualization performance, which reflects that the proposed method can effectively improve the fusion performance by enhancing the potential information neglected in the spatial–spectral domain and symmetric fusion.

E. Ablation Analysis

In this section, we perform ablation experiments on the CTBs in the spatial and spectral components of the proposed LGCT to assess their specific contribution. In addition, we analyze the effectiveness of symmetric fusion.

1) Spa-CTB and Spe-CTB: In detail, we perform three ablation experiments in the simulated dataset Houston and the real dataset YRE. The first row in Table V presents the quantitative results obtained from feature extraction between spatial and spectral domains using spatial window self-attention and channel-wise self-attention, respectively; this method is called SW-CW. The second and third rows indicate the quantitative results obtained by replacing the spatial window and channel-wise attention in the MSI and HSI feature streams. As shown in Table V, compared to SW-CW, focusing on both local details and global context can significantly enhance the

TABLE IV

Quantitative Indexes of the Different Methods on the YRE Dataset. The Optimal Values Are Shown in **Bold**

Index	Prior-based methods			CNN-based methods				Transformer-based methods			
	Hysure	CSTF	LTMR	ResTFNet	SSRNet	MSDCNN	SCPNet	MCT	PSRT	DCT	**LGCT**
QNR↑	0.6454	0.7233	0.7441	0.8362	0.7726	0.8227	0.8223	0.7686	0.7801	0.8236	**0.8492**
$D_\lambda\downarrow$	0.1143	0.0807	0.0700	0.0342	0.0492	0.0504	0.0319	0.0526	0.0646	0.0414	**0.0223**
$D_s\downarrow$	0.2713	0.2132	0.1998	0.1341	0.1874	0.1336	0.1506	0.1887	0.1661	0.1409	**0.1315**

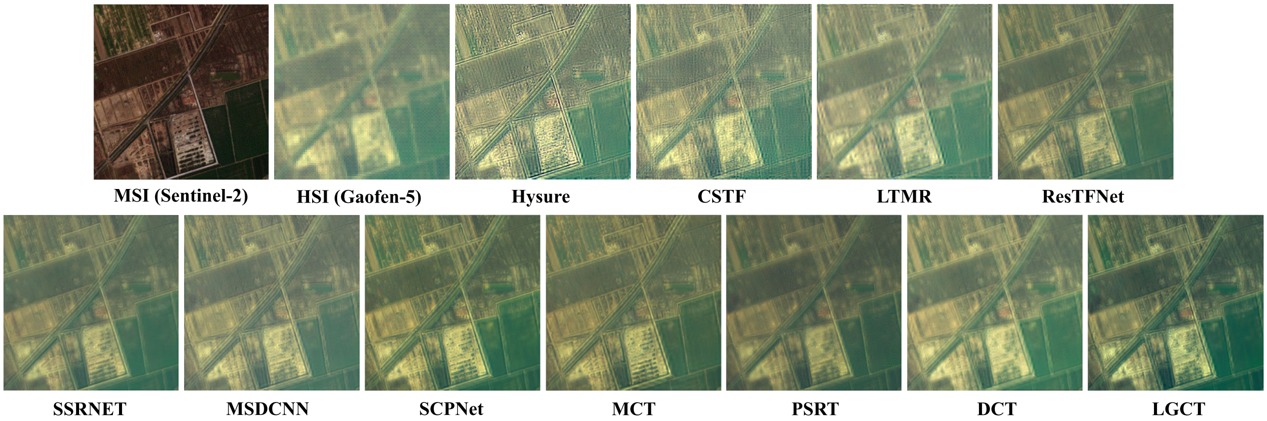


Fig. 9. Visual comparison of the different methods on the YRE real dataset. The presented fusion images are created by combining the 60th, 29th, and 7th channels.

TABLE V

Performance Contribution of the Proposed Spa-CTB and Spe-CTB on the Simulated Dataset and the Real Dataset. Values in Brackets Indicate the Difference From the First Row. The Optimal Values Are Shown in **Bold**

Methods	Simulated dataset (Houston)						Real dataset (YRE)		
	PSNR↑	RMSE↓	ERGAS↓	SAM↓	CC↑	SSIM↑	QNR↑	$D_\lambda\downarrow$	$D_s\downarrow$
SW-CW	49.69	0.80	1.166	1.188	0.9925	0.9940	0.8424	0.0265	0.1347
+SpaCTB	50.21 (+0.52)	0.75 (+0.05)	1.118 (+0.048)	1.175 (+0.013)	0.9930 (+0.0005)	0.9942 (+0.0002)	0.8462 (+0.0038)	0.0236 (+0.0029)	0.1333 (+0.0014)
+SpeCTB	49.89 (+0.20)	0.78 (+0.02)	1.126 (+0.040)	1.173 (+0.015)	0.9930 (+0.0005)	0.9942 (+0.0002)	0.8454 (+0.0030)	0.0232 (+0.0033)	0.1346 (+0.0001)
LGCT	**50.60** **(+0.91)**	**0.72** **(+0.08)**	**1.083** **(+0.083)**	**1.153** **(+0.035)**	**0.9933** **(+0.0008)**	**0.9944** **(+0.0004)**	**0.8492** **(+0.0068)**	**0.0223** **(+0.0042)**	**0.1315** **(+0.0032)**

TABLE VI

Performance of the Symmetric Fusion Strategy on the Houston Dataset. "w/o Sym" Indicates That Symmetric Fusion Is Not Used. Optimal Values Are Shown in **Bold**

Methods	PSNR↑	SSIM↑	SAM↓	#Params(M)	Flops(G)
w/o sym	48.35	0.9921	1.411	**5.12**	**17.23**
LGCT	**50.60**	**0.9944**	**1.153**	5.40	18.42

fusion performance, which is demonstrated in both simulated and real datasets. This indicates the importance of global and local processing mechanisms. Notably, applying SpeCTB results in a greater improvement in SAM and D_λ compared to applying SpaCTB. These two metrics are used to assess the spectral fidelity. Conversely, applying SpaCTB has a greater advantage over SpeCTB in terms of PSNR and D_s, which are metrics used to measure the spatial reconstruction quality. These results indicate the effectiveness of CTBs specifically designed for spatial and spectral domains in emphasizing spatial long-range dependencies and spectral local details. Applying CTBs to both spatial and spectral components can achieve optimal performance. This demonstrates the effectiveness of the proposed CTBs in complementing and enhancing overlooked potential spectral-spatial features, thereby contributing to achieving higher-quality fusion results.

2) Symmetric Fusion Strategy: To verify the effectiveness of the bimodal multiscale feature symmetric strategy, we conducted experiments without considering symmetric fusion and presented the experimental results in Table VI, where the first row indicates the results without symmetric fusion. The results show that while symmetric fusion leads to a limited increase in computational and parameter load, it significantly boosts the fusion performance of the model. This reflects the effectiveness of the symmetric fusion strategy, which achieves detail enhancement and spectral fidelity in the multiscale feature hierarchical fusion process of encoding and decoding stages by improving fusion feature reuse.

3) Spectral Multiscale Inputs: To evaluate the contribution of spectral information to HSI reconstruction, we performed an ablation study during the multiscale feature symmetric fusion

TABLE VII

CONTRIBUTION OF SPECTRAL MULTISCALE FEATURES TO IMAGE RECONSTRUCTION ON THE HOUSTON DATASET. "W/O SPE" INDICATES THE USE OF SPATIAL INFORMATION ONLY. "G" STANDS FOR GIGA. THE OPTIMAL VALUES ARE SHOWN IN **BOLD**

	w/o Spe	$+F_H^3$	$+F_H^3+F_H^2$	+All Spe
F_H^3	×	✓	✓	✓
F_H^2	×	×	✓	✓
F_H^1	×	×	×	✓
PSNR↑	48.96	49.64	49.92	**50.60**
RMSE↓	0.87	0.80	0.78	**0.72**
ERGAS↓	1.212	1.164	1.140	**1.083**
SAM↓	1.245	1.191	1.184	**1.153**
CC↑	0.9922	0.9925	0.9926	**0.9933**
SSIM↑	0.9936	0.9941	0.9942	**0.9944**
#Params(M)	**3.433**	5.157	5.240	5.406
Flops(G)	**14.156**	17.906	18.246	18.416

TABLE VIII

ANALYZE THE IMPACT OF WINDOW SIZE AND NUMBER OF GROUPS TO IMAGE RECONSTRUCTION ON THE HOUSTON DATASET. THE OPTIMAL VALUES ARE SHOWN IN **BOLD**

Window size (M)						
	PSNR↑	RMSE↓	ERGAS↓	SAM↓	CC↑	SSIM↑
M=4	50.34	0.74	1.112	1.165	0.9929	0.9943
M=8	**50.60**	**0.72**	**1.083**	**1.153**	**0.9933**	**0.9944**
M=16	50.32	0.74	1.115	1.159	0.9928	0.9942
M=32	49.97	0.77	1.147	1.196	0.9925	0.9940
Number of groups (N)						
	PSNR↑	RMSE↓	ERGAS↓	SAM↓	CC↑	SSIM↑
N=4	50.22	0.75	1.107	1.193	0.9931	0.9941
N=8	**50.60**	**0.72**	**1.083**	1.153	**0.9933**	**0.9944**
N=16	50.51	0.73	1.098	**1.140**	0.9931	**0.9944**
N=32	50.16	0.76	1.102	1.173	0.9932	0.9943

stage, specifically analyzing the impact of spectral multiscale features. The detailed results are presented in Table VII, where the first column represents the case where only spatial multiscale features are considered, serving as the control group. The quantitative results indicate that spectral information positively contributes to image reconstruction. The fusion of multiscale spectral features further enhances the quality of the reconstructed images, which clearly demonstrating the significant role of spectral information in image fusion.

4) Window Size and Number of Groups: To analyze the impact of window size M and the number of groups N on image reconstruction, we conducted ablation experiments on the Houston dataset. For the hyperparameter window size M in Spa-CTB, we trained models by setting M to 4, 8, 16, and 32. As shown in Table VIII, the results indicate that reconstruction performance initially increases with M, peaking at $M = 8$, before declining as M increases further. This confirms that a larger receptive field can enhance performance, but excessively large window may result in the loss of local details and introduce unnecessary complexity. The proposed Spa-CTB can implicitly improve the receptive field of the network when it has smaller windows, thereby complementing and enhancing global modeling capabilities. Consequently, we set the window size M to 8 in this study.

For the hyperparameter group number N in Spe-CTB, we keep the same settings as M to train the model. The results are shown in Table VIII, where the performance shows a similar trend to M as N increases. This indicates that fewer groups (which implies a larger window size for each group) generally result in better reconstruction performance. However, when the group number is too small, the lack of sufficient local details within each group may limit the performance gain. The optimal reconstruction performance is achieved at $N = 8$, so we set the number of groups N to 8 in this study.

F. Analysis of Model Complexity

In this section, model complexity is analyzed, and Fig. 1 compares the computational efficiency and fusion performance of the deep learning-based methods on the Houston dataset. Generally, Transformer-based methods exhibit lower efficiency compared to CNN-based methods, but they deliver superior fusion performance. While Transformer models are superior in capturing global contextual information, their fully connected attentional mechanisms typically involve more significant computational effort. Although dividing the Transformer into different windows can reduce model complexity, the global perceptual capability is limited, resulting in a performance bottleneck. As shown in Fig. 1, compared to other Transformer-based methods such as MCT and DCT, the proposed LGCT shows higher PSNR with significantly fewer parameters and Flops. Despite the lower computational cost of PSRT compared to the proposed LGCT, its fusion performance lags behind. These results demonstrate that our model effectively enhances global dependency capability through the proposed Spa-CTB and Spe-CTB without incurring additional computational costs. At the same time, it maintains attention to local semantic details, thus achieving a good balance between fusion performance and computational efficiency, making real-time applications feasible.

V. CONCLUSION

In this article, we proposed a Transformer-based model for HSI and MSI fusion called LGCT. The proposed LGCT consists of two separate feature extractors for spatial and spectral domains and a multiscale feature symmetric fusion network for fusing spectral–spatial deep features. With the carefully designed collaborative Transformer blocks (CTBs) for both spatial and spectral, two parallel feature extractors can effectively focus on and enhance overlooked potential crucial features in spatial and spectral domains, thus achieving efficient modeling of LR-HSI and HR-MSI from local details to global contexts at different scales. In addition, the proposed method progressively fuses multiscale enhanced features in a hierarchical and symmetrical manner to improve the quality of HSI reconstruction. Experimental results on simulated and real datasets show that LGCT has significant advantages in spatial and spectral metrics over existing fusion SOTA methods while maintaining high computational efficiency.

The purpose of fusing multimodal images to enhance image quality is to improve the performance of downstream tasks. Current approaches typically separate image enhancement and downstream tasks into two independent stages, which may limit the performance and efficiency of system. Moving forward, we plan to employ a multitask learning framework to integrate image fusion and downstream tasks in an end-to-end manner. This approach aims to enhance the generalization ability of model and expand its applicability.

REFERENCES

[1] M. Shimoni, R. Haelterman, and C. Perneel, "Hypersectral imaging for military and security applications: Combining myriad processing and sensing techniques," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 101–117, Jun. 2019.

[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] W. Sun, K. Ren, X. Meng, C. Xiao, G. Yang, and J. Peng, "A band divide-and-conquer multispectral and hyperspectral image fusion method," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5502113.

[4] H. Gao, S. Li, J. Li, and R. Dian, "Multispectral image pan-sharpening guided by component substitution model," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5406413.

[5] G. Vivone, "Multispectral and hyperspectral image fusion in remote sensing: A survey," *Inf. Fusion*, vol. 89, pp. 405–417, Jan. 2023.

[6] H. Xu, J. Ma, Z. Shao, H. Zhang, J. Jiang, and X. Guo, "SDPNet: A deep network for pan-sharpening with enhanced information representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4120–4134, May 2021.

[7] B. Tu, Q. Ren, J. Li, Z. Cao, Y. Chen, and A. Plaza, "NCGLF2: Network combining global and local features for fusion of multisource remote sensing data," *Inf. Fusion*, vol. 104, Apr. 2024, Art. no. 102192.

[8] R. Dian, S. Li, B. Sun, and A. Guo, "Recent advances and new guidelines on hyperspectral and multispectral image fusion," *Inf. Fusion*, vol. 69, pp. 40–51, May 2021.

[9] S. Jia, S. Jiang, Z. Lin, N. Li, M. Xu, and S. Yu, "A survey: Deep learning for hyperspectral image classification with few labeled samples," *Neurocomputing*, vol. 448, pp. 179–204, Aug. 2021.

[10] D. Zhu, B. Du, and L. Zhang, "Two-stream convolutional networks for hyperspectral target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6907–6921, Aug. 2021.

[11] N. Li, S. Jiang, J. Xue, S. Ye, and S. Jia, "Texture-aware self-attention model for hyperspectral tree species classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5502215.

[12] B. Tu, W. He, Q. Li, Y. Peng, and A. Plaza, "A new context-aware framework for defending against adversarial attacks in hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5505114.

[13] K. Wang, Y. Wang, X. Zhao, J. C. Chan, Z. Xu, and D. Meng, "Hyperspectral and multispectral image fusion via nonlocal low-rank tensor decomposition and spectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7654–7671, Nov. 2020.

[14] J. W. P. Sun, H. Li, W. Li, X. Meng, C. Ge, and Q. Du, "Low-rank and sparse representation for hyperspectral image processing: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 1, pp. 10–43, Jun. 2021.

[15] X. Fu, S. Jia, M. Xu, J. Zhou, and Q. Li, "Fusion of hyperspectral and multispectral images accounting for localized inter-image changes," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5517218.

[16] T. Huang, W. Dong, J. Wu, L. Li, X. Li, and G. Shi, "Deep hyperspectral image fusion network with iterative spatio-spectral regularization," *IEEE Trans. Comput. Imag.*, vol. 8, pp. 201–214, 2022.

[17] W. Dong, J. Qu, T. Zhang, Y. Li, and Q. Du, "Context-aware guided attention based cross-feedback dense network for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5530814.

[18] C. Zhou, Z. He, A. Lou, and A. Plaza, "RGB-to-HSV: A frequency-spectrum unfolding network for spectral super-resolution of RGB videos," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5609318.

[19] L. He, J. Zhu, J. Li, A. Plaza, J. Chanussot, and Z. Yu, "CNN-based hyperspectral pansharpening with arbitrary resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518821.

[20] P. Guan and E. Y. Lam, "Multistage dual-attention guided fusion network for hyperspectral pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5515214.

[21] J. Li, K. Zheng, W. Liu, Z. Li, H. Yu, and L. Ni, "Model-guided coarse-to-fine fusion network for unsupervised hyperspectral image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[22] Q. Yuan, Y. Wei, X. Meng, H. Shen, and L. Zhang, "A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 3, pp. 978–989, Mar. 2018.

[23] F. Zhou, R. Hang, Q. Liu, and X. Yuan, "Pyramid fully convolutional network for hyperspectral and multispectral image fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 5, pp. 1549–1558, May 2019.

[24] R. Ran, L.-J. Deng, T.-X. Jiang, J.-F. Hu, J. Chanussot, and G. Vivone, "GuidedNet: A general CNN fusion framework via high-resolution guidance for hyperspectral image super-resolution," *IEEE Trans. Cybern.*, vol. 53, no. 7, pp. 4148–4161, Jul. 2023.

[25] K. Li, W. Zhang, D. Yu, and X. Tian, "HyperNet: A deep network for hyperspectral, multispectral, and panchromatic image fusion," *ISPRS J. Photogramm. Remote Sens.*, vol. 188, pp. 30–44, Jun. 2022.

[26] J. Yao, D. Hong, J. Chanussot, D. Meng, X. Zhu, and Z. Xu, "Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 208–224.

[27] J. Hu, Y. Tang, Y. Liu, and S. Fan, "Hyperspectral image super-resolution based on multiscale mixed attention network fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[28] X. Zhang, W. Huang, Q. Wang, and X. Li, "SSR-NET: Spatial–spectral reconstruction network for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5953–5965, Jul. 2021.

[29] S. Liu, S. Miao, S. Liu, B. Li, W. Hu, and Y. Zhang, "Circle-net: An unsupervised lightweight-attention cyclic network for hyperspectral and multispectral image fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4499–4515, 2023.

[30] W. Wei, J. Nie, Y. Li, L. Zhang, and Y. Zhang, "Deep recursive network for hyperspectral image super-resolution," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 1233–1244, 2020.

[31] Q. Xie, M. Zhou, Q. Zhao, Z. Xu, and D. Meng, "MHF-Net: An interpretable deep network for multispectral and hyperspectral image fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1457–1473, Mar. 2020.

[32] R. Dian, A. Guo, and S. Li, "Zero-shot hyperspectral sharpening," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12650–12666, Oct. 2023.

[33] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[34] B. Tu, X. Liao, Q. Li, Y. Peng, and A. Plaza, "Local semantic feature aggregation-based transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5536115.

[35] W. G. C. Bandara and V. M. Patel, "HyperTransformer: A textural and spectral feature fusion transformer for pansharpening," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 1767–1777.

[36] Y. Gao, M. Zhang, J. Wang, and W. Li, "Cross-scale mixing attention for multisource remote sensing data fusion and classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5507815.

[37] M. Jiang et al., "GraphGST: Graph generative structure-aware transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5504016.

[38] S. Deng, L.-J. Deng, X. Wu, R. Ran, and R. Wen, "Bidirectional dilation transformer for multispectral and hyperspectral image fusion," in *Proc. 32nd Int. Joint Conf. Artif. Intell.*, Aug. 2023, pp. 3633–3641.

[39] L. Chen, G. Vivone, J. Qin, J. Chanussot, and X. Yang, "Spectral–spatial transformer for hyperspectral image sharpening," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 11, pp. 16733–16747, Nov. 2024.

[40] X. Wang, X. Wang, R. Song, X. Zhao, and K. Zhao, "MCT-net: Multi-hierarchical cross transformer for hyperspectral and multispectral image fusion," *Knowl.-Based Syst.*, vol. 264, Mar. 2023, Art. no. 110362.

[41] S. Jia, Z. Min, and X. Fu, "Multiscale spatial–spectral transformer network for hyperspectral and multispectral image fusion," *Inf. Fusion*, vol. 96, pp. 117–129, Aug. 2023.

[42] Y. Sun et al., "Dual spatial–spectral pyramid network with transformer for hyperspectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5526016.

[43] S.-Q. Deng, L.-J. Deng, X. Wu, R. Ran, D. Hong, and G. Vivone, "PSRT: Pyramid shuffle-and-reshuffle transformer for multispectral and hyperspectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5503715.

[44] Q. Ma, J. Jiang, X. Liu, and J. Ma, "Reciprocal transformer for hyperspectral and multispectral image fusion," *Inf. Fusion*, vol. 104, Apr. 2024, Art. no. 102148.

[45] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 2, pp. 528–537, Feb. 2011.

[46] C. Lanaras, E. Baltsavias, and K. Schindler, "Hyperspectral super-resolution by coupled spectral unmixing," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3586–3594.

[47] R. Wu, W.-K. Ma, X. Fu, and Q. Li, "Hyperspectral super-resolution via global–local low-rank matrix estimation," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7125–7140, Oct. 2020.

[48] R. Dian and S. Li, "Hyperspectral image super-resolution via subspace-based low tensor multi-rank regularization," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 5135–5146, Oct. 2019.

[49] Z. Zhu, J. Hou, J. Chen, H. Zeng, and J. Zhou, "Hyperspectral image super-resolution via deep progressive zero-centric residual learning," *IEEE Trans. Image Process.*, vol. 30, pp. 1423–1438, 2021.

[50] W. Dong, C. Zhou, F. Wu, J. Wu, G. Shi, and X. Li, "Model-guided deep hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 30, pp. 5754–5768, 2021.

[51] C. Zhu, S. Deng, Y. Zhou, L.-J. Deng, and Q. Wu, "QIS-GAN: A lightweight adversarial network with quadtree implicit sampling for multispectral and hyperspectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5531115.

[52] K. Zhang et al., "Spectral–spatial dual graph unfolding network for multispectral and hyperspectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5508718.

[53] Z. Wang, B. Chen, R. Lu, H. Zhang, H. Liu, and P. K. Varshney, "FusionNet: An unsupervised convolutional variational network for hyperspectral and multispectral image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 7565–7577, 2020.

[54] K. Han et al., "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2022.

[55] J.-F. Hu, T.-Z. Huang, L.-J. Deng, H.-X. Dou, D. Hong, and G. Vivone, "Fusformer: A transformer-based fusion network for hyperspectral image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[56] W. Shi et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.

[57] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.

[58] Y. Cai et al., "Degradation-aware unfolding half-shuffle transformer for spectral compressive imaging," in *Proc. Adv. Neural. Inf. Process. Syst.*, vol. 35, 2022, pp. 37749–37761.

[59] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5728–5739.

[60] K. Ren, W. Sun, X. Meng, G. Yang, J. Peng, and J. Huang, "A locally optimized model for hyperspectral and multispectral images fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5519015.

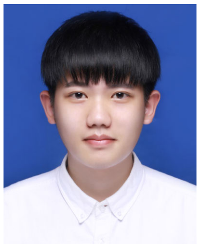
[61] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva, "Multispectral and panchromatic data fusion assessment without reference," *Photogramm. Eng. Remote Sens.*, vol. 74, no. 2, pp. 193–200, Feb. 2008.

[62] S. Li, R. Dian, L. Fang, and J. M. Bioucas-Dias, "Fusing hyperspectral and multispectral images via coupled sparse tensor factorization," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 4118–4130, Aug. 2018.

[63] M. Simoes, J. Bioucas-Dias, L. B. Almeida, and J. Chanussot, "A convex formulation for hyperspectral image superresolution via subspace-based regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3373–3388, Jun. 2015.

[64] X. Liu, Q. Liu, and Y. Wang, "Remote sensing image fusion based on two-stream fusion network," *Inf. Fusion*, vol. 55, pp. 1–15, Mar. 2020.

[65] B. Pan, Q. Qu, X. Xu, and Z. Shi, "Structure–color preserving network for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5520512.



Wangquan He (Student Member, IEEE) received the M.S. degree in information and communication engineering from Hunan Institute of Science and Technology, Yueyang, China, in 2022. He is currently pursuing the Ph.D. degree with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China.

His research interests include hyperspectral image processing and super-resolution.

Xiyou Fu (Member, IEEE) received the B.S. degree from Wuhan University, Wuhan, China, in 2012, and the M.S. and Ph.D. degrees from the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China, in 2015 and 2019, respectively.

He is currently an Assistant Professor with Shenzhen University, Shenzhen, China. His research interests include hyperspectral image restoration, anomaly detection, and super-resolution.

Nanying Li (Student Member, IEEE) received the B.E. degree in automation and the M.E. degrees in information and communication engineering from Hunan Institute of Science and Technology, Yueyang, China, in 2017 and 2021, respectively. She is currently pursuing the Ph.D. degree in computer science and technology with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China.

Her research interests include hyperspectral image classification and image segmentation.

Qi Ren (Student Member, IEEE) received the B.S. degree in computer science and technology from Shanxi Datong University, Datong, China, in 2020, and the M.S. degree in information and communication engineering from Hunan Institute of Technology, Yueyang, China, in 2023. He is currently pursuing the Ph.D. degree with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China.

His research interests include hyperspectral image processing and spectral construction.

Sen Jia (Senior Member, IEEE) received the B.E. and Ph.D. degrees from the College of Computer Science, Zhejiang University, Hangzhou, China, in 2002 and 2007, respectively.

Since 2008, he has been with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China, where he is currently a Full Professor. His research interests include remote sensing image processing, signal and image processing, and machine learning.