remote sensing

MDPI

Article

Hierarchical Spectral–Spatial Transformer for Hyperspectral and Multispectral Image Fusion

Tianxing Zhu [1], Qin Liu [1] and Lixiang Zhang [2,*]

[1] School of Software Engineering, Tongji University, Shanghai 200070, China; txzhu@tongji.edu.cn (T.Z.); qin.liu@tongji.edu.cn (Q.L.)
[2] School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China
* Correspondence: zhanglixiang@mail.nwpu.edu.cn

Abstract: This paper presents the Hierarchical Spectral–Spatial Transformer (HSST) network, a novel approach applicable to both drone-based and broader remote sensing platforms for integrating hyperspectral (HSI) and multispectral (MSI) imagery. The HSST network improves upon conventional multi-head self-attention transformers by integrating cross attention, effectively capturing spectral and spatial features across different modalities and scales. The network's hierarchical design facilitates the extraction of multi-scale information and employs a progressive fusion strategy to incrementally refine spatial details through upsampling. Evaluations on three prominent hyperspectral datasets confirm the HSST's superior efficacy over existing methods. The findings underscore the HSST's utility for applications, including drone operations, where the high-fidelity fusion of HSI and MSI data is crucial.

Keywords: hyperspectral image; image fusion; spectral–spatial transformer; feature fusion

Citation: Zhu, T.; Liu, Q.; Zhang, L. Hierarchical Spectral–Spatial Transformer for Hyperspectral and Multispectral Image Fusion. *Remote Sens.* **2024**, *16*, 4127. https://doi.org/10.3390/rs16224127

Academic Editor: Farid Melgani

Received: 18 September 2024
Revised: 26 October 2024
Accepted: 4 November 2024
Published: 5 November 2024

1. Introduction

Hyperspectral imagery (HSI), embodying an array of slender spectral bands that span from visible to near-infrared wavelengths, enables the meticulous discernment of terrestrial object compositions. This proficiency proves to be especially beneficial in fields such as environmental surveillance and precision agriculture [1]. The utility of HSI spans across various domains, including land cover classification and anomaly detection, thanks to their detailed object attribute characterization [2–5]. Nevertheless, the high spectral resolution inherent in HSI often comes at the expense of spatial resolution, a trade-off imposed by the constraints of imaging platforms, including those mounted on drones. The scattering of electromagnetic waves with narrow bandwidths into the instantaneous field of view typically requires a compromise in spatial resolution to maintain an acceptable signal-to-noise ratio, thereby limiting the broader application of HSI [6]. Consequently, there is a pressing need for research aimed at algorithmically fusing low-resolution hyperspectral images (LR-HSI) with high-resolution multispectral images (HR-MSI) to generate high-resolution hyperspectral images (HR-HSI), particularly for use in drone-based platforms where enhanced spatial detail is critical [7]. This integration is essential for maximizing the potential of HSI in both aerial and terrestrial applications, ensuring that drones can effectively utilize the resulting HR-HSI for a wide range of tasks.

Several machine learning techniques have been proposed for the integration of LR-HSI (low-resolution hyperspectral imaging) and HR-MSI (high-resolution multispectral imaging), including methods based on matrix and tensor factorization [8–12]. However, many of these techniques rely on manually designed priors, which not only limit their representational capacity but also prove to be time-consuming. The swift proliferation of deep learning has illustrated its immense potential for the fusion of hyperspectral and multispectral imaging, owing to its sturdy capabilities in the extraction of features.

This technique employs a multi-tiered, deep-seated neural network to discern the correlation among high-resolution multispectral imaging, low-resolution hyperspectral imaging, and the corresponding high-resolution hyperspectral imaging, thereby aiding the accomplishment of fusion tasks. Fusion methodologies predicated on deep learning not only extrapolate information from the input imagery but also utilize learned correlations as prior cognizance to reconstruct spectral and spatial details that are absent in multispectral and hyperspectral imaging. As a result, in contrast to conventional fusion techniques that rely on manually stipulated prior information, fusion methods underpinned by deep learning exhibit superior efficacy. In recent epochs, several fusion techniques predicated on convolutional neural networks have been conceived. To give an example, Yang et al. [13] proposed a fusion technique that amalgamates convolutional neural networks and spatial attention to extract intricate textures and enhance spatial structure. Cai et al. [14] incorporated a super-resolution module and progressive learning into their network, which enables the capture of spatial details at varying scales and their integration into upsampled multispectral images. Despite the enhancements in generalization capabilities by current convolutional neural network-based fusion algorithms [15–17], these algorithms fall short in effectively exploiting spatial location information and extracting long-range dependencies in images, thereby resulting in a deficiency of global context information. Inspired by the triumphant implementation of transformers in NLP, scholars have begun to suggest the use of vision transformers for fusion tasks. However, most extant transformer-based hyperspectral and multispectral imaging fusion methods extract features from a singular modality. This approach regrettably neglects the interplay between spatial and spectral modalities.

The observations previously discussed lay the groundwork for the introduction of a novel Hierarchical Spectral–Spatial Transformer (HSST) network. The HSST is a two-branch network that incorporates the self-attention mechanism of the transformer to extract and merge spectral features of HSI with spatial features of MSI. To effectively leverage the extensive spatial information in remote sensing images, a hierarchical structure is employed to facilitate multi-scale information extraction. For the reconstruction of HR-HSI, a hierarchical progressive fusion is employed to gradually restore spatial detail information through progressive upsampling, thereby harnessing the acquired multi-level feature representation. The main contributions of this paper are described in detail as follows:

- We introduce a Hierarchical Spectral–Spatial Transformer network (HSST) for the fusion of HSI and MSI. The HSST is designed to extract and merge deep spectral and spatial features via hierarchical Spectral–Spatial Transformers and subsequently reconstruct HR-HSI through a process of hierarchical progressive fusion.
- We also propose the use of a Hierarchical Spectral–Spatial Transformer to more effectively capture cross-modality spectral and spatial features at multiple scales. In addition to the traditional multi-head self-attention transformers, cross attention is incorporated to enhance the extraction of cross-modality features.
- To optimize the spatial details of the reconstructed HR-HSI, hierarchical progressive fusion is proposed to gradually recover spatial detail information through progressive upsampling and fusion. This cumulative process facilitates the gradual reconstruction of the HR-HSI result.

The remainder of this article is as follows: Section 2 provides an overview of relevant HSI and MSI fusion methods and a detailed description of the proposed HSST. Section 3 presents and analyses the experimental results on three datasets. Section 4 consists of discussions and ablation studies. Finally, Section 5 concludes the paper.

2. Materials and Methods

2.1. Related Works

2.1.1. HSI and MSI Fusion

The current fusion strategies for hyperspectral images (HSI) and multispectral images (MSI) primarily fall into four categories: matrix factorization, tensor factorization, pan-sharpening, and deep learning.

Matrix factorization-based techniques transform the 3D HSI into a 2D matrix. Thereafter, an HR-HSI is concocted utilizing the endmember matrix and abundance matrix extricated from the LR-HSI and HR-MSI. For example, the CNMF [18] method employs non-negative matrix factorization to disintegrate HR-MSI and LR-HSI into mixed pixels, thereby engendering a superior HR-HSI through the utilization of the abundance matrix of LR-HSI and the endmember matrix of HR-MSI.

Tensor factorization-based methods are utilized to maintain the spatial and spectral structure of images, as opposed to reshaping them into matrices. Two frequently used decompositions in the fusion of HSI and MSI are the Tucker decomposition and the Canonical Polyadic decomposition. Dian et al. [19] introduced a nonlocal sparse tensor factorization method for semi-blind fusion of HSI and MSI. To mitigate computational strain, Kanatsoulis et al. [20] implemented CP decomposition on the HR-HSI, obtaining each factor matrix by resolving the least squares equation.

Pan-sharpening methods integrate the fusion of HSI and MSI images. Grohnfeldt et al. [21] introduce a sparse representation (SR)-based pan-sharpening method for HSI and MSI fusion. Although it achieves satisfactory performance with a limited number of MSI bands, it fails to yield desirable outcomes when the number of bands increases due to the diminished correlation between the missing bands of MSI and the high-resolution images.

With the emergence of deep learning, it is now feasible to learn all parameters from training data using deep learning networks, eliminating the necessity for assumptions about the images [22]. Bearing this in mind, Dian et al. [23] propose a novel HSI sharpening method called DHSI for the fusion of HSI and MSI data. The method employs a deep residual network to learn image priors, thereby obviating the need for manually crafted priors. Palsson et al. [24] have proposed a method that utilizes a trained three-dimensional convolutional neural network to acquire filters for effectively merging MSI and HSI. To address the computational complexity associated with the 3D CNN, they have employed principal component analysis (PCA) as a means of reducing the dimensionality prior to fusion [25]. In a similar vein, Zheng et al. [26] have developed EC-FTN that aims to preserve low-level structural details, including sharp edges. However, it is worth noting that many of the existing learning-based approaches are supervised in nature, requiring a substantial amount of training data, which can pose challenges in real-world applications.

2.1.2. Hyperspectral Image Transformer

The proliferation of convolutional neural network (CNN)-based fusion methodologies for amalgamating hyperspectral and multispectral images is noteworthy. Nonetheless, the limiting receptive field of CNNs poses a challenge in extracting comprehensive information from images. This is particularly crucial in hyperspectral imaging (HSI) where strong correlations in the spectral dimension necessitate the extraction of global characteristics for enhanced fusion performance. The transformer model, renowned for its self-attention mechanism that accommodates long-range information, has gained momentum in various applications. In the domain of HSI processing, the transformer model has demonstrated its prowess in handling sequential data. For example, Hong et al. [27] proposed SpectralFormer, a primary network for HSI classification, adept at learning locally sequential spectral information from adjacent HSI bands and generating group-wise spectral embeddings. He et al. [28] introduced a spectral–spatial transformer classification network, which combines a well-structured CNN for spatial feature extraction with an improved transformer to capture sequential spectral relationships. Selen et al. [29] developed a spectral-swin transformer (SpectralSWIN) classification network, employing a spectral-swin module to

simultaneously process spatial and spectral features. Transformers have gained significant traction in the realm of HSI reconstruction. Cai et al. [30] were the pioneers in proposing a transformer-based method for HSI reconstruction. They utilized the feature map of each spectral channel to compute self-attention. Similarly, Bandara et al. [31] employed the self-attention mechanism of the transformer to transfer high-resolution textual features to low-resolution features for pan-sharpening.

2.2. Proposed Method

Our network utilizes a bifurcated structure, as illustrated in Figure 1. The input HSI and MSI experience a corresponding upsampling and downsampling, respectively, facilitated by linear and convolutive procedures. This process extracts spatial information from low-level detailed features alongside semantic information from high-level semantic features. The Spectral–Spatial Transformer is harnessed to execute comprehensive feature extraction, foster meaningful interplay of information, and facilitate the amalgamation of feature maps on a homogeneous scale. The image reconstruction incorporates a hierarchical progressive fusion procedure, which is designed to incrementally combine extracted features across varying scales.

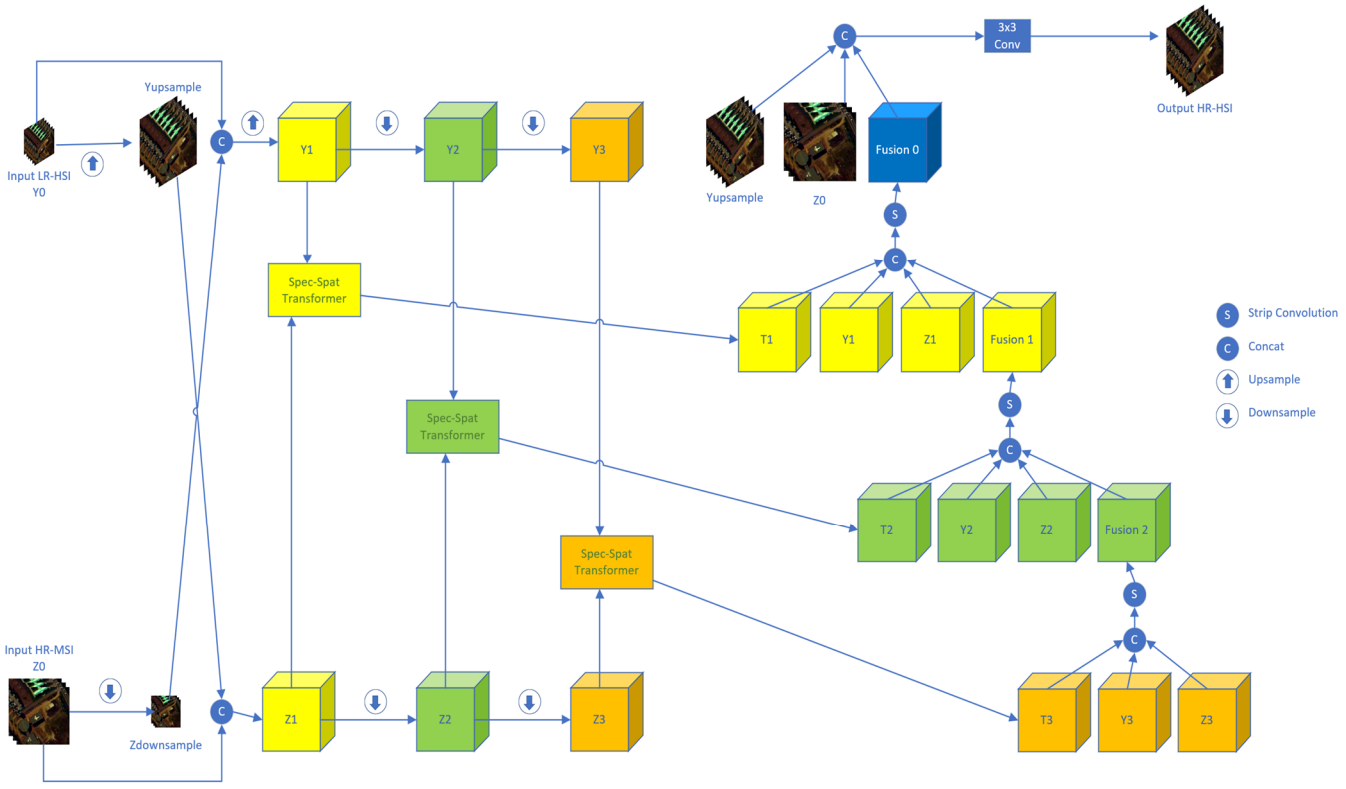

Figure 1. Framework of the proposed network.

Let $Y_0 \in \mathbb{R}^{r \times c \times S}$ represent the input LR-HSI, where r, c, and S denote the number of rows, columns, and spectral bands respectively. Similarly, let $Z_0 \in \mathbb{R}^{R \times C \times s}$ denote the input HR-MSI, where R, C, and s correspond to the number of rows, columns, and spectral bands respectively. Initially, we perform upsampling on the LR-HSI and downsampling on the HR-MSI to generate $Y_{upsample} \in \mathbb{R}^{R \times C \times S}$ and $Z_{downsample} \in \mathbb{R}^{r \times c \times s}$, respectively, employing the bilinear interpolation method, which can be expressed as follows:

$$Y_{upsample} \in UPS(Y_0) \tag{1}$$

$$Z_{downsample} \in DOWNS(Z_0) \tag{2}$$

where *UPS* and *DOWNS* denote the functions of upsampling and downsampling. Subsequently, we perform concatenation of Y_0 and $Z_{downsample}$ to get $Y_{concat} \in \mathbb{R}^{h \times w \times (S+s)}$, and concatenate Z and $Y_{upsample}$ to get $Z_{concat} \in \mathbb{R}^{H \times W \times (s+S)}$, which can be written as

$$Y_{concat} = concat\left(Y_0, Z_{downsample}\right) \tag{3}$$

$$Z_{concat} = concat\left(Z_0, Y_{upsample}\right) \tag{4}$$

where *concat* refers to the concatenation operation within the channel dimension. This cross-modality concatenation enables the interaction of cross-modality information between the two branches.

The intricate computational convolutions of the transformer are commensurate with the sequence length, rendering it impractical to condense the input image into a sequence for transformer assimilation. To mitigate this predicament, the Vision Transformer (ViT) [32] advocates dividing the image into static-size segments. Initially, LR-HSI Y_{concat} is upsampled to generate $Y_1 \in \mathbb{R}^{H \times W \times C}$, which is of the same scale as the HR-MSI Z_1. The same downsampling operation is then performed on Y_1 and Z_1 to produce a pair of feature maps of identical scale, $Y_2 \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 2C}$ and $Z_2 \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 2C}$. The second downsampling can be represented as $Y_3 \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 4C}$ and $Z_3 \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 4C}$. Throughout the gradual downscaling procedure, numerous local spatial features are harnessed. Subsequently, Y_n and Z_n are inputted into Spectral–Spatial Transformers to further apprehend the long-range correlation in a global context.

2.2.1. Spectral–Spatial Transformer

In order to integrate data from the spectral and spatial modalities and construct a comprehensive image representation, a Spectral–Spatial Transformer (SST) was developed. This concept is visually depicted in Figure 2. In the preliminary stage, the LR-HSI and HR-MSI feature maps are subjected to a linear mapping procedure. Following this, the overall dependency of features on both modalities is modeled to facilitate the fusion of cross-modality data, utilizing a fusion attention block. This block serves to coalesce the interactive data interwoven between the two modalities. Subsequently, the Multilayer Perceptron (MLP) is structured as a bi-layer perceptron, supplemented with a hidden layer expansion ratio. To conclude, Layer Normalization (LN) is executed.

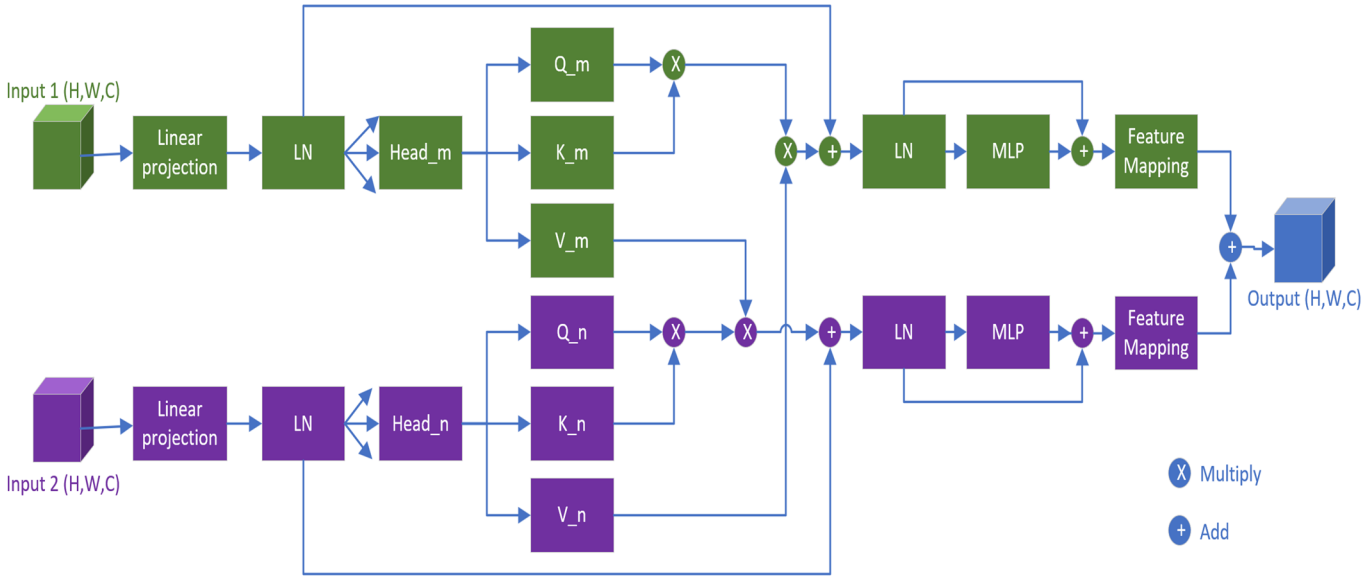

Figure 2. The schematic diagram of a multi-head Spectral–Spatial Transformer feature fusion block.

We began with two feature maps of identical dimensions, denoted as $F \in \mathbb{R}^{h \times w \times C}$. Subsequently, an *LN* was carried out on these features:

$$x_1 = LN(LP(F_1)), x_2 = LN(LP(F_2)) \tag{5}$$

where *LP* denotes the linear projection operation, and $x_1, x_2 \in \mathbb{R}^{d \times h \times w}$ signify the feature embeddings. *LN* signifies the layer normalization.

The Spectral–Spatial Fusion Self-Attention block is crafted with the intention of capturing the intricate interdependence of spectral and spatial feature maps within an image, wherein each map contains distinct semantic information at diverse scales. This design is specifically customized by calculating self-attention across spatial and spectral dimensions, thereby revealing the correlations present. Initially, we feed the feature embedding Y_n into the system to obtain the query matrix Q, the key matrix K, and the value matrix V through a trainable linear projection:

$$Q = x_n W_Q, K = x_n W_K, V = x_n W_V \tag{6}$$

where W_q, W_k, W_v are learnable projection matrices. Then, given two feature vectors x_1 and x_2, the fusion attention block can be expressed as:

$$y_1 = attention(Q_1, K_1, V_2) = softmax\left(\frac{Q_1 K_1^T}{\sqrt{d_k}}\right) V_2 \tag{7}$$

$$y_2 = attention(Q_2, K_2, V_1) = softmax\left(\frac{Q_2 K_2^T}{\sqrt{d_k}}\right) V_1 \tag{8}$$

where Q, K, and V denote the query, the key, and the value respectively, while y_1 and y_2 represent the output feature maps. This cross-attention mechanism encourages a more effective interaction between the two feature maps in the fusion task, resulting in an improved fusion outcome. Once the feature map is processed through another layer of normalization and MLP, its dimensions cannot align with the subsequent network structure. To rectify this, we employ a feature mapping module to reconfigure the output sequences into a standard 3D feature map of dimensions H × W × d. Subsequently, the channel count of the feature maps is reduced by the convolution operation, resulting in feature maps z_1 and z_2 that share the same dimensions as feature maps F_1 and F_2. The final step involves combining the two feature maps to generate the feature map F_{out}, thereby enhancing information fusion. The above process can be described as:

$$z_1 = FM(LN(y_1) + MLP(LN(y_1))), z_2 = FM(LN(y_2) + MLP(LN(y_2))) \tag{9}$$

$$F_{out} = z_1 + z_2 \tag{10}$$

where *LN* represents the layer normalization, *FM* represents the feature mapping, and *MLP* refers to the multilayer perceptron network.

2.2.2. Hierarchical Progressive Fusion

The quality of reconstructed images is profoundly influenced by the precise restoration of spatial intricacies. Furthermore, the efficacy of feature fusion utilization is pivotal for acquiring a comprehensive multi-level feature representation. Consequently, we have engineered a feature amalgamation reconstruction module to synchronously merge these three distinctive feature maps. This method incrementally recovers spatial detail intelligence via sequential elevation in sampling. Our methodology merges the spatial detail data from the reduction layer with the introductory information from the amplified layer. This initiative enhances the image's semantic features while concurrently maintaining the

spatial particulars of each band, thereby aiding the systematic reconstruction of the HR-HSI result.

$$fusion_n = \text{StripConv}(Concat(Y_n, Z_n, SST(Y_n, Z_n))) \quad (11)$$

where StripConv is the Strip Convolution Block, the *Concat* is the concatenation operation, and *SST* is the Spectral–Spatial Transformer.

The strip convolution block captures long-range context information from four different directions: horizontal, vertical, left diagonal, and right diagonal. In the strip convolution block, F is input to four different shapes of strip convolution paths after a 1×1 convolution. The output feature maps of the four paths are concatenated. Then, the upsampling operation and a 1×1 convolution are performed to obtain the final output of the strip convolution block. Let $w \in \mathbb{R}^{2k+1}$ be a strip convolution filter of size $2k + 1$, and let $D = (D_h, D_w)$ represent the direction of filter w. Let H be the result of strip convolution. The strip convolution can be defined as follows:

$$H_D[\mathrm{i}, \mathrm{j}] = (F * w)_D[\mathrm{i}, \mathrm{j}] \quad (12)$$

3. Results

3.1. Experimental Settings

The performance of the HSST network was evaluated on three openly accessible remote sensing datasets. Specifically, we employed the Pavia Center dataset, the Botswana dataset, and the Urban dataset for conducting experiments. We conducted comparative experiments between HSST and five SOTA fusion models. The five fusion algorithms were CNMF [18], MSD-CNN [33], TFNET [34], SSF-CNN [35], and MCT-NET [36]. CNMF is a matrix factorization-based method, while the other models are deep learning methods. MSD-CNN proposes a multi-scale and multi-depth CNN for remote sensing image fusion, which is based on residual learning and multiscale feature extraction. TFNet is a two-stream network that encodes spatial and spectral features independently, and then decodes the HR-HSI using the fusion of spatial and spectral features. SSF-CNN utilizes the direct concatenation of LR-HSI and HR-MSI to predict the HR-HSI, with the HR-MSI being concatenated in each convolutional layer. Lastly, MCT-NET employs a cross transformer to fuse spatial and spectral features.

The experimental procedure was carried out using LR-HSI, obtained from HR-HSI through the application of Gaussian blur and a subsequent four times downsampling. HR-MSI were generated by extracting red–green–blue bands at regular intervals that corresponded to their physical wavelengths from using specific dataset or sensors. For testing purposes, subsets with dimensions of 128×128 were extracted from the data center while the remaining components served as the training set. In each iterative cycle, the training area was randomly cropped to the size of 128×128 as part of the training process. The procedure was performed on a computational system equipped with an Intel i7-11700K 3.60 GHz CPU and an Nvidia RTX 3080Ti 12 GB GPU, with the assistance of PyTorch 2.2.2. The Adam optimization algorithm was selected for this particular experiment.

3.2. Evaluation Metrics

This paper utilizes four prominent indexes to thoroughly assess the quality of the reconstructed HR-HSI at a reduced resolution, as subsequently detailed.

To appraise the spatial quality, the Peak Signal-to-Noise Ratio (PSNR) is employed. PSNR serves as an objective evaluation metric, gauging the noise level or image distortion. A higher PSNR value indicates less distortion and superior image quality. The definition of PSNR is as follows:

$$PSNR(X, X') = 10\log_{10}\left(\frac{max(X_k)^2}{\frac{1}{HW}\|X_k - X'_k\|_2^2}\right) \quad (13)$$

where X' represents the estimated HR-HSI, X represents the ground truth HR-HSI, and X_k and X'_k denote the k_{th} band of the reference HR-HSI and the estimated HR-HSI, respectively.

We evaluated the spectral quality utilizing the Spectral Angle Mapper (SAM). SAM serves as a metric to estimate an image's spectral quality, achieved through the computation of the average spectral angle across the entire spatial domain, as defined below:

$$SAM(X, X') = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} arccos\left(\frac{X^T(i,j)X'^{(i,j)}}{\|X(i,j)\|_2\|X'^{(i,j)}\|_2}\right) \tag{14}$$

where H and W represent the number of rows and columns in the HR-HSI. The pixel vector of the reference HR-HSI is represented by $X(i,j)$, while the estimated HR-HSI at the same position (i,j) is represented by $X'(i,j)$. Spectral Angle Mapper (SAM) is a measure of spectral distortion, with a lower value indicating less distortion.

Erreur relative globale adimensionnelle de synthèse (ERGAS): The ERGAS index is specifically crafted to assess the comprehensive quality of fused images, defined as follows:

$$ERGAS(X, X') = \frac{100}{r}\sqrt{\frac{1}{S}\sum_{k=1}^{S}\frac{\|X_k - X'_k\|_2^2}{\mu^2(X_k)}} \tag{15}$$

where r signifies the downsampling ratio, and μ represents the mean value. A lower ERGAS value indicates a more favorable fusion outcome.

The Root Mean Squared Error (RMSE) is a statistical measure that represents the discrepancy between the values of X and X'. A lower value of RMSE indicates fewer reconstruction errors, thus implying superior quality of reconstruction. It is defined as follows:

$$RMSE = \sqrt{\frac{\sum_{k=1}^{S}\sum_{i=1}^{H}\sum_{j=1}^{W}(X_k(i,j) - X'_k(i,j))^2}{HWS}} \tag{16}$$

where $X_k(i,j)$ and $X'_k(i,j)$ represent the values at the position (i, j) in the kth band of the reference HR-HSI and the estimated HR-HSI.

3.3. Experimental Results on the Pavia Center Dataset

The Pavia Center dataset was obtained using drone-based ROSIS sensors. The sensor initially comprised 115 bands, which, post-processing, were reduced to 102. The dimensions of the Pavia Center dataset are 1096 × 715 pixels.

The fusion performances and quantitative evaluations are illustrated in Figure 3 and Table 1. Overall, the deep learning methods outperformed the traditional methods. The proposed HSST achieved three optimal results for PSNR, SAM, and RMSE, and a suboptimal result for ERGAS. The best values of PSNR and RMSE demonstrated the superior elementwise reconstruction quality of HSST, while the best value of SAM showed the spectral reconstruction quality of HSST. First place for ERGAS, which indicates the spatial reconstruction quality, was achieved by TFNet. TFNET has the deepest network, which gives it an advantage in extracting nonlinear deep features that are beneficial for spatial reconstruction.

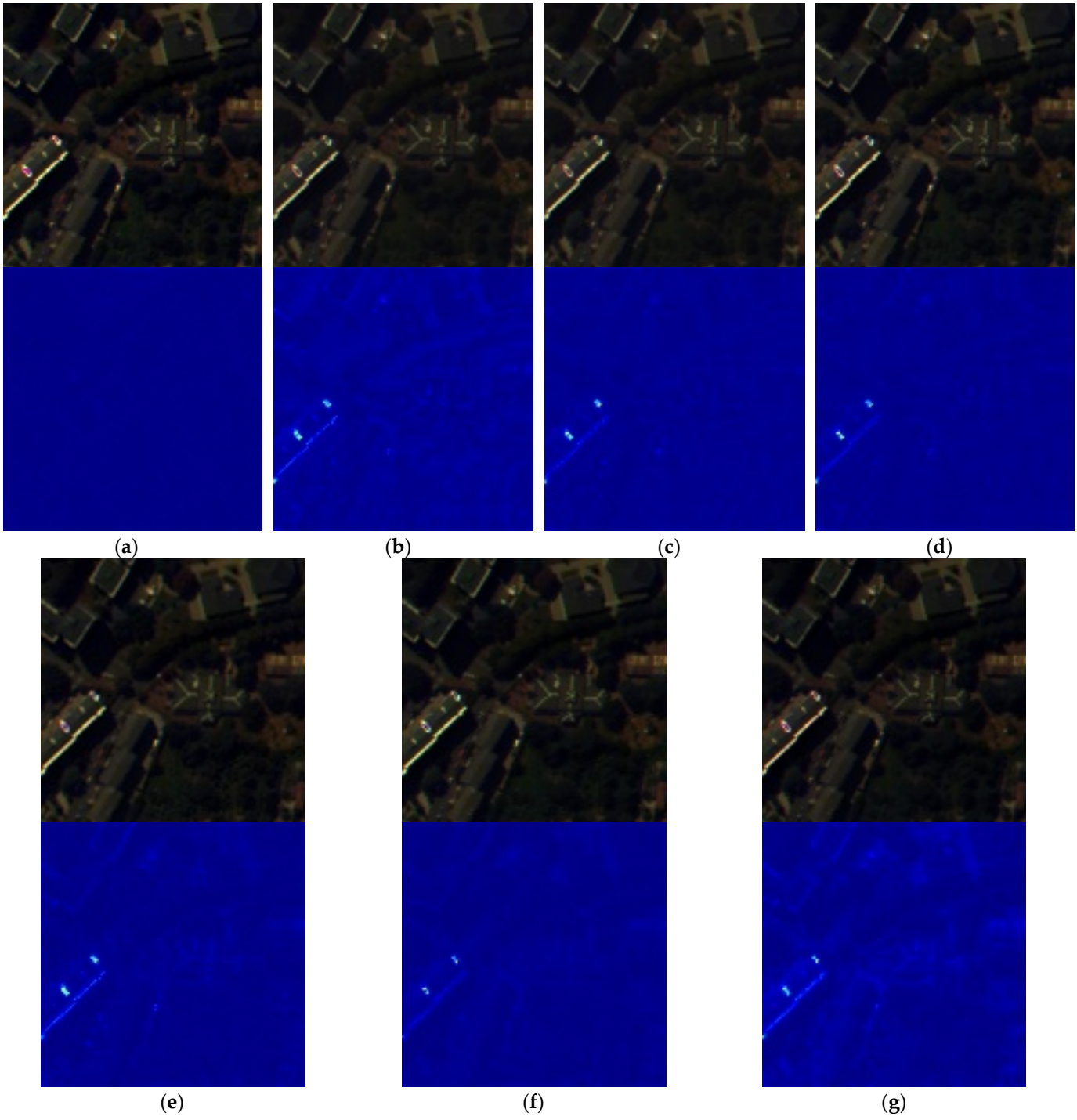

Figure 3. The fusion outcomes of various models on the Pavia Center dataset. The first row shows the R-G-B images after fusion, and the second row shows the pseudo-color processed differential images between the fused and reference images. (**a**) Original image; (**b**) CNMF; (**c**) MSD_CNN; (**d**) TFNET; (**e**) SSF-CNN; (**f**) MCT-NET; (**g**) HSST.

Table 1. Quantitative evaluations for the Pavia Center dataset. The optimal and sub-optimal values are **bolded** and underlined, respectively.

Methods	PSNR	SAM	ERGAS	RMSE
CNMF	25.2221	4.3635	11.4361	13.9777
MSD-CNN	35.7566	5.2540	4.7125	4.1563
TFNET	35.6575	4.8931	**3.7617**	4.2040
SSF-CNN	34.9898	4.7308	4.9807	4.5401
MCT-NET	36.9809	4.1504	4.1325	3.6099
HSST	**37.5183**	**4.1338**	3.8451	**3.3933**

3.4. Experimental Results on the Botswana Dataset

The Botswana dataset comprises a sequence of datasets procured by NASA satellites during the period from 2001 to 2004. The dataset encompasses a total of 145 spectral bands, subsequent to the exclusion of the uncalibrated band and the noise band, which encompasses the water absorption characteristic. The dimensions of the dataset amount to 1476 $\times$ 256 pixels.

The fusion performances and quantitative evaluations are depicted in Figure 4 and Table 2. Our HSST experimental results are in the second tier, trailing slightly behind MCT-NET. In the Botswana dataset, the pixel's spatial resolution reaches up to 30 m, making the spatial information more intricate compared to other datasets that have a higher requirement for feature extraction. The smaller size of HSST, in contrast to MCT-NET, constrains the performance of our model in extracting spatial features from a large receptive field.

Table 2. Quantitative evaluations for the Botswana dataset. The optimal and sub-optimal values are **bolded** and underlined, respectively.

Methods	PSNR	SAM	ERGAS	RMSE
CNMF	26.3457	2.4866	9.4849	26.3457
MSD-CNN	35.7160	2.7977	3.2249	0.5964
TFNET	36.5435	2.4479	2.9630	0.5422
SSF-CNN	30.0626	5.1641	16.2764	1.1434
MCT-NET	**37.8955**	**2.1803**	2.6303	**0.4640**
HSST	37.1824	2.2274	**2.4898**	0.5037

3.5. Experimental Results on the Urban Dataset

The Urban dataset, acquired in 1995 using a drone-mounted HYDICE sensor, was centered on Copper Tree Bay, located in the state of Texas, USA. This dataset, characterized by image dimensions of 307 $\times$ 307, originally consisted of 210 spectral bands. However, after eliminating the bands affected by noise and water absorption, only 162 bands were preserved for further processing and analysis.

The fusion performances and quantitative evaluations are presented in Figure 5 and Table 3. It is evident that the proposed HSST outperforms all other comparative methods in terms of all four evaluation metrics. The superiority of the proposed HSST can be observed from the perspective of elementwise reconstruction quality (RMSE and PSNR), spectral reconstruction quality (SAM), and spatial reconstruction quality (ERGAS).

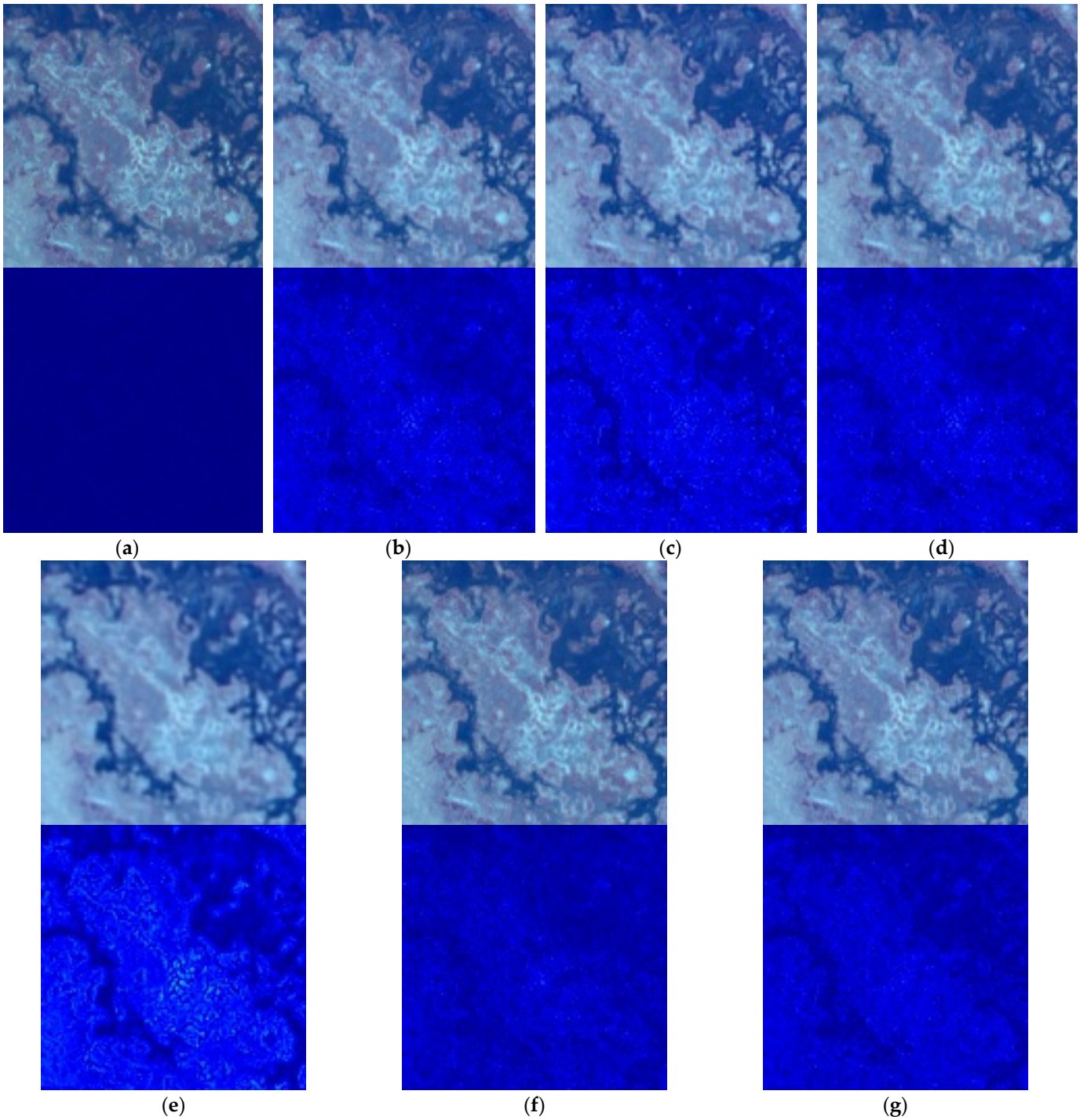

Figure 4. The fusion outcomes of various models on the Botswana dataset. The first row shows the R-G-B images after fusion, and the second row shows the pseudo-color processed differential images between the fused and reference images. (**a**) Original image; (**b**) CNMF; (**c**) MSD_CNN; (**d**) TFNET; (**e**) SSF-CNN; (**f**) MCT-NET; (**g**) HSST.

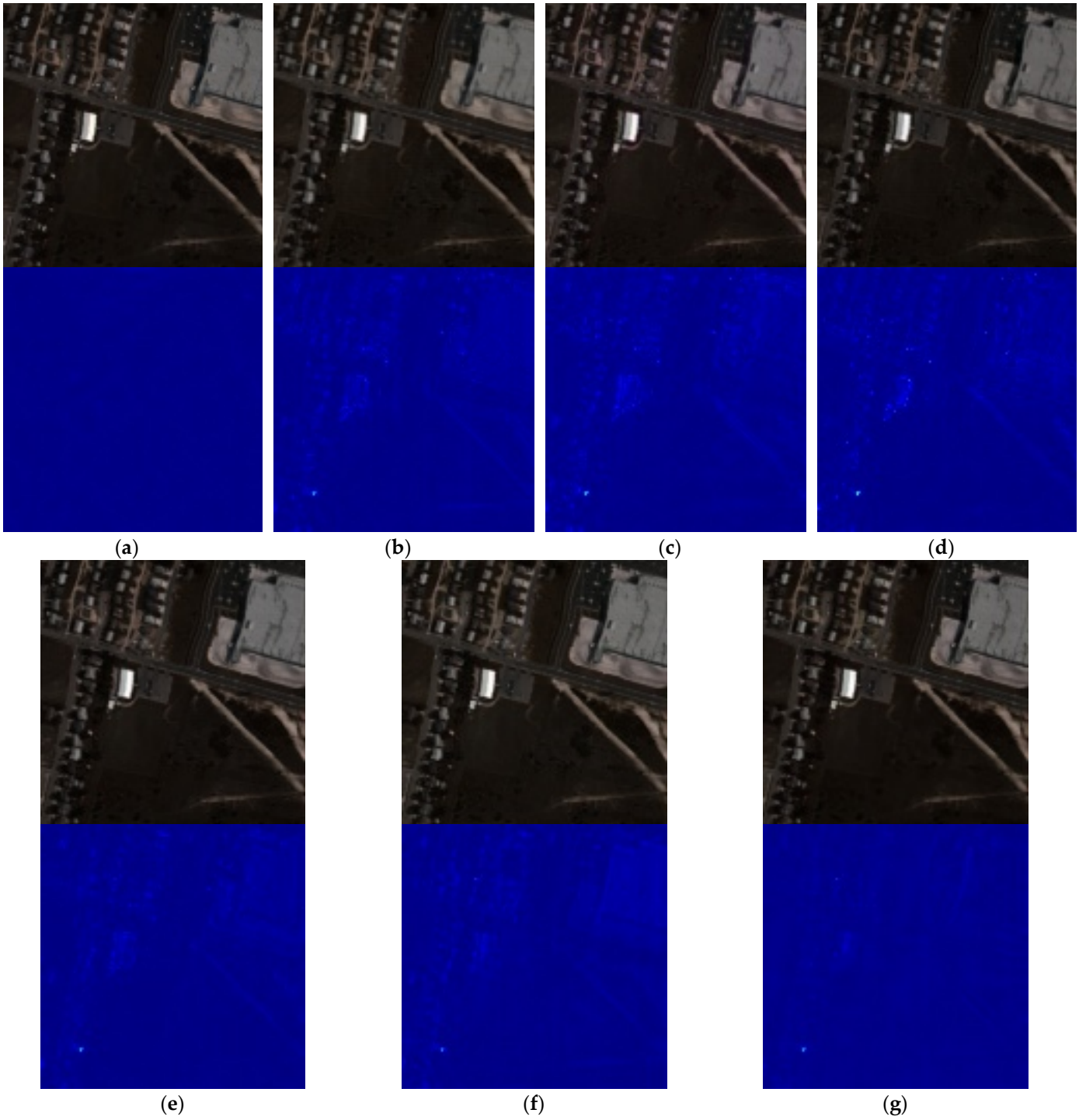

Figure 5. The fusion outcomes of various models on the Urban dataset. The first row shows the R-G-B images after fusion, and the second row shows the pseudo-color processed differential images between the fused and reference images. (**a**) Original image; (**b**) CNMF; (**c**) MSD_CNN; (**d**) TFNET; (**e**) SSF-CNN; (**f**) MCT-NET; (**g**) HSST.

Table 3. Quantitative evaluations for the Urban dataset. The optimal and sub-optimal values are **bolded** and underlined, respectively.

Methods	PSNR	SAM	ERGAS	RMSE
CNMF	36.0468	2.4971	1.5612	4.0198
MSD-CNN	35.7440	3.2117	1.8428	3.2133
TFNET	35.9584	3.0255	1.7885	3.1350
SSF-CNN	37.3912	2.5904	1.4638	2.6582
MCT-NET	37.3294	2.7076	1.4312	2.6772
HSST	**37.6249**	**2.4644**	**1.3954**	**2.5877**

4. Discussion

The Pavia Center dataset, obtained via drone-based ROSIS sensors, presents a challenging platform for evaluating the effectiveness of various fusion algorithms. This dataset is particularly beneficial for urban analysis, wherein detailed spatial information is crucial for applications like infrastructure monitoring, land use classification, and urban planning. Our proposed HSST model showcased optimal results in three out of the four evaluation metrics, namely PSNR, RMSE, and SAM, against this dataset. A close inspection of the fusion outcomes in Figure 3 suggests that deep learning-based approaches, including HSST, are superior at preserving spectral and spatial details compared to traditional methods like CNMF. The high PSNR and low RMSE values achieved by HSST indicate a high degree of fidelity in the reconstructed images, which is essential for accurately identifying and monitoring minor urban features like road networks, building footprints, and vegetation patches. The best SAM score of HSST underlines its capability in spectral reconstruction, preserving the spectral signatures necessary for distinguishing between different land cover types and materials in remote sensing applications. TFNet, another deep learning-based method, achieved the best ERGAS score, highlighting the advantage of deeper networks in extracting spatial features that maintain spatial resolution and contextual information in the fused image. However, despite TFNet's deeper architecture, HSST outperformed it in other metrics, suggesting that HSST strikes a balance between network depth and efficiency in feature extraction and reconstruction.

The Botswana dataset, taken from NASA satellite imagery, posed a unique challenge due to its high spatial resolution, which increases the complexity of spatial information. This complexity requires a model capable of effectively extracting features from a larger receptive field. In this context, MCT-NET outperformed HSST, which may be due to its larger size and enhanced capacity for spatial feature extraction. Despite having a smaller model size, HSST's performance was commendable, achieving sub-optimal results across all metrics. This suggests that HSST offers computational efficiency without significant compromises in fusion quality. However, the results imply that for datasets with higher spatial resolution, like the Botswana dataset, the model's complexity might need to be increased to improve feature extraction capabilities. The Botswana dataset is rich in spectral information, which is pivotal for environmental monitoring and ecosystem health assessment. HSST's strength in spectral reconstruction, as evidenced by its SAM score, is particularly relevant for these applications. The accurate spectral information provided by HSST can facilitate more precise monitoring of vegetation health, which is essential for understanding the effects of climate change, identifying drought stress, and assessing the proliferation of diseases among plant populations. The slightly sub-optimal spatial reconstruction performance of HSST, likely due to its smaller size, suggests that the model could potentially be enhanced for satellite-based datasets with high spatial resolution.

The Urban dataset presented a scenario that played to the strengths of HSST, where it excelled over all other methods across all evaluation metrics, including RMSE, PSNR, SAM, and ERGAS. The dataset's characteristics, featuring smaller dimensions and a significant number of spectral bands, appear to be ideally matched to HSST's capabilities. HSST's consistently optimal performance on the Urban dataset highlights its comprehensive skill in addressing the multifaceted challenges of image fusion, including elementwise, spectral,

and spatial reconstruction. This robust and versatile performance indicates that HSST is well-suited for a broad spectrum of hyperspectral image fusion tasks. With a focus on industrial and urban land cover, the Urban dataset greatly benefits from HSST's proficiency in capturing both spatial and spectral details. In applications such as pollution monitoring, HSST's superior reconstruction quality is instrumental in pinpointing pollution hotspots and tracking the spread of pollutants over time. For urban planning purposes, the detailed spatial and spectral information provided by HSST is invaluable for informed decision-making regarding land use, infrastructure development, and environmental impact assessments. The outstanding performance of HSST on the Urban dataset underscores its potential role in managing the intricate and dynamic interactions within urban environments, where the relationship between human activities and the environment is particularly complex.

The experimental outcomes across the Pavia Center, Botswana, and Urban datasets strongly support the efficacy and robustness of the proposed HSST model for hyperspectral image fusion. HSST's capability to maintain a balance between elementwise, spectral, and spatial reconstruction qualities positions it as a promising tool for applications in remote sensing and other hyperspectral imagery-dependent fields. The depth of networks like TFNet may enhance spatial reconstruction, but HSST's overall balance and efficiency suggest that increased network depth does not always equate to optimal performance. The results also underscore the significance of model complexity in relation to dataset characteristics such as spatial resolution and the number of spectral bands.

4.1. Ablation Studies

In order to explore the role of the components of HSST in HSI and MSI fusion, some ablation experiments were conducted on the Urban dataset based on the three model variations of a model with only a spectral transformer, a model with only a spatial transformer, and a model without progressive fusion.

The fusion performances and quantitative evaluations of the ablation experiments are presented in Figure 6 and Table 4. When comparing the model with only the spectral transformer and the model with only the spatial transformer, it can be observed that the first model, which reconstructs the spectral information, performed significantly better than the second model, which reconstructs the spatial information. These results suggest that the reconstruction of spectral information is easier compared to spatial information, indicating that spatial information is more complex. Ablation studies on progressive fusion demonstrate that the use of progressive fusion significantly improves the accurate reconstruction of spatial details, thereby affecting the quality of the final reconstructed images.

Table 4. Quantitative evaluations for the ablation studies on the Urban dataset. The optimal values are **bolded**.

Methods	PSNR	SAM	ERGAS	RMSE
Spectral Transformer only	34.3125	3.4971	2.9380	3.7891
Spatial Transformer only	21.2898	8.0370	9.4513	16.9695
Without progressive fusion	37.3294	2.7076	1.4312	2.6772
HSST	**37.6249**	**2.4644**	**1.3954**	**2.5877**

4.2. Classification Performance Studies

To further evaluate the advantages of HSI-MSI image fusion, we conducted additional classification experiments on the Pavia Center dataset using unfused LR-HSI and fused HR-HSI. The Pavia Center dataset is a high-resolution city scene located in Pavia, northern Italy. For this dataset, we utilized the Self-Adaptive 3D ASPP Multi-Scale Feature Fusion Network (SAAFN) [37] for HSI classification. For all experiments, only 10% of labeled samples were randomly selected for training, with the rest used for testing. Both qualitative maps and quantitative evaluations comprehensively analyzed performance using four common metrics: producer's accuracy, overall accuracy, average accuracy, and Kappa.

The classification maps and the corresponding quantitative evaluations for the classification experiments using SAAFN conducted at the Pavia Center are presented in Figure 7 and Table 5, respectively. The optimal value for each line is highlighted in bold, while the sub-optimal value is underlined.

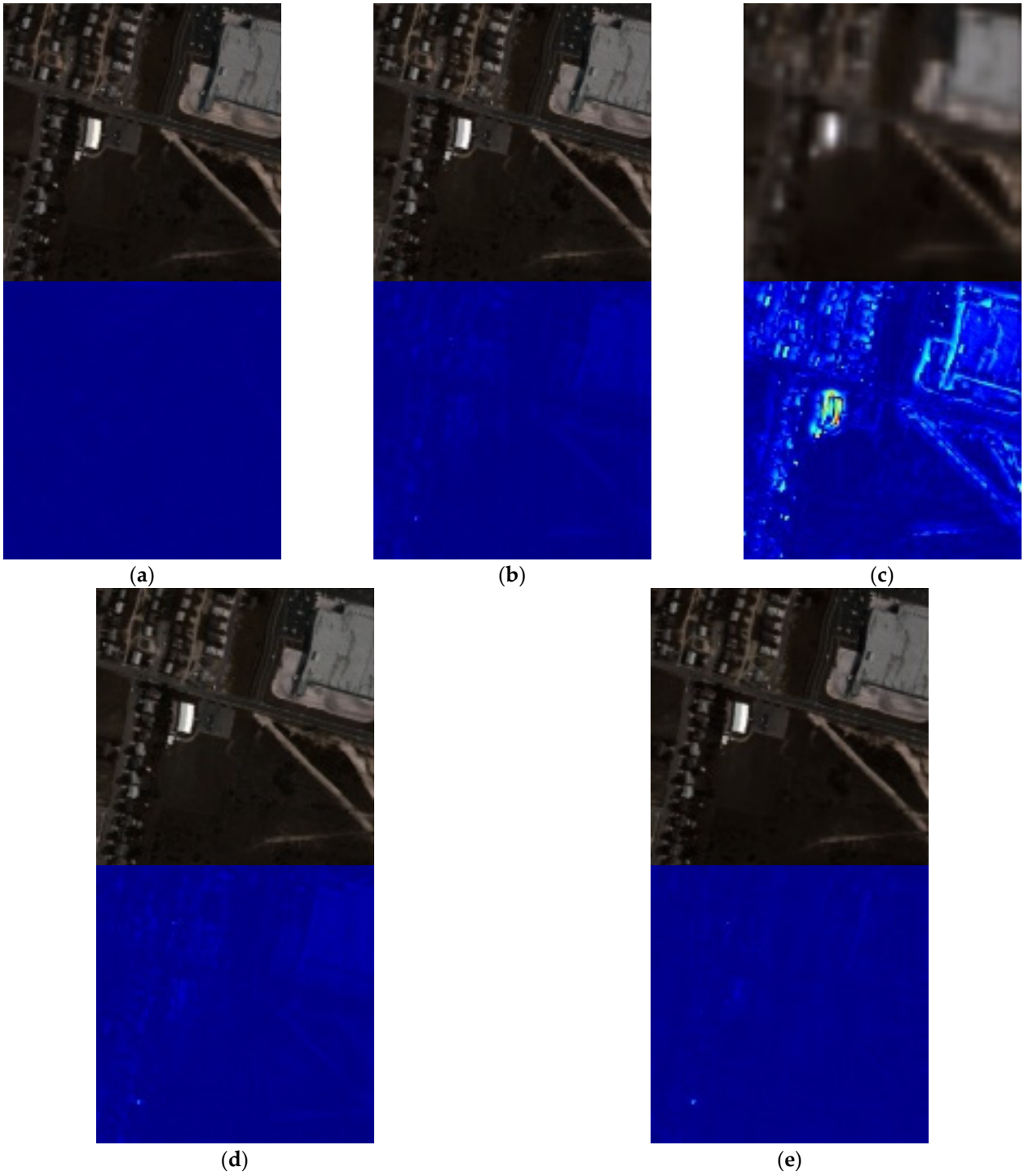

Figure 6. The fusion outcomes of ablation studies on the Urban dataset. The first row shows the R–G–B images after fusion, and the second row shows the pseudo-color processed differential images between the fused and reference images. (**a**) Original image; (**b**) Spectral Transformer only; (**c**) Spatial Transformer only; (**d**) Without progressive fusion; (**e**) HSST.

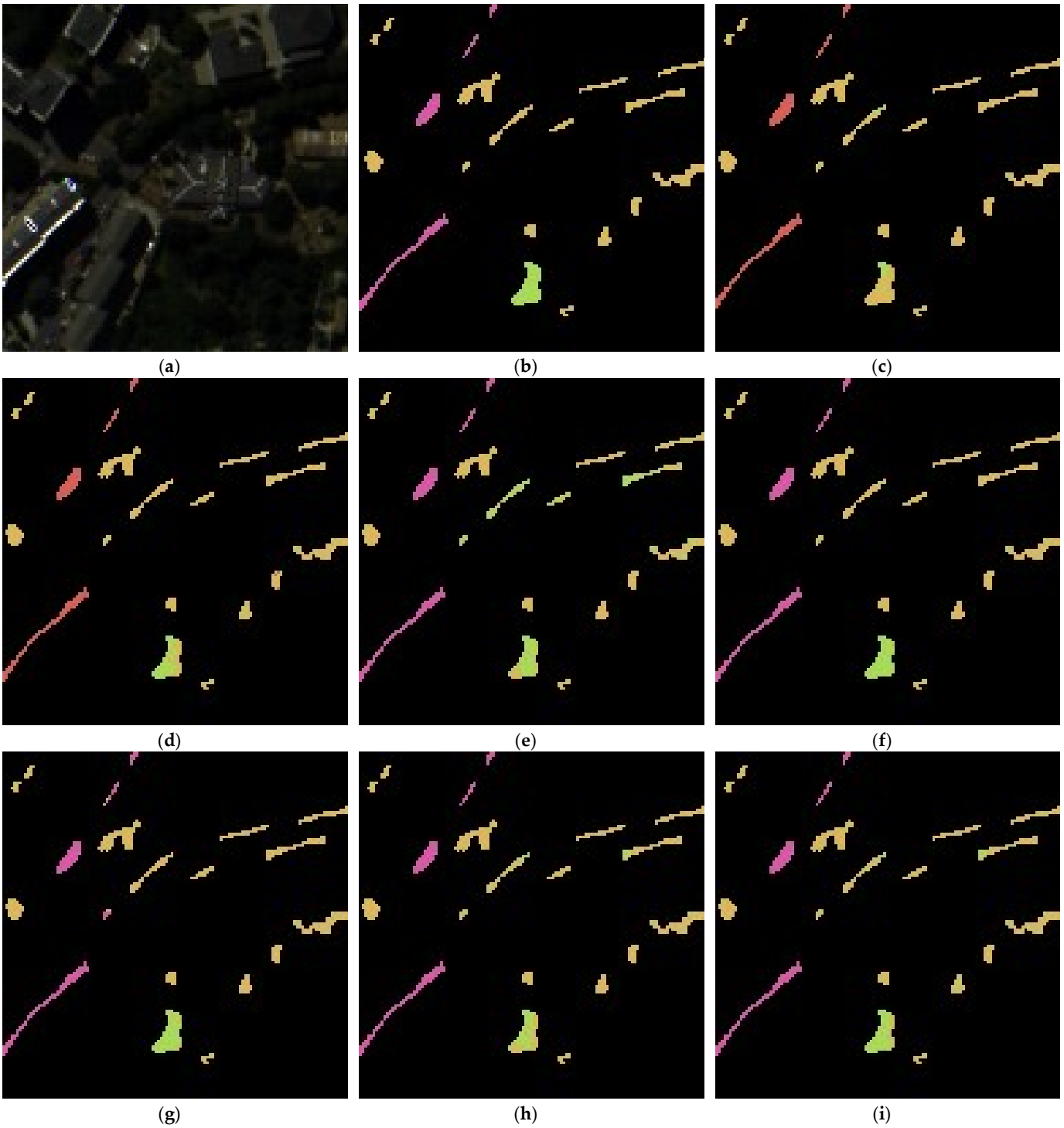

Figure 7. The results for the Pavia Center classification experiment: (**a**) Original image; (**b**) Ground truth; (**c**) LR-HSI; (**d**) CNMF; (**e**) MSD-CNN; (**f**)TFNET; (**g**) SSF-CNN; (**h**) MCT-NET; (**i**) HSST.

Table 5. Quantitative evaluations for Pavia Center classification experiment (%). The optimal and sub-optimal values are **bolded** and underlined, respectively.

	Unfused LR-HSI	Fused HR-HSI					
		CNMF	MSD-CNN	TFNET	SSF-CNN	MCT-NET	HSST
OA	81.93	82.54	94.37	96.68	94.25	96.11	**96.74**
AA	48.26	71.15	68.95	83.25	78.07	**91.40**	86.72
Kappa	0.73	0.76	0.92	**0.95**	0.92	0.94	**0.95**

The HSST method achieved the highest overall accuracy (OA) with a score of 96.74%, closely followed by the TFNET method at 96.68%. It is worth noting that the fused HR-HSI methods outperformed the unfused LR-HSI, which only achieved an OA of 81.93%. Both the TFNET and HSST methods achieved the highest Kappa score of 0.95. On the other hand, the CNMF method performed the worst among the fused HR-HSI methods, with an OA of 82.54%. Overall, the fused HR-HSI methods, particularly TFNET and HSST, demonstrated higher classification accuracy compared to the unfused LR-HSI, as evidenced by their higher OA, AA, and Kappa scores.

5. Conclusions

In this article, we introduced the innovative Hierarchical Spectral–Spatial Transformer (HSST) network, a technique particularly well-suited for enhancing the capabilities of drone-based imaging systems. The HSST network comprises two branches that harness the self-attention mechanism inherent in transformers, enabling the extraction and integration of spectral details from hyperspectral images (HSI) with the spatial details from multispectral images (MSI). To leverage the rich spatial information inherent in remote sensing imagery, including that captured by drones, we incorporated a hierarchical structure that captures multi-scale information. Furthermore, our hierarchical progressive fusion strategy is designed for the reconstruction of high-resolution hyperspectral images (HR-HSI), progressively restoring spatial detail through upsampling and effectively utilizing the multi-level feature representation. Comparative experiments of the proposed HSST and five state-of-the-art methods were conducted on three widely used remote sensing hyperspectral datasets, including the Pavia Center, the Botswana, and the Urban. The superior experimental results of HSST demonstrate the effectiveness of the proposed method, and can be a competitive method for practical applications.

Author Contributions: Conceptualization, T.Z.; methodology, T.Z. and L.Z.; software, T.Z.; writing—original draft preparation, T.Z.; writing—review and editing, Q.L. and L.Z.; supervision, Q.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original data presented in the study are openly available at http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes (accessed on 10 December 2023).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhuang, L.; Ng, M.K.; Fu, X.; Bioucas-Dias, J.M. Hy-Demosaicing: Hyperspectral Blind Reconstruction from Spectral Sub-Sampling. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
2. Bian, J.; Li, A.; Zhang, Z.; Zhao, W.; Lei, G.; Yin, G.; Jin, H.; Tan, J.; Huang, C. Monitoring Fractional Green Vegetation Cover Dynamics over a Seasonally Inundated Alpine Wetland Using Dense Time Series HJ-1A/B Constellation Images and an Adaptive Endmember Selection LSMM Model. *Remote Sens. Environ.* **2017**, *197*, 98–114. [CrossRef]
3. Jia, S.; Shen, L.; Zhu, J.; Li, Q. A 3-D Gabor Phase-Based Coding and Matching Framework for Hyperspectral Imagery Classification. *IEEE Trans. Cybern.* **2018**, *48*, 1176–1188. [CrossRef]
4. Zhao, J.; Zhong, Y.; Hu, X.; Wei, L.; Zhang, L. A Robust Spectral-Spatial Approach to Identifying Heterogeneous Crops Using Remote Sensing Imagery with High Spectral and Spatial Resolutions. *Remote Sens. Environ.* **2020**, *239*, 111605. [CrossRef]

5. Fu, X.; Jia, S.; Zhuang, L.; Xu, M.; Zhou, J.; Li, Q. Hyperspectral Anomaly Detection via Deep Plug-and-Play Denoising CNN Regularization. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 9553–9568. [CrossRef]
6. Zhuang, L.; Fu, X.; Ng, M.K.; Bioucas-Dias, J.M. Hyperspectral Image Denoising Based on Global and Nonlocal Low-Rank Factorizations. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 10438–10454. [CrossRef]
7. Ghassemian, H. A Review of Remote Sensing Image Fusion Methods. *Inf. Fusion* **2016**, *32*, 75–89. [CrossRef]
8. Wei, Q.; Dobigeon, N.; Tourneret, J.-Y. Fast Fusion of Multi-Band Images Based on Solving a Sylvester Equation. *IEEE Trans. Image Process.* **2015**, *24*, 4109–4121. [CrossRef] [PubMed]
9. Dian, R.; Li, S.; Fang, L.; Wei, Q. Multispectral and Hyperspectral Image Fusion with Spectral-Spatial Sparse Representation. *Inf. Fusion* **2019**, *49*, 262–270. [CrossRef]
10. Fu, X.; Jia, S.; Xu, M.; Zhou, J.; Li, Q. Fusion of Hyperspectral and Multispectral Images Accounting for Localized Inter-Image Changes. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–18. [CrossRef]
11. Li, S.; Dian, R.; Fang, L.; Bioucas-Dias, J.M. Fusing Hyperspectral and Multispectral Images via Coupled Sparse Tensor Factorization. *IEEE Trans. Image Process.* **2018**, *27*, 4118–4130. [CrossRef] [PubMed]
12. Dian, R.; Fang, L.; Li, S. Hyperspectral Image Super-Resolution via Non-Local Sparse Tensor Factorization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5344–5353.
13. Yang, Q.; Xu, Y.; Wu, Z.; Wei, Z. Hyperspectral and Multispectral Image Fusion Based on Deep Attention Network. In Proceedings of the 2019 10th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS), Amsterdam, The Netherlands, 24–26 September 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–5.
14. Cai, J.; Huang, B. Super-Resolution-Guided Progressive Pansharpening Based on a Deep Convolutional Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 5206–5220. [CrossRef]
15. Wang, X.; Wang, X.; Zhao, K.; Zhao, X.; Song, C. Fsl-Unet: Full-Scale Linked Unet with Spatial–Spectral Joint Perceptual Attention for Hyperspectral and Multispectral Image Fusion. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
16. Dong, M.; Li, W.; Liang, X.; Zhang, X. MDCNN: Multispectral Pansharpening Based on a Multiscale Dilated Convolutional Neural Network. *J. Appl. Remote Sens.* **2021**, *15*, 036516. [CrossRef]
17. Benzenati, T.; Kessentini, Y.; Kallel, A. Pansharpening Approach via Two-Stream Detail Injection Based on Relativistic Generative Adversarial Networks. *Expert Syst. Appl.* **2022**, *188*, 115996. [CrossRef]
18. Yokoya, N.; Yairi, T.; Iwasaki, A. Coupled Nonnegative Matrix Factorization Unmixing for Hyperspectral and Multispectral Data Fusion. *IEEE Trans. Geosci. Remote Sens.* **2011**, *50*, 528–537. [CrossRef]
19. Dian, R.; Li, S.; Fang, L.; Lu, T.; Bioucas-Dias, J.M. Nonlocal Sparse Tensor Factorization for Semiblind Hyperspectral and Multispectral Image Fusion. *IEEE Trans. Cybern.* **2020**, *50*, 4469–4480. [CrossRef]
20. Kanatsoulis, C.I.; Fu, X.; Sidiropoulos, N.D.; Ma, W.-K. Hyperspectral Superresolution: A Coupled Tensor Factorization Approach. *IEEE Trans. Signal Process.* **2018**, *66*, 6503–6517. [CrossRef]
21. Grohnfeldt, C.; Zhu, X.X.; Bamler, R. Jointly Sparse Fusion of Hyperspectral and Multispectral Imagery. In Proceedings of the 2013 IEEE International Geoscience and Remote Sensing Symposium-IGARSS, Melbourne, Australia, 21–26 July 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 4090–4093.
22. Xie, Q.; Zhou, M.; Zhao, Q.; Meng, D.; Zuo, W.; Xu, Z. Multispectral and Hyperspectral Image Fusion by MS/HS Fusion Net. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 1585–1594.
23. Dian, R.; Li, S.; Guo, A.; Fang, L. Deep Hyperspectral Image Sharpening. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *29*, 5345–5355. [CrossRef]
24. Palsson, F.; Sveinsson, J.R.; Ulfarsson, M.O. Multispectral and Hyperspectral Image Fusion Using a 3-D-Convolutional Neural Network. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 639–643. [CrossRef]
25. Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52. [CrossRef]
26. Zheng, Y.; Li, J.; Li, Y.; Guo, J.; Wu, X.; Shi, Y.; Chanussot, J. Edge-Conditioned Feature Transform Network for Hyperspectral and Multispectral Image Fusion. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
27. Hong, D.; Han, Z.; Yao, J.; Gao, L.; Zhang, B.; Plaza, A.; Chanussot, J. SpectralFormer: Rethinking Hyperspectral Image Classification with Transformers. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–15. [CrossRef]
28. He, X.; Chen, Y.; Lin, Z. Spectral-Spatial Transformer for Hyperspectral Image Classification. *Remote Sens.* **2021**, *13*, 498. [CrossRef]
29. Selen, A.; Esra, T.-G. SpectralSWIN: A Spectral-Swin Transformer Network for Hyperspectral Image Classification. *Int. J. Remote Sens.* **2022**, *43*, 4025–4044.
30. Cai, Y.; Lin, J.; Hu, X.; Wang, H.; Yuan, X.; Zhang, Y.; Timofte, R.; Van Gool, L. Maskguided Spectral-Wise Transformer for Efficient Hyperspectral Image Reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 17502–17511.
31. Bandara, W.G.C.; Patel, V.M. HyperTransformer: A Textural and Spectral Feature Fusion Transformer for Pansharpening. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1767–1777.
32. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.

33. Yuan, Q.; Wei, Y.; Meng, X.; Shen, H.; Zhang, L. A Multiscale and Multidepth Convolutional Neural Network for Remote Sensing Imagery Pan-Sharpening. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 978–989. [CrossRef]
34. Liu, X.; Liu, Q.; Wang, Y. Remote Sensing Image Fusion Based on Two-Stream Fusion Network. *Inf. Fusion* **2020**, *55*, 1–15. [CrossRef]
35. Han, X.H.; Shi, B.; Zheng, Y. SSF-CNN: Spatial and Spectral Fusion with CNN for Hyperspectral Image Super-Resolution. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 2506–2510.
36. Wang, X.; Wang, X.; Song, R.; Zhao, X.; Zhao, K. MCT-Net: Multi-Hierarchical Cross Transformer for Hyperspectral and Multispectral Image Fusion. *Knowl. Based Syst.* **2023**, *264*, 108630. [CrossRef]
37. Zhu, T.; Liu, Q.; Zhang, L. An Adaptive Atrous Spatial Pyramid Pooling Network for Hyperspectral Classification. *Electronics* **2023**, *12*, 5013. [CrossRef]