APSIPA Transactions on Signal and Information Processing, 2024, 13, e204

Original Paper

ESAFormer: Multi-resolution Fusion Network for Pansharpening

Xiangzeng Liu[1*], Rutao Li[2], Ziyao Wang[1], Ronghan Li[1], Qi Cheng[3] and Qiguang Miao[1]

[1]*School of Computer Science and Technology, Xi'dian University, Xi'an, 710071, China*

[2]*Guangzhou Institute of Technology, Xi'dian University, Guangzhou, 510555, China*

[3]*School of Mechanical and Electrical Engineering, China University of Mining and Technology, Beijing, 100083, China*

ABSTRACT

The pansharpening task is to fuse low-resolution multispectral (LRMS) images and high-resolution panchromatic (PAN) images to generate high-resolution multispectral images. Most of the existing methods do not preserve spatial and spectral details well, which is due to ignoring the difference in resolution between the two images. To address this issue, we propose a novel fusion network (ESAFormer) that effectively enhances the spatial and spectral information representation. In the proposed model, a hybrid multi-resolution structure of CNN and Transformer is deployed to allow the features of LRMS images and PAN images to fuse progressively. Subsequently, the enhanced spatial attention module is adopted to preserve spatial details and long-range information. Extensive experimental results indicate that the proposed method is superior to existing SOTA methods on World-View2 and IKONOS datasets.

Keywords: Pansharpening, feature integration, transformer, spatial attention

*Corresponding author: Xiangzeng Liu, xzliu@xidian.edu.cn.

Received 30 October 2023; Revised 21 December 2023
ISSN 2161-1823; DOI 10.1561/116.00000174

1 Introduction

With the development of satellite image sensors, the availability of remote-sensing images has increased in recent years. However, due to the technological limitations of existing sensors, images acquired by current remote sensing sensors have to make a trade-off between spectral and spatial resolution [20]. To meet the necessary signal-to-noise ratio (SNR), multispectral (MS) images with four or eight bands typically have a low spatial resolution, while PAN images tend to have a high spatial resolution but only consist of one band, as shown in Figure 1. However, a large number of high-resolution multispectral (HRMS) images are required for the interpretation of observed scenes. To create HRMS images, it is suggested to fuse low-resolution multispectral (LRMS) and PAN images using pansharpening [22].

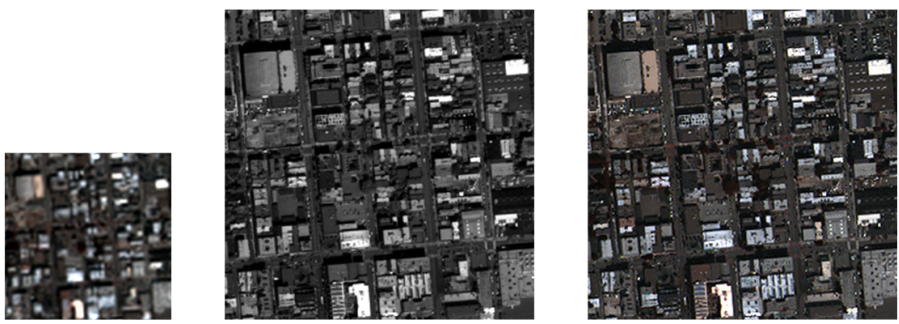

Figure 1: From left to right: MS (Multispectral) image, PAN (Panchromatic) image, and fused image. Due to sensor limitations, MS image contains richer spectral information but lower spatial resolution, while PAN image offers higher spatial resolution with a single spectral band. The fused image retains both spatial and spectral characteristics.

Numerous pansharpening techniques have been proposed in the past few decades, with promising outcomes. Existing pansharpening can be divided into traditional methods and deep learning-based methods. The traditional methods include component substitution, multi-resolution, and model-based methods [20]. Component substitution and multi-resolution methods obtain fusion image from the perspective of detail injection [26]. The former method is to completely replace one component of the MS with a PAN image, such as the BDSD-PC [11]. The fusion results of the component substitution method have a good effect on the preservation of spatial details. However, due to the difficulty in separating the spectral and spatial information in the multi-spectral image, the spectral distortion of those results is often large. The latter method adopts the spatial details extracted from PAN images to integrate into MS images to obtain high-resolution multi-spectral images, such as the generalized Laplacian Pyramid (GLP) [1]. Multiresolution-based methods

often achieve good results in spectral information retention. However, they model the spatial details of the source image and inject them into multispectral images, which tends to result in additional spatial artifacts in the fusion results.

Model-based methods focus on building feature representation models with appropriate regularizers to solve the fusion problem of multi-spectral and panchromatic images, then designing a module for the given model to reconstruct the high-resolution results [26]. The whole process can be regarded as the reconstruction of incomplete complementary observations of multi-channel data from a mathematical point of view. In general, it mainly includes Bayesian, sparse reconstruction (SR), and model-based optimization (MBO) techniques. For example, the P+XS method obtains the spectral information of a fused image by assuming that the PAN image can be approximated as a linear combination of high-resolution multispectral bands [5]. However, model-based methods often require hyperparameter adjustment, and the calculation burden is heavy.

Since these traditional methods rely on manually created features [6][15], their capacity to fuse images to preserve features is severely constrained. Due to CNN-based methods having strong feature representation ability, many CNN-based pansharpening techniques are superior to traditional methods [7] [14]. In the early stage, a three-layer convolutional neural network (CNN) was designed to process the pansharpening task of multi-spectral images, and better fusion results were obtained compared with the traditional existing methods [19]. After that, a deep CNN structure was developed to fuse the PAN and LRMS images by high-pass filtering technology [30]. Deng *et al.* [7] designed Fusion-Net with the idea of component substitution and multi-resolution. Recently, TF-Net [16] employed the deep convolution layer of residual connection to process the feature cascade, and performed image fusion in the feature domain. Zhang *et al.* designed TD-Net utilizing bi-directional information flow and multi-layer convolution processing to gradually generate fused images [32]. Nevertheless, CNN-based techniques still have certain drawbacks, such as the absence of long-range information modeling.

To address the lack of long-range dependency in CNN methods, Transformer-based pansharpening methods were developed and have achieved superior results in context presentation modeling [13, 26]. Dosovitskiy *et al.* firstly proposed a ViT-based approach for pansharpening with excellent results [9]. Subsequently, Meng *et al.* cut the source image into patches and used Transformer to learn the long-range dependency between these patches [21]. After that, Zhou *et al.* utilized reversible neural modules to represent and fuse features [35]. Zhang *et al.* [8] used the spatial and spectral attention mechanism to extract feature maps and used the graph attention mechanism to learn the similarity between feature maps. Although the global attention mechanism of Transformer allows the model to focus on feature dependencies over long distances, its ability is insufficient to represent local features that are very important in remote sensing image tasks.

In a word, the lack of synergy between long-range and short-range features at different scales in the above methods led to inadequate feature extraction and unsatisfactory pansharpening results. Meanwhile, the upsampled LRMS and PAN images are crudely layered in several deep-learning-based pansharpening approaches, while ignoring the resolution difference of the source images. As a result, the spatial features of the fused image are not significant enough.

To address the above issues, and inspired by recent work on image super-resolution [10, 23], we propose a novel multi-resolution fusion network for pansharpening, named ESAFormer. The overall architecture of the network uses a bi-directional information flow of LRMS and PAN images, where the spatial and spectral features of MS and PAN images are integrated progressively. To preserve as much spatial detail as possible, we introduce the enhanced spatial attention mechanism that uses as few parameters as possible to achieve better spatial detail retention. In the MS branch, we merge CNN with Swin-Transformer's attention mechanism to enable the model to combine their strengths in modeling both local and global information. The contribution of our work can be summarised as follows:

1) We propose a novel multi-resolution fusion network, which adopts a bi-directional structure to fully utilize the multi-resolution information of PAN image and spectral information of MS image.

2) We apply the enhanced spatial attention mechanism and Swin-Transformer to train the model, which enables ESAFormer to obtain richer short and long-range features at different scales.

3) Experiments on multiple datasets show that ESAFormer outperforms SOTA methods in both visual and quantitative comparisons. Ablation studies further demonstrate the effectiveness of the proposed method.

2 Multi-resolution Fusion Network for Pansharpening

In this section, we present the proposed method. Firstly, we give the problem formulation of the pansharpening task. Subsequently, the overall structure of the proposed approach is introduced. After that, we describe the implementation details of ESAFormer.

2.1 Problem Formulation

For convenience, the notation used throughout this paper is presented first. Let $P \in \mathbb{R}^{H \times W \times 1}$ denotes a high-resolution PAN image with the spatial size of $H \times W$ and $ms \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times B}$ denotes a low-resolution multi-spectral image with B spectral bands and spatial size $\frac{H}{4}$. To preserve the spectral information of MS images, the upsampled MS images $MS \in \mathbb{R}^{H \times W \times B}$ are added to the final output. The whole process can be formulated as:

$$Fused = MS + H_{net}(P, ms, \theta), \tag{1}$$

where $Fused \in \mathbb{R}^{H \times W \times B}$ is the fused result, H_{net} denotes the whole network, and θ denotes the parameters.

2.2 Bi-directional Flow Network Structure

The overall network architecture is a bi-directional information flow structure including MS and PAN branches as shown in Figure 2. The overall process is to extract the spatial information of PAN images and inject it into MS images progressively. The fused image is generated by integrating multi-spectral and spatial information from MS and PAN images gradually.

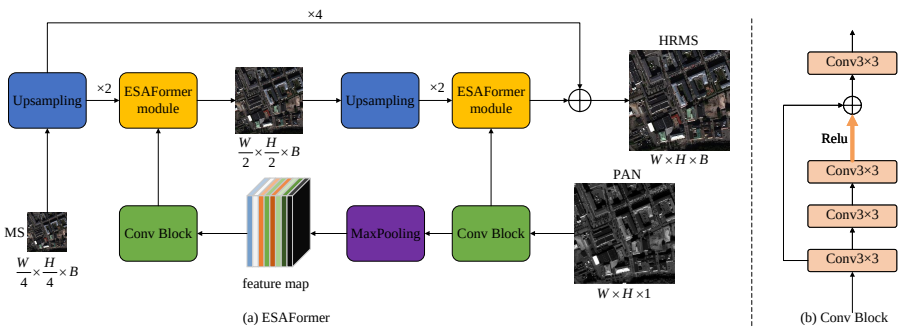

Figure 2: ESAFormer: Multi-resolution Fusion Network for Pansharpening.

The PAN branch adopts two Conv Blocks consisting of four convolutional layers and a residual concatenation to generate different resolution spatial features. The MS branch consists of two steps of upsampling and ESAFormer, corresponding to different resolution features of PAN. In each step, the ESAFormer module fuses spectral and spatial features to generate the fusion results with the corresponding resolution. Finally, the HRMS image is obtained by adding the multi-resolution fusion result with the MS image after 4-fold upsampling.

2.3 ESAFormer

This section introduces the workflow of ESAFormer, whose structure is shown in Figure 3. The ESAFormer module is the core module of the proposed network architecture. This module is adopted for integrating the spatial information of PAN and the spectral information of MS and generating the fused features.

Firstly, the upsampled MS image and the feature maps of PAN are directly stacked together to obtain the fused shallow features. Then, ESA is deployed to

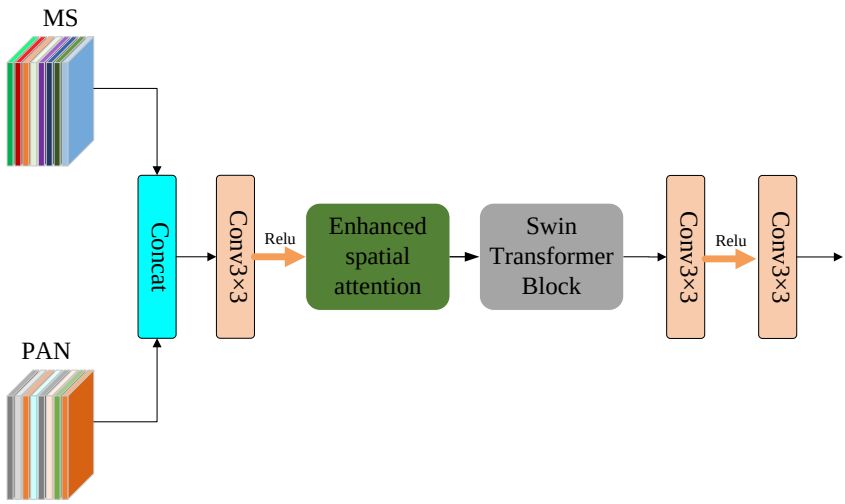

Figure 3: ESAFormer module.

increase the receptive field of the model. After that, the intermediate features are fed into the Swin-Transformer Block to further integrate the spectral and spatial information of the whole fused features. Finally, the two-layer convolution generates the initial fusion results while avoiding excessive spatial artifacts. The process of the ESAFormer can be described as:

$$\begin{aligned} F_{out} &= f_{ESAFormer}(F_{in}) \\ &= Conv_{group1}(H_{STB}(H_{ESA}(F_{in}))), \end{aligned} \tag{2}$$

where $H_{STB}(\cdot)$ denotes a Swin-Transformer Block, $Conv_{group1}(\cdot)$ is the operation of the convolution layer group, and $H_{ESA}(\cdot)$ denotes the ESA module. F_{in} denotes the shallow feature of the fused image, which could be formulated as :

$$F_{in} = Conv(Concat_{3\times3}(PAN_f, ms_{up})), \tag{3}$$

where PAN_f denotes the feature maps of PAN image, m_{up} denote the upsampled MS image.

The enhanced spatial attention (ESA) module is shown in Figure 4. Given an input F_{in} , ESA firstly obtains features F_1 as follows:

$$F_1 = Conv_1(F_{in}), \tag{4}$$

where F_1 is to reduce embedding dimension, and $Conv_1$ is a 1×1 convolution.Then ESA further calculates features F_2 as follows:

$$F_2 = Up(Conv_{group2}(Pooling(Conv_2(F_1)))), \tag{5}$$

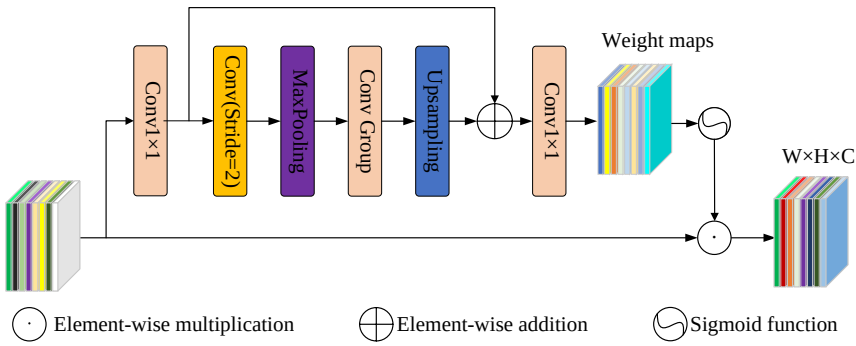

Figure 4: Enhanced Spatial Attention.

where $Up(\cdot)$ is the up-sampling function realized by bilinear interpolation, $Pooling$ is a max-pooling operation, $Conv_{group2}(\cdot)$ is a group composed of two 3×3 convolution layers with Relu, and $Conv_2$ is a 3×3 convolution with stride of 2. Both the pooling layer and the stridden convolutional layer diminish the spatial dimensions, which are subsequently recovered by the upsampling layer. Finally, the output of ESA module can be computed as:

$$F_{out}^{esa} = Sigmoid(Conv_3(F_1 + F_2)) \times F_{in}, \tag{6}$$

where $Conv_3$ is 1×1 convolutional layer used to recover the embedding dimension, $Sigmoid$ is the sigmoid function, and symbol $\times$ denotes element-wise multiplication operation. The features are more narrowly concentrated on the areas of interest at the start of the ESAFormer thanks to the operation of the ESA mechanism. After this operation, we aggregate the key areas of the image and greatly retain and enhance the spatial details, which is conducive to the subsequent work of pansharpening.

The architecture of Swin-Transformer Block (STB) is shown in Figure 5. It consists of layer-normalization (LN), window-based multi-head self-attention (W-MSA), shifted W-MSA (SW-MSA), and position-wise multilayer perception (MLP). STB can be explained by the following formulas:

$$F_3 = W\text{-}MSA(LN(F_{in}^{stb})) + F_{in}^{stb}, \tag{7}$$

$$F_4 = MLP(LN(F_3)) + F_3, \tag{8}$$

$$F_5 = SW\text{-}MSA(LN(F_4)) + F_4, \tag{9}$$

$$F_{out}^{stb} = MLP(LN(F_3)) + F_5. \tag{10}$$

In above formulas, F_3, F_4, F_5, F_{out}^{stb} indicate the feature maps in different layers. It is worth noting that F_{in}^{stb} equals to F_{out}^{esa}.

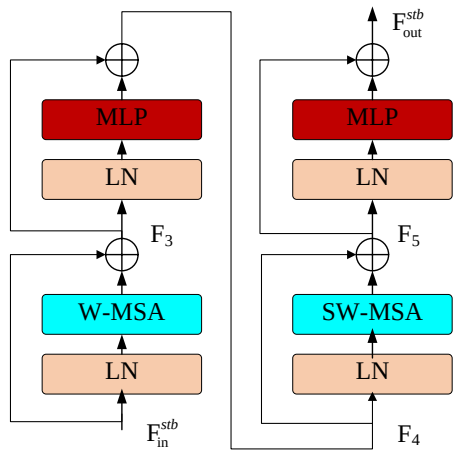

Figure 5: Swin Transformer Block.

In the flow of Swin-Transformer [17], the input image is first passed through an embedding layer to convert the pixel values into an embedded representation that can be processed by the Transformer model. Second, the image is divided into a series of windows, and there is some overlap between these windows. The self-attention mechanism within each window is adopted to capture local features, while the cross-attention mechanism between the windows is employed to capture long-range context information. In cross-attention, the shifted-window strategy is adopted to further improve the information transmission. Layer Normalization and MLP components are applied to normalize and map features for better feature presentation. Finally, residual connections are utilized to retain and convey information.

At the end stage of ESAFormer, the spectral and spatial information in the fused feature is integrated through two-layer convolution processing.

3 Experimental Results and Analysis

In this section, we conducted extensive comparison and ablation experiments to validate the performance of the ESAFormer. Firstly, we introduce the two datasets employed in our experiments. Subsequently, implementation details and evaluation metrics are given separately. After that, we indicate and analyze the visual and quantitative comparison results with other SOTA methods. Finally, the results of the ablation study are presented.

3.1 Datasets

WorldView-2 (WV2) and IKONOS (IK) datasets were adopted to verify the effectiveness of our method.

WorldView-2 (WV2): WorldView-2 dataset was captured from WorldView-2 satellite, which is the first commercial satellite to feature a Very-High-Resolution (VHR) sensor with more than four spectral bands. WV2 dataset has 500 pairs of images, each pair consisting of a PAN image with a resolution of 0.5 meters and a MS image with a resolution of 2 meters. The spatial sizes and radiometric resolutions of PAN/MS images are 1024 $\times$ 1024 / 256 $\times$ 256 and 11 bits respectively. The reduced-resolution (RR) data is obtained by degrading the MS and PAN images according to Wald's methodology [29]. As there is no GT, the original MS images are then used as GT. In the reduced resolution test, we used the 8:2 distribution to set the training set and the test set. In the full-resolution test, we randomly selected 50 pairs of images to test.

IKONOS (IK): The IK consists of 200 pairs of images with 1 meter PAN images and 4 meters MS spatial resolutions. As same as WV2 dataset setting, we applied 8:2 distribution for training and testing. 40 pairs of images were selected for testing in our experiments.

3.2 Implementation Details

Training Platform and Parameter Configuration: The proposed network was coded using Pytorch 1.7.0 and Python 3.8.0, and trained with an NVIDIA GPU GeForce RTX 3090. To minimize the loss function, we used the Adam optimizer with the betas of (0.9, 0.999) and weight decay of 0. We set the initial learning rate of 0.0004 and batch size of 4. To achieve better performance, every model underwent 1000 training epochs, with half-rate drop-in learning rate taking place every 200 epochs.

3.3 Evaluation Metrics

The similarity between the fused image and the ideal reference image (the original MS image) is measured by the reduced-resolution (RR) and full-resolution (FR) assessments. Multiple assessment metrics can be calculated to find the similarity. In the RR experiments, the Q2n (Q8 for 8-band datasets and Q4 for 4-band datasets) [4], the spectral angle mapper (SAM) [31], the dimensionless global error in synthesis (ERGAS) [29], and the spatial correlation coefficient (SCC) [34] are used. For SAM and ERGAS, the optimum values are 0, while for Q2n and SCC, they are 1. FR experiments are also required to validate the fusion performance. In contrast to the test scenarios with RR, there is no reference (GT) image. Therefore, three metrics without

GT involved are applied, including the quality with no reference (QNR), the spectral distortion D_λ, and the spatial distortion D_s [25]. We use the residual standard error (RSE) as a measure of the difference in the spectra between the fused image and the GT.

3.4 Visual and Quantitative Assessments

To verify the effectiveness of the proposed method, extensive comparative experiments were conducted between ESAFormer and twelve state-of-the-art pansharpening methods over two datasets. To be specific, our method was compared with seven traditional methods, including EXP [2], BT-H [18], BDSD-PC [11], C-GSA [3], SR-D [24], MTF-GLP-HPM-R [27], and MTF-GLP-FS [28]. Five representative deep learning-based methods were selected, namely, PNN [19], DiCNN [12], Fusion-Net [7], TD-Net [32], and TF-Net [16].

WV2 RR: Figure 6 shows a scenario of urban buildings from WV2. For better visibility, we presented the RSE between GT and the fused images across the eight spectra as shown in Figure 7. In Figure 7, the lower RSE result proves the better performance. As can be seen from these two figures, traditional pansharpening methods have poor result, such as SR-D and C-GSA suffer from certain spectral distortion and spatial detail blurring. However, deep-learning-based methods retain the details well, such as Fusion-Net, TF-Net have less difference between GT and fused results than traditional methods. Compared to the above methods, the visualization results of our method have the best performance. Furthermore, we did quantitative comparison experiments to further illustrate the superiority of our method, as shown in Table 1 (Q8, SAM, ERGAS, and SCC). Our method achieved the best values in all 4 metrics in RR experiments. Especially in Q8, our method outperforms traditional methods over 5% and outperforms current deep learning methods such as TF-Net over 1%. The quantitative and visual results demonstrate the effectiveness of our method.

IK RR: Figure 8 is a scenario including different geographical instances such as land, terraces, and rivers. Similar to WV2 RR experiments, we presented the results of RSE for this scene, as shown in Figure 9. As can be seen in Figure 9, the RSE of traditional methods such as I-MTF-GLP-FS, I-MTF-GLP-HPM-R is higher, while the deep learning-based methods such as TD-Net, PNN is similarly higher. This means that their fusion is less effective. As shown in Table 2 (Q8, SAM, ERGAS, and SCC), our method achieved best values on all 4 metrics. Especially in ERGAS, our method significantly outperforms other methods. These quantitative and visual results prove the superiority of our method.

WV2 FR: Figure 10 shows the full resolution experiment performed on WV2 dataset. As shown in Figure 10, some methods have severe spectral distortion of the vegetation in BT-H, C-GSA, and PNN. Some spatial structure

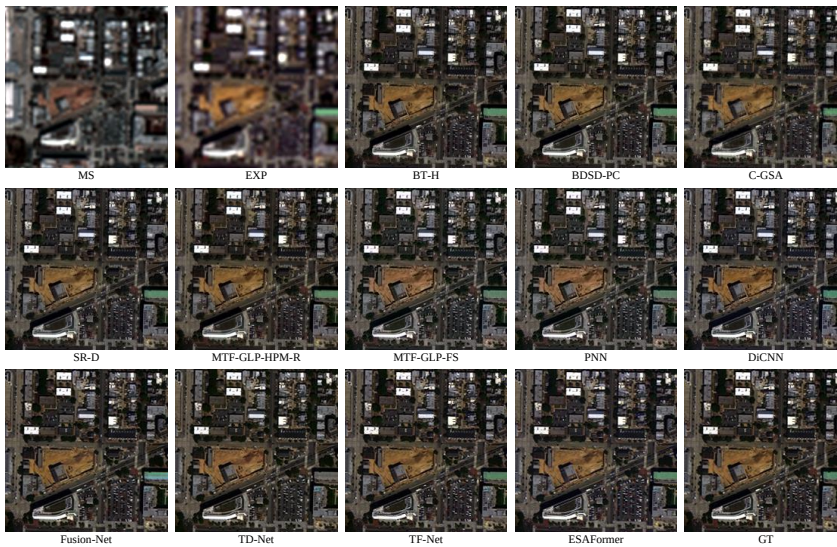

Figure 6: Visualization results of RR fusion experiments on WorldView-2 dataset.

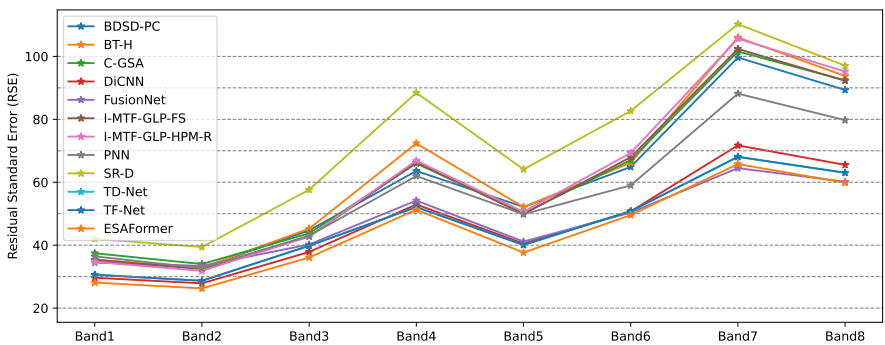

Figure 7: Spectral difference visualization of RR experiments on WorldView-2 dataset.

of fused images of BDSD-PC, MTF-GLP-HPM-R, and MTF-GLP-FS are lost. With full-resolution experiments at WV2, our method is the best in both spectral fidelity and spatial detail. As can be seen from Table 1 (D_λ, D_s, and QNR), our method achieved the best values for the QNR metric and ranks among the top of all methods for both the D_λ and D_s metrics. The above data fully demonstrate the superiority of our method in full-resolution experiments.

Table 1: Mean quantitative evaluation of results in WorldView-2 dataset.

Method	Q8	SAM	ERGAS	SCC	D_λ	D_s	QNR
EXP	0.4940	5.4002	7.2886	0.7595	0.0539	0.1393	0.8144
BT-H	0.7285	4.3107	3.6495	0.9590	0.1101	0.0765	0.8227
BDSD-PC	0.7267	4.8560	3.7921	0.9444	0.1753	**0.0236**	0.8054
C-GSA	0.7212	4.9467	4.0890	0.9334	0.1035	0.0713	0.8334
SR-D	0.7040	4.7177	4.1292	0.9391	0.0182	0.0706	0.9125
MTF-GLP-HPM-R	0.7170	4.9743	4.5882	0.9200	0.0373	0.0430	0.9214
MTF-GLP-FS	0.7205	4.8631	3.9598	0.9323	**0.0348**	0.0494	0.9176
PNN	0.7287	4.3602	3.5196	0.9642	0.0882	0.0404	0.8755
DiCNN	0.7533	3.5695	2.9225	0.9758	0.0729	0.0264	0.9029
Fusion-Net	0.7479	3.3356	2.7297	0.9816	0.1423	0.0346	0.8278
TD-Net	0.7501	3.4814	2.8192	0.9796	0.1233	0.0338	0.8477
TF-Net	0.7542	3.5023	2.9471	0.9778	0.0722	0.0569	0.8749
ESAFormer	**0.7666**	**3.2497**	**2.6134**	**0.9829**	0.0498	0.0292	**0.9224**
Idea Value	1	0	0	1	0	0	1

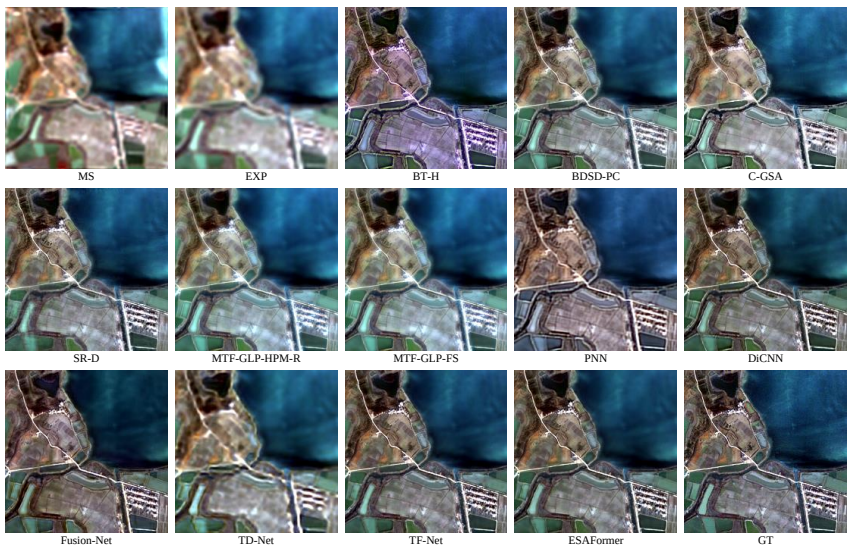

Figure 8: Visualization results of RR fusion experiments on IKONOS dataset.

IK FR: Figure 11 presents the full-resolution experiments performed on the IK dataset. From the visualization results, we can see that the spatial details are slightly worse in the traditional methods such as BDSD-PC, SR-D, and MTF-GLP-FS. The spectral distortion is worse in PNN and Fusion-Net,

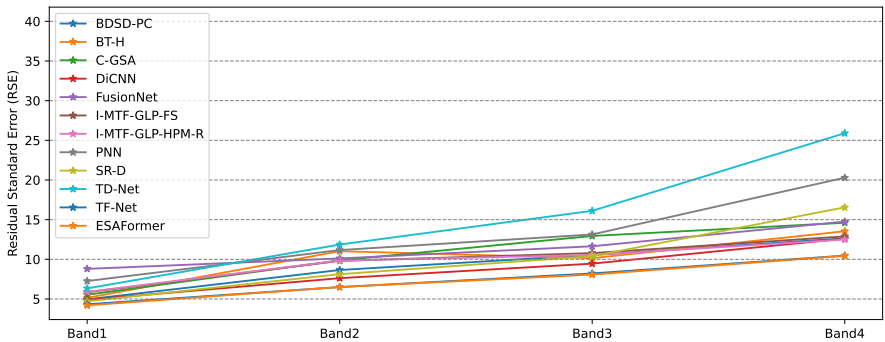

Figure 9: Spectral difference visualization of RR experiments on IKONOS dataset.

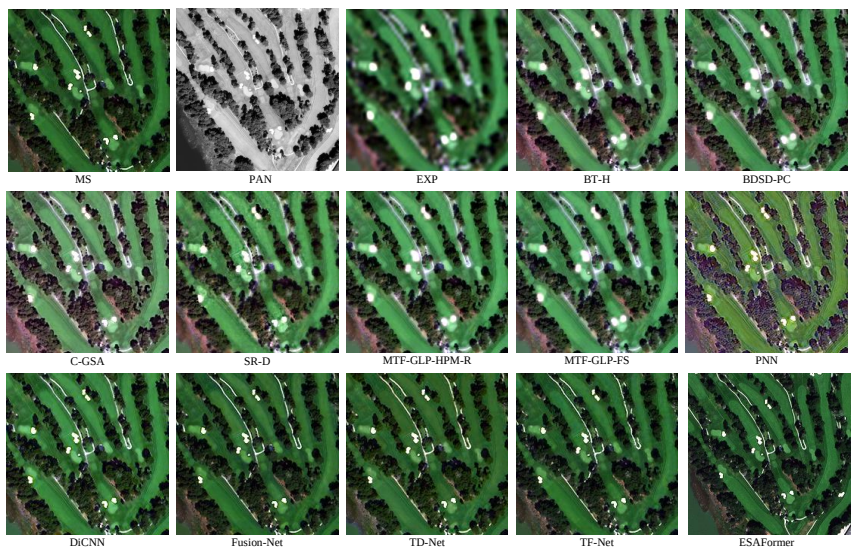

Figure 10: Visualization results of FR fusion experiments on WorldView-2 dataset.

but the overall results are still good in our method and TF-Net. As can be seen from Table 2 (D_λ, D_s, and QNR), our method achieved top-rank in all metrics, which demonstrates the excellent performance of our method in full-resolution experiments.

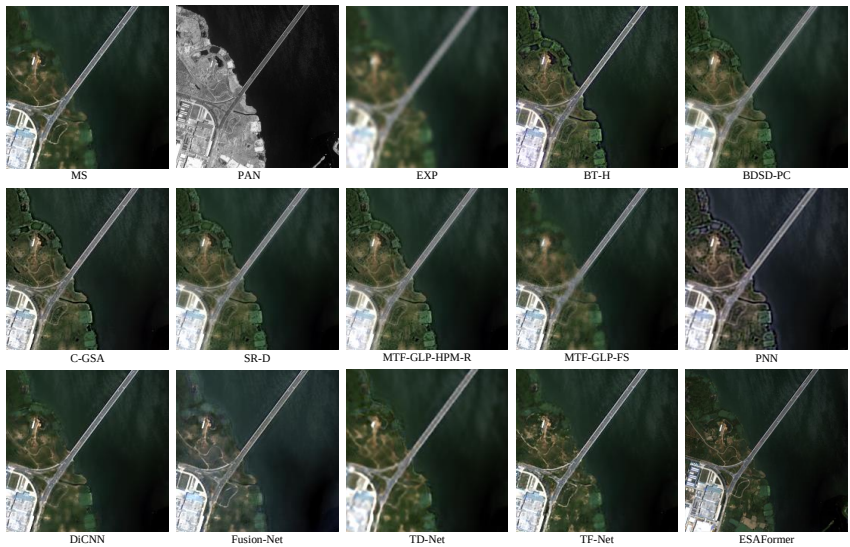

Figure 11: Visualization results of FR fusion experiments on IKONOS dataset.

Table 2: Mean quantitative evaluation of results in IKONOS dataset.

Method	Q8	SAM	ERGAS	SCC	D_λ	D_s	QNR
EXP	0.6386	2.4664	2.5898	0.8076	0.0806	0.1511	0.7818
BT-H	0.8191	2.0141	1.4239	0.9570	0.1678	0.1649	0.7117
BDSD-PC	0.8396	1.8871	1.3717	0.9622	0.1332	0.1046	0.7809
C-GSA	0.8356	1.8619	1.3643	0.9620	0.1377	0.1491	0.7443
SR-D	0.8419	1.8017	1.3735	0.9618	**0.0418**	0.0872	**0.8752**
MTF-GLP-HPM-R	0.8413	1.8185	1.3385	0.9639	0.0820	0.1323	0.8000
MTF-GLP-FS	0.8382	1.8774	1.3828	0.9599	0.0856	0.1354	0.7939
PNN	0.7908	2.2485	1.7463	0.9429	0.2453	0.0900	0.6928
DiCNN	0.8582	1.5829	1.1853	0.9749	0.1275	0.1234	0.7746
Fusion-Net	0.7807	1.5514	1.4539	0.9783	0.3626	**0.0721**	0.5987
TD-Net	0.7353	2.1117	2.1117	0.8881	0.0853	0.1907	0.7401
TF-Net	0.8730	1.4878	1.1422	0.9792	0.0810	0.0885	0.8411
ESAFormer	**0.8839**	**1.3665**	**1.0300**	**0.9823**	0.0729	0.0829	0.8553
Idea Value	1	0	0	1	0	0	1

3.5 Ablation Study

To verify the effectiveness of the components of our network, we did ablation study. Multiple experiments with different Network configurations were performed. Configurations include baseline [33], w/o ESA, ESA $\times$ 2, w/o branch, and ESAFormer.

baseline: We use a bi-directional flow network without ESAFormer module as our baseline. As shown in Table 3, our method is higher than the baseline in all the metrics. This greatly demonstrates that ESAFormer module has greatly improved our spectral preservation and spatial enhancement ability.

Table 3: Results of ablation study.

Method	Q8	SAM	ERGAS	SCC	D_λ	D_s	QNR
baseline	0.7609	3.4626	2.8195	0.9785	0.0698	0.0276	0.9047
w/o ESA	0.7663	3.2480	2.6251	0.9825	0.0574	0.0448	0.9003
ESA $\times 2$	0.7659	3.2210	2.5559	0.9836	0.0519	0.0667	0.8849
w/o branch	0.7632	3.3765	2.7289	0.9143	0.0565	0.0461	0.9000
ESAFormer	**0.7666**	**3.1497**	**2.5134**	**0.9829**	**0.0498**	**0.0232**	**0.9224**
Idea Value	1	0	0	1	0	0	1

w/o ESA: In this experiment, we removed the ESA of the ESAFormer module and retain the original STB and convolutional layer structure. The experimental results reported that all the metrics of our method is higher than the model without ESA. Obviously utilizing the ESA to increase the receptive field is necessary for the pansharpening task. This further demonstrates that the proposed method is rational and effective.

ESA $\times$ 2: In this experiment, we explored whether multiple ESA components are necessary to achieve better results for the pansharpening task. We added an extra ESA after the STB in the ESAFormer module. The experimental results show that the model with additional ESA leads to performance decline instead. Therefore, although the strategy of using ESA to increase the receptive field is effective, too many ESAs stacked can lead to dilution of local features by overly dispersed contextual weights. Thus it is important to set an appropriate number of ESAs for the pansharpening task.

w/o branch: To verify the effectiveness of the bi-directional structure, we removed the PAN branch. We fed the stacked 4-fold upsampled MS and PAN images directly to the network instead of upsampling them step by step. From the results, we can see that the results of the single-branch structure on all the metrics are lower than our method. This suggests that the bi-directional structure plays an important role in the pansharpening task, in other words, the cascading injection of different resolution features is necessary.

4 Conclusions

In this paper, we propose a multi-resolution network that combines CNN and Transformer to perform a multi-scale fusion by a bi-directional structure. This structure makes full use of the spatial details of PAN images and gradually

injects them into MS images. We innovatively proposed the ESAFormer module, which utilizes the enhanced spatial attention mechanism and the Swin-Transformer Block. In the proposed method, the spectral features can be better preserved and the spatial details can be enhanced. In comparison with other SOTA methods, the superiority of our method is demonstrated. Through ablation study, the effectiveness of the proposed components is verified.

Acknowledgements

This work was supported by the Fundamental Research Funds for the Central Universities (20199236855), the Aeronautical Science Foundation of China (D023030002), and the National Key Research and Development Program of China, project 3 (2022YFB4703703). We acknowledge the authors of WorldView-2 and IKONOS datasets used for this study.

References

[1] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, "An MTF-based spectral distortion minimizing model for pan-sharpening of very high resolution multispectral images of urban areas", in *2003 2nd GRSS/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas*, IEEE, 2003, 90–4.

[2] B. Aiazzi, L. Alparone, S. Baronti, and A. Garzelli, "Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis", *IEEE Transactions on geoscience and remote sensing*, 40(10), 2300–12.

[3] B. Aiazzi, S. Baronti, and M. Selva, "Improving component substitution pansharpening through multivariate regression of MS + Pan data", *IEEE Transactions on Geoscience and Remote Sensing*, 45(10), 3230–9.

[4] L. Alparone, S. Baronti, A. Garzelli, and F. Nencini, "A global quality measurement of pan-sharpened multispectral imagery", *IEEE Geoscience and Remote Sensing Letters*, 1(4), 313–7.

[5] C. Ballester, V. Caselles, L. Igual, J. Verdera, and B. Rougé, "A variational model for P+ XS image fusion", *International Journal of Computer Vision*, 69, 43–58.

[6] J. Choi, K. Yu, and Y. Kim, "A new adaptive component-substitution-based satellite image fusion by using partial replacement", *IEEE transactions on geoscience and remote sensing*, 49(1), 295–309.

[7] L.-J. Deng, G. Vivone, C. Jin, and J. Chanussot, "Detail injection-based deep convolutional neural networks for pansharpening", *IEEE Transactions on Geoscience and Remote Sensing*, 59(8), 6995–7010.

[8] W. Diao, F. Zhang, H. Wang, J. Sun, and K. Zhang, "Pansharpening via triplet attention network with information interaction", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15, 3576–88.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale", *arXiv preprint arXiv:2010.11929*.

[10] J. Fang, H. Lin, X. Chen, and K. Zeng, "A hybrid network of cnn and transformer for lightweight image super-resolution", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, 1103–12.

[11] A. Garzelli, F. Nencini, and L. Capobianco, "Optimal MMSE pan sharpening of very high resolution multispectral images", *IEEE Transactions on Geoscience and Remote Sensing*, 46(1), 228–36.

[12] L. He, Y. Rao, J. Li, J. Chanussot, A. Plaza, J. Zhu, and B. Li, "Pansharpening via detail injection based convolutional neural networks", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(4), 1188–204.

[13] L. Hou, B. Zhang, and B. Wang, "PAN-guided multiresolution fusion network using Swin transformer for pansharpening", *IEEE Geoscience and Remote Sensing Letters*, 20, 1–5.

[14] S. Jian, H. Kaiming, R. Shaoqing, and Z. Xiangyu, "Deep residual learning for image recognition", in *IEEE Conference on Computer Vision & Pattern Recognition*, 2016, 770–8.

[15] C. A. Laben and B. V. Brower, "Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening", US Patent 6,011,875, January 2000.

[16] X. Liu, Q. Liu, and Y. Wang, "Remote sensing image fusion based on two-stream fusion network", *Information Fusion*, 55, 1–15.

[17] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows", in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, 10012–22.

[18] S. Lolli, L. Alparone, A. Garzelli, and G. Vivone, "Haze correction for contrast-based multispectral pansharpening", *IEEE Geoscience and Remote Sensing Letters*, 14(12), 2255–9.

[19] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks", *Remote Sensing*, 8(7), 594.

[20] X. Meng, H. Shen, H. Li, L. Zhang, and R. Fu, "Review of the pansharpening methods for remote sensing images based on the idea of meta-analysis: Practical discussion and challenges", *Information Fusion*, 46, 102–13.

[21] X. Meng, N. Wang, F. Shao, and S. Li, "Vision transformer for pansharpening", *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–11.

[22] X. Meng, Y. Xiong, F. Shao, H. Shen, W. Sun, G. Yang, Q. Yuan, R. Fu, and H. Zhang, "A large-scale benchmark data set for evaluating pansharpening performance: Overview and implementation", *IEEE Geoscience and Remote Sensing Magazine*, 9(1), 18–52.

[23] Z. Su, Y. Yang, S. Huang, W. Wan, J. Sun, W. Tu, and C. Chen, "STCP: Synergistic Transformer and Convolutional Neural Network for Pansharpening", *IEEE Transactions on Geoscience and Remote Sensing*.

[24] M. R. Vicinanza, R. Restaino, G. Vivone, M. Dalla Mura, and J. Chanussot, "A pansharpening method based on the sparse representation of injected details", *IEEE Geoscience and Remote Sensing Letters*, 12(1), 180–4.

[25] G. Vivone, L. Alparone, J. Chanussot, M. Dalla Mura, A. Garzelli, G. A. Licciardi, R. Restaino, and L. Wald, "A critical comparison among pansharpening algorithms", *IEEE Transactions on Geoscience and Remote Sensing*, 53(5), 2565–86.

[26] G. Vivone, M. Dalla Mura, A. Garzelli, R. Restaino, G. Scarpa, M. O. Ulfarsson, L. Alparone, and J. Chanussot, "A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods", *IEEE Geoscience and Remote Sensing Magazine*, 9(1), 53–81.

[27] G. Vivone, R. Restaino, and J. Chanussot, "A regression-based high-pass modulation pansharpening approach", *IEEE Transactions on Geoscience and Remote Sensing*, 56(2), 984–96.

[28] G. Vivone, R. Restaino, and J. Chanussot, "Full scale regression-based injection coefficients for panchromatic sharpening", *IEEE Transactions on Image Processing*, 27(7), 3418–31.

[29] L. Wald, "Quality of high resolution synthesised images: Is there a simple criterion?", in *Third conference" Fusion of Earth data: merging point measurements, raster maps and remotely sensed images"*, SEE/URISCA, 2000, 99–103.

[30] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, "PanNet: A deep network architecture for pan-sharpening", in *Proceedings of the IEEE international conference on computer vision*, 2017, 5449–57.

[31] R. H. Yuhas, A. F. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm", in *JPL, Summaries of the Third Annual JPL Airborne Geoscience Workshop. Volume 1: AVIRIS Workshop*, 1992.

[32] T.-J. Zhang, L.-J. Deng, T.-Z. Huang, J. Chanussot, and G. Vivone, "A triple-double convolutional neural network for panchromatic sharpening", *IEEE Transactions on Neural Networks and Learning Systems*.

[33] Y. Zhang, C. Liu, M. Sun, and Y. Ou, "Pan-sharpening using an efficient bidirectional pyramid network", *IEEE Transactions on Geoscience and Remote Sensing*, 57(8), 5549–63.

[34] J. Zhou, D. L. Civco, and J. A. Silander, "A wavelet transform method to merge Landsat TM and SPOT panchromatic data", *International journal of remote sensing*, 19(4), 743–57.

[35] M. Zhou, J. Huang, Y. Fang, X. Fu, and A. Liu, "Pan-sharpening with customized transformer and invertible neural network", in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, No. 3, 2022, 3553–61.