Spatial-Spectral Similarity-Guided Fusion Network for Pansharpening

Jiazhuang Xiong[1], **Yongshan Zhang**[1*], **Xinxin Wang**[2] and **Lefei Zhang**[3]
[1]School of Computer Science, China University of Geosciences
[2]Department of Computer and Information Science, University of Macau
[3]School of Computer Science, Wuhan University
{xjz19980901, yszhang.cug}@gmail.com, xinxinwang1024@gmail.com, zhanglefei@whu.edu.cn

Abstract

Pansharpening fuses lower-resolution multispectral (LRMS) images with high-resolution panchromatic (PAN) images to generate high-resolution multispectral (HRMS) images that preserves both spatial and spectral information. Most deep pansharpening methods face challenges in cross-modal feature extraction and fusion, as well as in exploring the similarities between the fused image and both PAN and LRMS images. In this paper, we propose a spatial-spectral similarity-guided fusion network (S^3FNet) for pansharpening. This architecture is composed of three parts. Specifically, a shallow feature extraction layer learns initial spatial, spectral and fused features from PAN and LRMS images. Then, a multi-branch asymmetric encoder, consisting of spatial, spectral and fusion branches, generates corresponding high-level features at different scales. A multi-scale reconstruction decoder, equipped with a well-designed cross-feature multi-head attention fusion block, processes the intermediate feature maps to generate HRMS images. To ensure HRMS images retain maximum spatial and spectral information, a similarity-constrained loss is defined for network training. Extensive experiments demonstrate the effectiveness of our S^3FNet over state-of-the-art methods. The code is released at https://github.com/ZhangYongshan/S3FNet.

1 Introduction

Modern imaging sensors often face a trade-off between capturing high-resolution images with limited spectral information and low-resolution images rich in spectral data due to hardware limitations. It is challenging to acquire images with both high spatial resolution and abundant spectral information from a single sensor. To address this problem, satellites are typically equipped with two distinct sensors: one for capturing high-resolution panchromatic (PAN) images and another for low-resolution multispectral (LRMS) images [Zhou *et al.*, 2025]. These complementary images are then fused through pansharpening techniques to generate high-resolution multispectral (HRMS) images for various tasks, such as land-cover recognition [Zhang *et al.*, 2024] and anomaly detection [Shikhar and Sobti, 2024].

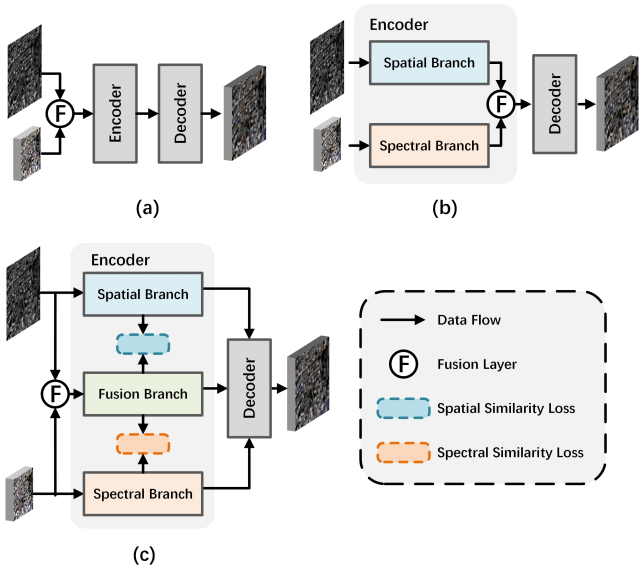

Figure 1: Different structures of existing deep pansharpening methods. (a) Single-branch structure. (b) Dual-branch structure. (c) Our multi-branch structure with similarity-constrained loss.

Pansharpening is a promising research topic in the remote sensing community, as it enhances the spatial resolution of multispectral (MS) images while preserving their spectral information through the integration of high-resolution single-band PAN images. There are numerous effective traditional and deep pansharpening methods [Deng *et al.*, 2022]. Traditional pansharpening methods are mainly divided into three categories: component substitution (CS) methods [Choi *et al.*, 2010; Vivone, 2019], multiresolution analysis (MRA) methods [Vivone *et al.*, 2013; Vivone *et al.*, 2018], and variational optimization (VO) methods [Fu *et al.*, 2019; Tian *et al.*, 2021]. These methods are well-suited for a range of satellite images. However, their performance is limited by a heavy reliance on hand-crafted features.

Due to the remarkable feature extraction capabilities of

*Corresponding author.

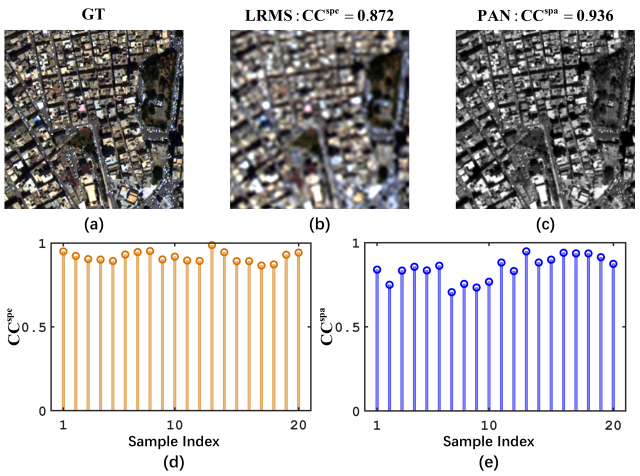

Figure 2: Correlation exploration between diverse images on the WorldView-3 dataset. (a) Ground-truth (GT) image. (b) LRMS image showing a spectral correlation coefficient ($CC^{spe} = 0.872$) between GT image. (c) PAN image showing a spatial correlation coefficient ($CC^{spa} = 0.936$) between GT image. (d) Histogram of spectral correlation coefficients between LRMS and GT images across 20 samples. (e) Histogram of spatial correlation coefficients between PAN and GT images across 20 samples.

neural networks, numerous deep pansharpening methods have been proposed. Based on their network architecture, these methods are primarily categorized into single-branch and dual-branch structures. As shown in Figure 1 (a), single-branch networks concatenate both PAN and LRMS images through a single pathway to obtain fusion images [Masi *et al.*, 2016; Deng *et al.*, 2021]. They neglect the unique characteristics of features from different modalities, leading to inadequate feature extraction. In contrast, as shown in Figure 1 (b), dual-branch networks process PAN and MS images through two separated branches, and then fuse them via concatenation before passing them to the subsequent decoder [Zhang *et al.*, 2022; Zhou *et al.*, 2022]. They enable complementary feature extraction and facilitate high-quality image fusion.

Although deep pansharpening methods yield impressive fusion results, two potential issues hinder their performance. First, most existing methods are ineffective for cross-modal feature extraction and fusion [Zhang *et al.*, 2022; Xing *et al.*, 2024]. The differences in spatial and spectral information between PAN and LRMS images make them challenging to extract and combine features effectively in single-branch or dual-branch structures with simple concatenation, limiting the full utilization of complementary modalities. To solve this issue, as shown in Figure 1 (c), a multi-branch structure with an additional fusion branch should be considered to better extract and integrate features from both input images, enabling more effective fusion and enhancing overall performance. Second, most previous methods neglect the similarities between the fused image and both PAN and LRMS images [Peng *et al.*, 2023; Duan *et al.*, 2024]. As shown in Figure 2, the correlation analysis reveals that an LRMS image exhibits a strong spectral correlation with the ground-truth (GT) image, while a PAN image displays a strong spatial correlation with the GT image. The absence of similarity constraints between the fused image and input images in existing methods may result in the loss of important spatial and spectral details, thereby compromising the pansharpening quality. To solve this issue, similarity guidance between the fused features and those extracted from input images should be introduced during network training.

Motivated by the above observations, in this paper, we propose a spatial-spectral similarity-guided fusion network (S^3FNet) for pansharpening. As shown in Figure 3, there are three key components in this architecture. Specifically, a shallow feature extraction layer processes PAN and LRMS images through Transformer blocks and convolution to learn initial spatial, spectral and fused features. To enable high-level feature learning at different scales, a multi-branch asymmetric encoder is presented by incorporating distinct spatial, spectral and fusion branches. Based on a well-designed cross-feature multi-head attention fusion block, a multi-scale reconstruction decoder is able to generate high-quality HRMS images by fusing diverse intermediate feature maps. To preserve as much spatial and spectral information as possible in HRMS images, a similarity-constrained loss is formulated to facilitate the network training and parameter update. Our main contributions are as follows:

- We propose a similarity-guided fusion network designed with a multi-branch structure for PAN and LRMS image fusion. This framework facilitates the effective exploitation of complementary information while capturing both spatial and spectral correlations.
- We present spatial, spectral and fusion branches with distinct designs in the multi-branch asymmetric encoder to hierarchically learn high-level features with discriminative information.
- We design a cross-feature multi-head attention fusion block with spatial and spectral cross-attention mechanisms in the multi-scale reconstruction decoder to facilitate effective cross-modal feature fusion.
- We formulate a similarity-constrained loss for network training that integrates spatial and spectral correlation measures to ensure the preservation of maximum details in the fused images. Experimental results demonstrate the effectiveness of our proposed method.

2 Related Works

In recent years, numerous deep pansharpening methods have achieved remarkable results, owing to their exceptional feature extration capabilities. Pansharpening neural network (PNN) [Masi *et al.*, 2016] is a pioneer to perform image fusion using a three-layer convolutional network. Further, detail injection based convolutional neural network (DiCNN) [He *et al.*, 2019] injects the details of PAN images into LRMS images to provide explicit physical interpretations. Considering the complementarity of input images, bidirectional pyramid network (BDPN) [Zhang *et al.*, 2019] processes them in two separated branches to reconstruct fusion images. To explore local and global context, bidomain modeling pansharpening (BiMPan) method [Hou *et al.*, 2023]

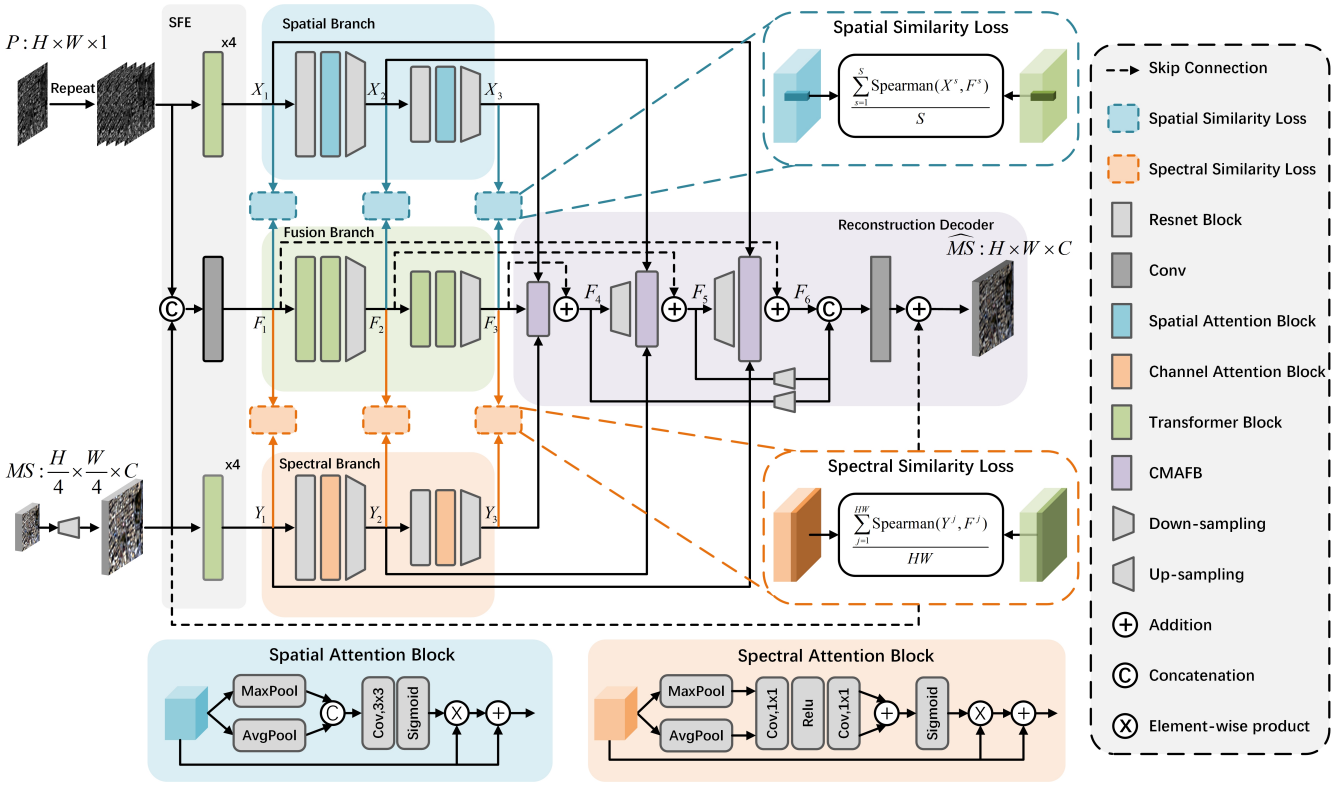

Figure 3: Overall structure of our S^{3}FNet. It contains three key parts. PAN and LRMS images firstly pass through a shallow feature extraction (SFE) layer. Next, the extracted features enter a multi-branch asymmetric encoder with spatial, spectral and fusion branches. Finally, the features are processed through a multi-scale reconstruction decoder incorporated with a well-designed cross-feature multi-head attention fusion block (CMAFB) to generate the HRMS image.

presents local adaptive and global detail branches for better image fusion. Additionally, double U-shape network (U2Net) [Peng *et al.*, 2023] processes PAN and LRMS images through spatial and spectral U-Nets, and performs fusion via the S2Block. To solve edge blur and spectral distortion problems, multi-supervised mask protection network (MMPN) [Chen *et al.*, 2023] utilizes a dual-stream multi-scale feature fusion module to fuse input images and their masked data. To improve the interpretability, deep intrinsic supervision pansharpening network (DISPNet) [Wang *et al.*, 2024] incorporates spatial consistency and spectral projection priors into the formulated variational minimization model. In addition, there are many other effective methods by introducing diverse techniques [Duan *et al.*, 2024; Xing *et al.*, 2024; He *et al.*, 2025].

Although great success has been achieved by previous single-branch or dual-branch networks, cross-modal feature fusion and similarity guidance between fusion image and input images are not well taken into consideration. To handle these issues, we will introduce our proposed method in detail.

3 Methodology

Figure 3 illustrates the overall structure of S^{3}FNet. The proposed network mainly consists of three parts, including a shallow feature extraction (SFE) layer, a multi-branch asymmetric encoder and a multi-scale reconstruction decoder with the well-designed cross-feature multi-head attention fusion block (CMAFB). The proposed network is trained using a similarity-constrained loss to enhance image fusion.

3.1 Shallow Feature Extraction Layer

To ensure the consistent feature map size, we first replicate the PAN image $P \in R^{H\times W}$ C times along the spectral dimension, resulting in $P_{repeat} \in R^{H\times W\times C}$. Then, we perform bicubic interpolation on the MS image $MS \in R^{\frac{H}{4}\times\frac{W}{4}\times C}$ to obtain $MS_{\times 4} \in R^{H\times W\times C}$. Both preprocessed images are input to the shallow feature extraction layer consisting of four cascaded Transformer blocks [Zamir *et al.*, 2022], yielding spatial features $X_1 \in R^{H\times W\times S}$ from P_{repeat} and spectral features $Y_1 \in R^{H\times W\times S}$ from $MS_{\times 4}$. Besides, we concatenate P_{repeat} and $MS_{\times 4}$ to produce the initial fused features F_1 via a convolutional layer. The shallow feature extraction layer is designed to capture initial spatial, spectral and fusion features from PAN and LRMS images. They are served as input for the subsequent encoder-decoder network to enable further feature extraction and fusion.

3.2 Multi-Branch Asymmetric Encoder

Unlike existing single-branch or dual-branch structures, the multi-branch asymmetric encoder adopts a three-branch design, consisting of a spatial branch, a spectral branch and a fusion branch. Each branch is specifically tailored with distinct

designs to learn high-level spatial, spectral and fused features at different scales for high-quality image fusion.

Spatial Branch. The spatial branch takes X_1 as input and enhances spatial features X_2 and X_3 through a cascaded process using a combination of ResNet block (ResBlock) [Szegedy *et al.*, 2017], spatial attention block (SpaAttBlock), and downsampling operation. The ResBlock focuses on local feature extraction, the SpaAttBlock emphasizes important local features across regions, and the downsampling operation reduces spatial resolution. The process for learning the enhanced spatial features is as follows:

$$X_n = \text{Down}(\text{SA}(\text{ResBlock}(X_{n-1}))), \quad (1)$$

where n represents the stage index ($n \in \{2, 3\}$). ResBlock$(\cdot)$ denotes the ResBlock, comprising two 3×3 convolutional layers followed by a leaky ReLU activation function. SA$(\cdot)$ represents the SpaAttBlock, composing two parallel pooling operations to aggregate channel information, followed by a 3×3 convolutional layer and a sigmoid activation function. Input features of the SpaAttBlock are processed through element-wise multiplication with attention features, followed by element-wise addition. Down$(\cdot)$ represents the downsampling operation, involving a 2×2 kernel with stride 2 and a depth-wise convolutional layer to increase the feature map channels.

Spectral Branch. The spectral branch takes Y_1 as input and learns advanced spectral features Y_2 and Y_3 through a cascaded process using a combination of ResBlock, spectral attention block (SpeAttBlock) [Woo *et al.*, 2018] and downsampling operation. The ResBlock and downsampling operation are the same as those in the spatial branch, while the SpeAttBlock highlights key features by learning the importance of each spectral channel. The process for learning advanced spectral features is as follows:

$$Y_n = \text{Down}(\text{CA}(\text{ResBlock}(Y_{n-1}))), \quad (2)$$

where CA$(\cdot)$ represents the SpeAttBlock, comprising two parallel pooling operations to aggregate spatial information and generate spatial context descriptors, followed by two multi-layer perceptrons (MLPs) with 1×1 convolution and ReLU activation function, and a sigmoid activation function to obtain attention features. The input features of the SpeAttBlock are combined with attention features through element-wise multiplication and addition.

Fusion Branch. The fusion branch takes F_1 as input and learns complex fusion features F_2 and F_3 in a cascaded manner using a combination of two Transformer blocks and a downsampling operation. The Transformer block is the same as in the shallow feature extraction layer, while the downsampling operation mirrors those in the spatial and spectral branches. The process for learning fusion features is as follows:

$$F_n = \text{Down}(\text{TB}(\text{TB}(F_{n-1}))), \quad (3)$$

where TB$(\cdot)$ represents the Transformer block. The use of the Transformer block is motivated by its self-attention mechanism, enabling global feature extraction. This mechanism allows the model to capture long-range dependencies and relationships between different parts of input data, enhancing its ability to understand complex patterns.

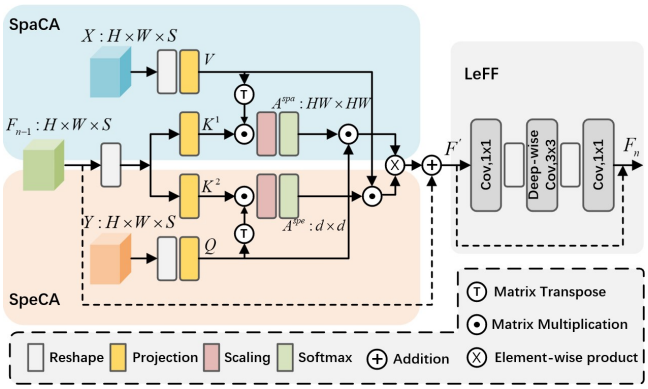

Figure 4: Structure of CMAFB. It takes the features generated by the three branches in the encoder as input and produces fused features using spatial cross-attention (SpaCA), spectral cross-attention (SpeCA) mechanisms, and a locally-enhanced feed-forward layer (LeFF).

3.3 Multi-Scale Reconstruction Decoder

To facilitate cross-modal image fusion, the outputs from the three branches of the encoder are fed into a multi-scale reconstruction decoder. This decoder consists of three stages. Each stage contains one upsampling operation (except for the first stage) and a multi-head attention fusion block (CMAFB). In each stage, the spatial and spectral features extracted from the corresponding branches, along with the fused features from the previous stage, are input to the CMAFB. The resulting features are then added to the skip-connected fusion features from the encoder (F_3, F_2, and F_1) to generate the fused features for the next stage (F_4, F_5, and F_6). Finally, F_4 and F_5 are upsampled to the spatial resolution of F_6, and all three are concatenated and processed by a convolutional layer. The output is added to the $MS_{\times 4}$ to produce the final fused result $\widehat{MS}$, as follows:

$$\widehat{MS} = \text{Conv}\big(\text{Concat}(\text{Up}(F_4), \text{Up}(F_5), F_6)\big) + MS_{\times 4}, \quad (4)$$

where Up$(\cdot)$ denotes the upsampling operation used to align the spatial dimensions of feature maps at different scales.

3.4 Cross-Feature Multi-Head Attention Fusion

Unlike traditional fusion strategies that directly concatenate features, as illustrated in Figure 4, the CMAFB incorporates spatial cross-attention (SpaCA) and spectral cross-attention (SpeCA) mechanisms to enhance the cross-modal fusion of shallow and deep features by highlighting their distinctions. Specifically, the SpaCA focuses on the differences between spatial and fusion features along the spatial dimension, while the SpeCA emphasizes the differences between spectral and fusion features along the spectral dimension. For each CMAFB, spatial, spectral and fusion feature maps of corresponding scale are taken as inputs. For simplicity, we ignore the subscripts of these feature maps and reshape them to matrix representations. Then, they are linearly mapped to the query Q and key K^1 in the speCA, and to the key K^2 and

value V in the SpaCA as follows:

$$V = W_V X, \tag{5}$$

$$Q = W_Q Y, \tag{6}$$

$$K^1 = W_{K^1} F, \tag{7}$$

$$K^2 = W_{K^2} F, \tag{8}$$

where W_Q, W_V, W_{K^1} and W_{K^2} are learnable parameters in the linear mapping layer. Subsequently, K^1, K^2, Q and V are divided into m heads based on S channels of the features processed in the current stage. The dimension of each head is $d = S/m$. For simplicity, the operations of splitting and merging the multiple heads are omitted in the figure. Thus, the calculation of the spatial and spectral attention matrices of the j-th head, A_j^{spa} and A_j^{spe}, is as follows:

$$A_j^{spa} = \text{Softmax}\left(\frac{K_j^1 \cdot (V_j)^T}{\sqrt{(d)^2}}\right), \tag{9}$$

$$A_j^{spe} = \text{Softmax}\left(\frac{(Q_j)^T \cdot K_j^2}{\frac{\sqrt{(d)^3}}{HW}}\right), \tag{10}$$

where $(\cdot)^T$ denotes the matrix transpose operation.

After obtaining A_j^{spe} and A_j^{spa}, we perform matrix multiplication between them and V_j from spatial features and Q_j from spectral features, respectively. This results in the fusion output of the j-th head, $head_j$, combining both spatial and spectral information. The specific calculation is as follows:

$$head_j = \left(V_j \cdot A_j^{spe}\right) \cdot \left(A_j^{spa} \cdot Q_j\right). \tag{11}$$

Subsequently, $head_j$ from different heads are concatenated along the channel dimension to form the final fusion output $Head$ via the cross-attention mechanism. This fusion features integrate spatial and spectral information while employing attention mechanisms to capture long-range dependencies.

Followed by the cross-attention mechanisms, a locally-enhanced feed-forward (LeFF) network layer [Wang *et al.*, 2022] is introduced. It consists of a 1×1 convolution to increase the number of channels, a 3×3 convolution to capture local information, and another 1×1 convolution to reduce the number of channels. This design strengthens the representation of local information. The final fusion features F_n are derived as follows:

$$F' = \text{LN}(Head + F_{n-1}), \tag{12}$$

$$F_n = \text{LN}(\text{LeFF}(F') + F'), \tag{13}$$

where $\text{LN}(\cdot)$ represents layer normalization, and $\text{LeFF}(\cdot)$ denotes the locally-enhanced feed-forward layer. This process integrates spatial and spectral details through cross-attention mechanisms and local feature enhancements, resulting in generating the final fusion features.

3.5 Loss Function

Overall Loss. To achieve better fusion, we incorporate the similarity constraint into our loss function, i.e.,

$$L_{total} = L_1 + \alpha L_{sim}, \tag{14}$$

where $L_1 = \|GT - \widehat{MS}\|_1$ measures the ℓ_1 loss between the GT and reconstructed fusion images, and L_{sim} represents the multi-scale spatial-spectral similarity-constrained loss by imposing constraints on the correlations between the fused features and both the spatial and spectral features at different scales. α is a hyperparameter that balances the two loss terms.

Multi-Scale Spatial-Spectral Similarity-constrained Loss. To ensure the fusion image is closer to the GT image, similarity constraints from both spatial and spectral perspectives should be incorporated into the pansharpening process. Inspired by [Zhao *et al.*, 2023], we design a multi-scale spatial-spectral similarity-constrained loss based on spatial and spectral correlations between intermediate features from the spatial, spectral and fusion branches at different scales. Measured by the Spearman correlation coefficient, this loss ensures the fusion image maintains high spectral and spatial correlations with the GT image. It is formulated as follows:

$$L_{sim} = \frac{1}{3}\sum_{i=1}^{3}\left(\frac{\beta}{CC_i^{spa}} + \frac{1-\beta}{CC_i^{spe}}\right), \tag{15}$$

where β is a parameter to control the ratio of spatial and spectral similarities. CC_i^{spa} and CC_i^{spe} denote the spatial and spectral correlation coefficients at the i-th scale, respectively.

Spatial and Spectral Correlation Coefficients. Spatial correlation is calculated to measure the correlation between the fused and spatial feature maps in the spatial dimension. It is formulated as:

$$CC_i^{spa} = \frac{\sum_{s=1}^{S_i} \text{Spearman}(X_i^s, F_i^s)}{S_i}, \tag{16}$$

where $\text{Spearman}(\cdot)$ calculates the Spearman correlation coefficient, and S_i denotes the number of spectral channels at the i-th scale. X_i^s and F_i^s represent the s-th spectral channel of features X_i and F_i, respectively. Before calculating the Spearman correlation coefficient, they are flattened along the spatial positions of the pixels. Similarly, spectral correlation is computed to quantify the correlation between the fused and spectral feature maps in the spectral dimension. It is represented as:

$$CC_i^{spe} = \frac{\sum_{j=1}^{H_i W_i} \text{Spearman}(Y_i^j, F_i^j)}{H_i W_i}, \tag{17}$$

where H_i and W_i represent the height and width of features Y_i and F_i at the i-th scale, respectively. Y_i^j and F_i^j denote the j-th pixel of features Y_i and F_i, respectively.

4 Experiments

4.1 Experimental Setup

Datasets. Three datasets used in this study were obtained from the PanCollection repository [Deng *et al.*, 2022], including data from the WorldView-3 (WV3), GaoFen-2 (GF2), WorldView-2 (WV2) satellites. Both reduced-resolution and full-resolution datasets are adopted for experiments. The reduced-resolution data is synthesized using Wald's protocol, where LRMS and PAN images are downsampled by a factor of 4, with the original LRMS images serving as the ground truth.

Method	WorldView-3				GaoFen-2				WorldView-2			
	SAM↓	ERGAS↓	Q8↑	SCC↑	SAM↓	ERGAS↓	Q4↑	SCC↑	SAM↓	ERGAS↓	Q8↑	SCC↑
BT-H	4.8985	4.5150	0.8182	0.9240	1.6810	1.5524	0.9089	0.9508	12.4846	9.1758	0.5882	0.7694
MTF-GLP-FS	5.3233	4.6452	0.8177	0.8984	1.6757	1.6023	0.8914	0.9390	12.7166	9.2195	0.5840	0.7505
BDSD-PC	5.4675	4.6549	0.8117	0.9049	1.7110	1.7025	**0.9932**	0.9448	12.9399	9.1165	0.5971	0.7631
PNN	3.6798	2.6819	0.8929	0.9761	1.0480	1.0570	0.9600	0.9772	7.1158	5.6152	0.7619	0.8782
DiCNN	3.5929	2.6733	0.9004	0.9763	1.0525	1.0812	0.9594	0.9771	6.9216	6.2507	0.7205	0.8552
MSDCNN	3.7773	2.7608	0.8900	0.9741	1.0472	1.0413	0.9612	0.9782	6.0064	4.7438	0.8241	0.8972
BDPN	4.1646	3.0871	0.8581	0.9577	1.4158	1.4493	0.9255	0.9532	7.0934	4.8568	0.8235	0.9033
FusionNet	3.3252	2.4666	0.9044	0.9807	0.9735	0.9878	0.9641	0.9806	6.4257	5.1363	0.7961	0.8746
LAGNet	3.1117	2.3091	0.9102	0.9838	0.7859	0.6869	0.9804	0.9906	6.9545	5.3262	0.8054	0.9125
BiMPan	3.0141	2.2808	0.9145	0.9843	0.8871	0.8062	0.9728	0.9886	5.7496	4.5111	0.8271	0.9127
Pan-Mamba	2.8444	2.1937	0.9179	0.9855	0.7503	0.7053	0.9806	0.9890	6.0743	5.5888	0.8357	0.8672
Ours	**2.8194**	**2.1591**	**0.9192**	**0.9860**	**0.6605**	**0.6107**	0.9841	**0.9917**	**5.1873**	**4.0493**	**0.8490**	**0.9282**

Table 1: Quantitative comparison of different methods on reduced-resolution data. The best results are marked in bold. ↑ indicates that higher values correspond to better performance, while ↓ signifies the opposite.

Method	WorldView-3			GaoFen-2			WorldView-2		
	D_λ↓	D_s↓	QNR↑	D_λ↓	D_s↓	QNR↑	D_λ↓	D_s↓	QNR↑
BT-H	0.0574	0.0810	0.8670	0.0602	0.1313	0.8165	0.0786	0.0858	0.8428
MTF-GLP-FS	0.0354	0.0630	0.9043	0.0336	0.1404	0.8309	**0.0325**	0.0756	0.8948
BDSD-PC	0.0625	0.0730	0.8698	0.0759	0.1548	0.7812	0.1429	0.0386	0.8242
PNN	0.0213	0.0428	0.9369	0.0367	0.0943	0.8726	0.1484	0.0771	0.7869
DiCNN	0.0362	0.0462	0.9195	0.0413	0.0992	0.8636	0.1412	0.1023	0.7700
MSDCNN	0.0230	0.0467	0.9316	0.0269	0.0730	0.9020	0.0589	**0.0290**	0.9143
BDPN	0.0395	0.0459	0.9168	0.0326	0.0701	0.8994	0.1117	0.0328	0.8606
FusionNet	0.0239	0.0364	0.9406	0.0400	0.1013	0.8628	0.0519	0.0559	0.8948
LAGNet	0.0368	0.0418	0.9230	0.0324	0.0792	0.8910	0.1302	0.0547	0.8229
BiMPan	0.0196	0.0340	0.9467	0.0296	0.0528	0.9192	0.0468	0.0300	0.9247
Pan-Mamba	0.0226	0.0358	0.9425	0.0192	0.0364	0.9451	0.0436	0.0633	0.8956
Ours	**0.0175**	**0.0310**	**0.9520**	**0.0186**	**0.0147**	**0.9669**	0.0403	0.0319	**0.9292**

Table 2: Quantitative comparison of different methods on full-resolution data. The best results are marked in bold. ↑ indicates that higher values correspond to better performance, while ↓ signifies the opposite.

Compared Methods. To evaluate the proposed method, several state-of-the-art pansharpening methods are used for comparison, including three traditional methods (BT-H [Aiazzi *et al.*, 2006], MTF-GLP-FS [Vivone *et al.*, 2018], and BDSD-PC [Vivone, 2019]) and eight deep learning methods (PNN [Masi *et al.*, 2016], MSDCNN [Wei *et al.*, 2017], DiCNN [He *et al.*, 2019], BDPN [Zhang *et al.*, 2019], FusionNet [Deng *et al.*, 2020], LAGNet [Jin *et al.*, 2022] , BiMPan [Hou *et al.*, 2023] and Pan-Mamba [He *et al.*, 2025]).

Evaluation Metrics. To evaluate the quality of the fused images, the spectral angle mapper (SAM), relative global dimensional synthesis error (ERGAS), Q4/Q8 metrics, and structural content correlation (SCC) are employed for the reduced-resolution datasets. For the full-resolution datasets, D_λ, D_s, and quality with no reference (QNR) are adopted. Details of these metrics can be found in [Lu *et al.*, 2023].

Implementation Details. Our model was implemented using PyTorch on a mahine with Nvidia 4090 GPU. The Adam optimizer is used for network training over 300 epochs with a batch size of 16. The initial learning rate is set to 0.001 and halved every 100 epochs. For the loss function, $\alpha = 0.001$ and $\beta = 0.5$, with α decaying during training. For the shallow feature extraction layer, the output channels are set to $S = 32$ for the three datasets. Additionally, $C = 8$ for the WV3 and WV2 datasets, and $C = 4$ for the GF2 dataset. For BiMPan and Pan-Mamba, we use the default settings from their released code. For other deep learning methods, we adopt the settings specified in [Deng *et al.*, 2022].

4.2 Experimental Results

Comparison on Reduced-Resolution Data. The quantitative results of different methods on reduced-resolution datasets are presented in Table 1. The three traditional methods exhibit limited performance because they heavily rely on hand-crafted features. In contrast, the deep learning-based methods significantly outperform traditional methods owing to their superior feature extraction capabilities. Compared to other methods, our S^{3}FNet shows the best performance on reduced-resolution datasets, verifying its effectiveness in pansharpening tasks. Additionally, the visual results on a sample from the WV3 dataset are displayed in Figure 5. It is obvious that our S^{3}FNet obtains the fusion image most similar to the GT image. The absolute error maps show that S^{3}FNet has notably fewer residuals than other methods.

Comparison on Full-Resolution Data. To further verify the generalization ability of our S^{3}FNet, experiments on full-resolution datasets are conducted, and the quantitative results are presented in Table 2. Experimental reults demonstrate that our S^{3}FNet also achieves the best performance among all compared methods on full-resolution datasets to produce fusion images with low spectral and spatial distortion. This

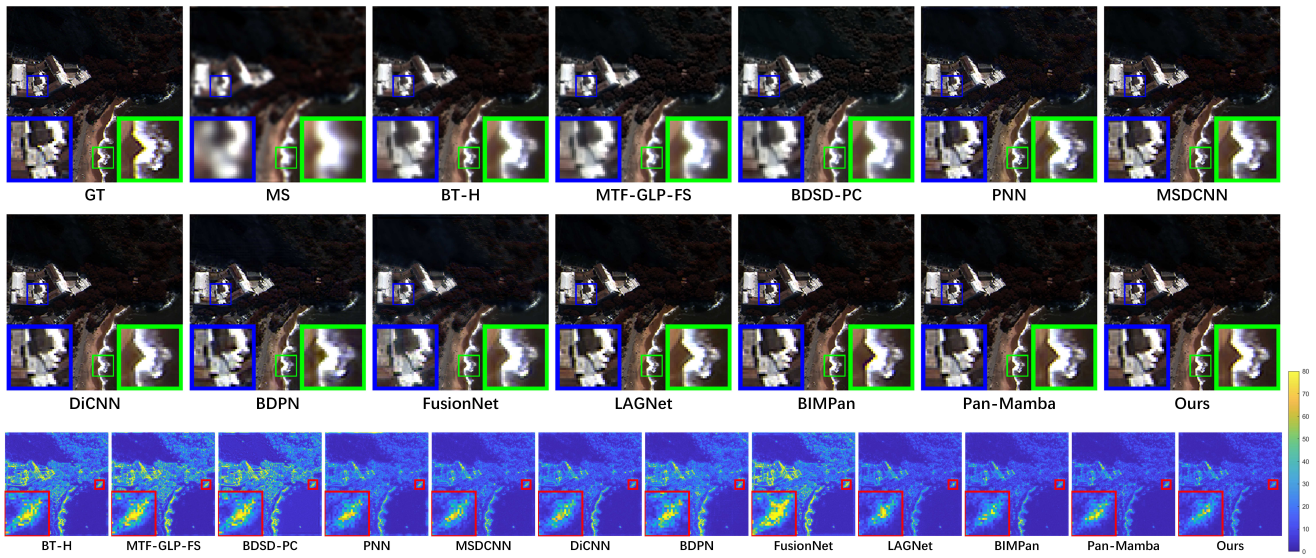

Figure 5: Visual comparison of different methods on a reduced-resolution sample from the WV3 dataset. The mean square error residuals between the fused and GT images are displayed in the last row.

Setting	MB	CMAFB	L_{sim}	SAM↓	ERGAS↓	Q8↑	SCC↑
(I)	×	×	×	3.7070	2.863	0.899	0.9698
(II)	√	×	×	2.8440	2.188	0.918	0.9857
(III)	√	√	×	2.8266	2.165	0.918	0.9859
Ours	√	√	√	**2.8194**	**2.159**	**0.919**	**0.9860**

Table 3: Ablation study on the WV3 dataset.

indicates its practicality and adaptability for remote sensing image fusion.

4.3 Ablation Study

To validate the effectiveness of key components in S^{3}FNet, ablation study is conducted on the WV3 dataset with three degraded variants. Specifically, the multi-branch (MB) encoder can be replaced with a dual-branch structure, the CMAFB in the decoder can be replaced with a standard convolution, and the similarity-constrained loss can be included or excluded. The corresponding results are shown in Table 3. Setting (I) shows the worst results because no key components are included. By introducing the multi-branch encoder, setting (II) outperforms setting (I) due to more effective feature processing across different branches. By retaining the CMAFB, setting (III) is superior to setting (II) because of more effective cross-modal feature fusion. Finally, our S^{3}FNet, with the multi-branch encoder, CMAFB and similarity-constrained loss, outperforms other degraded settings. These results confirm the effectiveness of the design in S^{3}FNet.

4.4 Parameter Study

In S^{3}FNet, α controls the impact of the similarity-constrained loss, while β adjusts the balance between spatial and spectral similarities. To investigate the effect of different parameter settings, α is varied over $\{0.0001, 0.001, 0.01, 0.1, 1, 10\}$ with β fixed at its default value of 0.5. Conversely, β is varied from 0.1 to 0.9 in increments of 0.1 with α fixed at its default value of 0.001. The quantitative results for different parameter settings are reported in Figure 6. When α is set too large, the two loss terms become unbalanced in magnitude. With $\alpha = 0.001$ and $\beta = 0.5$ as default values, S^{3}FNet achieves superior performance.

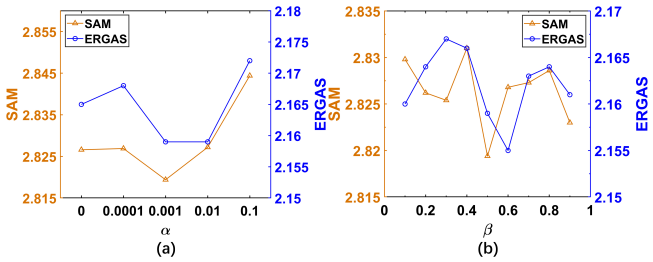

Figure 6: Quantitative results for different parameter settings on the WV3 dataset. (a) Effect of α. (b) Effect of β.

5 Conclusion

In this work, we proposed a spatial-spectral similarity-guided fusion network named S^{3}FNet for pansharpening. The S^{3}FNet is designed with a multi-branch architecture. Specifically, the multi-branch structure is implemented in the encoder to generate high-level spatial, spectral and fusion features from the initial features of input images obtained by the shallow feature extraction layer. Driven by the well-designed cross-feature multi-head attention fusion block, the decoder reconstructs the fused images by incorporating spatial and spectral details at different scales. The proposed network is trained using a similarity-constrained loss to ensure high-quality image fusion. Extensive experiments on three satellite datasets verify the effectiveness of the proposed method against state-of-the-art methods.

Acknowledgments

This work was funded by the China National Key R&D Program (Grant 2023YFF0807000).

References

[Aiazzi *et al.*, 2006] Bruno Aiazzi, Luciano Alparone, Stefano Baronti, Andrea Garzelli, and Massimo Selva. Mtf-tailored multiscale fusion of high-resolution ms and pan imagery. *Photogrammetric Engineering & Remote Sensing*, 72(5):591–596, 2006.

[Chen *et al.*, 2023] Changjie Chen, Yong Yang, Shuying Huang, Wei Tu, Weiguo Wan, and Shengna Wei. Mmpn: Multi-supervised mask protection network for pansharpening. In *IJCAI*, pages 573–580, 2023.

[Choi *et al.*, 2010] Jaewan Choi, Kiyun Yu, and Yongil 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, 2010.

[Deng *et al.*, 2020] Liang-Jian Deng, Gemine Vivone, Cheng Jin, and Jocelyn Chanussot. Detail injection-based deep convolutional neural networks for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 59(8):6995–7010, 2020.

[Deng *et al.*, 2021] Liang-Jian Deng, Gemine Vivone, Cheng Jin, and Jocelyn Chanussot. Detail injection-based deep convolutional neural networks for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 59(8):6995–7010, 2021.

[Deng *et al.*, 2022] Liang-Jian Deng, Gemine Vivone, Mercedes E Paoletti, Giuseppe Scarpa, Jiang He, Yongjun Zhang, Jocelyn Chanussot, and Antonio Plaza. Machine learning in pansharpening: A benchmark, from shallow to deep networks. *IEEE Geoscience and Remote Sensing Magazine*, 10(3):279–315, 2022.

[Duan *et al.*, 2024] Yule Duan, Xiao Wu, Haoyu Deng, and Liang-Jian Deng. Content-adaptive non-local convolution for remote sensing pansharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27738–27747, 2024.

[Fu *et al.*, 2019] Xueyang Fu, Zihuang Lin, Yue Huang, and Xinghao Ding. A variational pan-sharpening with local gradient constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10265–10274, 2019.

[He *et al.*, 2019] Lin He, Yizhou Rao, Jun Li, Jocelyn Chanussot, Antonio Plaza, Jiawei Zhu, and Bo Li. Pansharpening via detail injection based convolutional neural networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(4):1188–1204, 2019.

[He *et al.*, 2025] Xuanhua He, Ke Cao, Jie Zhang, Keyu Yan, Yingying Wang, Rui Li, Chengjun Xie, Danfeng Hong, and Man Zhou. Pan-mamba: Effective pan-sharpening with state space model. *Information Fusion*, 115:102779, 2025.

[Hou *et al.*, 2023] Junming Hou, Qi Cao, Ran Ran, Che Liu, Junling Li, and Liang-jian Deng. Bidomain modeling paradigm for pansharpening. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 347–357, 2023.

[Jin *et al.*, 2022] Zi-Rong Jin, Tian-Jing Zhang, Tai-Xiang Jiang, Gemine Vivone, and Liang-Jian Deng. Lagconv: Local-context adaptive convolution kernels with global harmonic bias for pansharpening. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 1113–1121, 2022.

[Lu *et al.*, 2023] Hangyuan Lu, Yong Yang, Shuying Huang, Xiaolong Chen, Biwei Chi, Aizhu Liu, and Wei Tu. Awfln: An adaptive weighted feature learning network for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.

[Masi *et al.*, 2016] Giuseppe Masi, Davide Cozzolino, Luisa Verdoliva, and Giuseppe Scarpa. Pansharpening by convolutional neural networks. *Remote Sensing*, 8(7):594, 2016.

[Peng *et al.*, 2023] Siran Peng, Chenhao Guo, Xiao Wu, and Liang-Jian Deng. U2net: A general framework with spatial-spectral-integrated double u-net for image fusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3219–3227, 2023.

[Shikhar and Sobti, 2024] Sambal Shikhar and Anupam Sobti. Label-free anomaly detection in aerial agricultural images with masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5440–5449, 2024.

[Szegedy *et al.*, 2017] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[Tian *et al.*, 2021] Xin Tian, Yuerong Chen, Changcai Yang, and Jiayi Ma. Variational pansharpening by exploiting cartoon-texture similarities. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2021.

[Vivone *et al.*, 2013] Gemine Vivone, Rocco Restaino, Mauro Dalla Mura, Giorgio Licciardi, and Jocelyn Chanussot. Contrast and error-based fusion schemes for multispectral image pansharpening. *IEEE Geoscience and Remote Sensing Letters*, 11(5):930–934, 2013.

[Vivone *et al.*, 2018] Gemine Vivone, Rocco Restaino, and Jocelyn Chanussot. Full scale regression-based injection coefficients for panchromatic sharpening. *IEEE Transactions on Image Processing*, 27(7):3418–3431, 2018.

[Vivone, 2019] Gemine Vivone. Robust band-dependent spatial-detail approaches for panchromatic sharpening. *IEEE transactions on Geoscience and Remote Sensing*, 57(9):6421–6433, 2019.

[Wang *et al.*, 2022] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference*

on computer vision and pattern recognition, pages 17683–17693, 2022.

[Wang *et al.*, 2024] Hebaixu Wang, Meiqi Gong, Xiaoguang Mei, Hao Zhang, and Jiayi Ma. Deep unfolded network with intrinsic supervision for pan-sharpening. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5419–5426, 2024.

[Wei *et al.*, 2017] Yancong Wei, Qiangqiang Yuan, Xiangchao Meng, Huanfeng Shen, Liangpei Zhang, and Michael Ng. Multi-scale-and-depth convolutional neural network for remote sensed imagery pan-sharpening. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3413–3416. IEEE, 2017.

[Woo *et al.*, 2018] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[Xing *et al.*, 2024] Yinghui Xing, Litao Qu, Shizhou Zhang, Kai Zhang, Yanning Zhang, and Lorenzo Bruzzone. Crossdiff: Exploring self-supervisedrepresentation of pansharpening via cross-predictive diffusion model. *IEEE Transactions on Image Processing*, 33:5496–5509, 2024.

[Zamir *et al.*, 2022] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022.

[Zhang *et al.*, 2019] Yongjun Zhang, Chi Liu, Mingwei Sun, and Yangjun Ou. Pan-sharpening using an efficient bidirectional pyramid network. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8):5549–5563, 2019.

[Zhang *et al.*, 2022] Tian-Jiang Zhang, Liang-Jian Deng, Ting-Zhu Huang, Jocelyn Chanussot, and Gemine Vivone. A triple-double convolutional neural network for panchromatic sharpening. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):9088–9101, 2022.

[Zhang *et al.*, 2024] Yongshan Zhang, Shuaikang Yan, Lefei Zhang, and Bo Du. Fast projected fuzzy clustering with anchor guidance for multimodal remote sensing imagery. *IEEE Transactions on Image Processing*, 33:4640–4653, 2024.

[Zhao *et al.*, 2023] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5906–5916, 2023.

[Zhou *et al.*, 2022] Huanyu Zhou, Qingjie Liu, and Yunhong Wang. Panformer: A transformer based model for pansharpening. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022.

[Zhou *et al.*, 2025] Man Zhou, Naishan Zheng, Xuanhua He, Danfeng Hong, and Jocelyn Chanussot. Probing synergistic high-order interaction for multi-modal image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2):840–857, 2025.