U2Net: A General Framework with Spatial-Spectral-Integrated Double U-Net for Image Fusion

Siran Peng*
School of Information and Communication Engineering, University of Electronic Science and Technology of China
Siran_Peng@163.com

Chenhao Guo*
School of Information and Communication Engineering, University of Electronic Science and Technology of China
carlguo508@gmail.com

Xiao Wu
School of Mathematical Sciences, University of Electronic Science and Technology of China
wxwsx1997@gmail.com

Liang-Jian Deng†
School of Mathematical Sciences, University of Electronic Science and Technology of China
liangjian.deng@uestc.edu.cn

ABSTRACT

In image fusion tasks, images obtained from different sources exhibit distinct properties. Consequently, treating them uniformly with a single-branch network can lead to inadequate feature extraction. Additionally, numerous works have demonstrated that multi-scaled networks capture information more sufficiently than single-scaled models in pixel-level computer vision problems. Considering these factors, we propose U2Net, a spatial-spectral-integrated double U-shape network for image fusion. The U2Net utilizes a spatial U-Net and a spectral U-Net to extract spatial details and spectral characteristics, which allows for the discriminative and hierarchical learning of features from diverse images. In contrast to most previous works that merely employ concatenation to merge spatial and spectral information, this paper introduces a novel spatial-spectral integration structure called S2Block, which combines feature maps from different sources in a logical and effective way. We conduct a series of experiments on two image fusion tasks, including remote sensing pansharpening and hyperspectral image super-resolution (HISR). The U2Net outperforms representative state-of-the-art (SOTA) approaches in both quantitative and qualitative evaluations, demonstrating the superiority of our method. The code is available at https://github.com/PSRben/U2Net.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

image fusion, pansharpening, hyperspectral image super-resolution, deep learning, U-Net

*Equal contribution.
†Corresponding author.

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada.

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612084

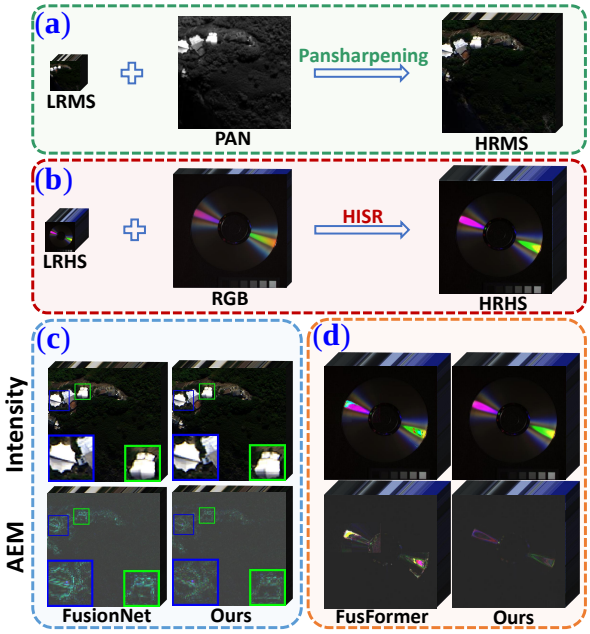


Figure 1: (a) The schematic diagram of pansharpening. (b) The schematic diagram of HISR. (c) The pansharpened images and their absolute error maps (AEMs) of FusionNet [3] and U2Net. (d) The acquired HRHS images and corresponding AEMs of Fusformer [11] and U2Net. It is obvious that our method yields the darker AEMs on both image fusion tasks, indicating its superiority over other competitors.

ACM Reference Format:
Siran Peng, Chenhao Guo, Xiao Wu, and Liang-Jian Deng. 2023. U2Net: A General Framework with Spatial-Spectral-Integrated Double U-Net for Image Fusion. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3612084

1 INTRODUCTION

Due to hardware limitations, sensors can only acquire high resolution images with sparse spectral information and low resolution images with copious spectral data. Image fusion aims to combine

these two kinds of images to produce high resolution results with a wealth of spectral information. Over recent years, image fusion algorithms have been widely used in fields such as remote sensing [22], medical imaging [12], and computer vision [39], proving high application values. This work mainly investigates two image fusion tasks: remote sensing pansharpening and hyperspectral image super-resolution (HISR). As illustrated in Fig. 1, pansharpening involves merging a panchromatic (PAN) image with a low resolution multispectral (LRMS) image to create a high resolution multispectral (HRMS) outcome, while HISR aims at generating a high resolution hyperspectral (HRHS) result from an RGB image and a low resolution hyperspectral (LRHS) image.

The traditional pansharpening works can be roughly divided into three categories [3], *i.e.*, the component substitution (CS) approaches, the multi-resolution analysis (MRA) methods, and the variational optimization-based (VO) techniques. The CS-based approaches [2, 21] project the LRMS image into a transformed domain, where the spatial information can be viewed as a component. By replacing this component with the PAN image, a desired HRMS result is generated. Although the CS-based approaches offer simple operation, low computational burden, and high spatial fidelity, they often suffer from significant spectral distortions. The MRA-based methods [24, 25] utilize a multi-resolution analysis framework to inject spatial details from the PAN image into the LRMS image, resulting in an HRMS output. While these methods maintain spectral characteristics effectively, they may encounter spatial distortion. The VO-based techniques [10, 18, 33, 38] exploit different optimization algorithms to solve the pansharpening issue and generally outperform CS-based and MRA-based approaches. Nevertheless, these techniques have problems such as high computational burden and complex parameterization, which restrict their practical implementation. As for the HISR task, conventional approaches focus on exploring the inherent relationship between the RGB and LRHS images and mainly establish models based on optimization to obtain the HRHS image.

Over the past few years, deep learning (DL) has emerged as a popular solution for image fusion problems. Thanks to the exceptional feature learning capacity of neural networks, numerous DL-based methods have yielded impressive outcomes. The classic DL-based approaches [9, 11, 13, 34] apply concatenation to combine images from different sources. The cascaded output is then fed into a single-branch, single-scale network to generate a desired outcome. However, this strategy suffers from several significant defects. Firstly, the concatenation operation fails to consider the distinctions between two types of images, causing insufficient information integration. Secondly, the single-branch design leads to inefficient feature extraction as it treats spatial and spectral characteristics equally. Thirdly, some deep-level information may be ignored due to single-scale image processing.

To address the abovementioned concerns, we propose a spatial-spectral-integrated double U-shape network called U2Net for image fusion. The U2Net utilizes a spatial U-Net to capture spatial details from the PAN/RGB image and employs a spectral U-Net to extract spectral characteristics from the LRMS/LRHS image. This enables our method to learn diverse features in a discriminative and hierarchical manner. Besides, a novel structure named S2Block is introduced to integrate the two kinds of information. In the S2Block,

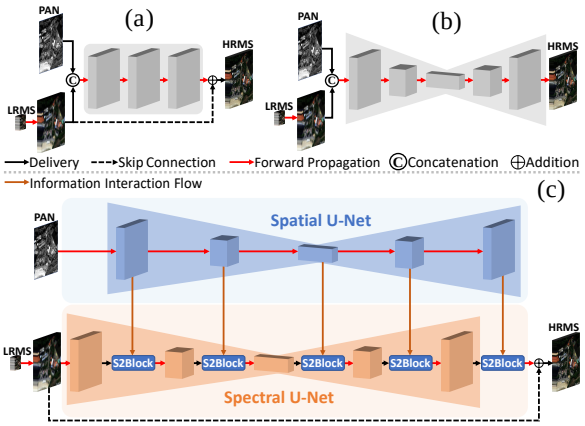


Figure 2: The structural comparison between existing DL-based image fusion works and U2Net (demonstrated with the pansharpening task). (a) The overall architecture of single-branch, single-scale methods, such as PanNet [34] and Fusformer [11]. These methods extract features at one specific scale, thus ignoring some deep-level information. (b) The overall architecture of single-branch, multi-scale approaches, including DCFNet [30] and MUCNN [27]. (c) The overall structure of the double-branch, multi-scale U2Net.

we first generate two sets of square matrices, namely *spatial self-correlation matrices* and *spectral self-correlation matrices*, to better describe the spatial and spectral information. Subsequently, a series of operators are applied to combine the square matrices, spatial feature maps, and spectral feature maps, producing a high-quality fusion result. The contributions of this work are as follows:

- A double U-shape network architecture consisting of a spatial U-Net and a spectral U-Net is created for image fusion tasks. This framework enables the effective learning of spatial details and spectral characteristics in a discriminative and hierarchical manner.
- A novel spatial-spectral integration structure called S2Block is designed to sufficiently merge feature maps from diverse images in a logical and comprehensive way.
- The spatial U-Net and spectral U-Net are connected through S2Blocks, composing our U2Net. The proposed method is tested on different image fusion tasks and achieves SOTA performance in quantitative and qualitative assessments.

2 RELATED WORK

DL-based Methods. In recent years, a number of DL-based image fusion methods have been proposed. These methods outperform traditional works due to the superior capabilities of DL in feature extraction and nonlinear fitting. For pansharpening, the pioneering work is the PNN [8] which utilizes three convolutional layers to achieve the best performance at that time. Since then, impressive methods such as PanNet [34], DiCNN [9], and FusionNet [3] have successively emerged, further validating the potential of DL in the field of pansharpening. There are also many exceptional DL-based works in the field of HISR, including ResTFNet [17], SSRNet

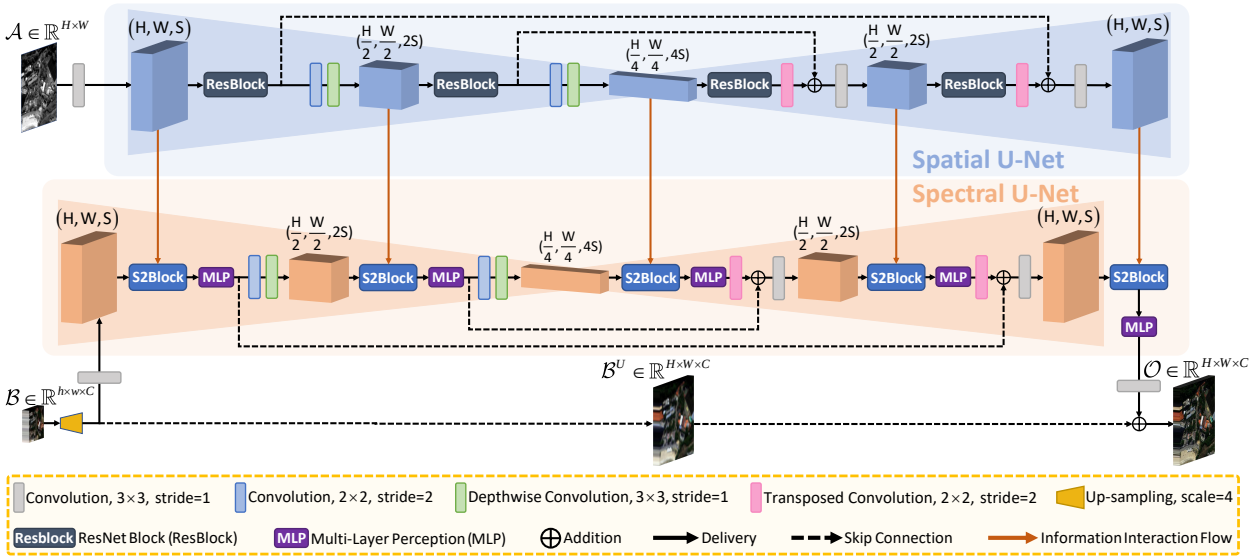


Figure 3: The overall structure of the proposed method. The U2Net employs a spatial U-Net and a spectral U-Net to extract spatial details and spectral characteristics, respectively. Besides, feature maps from different sources are integrated through the well-designed S2Blocks. The notations in this figure are explained in Section 3.1.

[36], and Fusformer [11]. However, due to unreasonable structural design, most DL-based image fusion approaches suffer from drawbacks such as spectral distortion and poor generalization ability.

U-Net. The U-shape network is put forward by [19] for the pixel-wise segmentation problem. It utilizes a series of symmetric down-sampling and up-sampling layers to capture information hierarchically. The U-shape network boasts a strong feature extraction capability, as evidenced by its extensive application in many pixel-level computer vision tasks. Recently, U-Net has also been introduced into image fusion tasks, and the representatives include DCFNet [30] and MUCNN [27]. It is worth noting that these methods only employ concatenation to merge images from different sources. Then, the cascaded output is fed into a single-branch U-shape network for feature extraction. This structural design can lead to insufficient information fusion and inefficient feature learning, thus requiring numerous parameters to attain satisfactory outcomes.

Motivation. Images obtained from different sensors possess unique properties, *e.g.*, PAN/RGB images exhibit rich spatial details, whereas LRMS/LRHS images contain a wealth of spectral information. Therefore, it is imperative to consider their differences when performing image fusion, which aims to produce a fused outcome from the two types of inputs. However, the majority of previous studies employ a single-branch network to uniformly extract spatial and spectral characteristics, as illustrated in Fig. 2. Furthermore, these approaches merely apply concatenation to combine the two kinds of images. Consequently, they suffer from significant issues, such as inefficient feature learning, poor generalization ability, and insufficient information integration. The above situation motivates us to propose U2Net, which captures spatial and spectral features discriminately and hierarchically using two U-shape networks. Besides, the well-designed S2Block enables the effective integration of feature maps from different sources.

3 THE PROPOSED METHOD

3.1 Notations

The PAN/RGB image is represented as $\mathcal{A} \in \mathbb{R}^{H\times W\times c}$, where H, W, and c denote height, width, and input channel, respectively. $\mathcal{B} \in \mathbb{R}^{h\times w\times C}$ represents the LRMS/LRHS image, in which $h = \frac{H}{4}$ and $w = \frac{W}{4}$. Besides, C denotes the spectral band. The up-sampled LRMS/LRHS image, desired HRMS/HRHS image, and ground-truth (GT) image are represented as $\mathcal{B}^U \in \mathbb{R}^{H\times W\times C}$, $\mathcal{O} \in \mathbb{R}^{H\times W\times C}$, and $\mathcal{X} \in \mathbb{R}^{H\times W\times C}$, respectively.

3.2 U2Net

To learn features from diverse images in a discriminative and hierarchical manner, we develop a double U-shape network architecture consisting of a spatial U-Net and a spectral U-Net, as shown in Fig. 3. The spatial U-Net focuses on extracting spatial details from $\mathcal{A}$, while the spectral U-Net is designed to collect the spectral data in $\mathcal{B}$. In order to capture sufficient deep-level information under limited network parameters, we process features maps at three distinct scales, *i.e.*, $H \times W \times S$ (S denotes the channel number of input feature maps), $\frac{H}{2} \times \frac{W}{2} \times 2S$, and $\frac{H}{4} \times \frac{W}{4} \times 4S$. Thus, the learning process of our U-Net consists of five stages. Each stage employs a neural network to extract information from the feature map of a particular size. According to the structural symmetry of the U-shape network, feature maps of sizes $H \times W \times S$, $\frac{H}{2} \times \frac{W}{2} \times 2S$, and $\frac{H}{4} \times \frac{W}{4} \times 4S$ are processed in stages one and five, stages two and four, and stage three, respectively. Between each pair of adjacent stages, there exists a step that involves operations for down-sampling/up-sampling and dimension transformation. In the initial two steps, we utilize 2×2 convolution kernels with a stride of 2 for down-sampling and apply depth-wise convolutional layers to augment

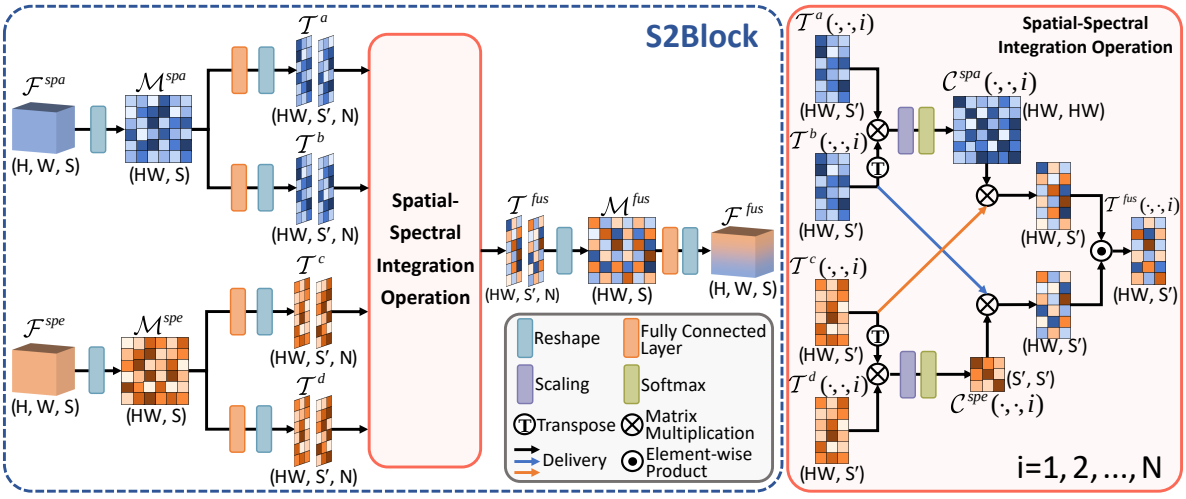


Figure 4: The structure of S2Block. To enhance the description of spatial details and spectral characteristics, we begin by producing the spatial self-correlation matrices C^{spa} and spectral self-correlation matrices C^{spe}. After that, a series of operators are employed to integrate spatial and spectral information. The notations in this figure are explained in Section 3.3.

the channel numbers of feature maps. For the final two steps, we use transposed convolutional layers with 2×2 kernels and a stride of 2 to achieve both up-sampling and channel reduction. In addition, the outputs of stages one and two are summed with the inputs of stages five and four, respectively.

This paragraph describes how the spatial U-Net captures spatial details. Firstly, a convolutional layer with 3×3 kernels is employed to increase the dimension of $\mathcal{A}$, generating the input for the spatial U-Net. Then, in the initial four stages of the spatial U-shape network, we utilize ResNet blocks (ResBlock) to extract spatial information. Each ResBlock comprises two convolutional layers with 3×3 kernels and a layer of leaky rectified linear units (LReLU). To avoid gradient disappearance, a skip connection is established between the input and output.

This paragraph explains how the spectral U-Net is utilized to construct O. Firstly, we up-sample $\mathcal{B}$ to obtain $\mathcal{B}^U$ with high spatial resolution. Next, we employ 3×3 convolution kernels to augment the channel number of $\mathcal{B}^U$, producing the input for the spectral U-Net. At each stage of the spectral U-shape network, the spectral feature map is combined with the spatial one using S2Block, resulting in a fused outcome. Then, we apply the multi-layer perception (MLP) to acquire spectral characteristics from the fused outcome while preserving spatial information. The MLP primarily consists of two fully connected layers and an LReLU layer. Similar to ResBlock, the input is linked to the output. Upon obtaining the output of the final stage, we utilize a convolutional layer with 3×3 kernels to reconstruct it into a feature map of the size $H\times W\times C$. The feature map is then added to $\mathcal{B}^U$, generating the desired O.

3.3 S2Block

Most existing image fusion approaches apply concatenation to integrate spatial and spectral information. However, this operation disregards the distinctions between the two types of images, resulting in unsatisfactory fusion outcomes. To overcome this limitation, we develop a novel structure called S2Block, as illustrated in Fig. 4, for effective spatial-spectral integration. Next, we will take the S2Block in the first stage of the spectral U-Net as an example to explain this structure.

The spatial and spectral feature maps input to the S2Block are denoted as $\mathcal{F}^{spa}\in\mathbb{R}^{H\times W\times S}$ and $\mathcal{F}^{spe}\in\mathbb{R}^{H\times W\times S}$. For $\mathcal{F}^{spa}$, we first reshape it into a matrix denoted $\mathcal{M}^{spa}\in\mathbb{R}^{HW\times S}$, with each row representing the feature vector of a particular spatial location. Then, the $\mathcal{M}^{spa}$ is simultaneously processed by two parallel fully connected layers, producing two matrices of the same size. To better utilize the information on feature vectors, we divide each matrix evenly into several smaller parts by column. Technically, the matrices are reshaped into two tensors, denoted as $\mathcal{T}^a\in\mathbb{R}^{HW\times S'\times N}$ and $\mathcal{T}^b\in\mathbb{R}^{HW\times S'\times N}$, where N is the number of small parts and $S'=\frac{S}{N}$. For $\mathcal{F}^{spe}$, we first reshape it into a matrix denoted $\mathcal{M}^{spe}\in\mathbb{R}^{HW\times S}$. Similarly, the $\mathcal{M}^{spe}$ is transformed into two tensors, denoted as $\mathcal{T}^c\in\mathbb{R}^{HW\times S'\times N}$ and $\mathcal{T}^d\in\mathbb{R}^{HW\times S'\times N}$.

This paragraph explains the spatial-spectral integration operation (SSIO), which is the core of S2Block. In SSIO, we first produce two sets of square matrices, namely *spatial self-correlation matrices* and *spectral self-correlation matrices*, to accurately depict spatial details and spectral characteristics. Next, a series of operators are utilized to merge the square matrices with the spatial and spectral data, resulting in a fusion outcome. Specifically, we represent the set of spatial self-correlation matrices as $C^{spa}\in\mathbb{R}^{HW\times HW\times N}$, and the production of its i^{th} square matrix is expressed as:

$$C^{spa}(\cdot,\cdot,i)=\text{Softmax}\Big(\frac{\mathcal{T}^a(\cdot,\cdot,i)\{\mathcal{T}^b(\cdot,\cdot,i)\}^T}{\sqrt{S'}}\Big), \tag{1}$$

where $C^{spa}(\cdot,\cdot,i)\in\mathbb{R}^{HW\times HW}$ denotes the i^{th} matrix of C^{spa}. $\mathcal{T}^a(\cdot,\cdot,i)\in\mathbb{R}^{HW\times S'}$ and $\mathcal{T}^b(\cdot,\cdot,i)\in\mathbb{R}^{HW\times S'}$ represent the i^{th}

Table 1: Quantitative results on 20 reduced-resolution and 20 full-resolution samples of WV3. (Red: best; Blue: second best).

Method	Reduced-Resolution				Full-Resolution		
	PSNR(±std)	Q8(±std)	SAM(±std)	ERGAS(±std)	D_λ(±std)	D_s(±std)	QNR(±std)
BT-H [1]	33.080±2.880	0.832±0.094	4.920±1.425	4.580±1.496	0.0574±0.0232	0.0810±0.0374	0.8670±0.0540
TV [18]	32.381±2.328	0.795±0.120	5.692±1.808	4.855±1.434	0.0234±0.0061	0.0393±0.0227	0.9383±0.0269
MTF-GLP-HPM [25]	33.095±2.800	0.835±0.092	5.333±1.761	4.616±1.503	0.0206±0.0082	0.0630±0.0284	0.9180±0.0346
MTF-GLP-FS [24]	32.963±2.753	0.833±0.092	5.315±1.765	4.700±1.597	0.0197±0.0078	0.0630±0.0289	0.9187±0.0347
BDSD-PC [21]	32.970±2.784	0.829±0.097	5.428±1.822	4.697±1.617	0.0625±0.0235	0.0730±0.0356	0.8698±0.0531
PNN [8]	37.313±2.646	0.893±0.092	3.677±0.762	2.681±0.647	0.0213±0.0080	0.0428±0.0147	0.9369±0.0212
PanNet [34]	37.346±2.688	0.891±0.093	3.613±0.766	2.664±0.688	**0.0165**±0.0074	0.0470±0.0210	0.9374±0.0271
MSDCNN [29]	37.068±2.686	0.890±0.090	3.777±0.803	2.760±0.689	0.0230±0.0091	0.0467±0.0199	0.9316±0.0271
DiCNN [9]	37.390±2.761	0.900±0.087	3.592±0.762	2.672±0.662	0.0362±0.0111	0.0462±0.0175	0.9195±0.0258
BDPN [37]	36.191±2.702	0.871±0.100	4.201±0.857	3.046±0.732	0.0364±0.0142	0.0459±0.0192	0.9196±0.0308
FusionNet [3]	38.047±2.589	0.904±0.090	3.324±0.698	2.465±0.644	0.0239±0.0090	0.0364±0.0137	0.9406±0.0197
MUCNN [27]	38.262±2.703	0.911±0.089	3.206±0.681	2.400±0.617	0.0258±0.0111	0.0327±0.0140	0.9424±0.0205
LAGNet [14]	38.592±2.778	0.910±0.091	3.103±0.558	2.292±0.607	0.0368±0.0148	0.0418±0.0152	0.9230±0.0247
PMACNet [16]	38.595±2.882	0.912±0.092	3.073±0.623	2.293±0.532	0.0540±0.0232	0.0336±0.0115	0.9143±0.0281
U2Net	**39.117**±3.009	**0.920**±0.085	**2.888**±0.581	**2.149**±0.525	0.0178±0.0072	**0.0313**±0.0075	**0.9514**±0.0115
Ideal value	+∞	**1**	**0**	**0**	**0**	**0**	**1**

matrices in $\mathcal{T}^a$ and $\mathcal{T}^b$, respectively. Besides, T defines the transpose operation and Softmax$(\cdot)$ stands for the Softmax function. The spatial self-correlation matrices offer a concrete and intuitive representation of spatial information, as each value of $\mathcal{C}^{spa}(\cdot,\cdot,i)$ signifies the similarity between two spatial locations in $\mathcal{A}$. The set of spectral self-correlation matrices is represented as $\mathcal{C}^{spe} \in \mathbb{R}^{S'\times S'\times N}$, and we express its i^{th} matrix as:

$$\mathcal{C}^{spe}(\cdot,\cdot,i) = \text{Softmax}\left(\frac{\{\mathcal{T}^c(\cdot,\cdot,i)\}^T\mathcal{T}^d(\cdot,\cdot,i)}{\frac{\sqrt{(S')^3}}{HW}}\right), \tag{2}$$

where $\mathcal{C}^{spe}(\cdot,\cdot,i) \in \mathbb{R}^{S'\times S'}$ denotes the i^{th} square matrix of $\mathcal{C}^{spe}$. $\mathcal{T}^c(\cdot,\cdot,i) \in \mathbb{R}^{HW\times S'}$ and $\mathcal{T}^d(\cdot,\cdot,i) \in \mathbb{R}^{HW\times S'}$ stand for the i^{th} matrices in $\mathcal{T}^c$ and $\mathcal{T}^d$. Since each value of $\mathcal{C}^{spe}(\cdot,\cdot,i)$ represents the similarity between two channels of $\mathcal{B}$, the spectral self-correlation matrices provide a tangible and intuitive description of spectral characteristics. Upon obtaining the $\mathcal{C}^{spa}(\cdot,\cdot,i)$ and $\mathcal{C}^{spe}(\cdot,\cdot,i)$, we combine them with the spatial and spectral data, expressed as:

$$\mathcal{T}^{fus}(\cdot,\cdot,i) = \{\mathcal{C}^{spa}(\cdot,\cdot,i)\mathcal{T}^c(\cdot,\cdot,i)\} \odot \{(\mathcal{T}^b(\cdot,\cdot,i)\mathcal{C}^{spe}(\cdot,\cdot,i)\}, \tag{3}$$

where $\mathcal{T}^{fus} \in \mathbb{R}^{HW\times S'\times N}$ denotes the fused output that contains both spatial and spectral information. $\mathcal{T}^{fus}(\cdot,\cdot,i) \in \mathbb{R}^{HW\times S'}$ represents the i^{th} matrix in $\mathcal{T}^{fus}$. Additionally, $\odot$ defines the element-wise multiplication. Compared with other fusion techniques like concatenation, the SSIO enables effective and comprehensive integration of spatial details and spectral characteristics.

After acquiring $\mathcal{T}^{fus}$, we reshape it into a fusion matrix, denoted as $\mathcal{M}^{fus} \in \mathbb{R}^{HW\times S}$. Subsequently, we employ a fully connected layer to process the $\mathcal{M}^{fus}$ and convert it into a spatial-spectral-integrated feature map, represented as $\mathcal{F}^{fus} \in \mathbb{R}^{H\times W\times S}$.

Besides, please refer to the *Sup. Mat.* for a comprehensive explanation regarding the relationship between our S2Block and multi-head attention in Transformer [20].

3.4 Loss Function

The main contributions of this work focus on the network architecture, thus we only employ the commonly used ℓ_1 loss function for network training, shown as follows:

$$\mathcal{L}oss = \frac{1}{M}\sum_{m=1}^{M} \|f_\Theta(\mathcal{A}^{\{m\}}, \mathcal{B}^{\{m\}}) - \mathcal{X}^{\{m\}}\|_1, \tag{4}$$

where $\mathcal{A}^{\{m\}}$, $\mathcal{B}^{\{m\}}$, and $\mathcal{X}^{\{m\}}$ represent the m^{th} PAN/RGB image, LRMS/LRHS image, and GT image in the training dataset. $f_\Theta(\cdot)$ denotes the U2Net with learnable parameters Θ, and M is the total number of training examples. Besides, $\|\cdot\|_1$ defines the ℓ_1 norm.

4 EXPERIMENTS FOR PANSHARPENING

To demonstrate the effectiveness of our method, we conduct a series of experiments on datasets acquired by WorldView-3 (WV3) and WorldView-2 (WV2) satellites. The U2Net is compared with several recent SOTA pansharpening approaches.

4.1 Experiment Settings

Datasets. For the pansharpening problem, we train the DL-based methods on a dataset acquired by WV3 which contains 10000 samples (90% for training and 10% for validation). Each sample consists of a PAN/LRMS/GT image pair of sizes 64×64, $16\times 16\times 8$, and $64\times 64\times 8$, respectively. The PAN images have a spatial resolution of 0.3m, while the LRMS images have a spatial resolution of 1.2m. Additionally, the LRMS bands comprise four standard colors (RGB and near-infrared 1) and four new bands (coastal, yellow, red edge, and near-infrared 2). We compare our U2Net with representative pansharpening approaches using various datasets acquired by WV3 and WV2. The testing datasets are categorized into two classes, *i.e.*, the reduced-resolution datasets and the full-resolution

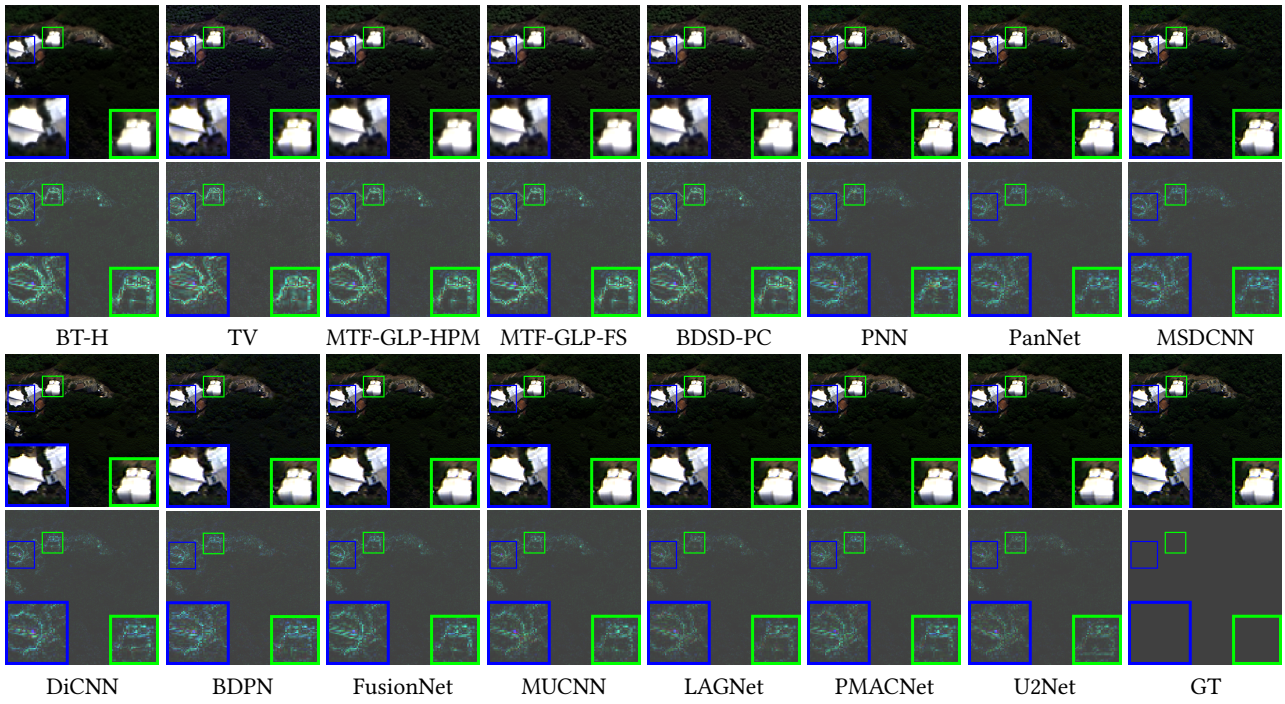


Figure 5: Qualitative evaluation results on a reduced-resolution sample acquired by WV3. The natural color maps are presented in the first and third rows, while the corresponding AEMs are listed in the second and fourth rows.

datasets. The former includes PAN/LRMS/GT image pairs with dimensions 256×256, $64 \times 64 \times 8$, and $256 \times 256 \times 8$, while the latter consists of PAN/LRMS image pairs of sizes 512×512 and $128 \times 128 \times 8$. All datasets used in this section are from the PanCollection proposed by [4]. The PanCollection offers multiple pansharpening datasets, accompanied by detailed descriptions of data simulation, and can be downloaded from this website[1].

Benchmarks. We compare our method with recent SOTA works consisting of five traditional approaches: BT-H [1], TV [18], MTF-GLP-HPM [25], MTF-GLP-FS [24], and BDSD-PC [21]; and nine DL-based methods: PNN [8], PanNet [34], MSDCNN [29], DiCNN [9], BDPN [37], FusionNet [3], MUCNN [27], LAGNet [14], and PMACNet [16]. To ensure fairness, we train DL-based methods using the same Nvidia GPU-3090 and PyTorch environment.

Evaluation Metrics. Following the research standard of pansharpening, we utilize four metrics to evaluate the results on reduced-resolution datasets, including PSNR, Q8 [7], SAM, and ERGAS [26]. As for full-resolution datasets, we apply D_λ, D_s, and QNR indexes [23] for evaluation.

Parameters Tuning. For the pansharpening task, we set the values of S and S' in our network to 32 and 16, respectively. Additionally, the value of N depends on the S and S'. On training the U2Net, the initial learning rate, epoch, and batch size are set to 0.001, 360, and 16, respectively. We select Adam as the optimizer, and the learning rate is reduced by half every 100 epochs. As for other DL-based methods, we utilize the default settings in related papers or codes to train the networks.

[1]https://github.com/liangjiandeng/PanCollection

4.2 Results on WV3 Datasets

Reduced-Resolution Assessment. We assess the performances of representative approaches and our method, using 20 reduced-resolution samples acquired by WV3. The quantitative evaluation outcomes are presented in Tab. 1, and the proposed method obtains the best average results on all quality indexes. Additionally, the qualitative evaluation outcomes on one of the 20 samples are shown in Fig. 5, alongside the GT. As the darker absolute error map (AEM) indicates a better result, our U2Net outperforms other approaches. The experimental outcomes above demonstrate that our method is superior to recent SOTA pansharpening works.

Full-Resolution Assessment. To prove the practical usefulness of our method, we conduct experiments on 20 full-resolution samples acquired by WV3. The quantitative evaluation results are presented in Tab. 1. The U2Net achieves the best overall performance, proving the high application value of our method.

4.3 Generalization

Generalization ability is a crucial concern for DL-based methods in the pansharpening task. If there is a significant difference between the testing and training datasets, some approaches may not perform well. We use 20 reduced-resolution samples acquired by WV2 to test all DL-based models trained on the WV3 dataset. The quantitative evaluation outcomes are presented in Tab. 2, and the U2Net yields the best results on all four metrics, indicating the strong generalization capability of our method. Notably, the inflexible and unreasonable structure of PMACNet [16] significantly

Table 2: Quantitative evaluation results of DL-based methods on 20 reduced-resolution samples acquired by WV2. Section 4.3 explains the unsatisfactory outcomes of the PMACNet. (Red: best; Blue: second best).

Method	PSNR(±std)	Q8(±std)	SAM(±std)	ERGAS(±std)
PNN	28.045±1.865	0.762±0.093	7.115±1.682	5.615±0.943
PanNet	30.276±2.290	0.840±0.080	5.495±0.713	4.337±0.520
DiCNN	27.200±2.327	0.721±0.075	6.921±0.788	6.251±0.574
MSDCNN	29.441±2.227	0.824±0.080	6.006±0.638	4.744±0.494
BDPN	28.973±1.714	0.824±0.093	7.089±0.864	4.856±0.570
FusionNet	28.735±2.460	0.796±0.074	6.426±0.860	5.136±0.515
MUCNN	27.839±2.328	0.777±0.088	7.504±0.539	5.517±0.299
LAGNet	28.050±2.239	0.805±0.084	6.955±0.474	5.326±0.318
PMACNet	19.160±4.512	0.509±0.128	15.95±3.329	15.69±3.307
U2Net	**30.740**±2.173	**0.849**±0.085	**5.250**±0.545	**4.070**±0.392

Table 3: Ablation study on 20 reduced-resolution samples acquired by WV2. (Red: best; Blue: second best).

Method	PSNR(±std)	Q8(±std)	SAM(±std)	ERGAS(±std)
V1	29.849±2.171	0.830±0.087	5.773±0.731	4.512±0.740
V2	30.295±2.324	0.839±0.083	5.520±0.634	4.281±0.380
V3	30.394±2.380	0.841±0.081	**5.165**±0.610	4.248±0.376
V4	30.104±2.246	0.848±0.086	5.575±0.691	4.380±0.535
U2Net	**30.740**±2.173	**0.849**±0.085	5.250±0.545	**4.070**±0.392

restricts its generalization ability, leading to extremely unsatisfactory outcomes.

4.4 Comparison of Parameter Numbers

We categorize the DL-based pansharpening methods into two groups based on their number of parameters (NoPs). Specifically, models with less than 1×10^5 parameters are considered lightweight networks, whereas those with more than 5×10^5 parameters are classified as heavyweight networks. The U2Net is a heavyweight network, which prompts us to develop a lightweight version called U2Net-L to demonstrate the superiority of our method more effectively. To ensure fairness, we compare U2Net-L with lightweight networks and U2Net with heavyweight networks. Fig. 6 shows the comparisons of NoPs on 20 reduced-resolution samples acquired by WV3. Both U2Net-L and U2Net achieve exceptional performance within their respective categories, demonstrating the superiority of our framework. For more details, kindly refer to the *Sup. Mat.*

4.5 Ablation Study

To validate the effectiveness of our method, we create four variants of the U2Net. In the first variant (V1), we employ a single-branch U-shape network to extract spatial and spectral features uniformly while maintaining the original structure of the S2Block. The purpose of V1 is to demonstrate that the double-branch network is more effective in capturing diverse information compared

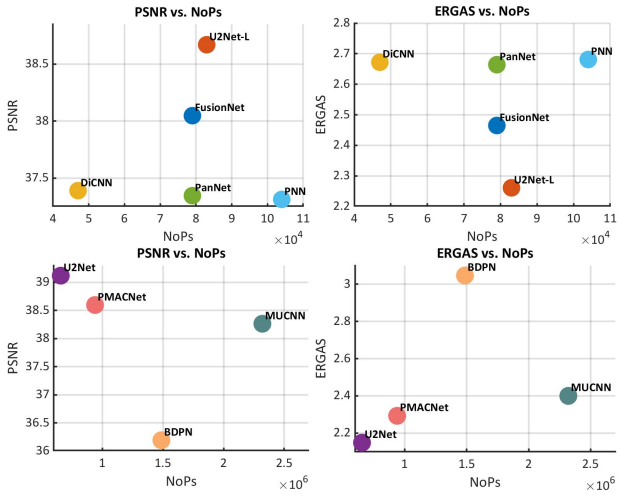


Figure 6: The comparisons of NoPs. The first row: comparisons of lightweight networks ($\leq 1 \times 10^5$ parameters) based on PSNR and ERGAS indexes. The second row: comparisons of heavyweight networks ($\geq 5 \times 10^5$ parameters) based on the same quality indexes.

to the single-branch one. The second variant (V2) retains the double U-shape network architecture but replaces the S2Blocks with concatenation operations. The V2 is designed to confirm the superiority of S2Blocks in information integration. In the third variant (V3), the S2Block only produces spatial self-correlation matrices and combines them with the spectral feature map. As for the fourth variant (V4), only spectral self-correlation matrices are generated and merged with the spatial feature map.

We perform experiments on 20 reduced-resolution samples acquired by WV2. The results are presented in Tab. 3, and the U2Net yields the best overall performance, proving the effectiveness of our method. Further explanation and discussion on the ablation study can be found in the *Sup. Mat.*

5 EXPERIMENTS FOR HISR

5.1 Experiment Settings

Datasets. For the HISR task, experiments are conducted on the CAVE dataset [35], which contains 31 RGB/LRHS image pairs with sizes $512 \times 512 \times 3$ and $512 \times 512 \times 31$. We select 20 samples for training, and the rest are for testing. The 20 training samples are made into 3920 overlapped RGB/LRHS/GT image pairs (80% for training and 20% for validation) with sizes $64 \times 64 \times 3$, $16 \times 16 \times 31$, and $64 \times 64 \times 31$, while the testing samples are processed as 11 RGB/LRHS/GT image pairs with sizes $512 \times 512 \times 3$, $128 \times 128 \times 31$, and $512 \times 512 \times 31$.

Benchmarks and Evaluation Metrics. We compare our method with some recent SOTA approaches, including five traditional methods: CSTF [15], LTMR [5], LTTR [6], UTV [32], and IR-TenSR [31]; and three DL-based works: ResTFNet [17], SSRNet [36], and Fusformer [11]. Four commonly used metrics are selected, including PSNR, SSIM [28], SAM, and ERGAS [26].

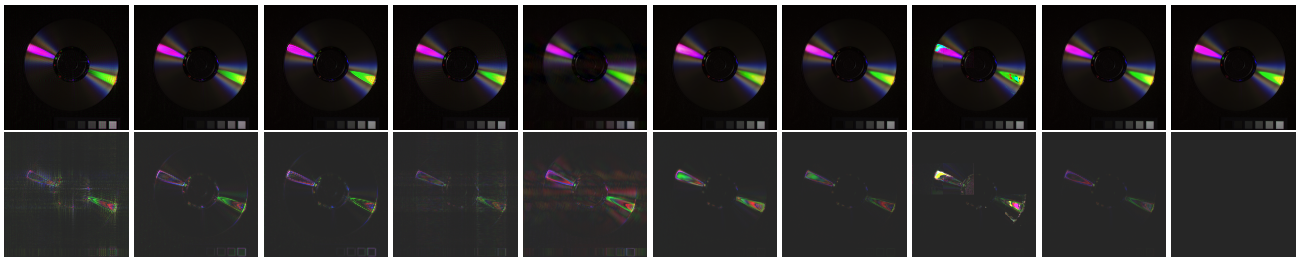
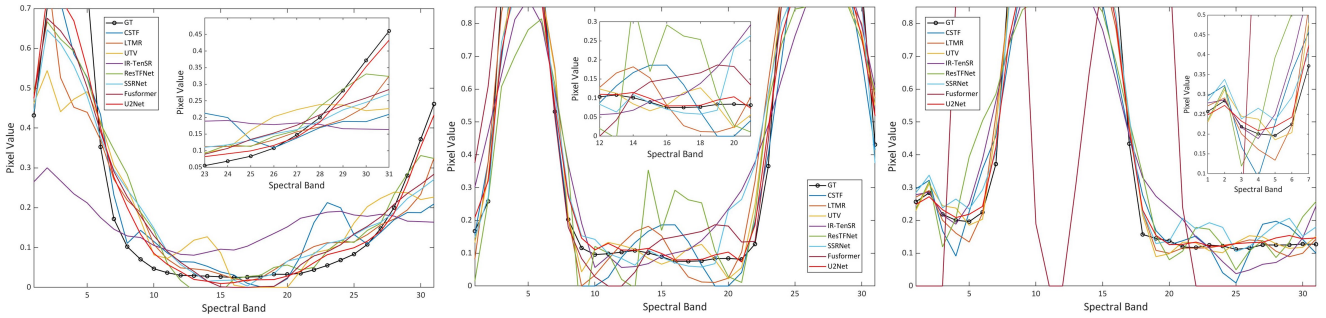


Figure 7: Qualitative evaluation results on a CAVE testing sample. The first row: natural color maps. The second row: AEMs.



Spectral vectors at (293, 233) Spectral vectors at (194, 172) Spectral vectors at (286, 403)

Figure 8: The comparisons of spectral vectors from three spatial locations of a CAVE testing sample.

Table 4: Quantitative evaluation results on 11 testing samples of the CAVE dataset. (Red: best; Blue: second best).

Method	PSNR(±std)	SSIM(±std)	SAM(±std)	ERGAS(±std)
CSTF	34.463±4.281	0.866±0.075	14.368±5.302	8.289±5.285
LTMR	36.543±3.300	0.963±0.021	6.711±2.193	5.387±2.529
LTTR	35.851±3.488	0.956±0.029	6.990±2.554	5.990±2.921
UTV	38.615±4.064	0.941±0.043	8.649±3.376	4.519±2.817
IR-TenSR	35.608±3.446	0.945±0.027	12.295±4.683	5.897±3.046
ResTFNet	45.584±5.465	0.994±0.006	2.764±0.699	2.313±2.438
SSRNet	48.620±3.918	0.995±0.002	2.542±0.837	1.636±1.219
Fusformer	49.983±8.097	0.994±0.011	2.203±0.851	2.534±5.305
U2Net	**50.441**±4.403	**0.997**±0.002	**2.164**±0.609	**1.267**±0.967
Ideal value	+∞	**1**	**0**	**0**

Parameters Tuning. For the HISR task, we set the values of S and S' in our network to 64 and 16, respectively. Upon training the U2Net, the initial learning rate, epoch, and batch size are set to 0.0003, 500, and 8, respectively. Additionally, we choose Adam as the optimizer, and the learning rate is halved every 50 epochs.

5.2 Results on the Cave Dataset

We assess the performance of recent SOTA approaches and our method on 11 testing samples of the CAVE dataset. The quantitative evaluation outcomes are presented in Tab. 4, and the U2Net achieves the best average results on all quality indicators. Additionally, the qualitative evaluation outcomes are shown in Fig. 7 together with the GT. Obviously, our method exhibits the darkest AEM, proving its superiority in the HISR task. Furthermore, in Fig. 8, we display the spectral vectors from three different spatial locations of a testing sample. The spectral vectors of U2Net are the closest to the GT, indicating that our method has a potent spectral preservation ability.

6 CONCLUSION

In this paper, we propose a spatial-spectral-integrated double U-shape network called U2Net for image fusion tasks. The U2Net employs a spatial U-Net and a spectral U-Net to extract spatial details and spectral characteristics discriminately and hierarchically. Besides, we create a novel structure named S2Block that sufficiently merges feature maps from diverse images in a logical and comprehensive manner. We compare our U2Net with several recent SOTA pansharpening and HISR approaches. The proposed method outperforms all others on a series of datasets, demonstrating its exceptional feature learning, information integration, and generalization capabilities. Therefore, we are confident that our method offers an effective solution for the image fusion problems.

ACKNOWLEDGMENTS

This research is supported by NSFC (12271083), Natural Science Foundation of Sichuan Province (2022NSFSC0501).

REFERENCES

[1] Bruno Aiazzi, L Alparone, Stefano Baronti, Andrea Garzelli, and Massimo Selva. 2006. MTF-tailored multiscale fusion of high-resolution MS and Pan imagery. *Photogrammetric Engineering & Remote Sensing* 72, 5 (2006), 591–596.

[2] Jaewan Choi, Kiyun Yu, and Yongil Kim. 2010. A New Adaptive Component-Substitution-Based Satellite Image Fusion by Using Partial Replacement. *IEEE Transactions on Geoscience and Remote Sensing* (2010).

[3] Liang-Jian Deng, Gemine Vivone, Cheng Jin, and Jocelyn Chanussot. 2021. Detail Injection-Based Deep Convolutional Neural Networks for Pansharpening. *IEEE Transactions on Geoscience and Remote Sensing* 59, 8 (2021), 6995–7010. https://doi.org/10.1109/TGRS.2020.3031366

[4] Liang-jian Deng, Gemine Vivone, Mercedes E. Paoletti, Giuseppe Scarpa, Jiang He, Yongjun Zhang, Jocelyn Chanussot, and Antonio Plaza. 2022. Machine Learning in Pansharpening: A benchmark, from shallow to deep networks. *IEEE Geoscience and Remote Sensing Magazine* 10, 3 (2022), 279–315. https://doi.org/10.1109/MGRS.2022.3187652

[5] Renwei Dian and Shutao Li. 2019. Hyperspectral image super-resolution via subspace-based low tensor multi-rank regularization. *IEEE Transactions on Image Processing* 28, 10 (2019), 5135–5146.

[6] Renwei Dian, Shutao Li, and Leyuan Fang. 2019. Learning a Low Tensor-Train Rank Representation for Hyperspectral Image Super-Resolution. *IEEE Transactions on Neural Networks and Learning Systems* 30, 9 (2019), 2672–2683. https://doi.org/10.1109/TNNLS.2018.2885616

[7] Andrea Garzelli and Filippo Nencini. 2009. Hypercomplex Quality Assessment of Multi/Hyperspectral Images. *IEEE Geoscience and Remote Sensing Letters* 6, 4 (2009), 662–665. https://doi.org/10.1109/LGRS.2009.2022650

[8] Masi Giuseppe, Cozzolino Davide, Verdoliva Luisa, and Scarpa Giuseppe. 2016. Pansharpening by Convolutional Neural Networks. *Remote Sensing* 8, 7 (2016), 594.

[9] Lin He, Yizhou Rao, Jun Li, Jocelyn Chanussot, Antonio Plaza, Jiawei Zhu, and Bo Li. 2019. Pansharpening via Detail Injection Based Convolutional Neural Networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12, 4 (2019), 1188–1204. https://doi.org/10.1109/JSTARS.2019.2898574

[10] Xiyan He, Laurent Condat, José M. Bioucas-Dias, Jocelyn Chanussot, and Junshi Xia. 2014. A New Pansharpening Method Based on Spatial and Spectral Sparsity Priors. *IEEE Transactions on Image Processing* 23, 9 (2014), 4160–4174. https://doi.org/10.1109/TIP.2014.2333661

[11] Jin-Fan Hu, Ting-Zhu Huang, Liang-Jian Deng, Hong-Xia Dou, Danfeng Hong, and Gemine Vivone. 2022. Fusformer: A Transformer-Based Fusion Network for Hyperspectral Image Super-Resolution. *IEEE Geoscience and Remote Sensing Letters* 19 (2022), 1–5. https://doi.org/10.1109/LGRS.2022.3194257

[12] Alex Pappachen James and Belur V Dasarathy. 2014. Medical image fusion: A survey of the state of the art. *Information fusion* 19 (2014), 4–19.

[13] Zi-Rong Jin, Liang-Jian Deng, Tian-Jing Zhang, and Xiao-Xu Jin. 2021. BAM: Bilateral Activation Mechanism for Image Fusion. In *Proceedings of the 29th ACM International Conference on Multimedia* (Virtual Event, China) *(MM '21)*. Association for Computing Machinery, New York, NY, USA, 4315–4323. https://doi.org/10.1145/3474085.3475571

[14] Zi-Rong Jin, Tian-Jing Zhang, Tai-Xiang Jiang, Gemine Vivone, and Liang-Jian Deng. 2022. LAGConv: Local-context Adaptive Convolution Kernels with Global Harmonic Bias for Pansharpening. *AAAI Conference on Artificial Intelligence (AAAI)* (2022).

[15] Shutao Li, Renwei Dian, Leyuan Fang, and José M. Bioucas-Dias. 2018. Fusing Hyperspectral and Multispectral Images via Coupled Sparse Tensor Factorization. *IEEE Transactions on Image Processing* 27, 8 (2018), 4118–4130. https://doi.org/10.1109/TIP.2018.2836307

[16] Yixun Liang, Ping Zhang, Yang Mei, and Tingqi Wang. 2022. PMACNet: Parallel Multiscale Attention Constraint Network for Pan-Sharpening. *IEEE Geoscience and Remote Sensing Letters* 19 (2022), 1–5. https://doi.org/10.1109/LGRS.2022.3170904

[17] Xiangyu Liu, Qingjie Liu, and Yunhong Wang. 2018. Remote Sensing Image Fusion Based on Two-stream Fusion Network. *Information Fusion* (2018).

[18] Frosti Palsson, Johannes R Sveinsson, and Magnus O Ulfarsson. 2013. A new pansharpening algorithm based on total variation. *IEEE Geoscience and Remote Sensing Letters* 11, 1 (2013), 318–322.

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[21] Gemine Vivone. 2019. Robust Band-Dependent Spatial-Detail Approaches for Panchromatic Sharpening. *IEEE Transactions on Geoscience and Remote Sensing* 57, 9 (2019), 6421–6433. https://doi.org/10.1109/TGRS.2019.2906073

[22] Gemine Vivone, Luciano Alparone, Jocelyn Chanussot, Mauro Dalla Mura, Andrea Garzelli, Giorgio A Licciardi, Rocco Restaino, and Lucien Wald. 2014. A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing* 53, 5 (2014), 2565–2586.

[23] Gemine Vivone, Luciano Alparone, Jocelyn Chanussot, Mauro Dalla Mura, Andrea Garzelli, Giorgio A. Licciardi, Rocco Restaino, and Lucien Wald. 2015. A Critical Comparison Among Pansharpening Algorithms. *IEEE Transactions on Geoscience and Remote Sensing* 53, 5 (2015), 2565–2586. https://doi.org/10.1109/TGRS.2014.2361734

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

[25] Gemine Vivone, Rocco Restaino, Mauro Dalla Mura, Giorgio Licciardi, and Jocelyn Chanussot. 2014. Contrast and Error-Based Fusion Schemes for Multispectral Image Pansharpening. *IEEE Geoscience and Remote Sensing Letters* 11, 5 (2014), 930–934. https://doi.org/10.1109/LGRS.2013.2281996

[26] Lucien Wald. 2002. Data Fusion. Definitions and Architectures - Fusion of Images of Different Spatial Resolutions. *Presses des MINES* (2002).

[27] Yudong Wang, Liang-Jian Deng, Tian-Jing Zhang, and Xiao Wu. 2021. SSconv: Explicit Spectral-to-Spatial Convolution for Pansharpening. In *Proceedings of the 29th ACM International Conference on Multimedia* (Virtual Event, China) *(MM '21)*. Association for Computing Machinery, New York, NY, USA, 4472–4480. https://doi.org/10.1145/3474085.3475600

[28] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612. https://doi.org/10.1109/TIP.2003.819861

[29] Yancong Wei, Qiangqiang Yuan, Xiangchao Meng, Huanfeng Shen, Liangpei Zhang, and Michael Ng. 2017. Multi-scale-and-depth convolutional neural network for remote sensed imagery pan-sharpening. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. 3413–3416. https://doi.org/10.1109/IGARSS.2017.8127731

[30] Xiao Wu, Ting-Zhu Huang, Liang-Jian Deng, and Tian-Jing Zhang. 2021. Dynamic Cross Feature Fusion for Remote Sensing Pansharpening. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 14687–14696.

[31] Ting Xu, Ting-Zhu Huang, Liang-Jian Deng, and Naoto Yokoya. 2022. An Iterative Regularization Method based on Tensor Subspace Representation for Hyperspectral Image Super-Resolution. *IEEE Transactions on Geoscience and Remote Sensing* (2022).

[32] Ting Xu, Ting-Zhu Huang, Liang-Jian Deng, Xi-Le Zhao, and Jie Huang. 2020. Hyperspectral Image Superresolution Using Unidirectional Total Variation With Tucker Decomposition. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13 (2020), 4381–4398. https://doi.org/10.1109/JSTARS.2020.3012566

[33] Keyu Yan, Man Zhou, Jie Huang, Feng Zhao, Chengjun Xie, Chongyi Li, and Danfeng Hong. 2022. Panchromatic and Multispectral Image Fusion via Alternating Reverse Filtering Network. *arXiv preprint arXiv:2210.08181* (2022).

[34] Junfeng Yang, Xueyang Fu, Yuwen Hu, Yue Huang, Xinghao Ding, and John Paisley. 2017. PanNet: A Deep Network Architecture for Pan-Sharpening. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 1753–1761. https://doi.org/10.1109/ICCV.2017.193

[35] Fumihito Yasuma, Tomoo Mitsunaga, Daisuke Iso, and Shree K. Nayar. 2010. Generalized Assorted Pixel Camera: Postcapture Control of Resolution, Dynamic Range, and Spectrum. *IEEE Transactions on Image Processing* 19, 9 (2010), 2241–2253. https://doi.org/10.1109/TIP.2010.2046811

[36] Xueting Zhang, Wei Huang, Qi Wang, and Xuelong Li. 2021. SSR-NET: Spatial–Spectral Reconstruction Network for Hyperspectral and Multispectral Image Fusion. *IEEE Transactions on Geoscience and Remote Sensing* 59, 7 (2021), 5953–5965. https://doi.org/10.1109/TGRS.2020.3018732

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

[38] Man Zhou, Jie Huang, Keyu Yan, Gang Yang, Aiping Liu, Chongyi Li, and Feng Zhao. 2022. Normalization-Based Feature Selection and Restitution for Pan-Sharpening. In *Proceedings of the 30th ACM International Conference on Multimedia* (Lisboa, Portugal) *(MM '22)*. Association for Computing Machinery, New York, NY, USA, 3365–3374. https://doi.org/10.1145/3503161.3547774

[39] Yabin Zhu, Chenglong Li, Bin Luo, Jin Tang, and Xiao Wang. 2019. Dense Feature Aggregation and Pruning for RGBT Tracking. In *Proceedings of the 27th ACM International Conference on Multimedia* (Nice, France) *(MM '19)*. Association for Computing Machinery, New York, NY, USA, 465–472. https://doi.org/10.1145/3343031.3350928