Fully-Connected Transformer for Multi-Source Image Fusion

Xiao Wu, Zi-Han Cao, Ting-Zhu Huang, *Member, IEEE*, Liang-Jian Deng, *Senior Member, IEEE*, Jocelyn Chanussot, *Fellow, IEEE*, and Gemine Vivone, *Senior Member, IEEE*

***Abstract*—Multi-source image fusion combines the information coming from multiple images into one data, thus improving imaging quality. This topic has aroused great interest in the community. How to integrate information from different sources is still a big challenge, although the existing self-attention based transformer methods can capture spatial and channel similarities. In this paper, we first discuss the mathematical concepts behind the proposed generalized self-attention mechanism, where the existing self-attentions are considered basic forms. The proposed mechanism employs multilinear algebra to drive the development of a novel fully-connected self-attention (FCSA) method to fully exploit local and non-local domain-specific correlations among multi-source images. Moreover, we propose a multi-source image representation embedding it into the FCSA framework as a non-local prior within an optimization problem. Some different fusion problems are unfolded into the proposed fully-connected transformer fusion network (FC-Former). More specifically, the concept of generalized self-attention can promote the potential development of self-attention. Hence, the FC-Former can be viewed as a network model unifying different fusion tasks. Compared with state-of-the-art methods, the proposed FC-Former method exhibits robust and superior performance, showing its capability of faithfully preserving information.**

***Index Terms*—Transformer, multilinear algebra, model-driven neural network, multi-source image fusion, multispectral and hyperspectral image fusion, remote sensing pansharpening, visible and infrared image fusion.**

Received 20 November 2023; revised 5 October 2024; accepted 12 December 2024. Date of current version 5 February 2025. This work was supported in part by the NSFC under Grant 12171072, Grant 12271083, in part by the Natural Science Foundation of Sichuan Province under Grant 2024NSFSC0038, and in part by the National Key Research and Development Program of China under Grant 2020YFA0714001. Recommended for acceptance by J. Wang. *(Corresponding authors: Ting-Zhu Huang; Liang-Jian Deng.)*

Xiao Wu, Zi-Han Cao, Ting-Zhu Huang, and Liang-Jian Deng are with the School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: wxwsx1997@gmail.com; iamzihan666@gmail.com; tingzhuhuang@126.com; liangjian.deng@uestc.edu.cn).

Jocelyn Chanussot is with the Inria, CNRS, Grenoble INP, LJK, Université Grenoble Alpes, 38000, Grenoble, France, and also with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100045, China (e-mail: jocelyn.chanussot@grenoble-inp.fr).

Gemine Vivone is with the Institute of Methodologies for Environmental Analysis, CNR-IMAA, 85050 Tito Scalo, Italy, and also with the National Biodiversity Future Center (NBFC), 90133 Palermo, Italy (e-mail: gemine.vivone@imaa.cnr.it).

Our code is available at https://github.com/XiaoXiao-Woo/FC-Former.

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2024.3523364, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2024.3523364

I. Introduction

THE use of deep learning technology for the analysis and processing of biomedical and image information has become an important research direction [1], [2], [3]. In the field of multi-source image fusion, there is widely application in various image processing problems, such as image fusion [4], [5], [6], image denoising and reconstruction [7], [8], [9], image enhancement [10], and further applied to high-level computer vision tasks, such as classification [11], object detection [12], [13], and medical diagnosis [14]. Differently from the incomplete information that can be captured by a single device, a multi-source imaging system can better describe the information in the scene, e.g., combining visible and hyperspectral images, or thermal infrared images in night scenes, as well as panchromatic and multispectral data, digital images, etc. Hence, multi-source image fusion (MSIF) can be divided into several research fields, such as multispectral and hyperspectral image fusion (MHIF) [15], [16], visible and infrared image fusion (VIS-IR) [17], [18], remote sensing pansharpening [19], [20], [21], [22], multi-focus image fusion, and multi-exposure image fusion. The fused image preserves spatial information and spectral images for MHIF, remote sensing pansharpening, while for VIS-IR, digital photographic image fusion, the complementary features of the two images are fused to avoid the influence of the shooting environment on the camera.

Recently, deep-learning techniques have obtained increasing attention, clearly outperforming the latest model-based methods [27], [28]. Classic CNN-based methods [29], [30] adopt single scale [31] or multi-scale structures [32], [33], [34] to learn high-quality information for various vision tasks. However, in the aforementioned approaches, the network structure determines whether the information in the data can be fully extracted.

Researchers have also devoted attention to model-driven neural network techniques that offer both good interpretability and superior generalization capabilities getting state-of-the-art results. Model unfolding methods [35], [36] represent an example in this class. These approaches involve the transformation of a linear observation model through a certain variant replacement (i.e., the half-quadratic splitting (HQS) [37], [38] and the alternate direction multiplier method (ADMM) algorithm [39]). Afterwards, the transformed model is converted into a learnable network structure, thus endowing the traditional method with a nonlinear representation [40], [41]. As prior knowledge, the deep [42] and autoencoder priors [43] impose local priors. A

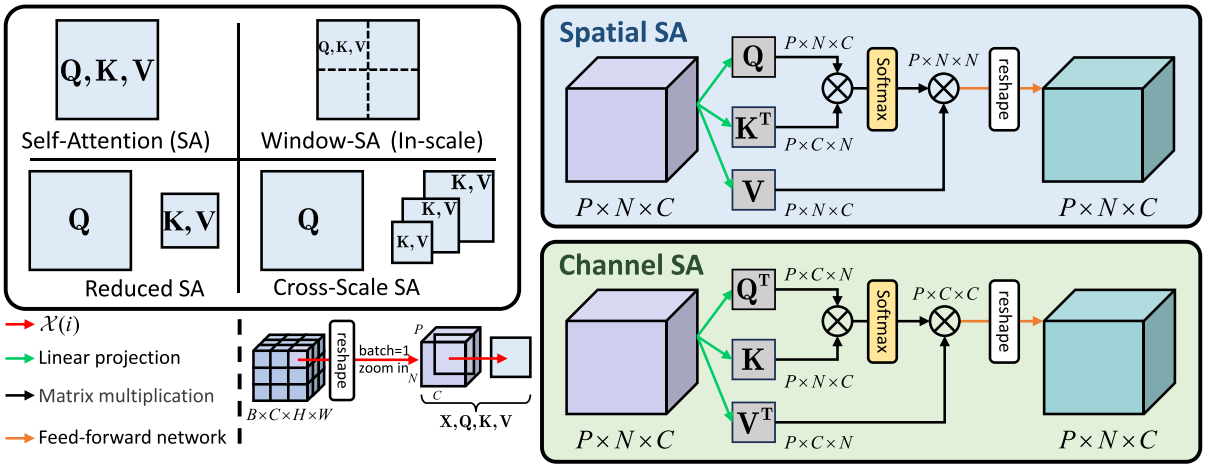


Fig. 1. The current existing forms for self-attention along spatial or channel modes. They are built by matrix multiplication, connecting all the other elements in one mode. To illustrate the information representation of self-attention, we show four key variants: self-attention (SA) [23], window-SA [24], reduced SA [25], and cross-scale SA [26]. For multi-source image fusion tasks, cross-scale SA can process mutual fusion at different scales. For example, the *Query*, $\mathbf{Q}$, can be the LR-HSI, then the matrices $\mathbf{K}$ and $\mathbf{V}$ can both be the HR-MSI. Hence, the SA can retain domain-specific information from different domains, while simultaneously disregarding the two internal source paradigms across scales.

non-local method has been proposed in [44] using non-local priors for model-driven neural networks.

However, the aforementioned single-scale networks lack contextual guidance for feature representations. In contrast, multi-scale networks always reduce the spatial resolution of features in the process of feature extractions using skip-connections to compensate for information loss, thus failing to achieve the expected feature representation for the multi-source image fusion task. Another shortcoming is that CNN-based methods have limited receptive fields and feature representation ability due to static kernels for feature extraction [45], [46]. Recent exploration into the self-attention (SA) mechanism within transformers, as elaborated by Vaswani et al. [23], seeks to unveil latent non-local relationships across specific dimensions (or modes). More specifically, transformer-based methods [24], [47], [48] exploit corresponding non-local information by computing the response of a given pixel along a specific dimension (or mode). Transformer methods, proposed in the field of multi-source image fusion [49], [50], capture domain-related non-local information in both spatial and spectral domains. However, the quality of fused images is limited due to a lack of multi-dimensional information. Therefore, researchers developed various forms of self-attention and performed matrix multiplication among three factors (*Query*, *Key*, and *Value*) along different dimensions (or modes) within intra-scales (aka in-scale) and cross-scales, i.e., spatial self-attention, channel self-attention, and hybrid self-attention, as shown in Fig. 1. Regarding spatial in-scale self-attention, each spatial element is connected to all other elements while integrating channel information, without being able to perceive the channel information of each element. Besides, some hybrid self-attention methods combined different vertical and horizontal self-attention paths to model pixel relations in all dimensions [51], [52]. Since the in-scale self-attention is intrinsic similarity, it cannot learn cross-scale patch similarity, leading to reduced accuracy. Accordingly, Mei et al. [53] explored in-scale and cross-scale self-attention in an independent connection module. Zhou et al. [54] proposed a cross-scale

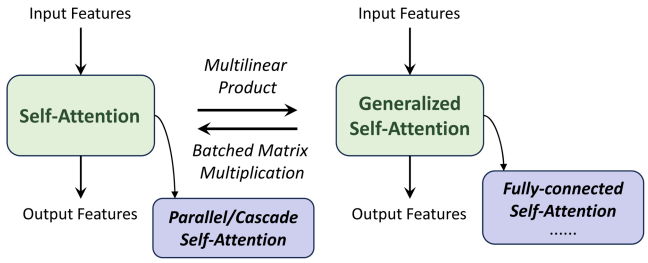


Fig. 2. The comparison between the existing self-attention mechanism and the proposed fully-connected self-attention framework based on the proposed generalized self-attention scheme.

self-attention to build spatial similarity in two feature resolutions of the image. Instead, NLRN [26] directly adopts a non-local framework as soft block matching, and euclidean distance with a kernel function to measure the spatial self-similarity. They just verify that the cross-scale patch similarity widely exists in a single dimension (mode) of the images.

Although the above-mentioned papers provided relevant contributions, they show some shortcomings in feature representation. On one hand, self-attention just achieves preliminary similarities for one or more unfolded dimensions (modes). This leads to a lack of multi-dimensional information. On the other hand, in-scale and cross-scale self-attentions are independent, and thus not unified in a mathematical mechanism.

In this work, we derive a generalized version of self-attention from the computational process of self-attention in terms of multilinear algebra [55], [56]. Based on the proposed generalized self-attention mechanism, the form of self-attention can be further extended by getting the so-called fully-connected self-attention (FCSA). Fig. 2 depicts the relationship between the proposal and the existing self-attention mechanism. Afterwards, we present a novel architecture for the task of multi-source image fusion (MSIF), i.e., the fully-connected transformer (FC-Former). The proposed FC-Former adopts three

Fig. 3. Schematic illustration of the different MSIF tasks, including multispectral and hyperspectral image fusion, visible and infrared image fusion, and remote sensing pansharpening.

parallel branches for cross-scale fusion, where one branch retains the same resolution as the HR-MSI and serves as the main branch. One of the remaining two branches has the same spatial resolution as the LR-HSI image, and the last one is twice that of the LR-HSI image. The FCSA module successively calculates the in-scale channel self-attention from each branch and performs cross-scale spatial self-attention among branches. Compared with previously developed advanced hybrid non-local self-attention and transformer methods [53], [57], the proposed FCSA method implements the characteristics of intra-branch and inter-branch non-local self-similarity (NSS) into feature maps connecting each dimension (mode) to learn multi-dimensional information from these feature maps via multilinear products. Moreover, it also achieves intra-scale and cross-scale feature aggregation for MSIF tasks. Overall, our FC-Former can fully consider the differences among features from multi-source images. The obtained feature extraction enables the network to capture more details, leading to more faithful, accurate, and high-quality reconstructions.

The main contributions of this paper are summarized as follows:

1) The proposed generalized self-attention provides a unified framework relied upon a multilinear product among three factors for the existing self-attention mechanisms.
2) We propose a novel fully-connected self-attention framework (FCSA). The FCSA framework overcomes the limitations of self-attention in terms of multi-dimensional and domain-related characterizations. The proposed FCSA method can fully exploit the characteristics of feature maps among parallel branches, such as cross-scale and intra-scale, local, and non-local self-similarity.
3) We propose a novel architecture, called FC-Former, which is the first fully-connected self-attention network with multi-scale feature representation. Benefiting from the information fidelity of high-resolution branches, our model achieves state-of-the-art performance for some MSIF tasks as shown in Fig. 3, i.e., multispectral and hyperspectral image fusion (MHIF), visible and infrared (VIS-IR) image fusion, and remote sensing pansharpening. Extensive ablation experiments corroborate the effectiveness of the proposed network. In addition, we also provide digital photographic image fusion results in the supplementary material.
4) The proposed multi-source image representation incorporates and unfolds the fusion problem into the FC-Former. The network can be considered interpretable thanks to the explicit characterization of both image priors and feature representation.

This paper is an extended version of the conference paper in [58], which is the first cross-scale parallel fusion network specifically designed for remote sensing pansharpening, called DCFNet. In this version, we extended the work in [58] from both methodological and application points of view. The related improvements are as follows:

1) The DCFNet shows a trade-off between parameter number and feature representation. To get a win-win situation, we propose the new idea of generalized self-attention, even developing the FCSA framework to fully exploit multiple sources of information.
2) The proposed FCSA framework explores self-attention along different unfolded dimensions (modes), fully considering the differences between spatial and channel features.
3) We develop a model-inspired FC-Former, where the prefusion design is replaced by a multi-source input representation embedded as a network prior that improves the outcomes using classical physical constraints.
4) Unlike DCFNet, which is a network tailored to the pansharpening problem, three different applications are considered in this work: multispectral and hyperspectral image fusion (MHIF), visible and infrared image fusion (VIS-IR), remote sensing pansharpening, and digital photographic image fusion.

The rest of the paper is organized as follows. Section II sequentially presents three related works: model-based methods, data-driven methods, and model-driven methods. This section also provides the motivation for the work. Section III introduces the proposed mathematical idea and framework as well as the overall network. In Section IV, we conduct extensive experiments on three MSIF tasks. Furthermore, additional discussions and ablation studies demonstrating the FC-Former's superior performance, efficiency, and low parameters are reported in Sections V and VI. Finally, concluding remarks are drawn in Section VII.

II. Related Work

A. Model-Based Methods

In the MSIF task, some early methods exploited domain-specific features of source images using linear transformations, see, e.g., component substitution (CS) [59] and multi-resolution analysis (MRA) [60], [61] approaches.

Other methods related to the MSIF problem belong to the variational optimization-based (VO) class. VO approaches yield the unknown fused image by minimizing a given domain-specific optimization problem involving the multi-source images in input. The advantages of VO methods are the better representation of the information and an elevated interpretation. Prior knowledge is introduced by adding a regularization term to address the ill-posed nature of the optimization problem. For example, sparse representation methods in the VO class are related to the building of a dictionary to model (as a prior) the sparsity for image patches [62], [63], [64]. To regularize image gradients,

spatial priors impose the first-order smoothness on the unknown (fused) image [65], [66]. Some other methods [27], [67] exploit the low-rank property. Subspace analysis [67] and matrix/tensor decomposition [67], [68] have also been used in conjunction with the low-rank property. However, handcrafted priors are not usually enough to represent real-world data accurately.

B. Data-Driven Deep Learning Methods

Deep learning (DL)-based methods have successfully exploited their powerful feature representation capability. DL-based methods can be roughly summarized as convolutional neural network (CNN)-based methods and transformer-based methods. Regarding pansharpening, PNN has been proposed in [69]. It is based on a three-layer CNN to obtain the pansharpened image (HR-MSI). To fuse useful high-frequency information based on physical constraints, in [70], the fusion process is formulated as a linear observed model in which deep and fusion networks are used to extract and fuse features from different source images. Some specialized modules, such as the multi-scale mechanism [49] and the spatial/channel attentions [71], have recently been proposed for the MSIF problem. To enlarge the receptive field, a pixel-adaptive convolution method has been proposed, the so-called LAGNet [72], to exploit pixel-to-pixel similarity in local windows to characterize content-aware features. Bandara et al. [73] designed a cross-attention mechanism to correlate pixel relations for multi-source images in MHIF. Besides, Ma et al. in [74] first presented a transformer-based framework for VIS-IR image fusion and digital photographic image fusion, explaining the significance of the transformer's long-distance dependency on image fusion tasks. The study in [75] first blends image matching, fusion, and semantic awareness into the same framework, yielding promising results. Wang et al. [76] leveraged domain knowledge to design a semi-supervised transfer learning method to fuse infrared and visible image fusion. However, the above-mentioned networks are limited by the use of multi-scale and multi-dimensional feature representation from the self-attention mechanism, often resulting in poor fusion performance.

C. Model-Driven Deep Learning Methods

Wang et al. [77] proposed the DBIN model, where the estimation of the exploited observation model and the related fusion process are optimized iteratively and alternatively during the reconstruction. Xie et al. [15] proposed the so-called MHFNet to combine a low-rank prior and a complete basis set of HR-HSIs to build the unfolding network. Guo et al. [2] designed a variational gate mechanism to fuse three different similarities of miRNAs via a novel contrastive cross-entropy function. As in the case of classical convolutional networks, where local information is extracted by convolutions, the deep [42] and autoencoder priors [43] also impose local priors for model-based methods.

Non-local self-similarity (NSS) priors have recently been explored in various research fields [23], [78]. The approaches based on the use of these priors consider similar pixels/patches of a given image to exploit the internal redundant information. The self-attention mechanism is a good instance of NSS methods based on long-range dependencies through matrix multiplication. Unlike feature representation of convolutions, transformer [79] can theoretically expand the receptive field infinitely, thereby correlating different pixels/patches to each other. Transformer methods often demonstrate a superior ability to learn intrinsic features compared to CNN-based approaches.

To date, non-local networks [80] and transformer methods [81] represent state-of-the-art mechanisms in computer vision. To encourage joint feature learning across two dimensions (modes), cross-modality transformers [57], [74] have recently been designed to learn better feature representations between two different domains. Wang et al. [44] integrated a data-driven NSS prior and the HQS method addressing the problem with an optimization-inspired deep neural network.

D. Motivation

In multi-source image fusion, the input data contains rich multi-dimensional information and domain-specific information, namely, local and non-local similarities within or across scales, as well as spatial and spectral information. To fully explore this potential information, we use multilinear algebra to develop the mathematical concepts behind the generalized self-attention mechanism and propose the FCSA framework.

The naive self-attention-based methods are often limited to a single dimension or a specific perspective, resulting in the loss of key information from different sources. To solve this problem, the FCSA mechanism can simultaneously integrate process multi-dimensional feature information from different scales and domains, and fully mine rich details in the image. Then, the FCSA framework can deeply explore the information interaction between various features in the image. By parallel branch design, the FCSA framework establishes a fully connected relationship, ensuring the maximum utilization of potential information in the input data.

MSIF networks do not often get contextual guidance for feature representation, even showing a feature extraction phase that usually reduces the features' spatial resolution. Hence, we cannot advise its use for MSIF tasks. Instead, starting from the promising results obtained in our conference paper, we develop, in this work, the so-called FC-former network based on the FCSA framework to consider feature similarity within and across scales while obtaining discriminative information from different sources.

III. Fully-Connected Transformer Model

A. Generalized Self-Attention Mechanism

In this section, we first summarize the necessary notations and give several new definitions used in this paper. For the MSIF task, an image and another image are defined as $\mathbf{I_1} \in \mathbb{R}^{H\times W\times c}$ and $\mathbf{I_2} \in \mathbb{R}^{h\times w\times C}$, respectively. The desired fused image is indicated as $\mathbf{I_f}$, where the scale ratio is $r = H/h$ (e.g., 4 or 8). For the MHIF task, the source images are the HR-MSI and the LR-HSI, respectively, while for the visible and infrared image fusion, are the infrared and the visible images, respectively, and, for remote sensing pansharpening, are the panchromatic image and the multispectral cube, respectively.

Before introducing the generalized self-attention and the FCSA method, we first describe the classic spatial self-attention (Spa-SA) based on the definition of the batched matrix multiplication. Given the input tensor $\mathcal{X} \in \mathbb{R}^{B\times P\times N\times C}$, the Spa-SA can be formulated as follows:

$$\begin{aligned}&\mathbf{Q} = \mathbf{X}\mathbf{W}_{\mathbf{Q}}^T, \mathbf{K} = \mathbf{X}\mathbf{W}_{\mathbf{K}}^T, \mathbf{V} = \mathbf{X}\mathbf{W}_{\mathbf{V}}^T,\\ &\mathbf{A} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right),\\ &\mathbf{Z} = \mathbf{A}\mathbf{V} + \mathbf{X},\end{aligned} \tag{1}$$

where $\mathbf{W}_{\mathbf{Q},\mathbf{K},\mathbf{V}}$ indicates the learnable parameters, $\mathbf{Z}$ represents the output features, the *Query* is $\mathbf{Q} \in \mathbb{R}^{N_Q\times C_Q}$ determining the spatial sizes of the output features and of the attention matrix, the *Key* and *Value* are $\mathbf{K} \in \mathbb{R}^{N_K\times C_Q}$ and $\mathbf{V} \in \mathbb{R}^{N_K\times C_V}$ defining the sizes of the attention matrix and of the output channel of $\mathbf{Z}$, respectively. $\mathbf{K}$ and $\mathbf{V}$ must have the same spatial size, i.e., $N_K = N_V = N$. We assume that $\mathbf{Q}$ and $\mathbf{K}$ are d-dimensional vectors. The attention, $\mathbf{A}$, relies upon the dot product between $\mathbf{Q}$ and $\mathbf{K}$ to get spatial self-similarity, which influences $\mathbf{V}$ and vice verse. Thanks to the self-attention mechanism, transformers can achieve self-similarity along a specific dimension (mode).

Next, we will introduce the generalized self-attention mechanism through the following new definitions and theorems.

Definition 1 (Tensor Blocking): For a 4th-order tensor, $\mathcal{X} \in \mathbb{R}^{I_1\times I_2\times I_3\times I_4}$, a window $(q \times q)$ is set to be centered at each spatial location. Tensor blocking with stride $(s \times s)$ generates a blocking tensor, $\mathcal{T} \in \mathbb{R}^{I_1\times P\times q\times q\times I_2}$. Thus, we have:

$$\mathcal{T} = \mathbf{unfold}_{(s\times s)}^{(q\times q)}(\mathcal{X}), \tag{2}$$

where P denotes the number of patches and satisfies $P = \prod_{i=3}^{4}\frac{I_i-q+2*p}{s}$, where p is the border padding. The **unfold** operator is implemented in Pytorch [82] with fast runtime.

Definition 2 (Batched Mode-k Unfolding): Given an Nth-order tensor, $\mathcal{X} \in \mathbb{R}^{I_1\times I_2\times\cdots\times I_N}$, $\mathbf{n} = (n_1, n_2, \ldots, n_N)$ is a vector reordering. The batched mode-k unfolding of $\mathcal{X}$ is defined as $\mathbf{X}_{[\mathbf{n};k]} \in \mathbb{R}^{\prod_{i=1}^{k-1} I_{n_i}\times\prod_{j=k+1}^{N} I_{n_j}\times I_k}$ $(1 < k \leq N, k \in \mathbb{Z})$,

$$\begin{cases}\mathbf{X}_{[\mathbf{n};k]}(i_{n_1}i_{n_2}\cdots i_{n_{k-1}}, i_{n_{k+1}}i_{n_{k+2}}\cdots i_{n_N}, i_{n_k}) = \\ \quad \text{reshape}(\mathcal{X}, [I_{n_1}I_{n_2}\cdots I_{n_{k-1}}, \\ \qquad I_{n_{k+1}}I_{n_{k+2}}\cdots I_{n_N}, I_{n_k}]),\ 1 < k < N, \\ \mathbf{X}_{[\mathbf{n};N]}(i_{n_1}i_{n_2}\cdots I_{n_{N-2}}, i_{n_{N-1}}, i_{n_N}) = \\ \quad \text{reshape}(\mathcal{X}, [I_{n_1}I_{n_2}\cdots I_{n_{N-2}}, I_{n_{N-1}}, I_{n_N}]),\ k = N,\end{cases} \tag{3}$$

and its inverse operator yields $\mathcal{X} = \text{reshape}(\mathbf{X}_{\mathbf{n}}, [I_{n_1}, I_{n_2}, \ldots, I_{n_N}])$ via indices $\mathbf{n} = (n_1, n_2, \ldots, n_N)$.

Definition 2 can have well tensor permutation, $\mathbf{X} = \mathbf{X}(i_{n_1}, i_{n_2}, \ldots, i_{n_N})$ based on vector $\mathbf{n}$.

Definition 3 (Batched Tensor Product): Supposing that an Mth-order $\mathcal{X} \in \mathbb{R}^{I_1\times I_2\times\cdots\times I_M}$ and an Nth-order $\mathcal{Y} \in \mathbb{R}^{J_1\times J_2\times\cdots\times J_N}$. Assume that two vectors $\mathbf{m} = (m_1, m_2, \ldots, m_M)$ and $\mathbf{n} = (n_1, n_2, \ldots, n_N)$ are vectors satisfying $I_{m_i} = J_{n_j}$ for $i = 1, 2, \ldots, k$. The batched tensor product between $\mathcal{X}$ and $\mathcal{Y}$ along mode k $(1 < k \leq \min(M, N), k \in \mathbb{Z})$ in matrix form is as follows:

$$\mathbf{Z} = \mathcal{X} \times_{n_1,n_2,\ldots,n_k}^{m_1,m_2,\ldots,m_k} \mathcal{Y}, \tag{4}$$

where the size of $\mathbf{Z}$ is $\prod_{i=1}^{k-1} I_{m_i} \times \prod_{j=k+1}^{M} I_{m_j} \times \prod_{j=k+1}^{N} J_{n_j}$ for $1 < k < \min(M, N), k \in \mathbb{Z}$, or $\prod_{i=1}^{M-2} I_{m_i} \times I_{N-1} \times J_{N-1}$ for $k = \min(M, N)$. The last dimensions of $\mathcal{X}$ and $\mathcal{Y}$ are contracted. The batched tensor product is the batched format of the multilinear product and requires $I_{m_i} = J_{n_i}$ for $i = 1, 2, \ldots, k$ when $k \neq \min(M, N)$, or $I_{m_i} = J_{n_i}$ for $i = 1, 2, \ldots, k-2, k$ when $k = \min(M, N)$. The associative and commutative properties are not satisfied.

In Fig. 4(a), we give an illustration of the above definitions. Below, we will introduce the theorems of the generalized self-attention mechanism.

Theorem 1: Supposing that $\mathcal{X} \in \mathbb{R}^{I_1\times I_2\times\cdots\times I_M}$ and $\mathcal{Y} \in \mathbb{R}^{J_1\times J_2\times\cdots\times J_N}$ are two tensors. Thus, we have:

1) $\mathbf{Y}^T(i_{n_{1:k-1}}, i_{n_{k+1:N}}, i_{n_k}) = \mathbf{Y}(i_{n_{1:k-1}}, i_{n_k}, i_{n_{k+1:N}})$,
2) $\mathbf{Z} = \mathcal{X} \times_{n_1,n_2,\ldots,n_k}^{m_1,m_2,\ldots,m_k} \mathcal{Y} \Leftrightarrow \mathbf{X}_{[\mathbf{m};k]}\mathbf{Y}_{[\mathbf{n};k]}^T$.

The interested readers can refer to the supplementary material to have a look at the proof of Theorem 1. Theorem 1 describes the relationship between the batched tensor product and the batched matrix multiplication. Below, Theorem 2 will consider the self-attention mechanism with two special tensor forms by using the above definitions.

Theorem 2 (Generalized Self-Attention Mechanism): Let us assume an Nth-order tensor, $\mathcal{X} \in \mathbb{R}^{I_1\times I_2\times\cdots\times I_N}$, the learnable parameters, $\mathbf{W} \in \mathbb{R}^{I_{i_k}\times J_3}$, and the reordering vector, $\mathbf{i} = (i_1, i_2, \ldots, i_N)$. The generalized self-attention of $\mathcal{X}$ has three reordering factors, $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V} \in \mathbb{R}^{J_1\times J_2\times J_3}$ along mode k, where $(J_1, J_2, J_3) = (\prod_{i=1}^{k-1} I_{i_i}, \prod_{j=k+1}^{N} I_{i_j}, I_{i_k})$. Let us define $\mathbf{m} = (m_1, m_2, \ldots, m_N)$ and $\mathbf{n} = (n_1, n_2, \ldots, n_N)$ as indexes of the batched tensor product. The generalized self-attention generates two matrices $\mathbf{A}$ and $\mathbf{Z}$ along the kth dimension (mode k), which have the following forms:

$$\begin{aligned}&\mathbf{Q} = \mathbf{X}_{[\mathbf{i};k]}\mathbf{W}_{\mathbf{Q}}^T,\ \mathbf{K} = \mathbf{X}_{[\mathbf{i};k]}\mathbf{W}_{\mathbf{K}}^T,\ \mathbf{V} = \mathbf{X}_{[\mathbf{i};k]}\mathbf{W}_{\mathbf{V}}^T,\\ &\mathbf{A} = \text{Softmax}\left(\frac{\mathcal{Q} \times_{n_1,n_2,\ldots,n_k}^{m_1,m_2,\ldots,m_k} \mathcal{K}}{\sqrt{d}}\right),\\ &\mathbf{Z} = \mathcal{A} \times_{n_1,n_2,\ldots,n_k}^{1,n_{k+1}-k+1,n_{k+2}-k+1,\ldots,n_N-k+1,2} \mathcal{V} + \mathbf{X}_{[\mathbf{i};k]},\end{aligned} \tag{5}$$

where matrices $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$, and $\mathbf{A}$ perform the inverse operator of batched mode-k unfolding to tensor format.

The interested readers can refer to the supplementary material to have a look at the proof of Theorem 2. Here, by utilizing the tensor blocking operator given in Definition 1 and the batched mode-k unfolding operator in Definition 2, we can sequentially obtain three factors, $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$, represented in the self-attention mechanism. A graphic illustration of the generalized self-attention is in Fig. 4(b). A special form of spatial self-attention is shown based on our generalized mechanism.

By using the proposed definitions and theorems, we can derive several forms of self-attention. For example, assuming that the input tensor $\mathcal{Y}$ is $\mathbb{R}^{B\times d\times C\times H\times W}$, transforming the dimensions $H \times W$ into the spatial size S, for multi-head spatial self-attention, the batched mode-3 unfolding is performed to generate $\mathbf{Q}, \mathbf{K}$, and $\mathbf{V} \in \mathbb{R}^{Bd\times S\times C}$, where $\mathbf{i} = (1, 2, 4, 5, 3)$. Afterwards, the batched tensor product is performed for $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$, where $\mathbf{m} = \mathbf{n} = (1, 2, 3)$. For the channel self-attention, we first merge the H and W dimensions to obtain $\mathcal{Y} \in \mathbb{R}^{B\times d\times C\times S}$,

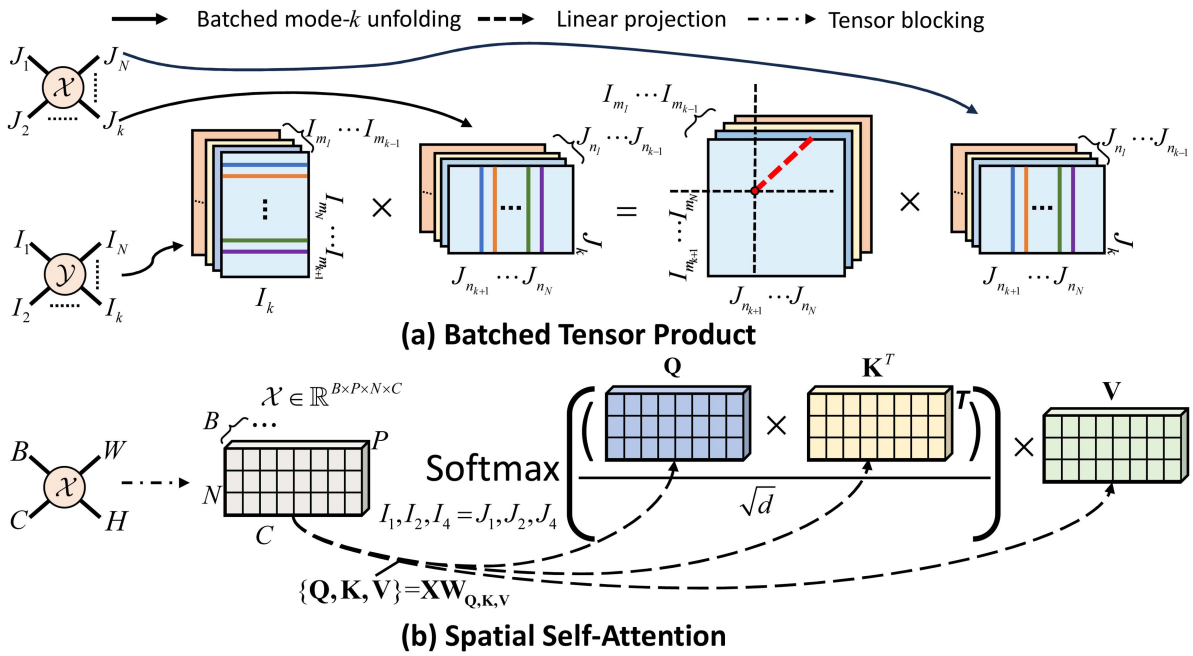


Fig. 4. Graphical illustration of the batched tensor product in Definition 3. Furthermore, we present the spatial self-attention based on the proposed definitions. The tensor blocking of Definition 1 takes precedence over the batched mode-k unfolding.

TABLE I
SOME NOTATIONS USED ARE SUMMARIZED AS FOLLOWS

Notation	Explanation
$\mathbf{unfold}_{(s\times s)}^{(q\times q)}(\cdot)$	Tensor blocking, see Definition 1.
$\mathbf{i}=(i_1,i_2,\cdots,i_N)$	Reordering vectors defined in Definition 2.
$\mathbf{X}^T$	Tensor transpose.
$\mathbf{m}=(m_1,m_2,\cdots,m_N)$	Index of the batched tensor product t defined in Theorem 1.
$\mathbf{n}=(n_1,n_2,\cdots,n_N)$	Index of the batched tensor product defined in Theorem 1.
$\times_{\mathbf{n}}^{\mathbf{m}}$	Batched tensor product between vector $\mathbf{m}$ and vector $\mathbf{n}$.
k	k-th dimension/mode of a N-th tensor ($1<k\leq N, k\in\mathbb{Z}$).
$\mathbf{X}_{[\mathbf{n};k]}$	Tensor permutation, see Definition 2.

then $\mathcal{Y}_{[\mathbf{i};4]}(1,2,3,4)$ yields $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}\in\mathbb{R}^{Bd\times C\times S}$, which is also derived from batched mode-4 unfolding. In addition, assuming that the input tensor $\mathcal{Y}$ is $\mathbb{R}^{B\times d\times P\times S\times C}$, the spatial and spectral multi-head self-attention forms are the same as above. Definition 2 gets three factors of self-attention with three dimensions of information, i.e., $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}\in\mathbb{R}^{Bd\times P\times SC}$, called patch self-attention [83], [84]. Further descriptions of the existing self-attention forms can be found in other related papers [85].

Remark 1: For $\mathcal{Q}$, $\mathcal{K}$ and $\mathcal{V}\in\mathbb{R}^{I_1\times I_2\times\cdots\times I_N}$, the generalized self-attention generates tensors $\mathcal{A}$ and $\mathcal{Z}$ along the kth dimension. In addition, Theorem 5 can have a simplified form that specifies the batched tensor product as $\mathbf{Q}\times_{n_1,n_2,n_3}^{m_1,m_2,m_3}\mathbf{K}$ and $\mathbf{A}\times_{n_1,n_2,n_3}^{1,m_3,2}\mathbf{V}$, where the inverse operator of batched mode-k unfolding is not used.

Previous works introduced different forms of self-attention and explored multi-dimensional information based on hybrid structures. In the paper, we exploit multilinear analysis in tensor algebra to generalize these self-attention forms.

B. Fully-Connected Self-Attention Framework

The separated matrices $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$, are multi-dimensional and domain-related. This information at different modes lacks an effective way to be combined. Based on the generalized self-attention mechanism, we can develop the FCSA framework. The FCSA framework is depicted in Fig. 5. More specifically, we use cross-scale and intra-scale (aka in-scale) self-attention to transfer features among features at different or at same resolutions. Following Theorem 2 and the previous self-attention mechanisms, the FCSA framework transforms three 1×1 Conv2D layers obtaining $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$, to calculate the response in the same resolution branch. Afterwards, we adopt cross-scale self-attention, which is defined in Theorem 2. Finally, these features are transformed into new features along different modes with different resolutions and channels.

The inputs of the FCSA framework are the high-resolution (HR) feature maps, $\mathcal{F}_H$ and $\mathcal{I}_H$, the medium-resolution (MR) feature maps, $\mathcal{F}_M$ and $\mathcal{I}_M$, and the low-resolution (LR) feature maps, $\mathcal{F}_L$ and $\mathcal{I}_L$. The tensors $\mathcal{I}_H$, $\mathcal{I}_M$, and $\mathcal{I}_L$ represent important source images, such as multispectral and thermal images, etc, which are downsampled to a lower resolution. Afterwards, the FCSA model calculates self-attention along each of their modes. By employing the proposed idea of generalized self-attention, the fully-connected self-attention scheme is as follows:

$$(\mathcal{Z}_H,\mathcal{Z}_M,\mathcal{Z}_L)=\mathbf{FCSA}_{[\mathbf{m}_k;\mathbf{n}_k]}(\mathcal{X}_H,\mathcal{X}_M,\mathcal{X}_L). \quad (6)$$

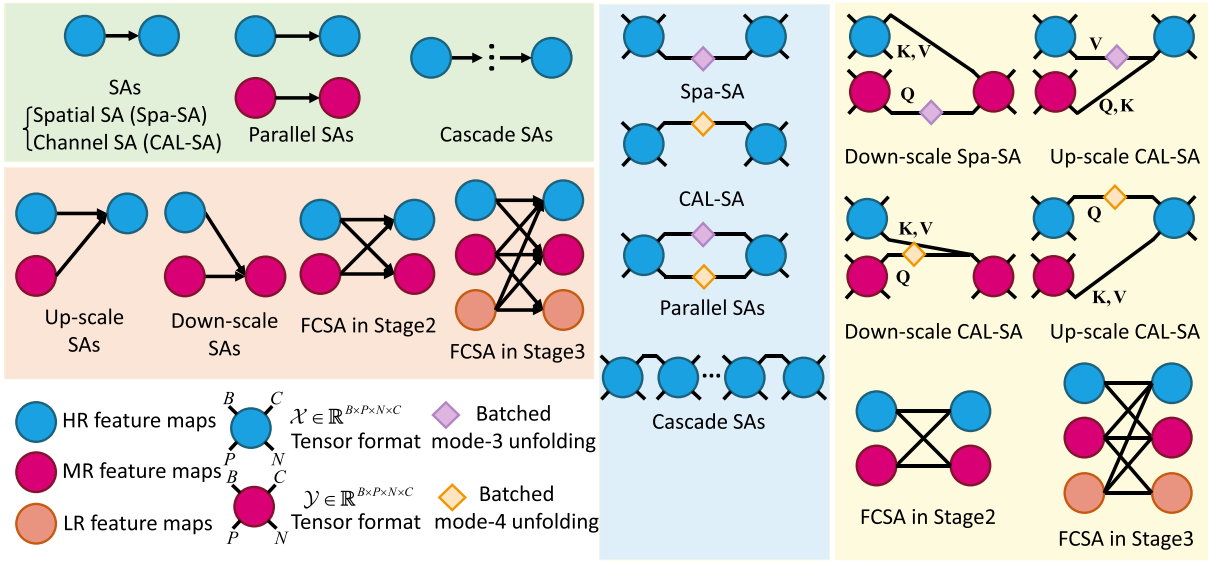


Fig. 5. Illustration of the FCSA framework. The proposed FCSA framework unifies several self-attention mechanisms, such as [49] and [57], and includes their corresponding multilinear product representation. The FCSA framework can facilitate the fusion of local and non-local prior information within and across images from different sources. Note that stages 2 and 3 of the FCSA are simply plotted, not affecting the required tensor format.

Here, we further use feature branches to represent input tensors, i.e., $(\mathcal{X}_H, \mathcal{X}_M, \mathcal{X}_L)$. Then, $\mathcal{X}_H = (\mathcal{F}_H, \mathcal{I}_M, \mathcal{I}_L)$, $\mathcal{X}_M = (\mathcal{F}_M, \mathcal{I}_H, \mathcal{I}_L)$, and $\mathcal{X}_L = (\mathcal{F}_L, \mathcal{I}_H, \mathcal{I}_M)$ denote the three factors (i.e., *Query*, *Key* and *Value*), respectively. $[\mathbf{m}_k; \mathbf{n}_k]$ is one of the reordering vectors of $\mathbf{m}$ and $\mathbf{n}$ at mode k. For a better understanding of (6), the detailed algorithm is reported in Algorithm 1.

Equation (5) performs the transfer of feature maps at different resolutions. When transferring a lower resolution branch to a higher resolution branch, the *Query* represents higher resolution feature maps, while *Key* and *Value* denote lower resolution feature maps. Following Definitions and Theorem 2, lower-resolution feature maps influence higher-resolution feature maps according to the reordering vector $\mathbf{m}_k$. Furthermore, these different resolution feature maps can progressively aggregate new feature maps from high-to-low and low-to-high branches and transfer the cross-scale feature maps back to high-resolution branches. In summary, the proposed scheme can enhance feature representation and achieve higher performance.

Remark 2: It is worth remarking that the FCSA framework conducts multi-dimensional self-attention using the generalized self-attention mechanism. The separated matrices, $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$, are used to calculate two different unfolded self-attentions. This induces both the long-range spatial and the global channel responses. The FCSA framework retains the transformer's solution while reducing the computational cost and increasing non-local information. The FCSA can improve the performance of MSIF, as reported in Table VI.

Algorithm 1: One Stage of FCSA.

Input: The feature maps, $\{\mathcal{F}\}$, with N dimensions $I_1 \times I_2 \times \cdots \times I_N$; Source images $\{\mathcal{I}\}$.
Initialization: K reordering vectors $\mathbf{m} = \{\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_k\}$, $\mathbf{n} = \{\mathbf{n}_1, \mathbf{n}_2, \cdots, \mathbf{n}_k\}$ $(1 < k \leq N, k \in \mathbb{Z})$.
foreach $\mathcal{X}$ *in feature branches* **do**
  $\mathcal{Z} = 0$;
  for $\mathbf{m}_k$, $\mathbf{n}_k$ *in reordering vectors* **do**
    Obtain $\mathcal{Z}_k$ from the generalized self-attention by Theorem 2; // Eq. (5)
    $k = k + 1$;
    $\mathcal{Z} = \mathcal{Z} + \mathcal{Z}_k$;
  end
  $\{\mathcal{Z}\} \leftarrow \mathcal{Z}$;
end
Result: The feature map, $\{\mathcal{Z}\}$, for each branch.

C. Complexity Analysis

Let us transform an Nth-order tensor, $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_k \times \cdots \times I_N}$, into $I_1 \times \cdots \times I_{k-1} \times S \times C$, where $I_k = S$ or $I_k = C$. Then, we have batched operations for $i < k$, and multilinear product operations for $i \leq k$, $1 < i \leq N$. Therefore, the computational complexity of the FCSA is $\mathcal{O}(\prod_{i=1}^{k-1} I_i I_k^2 I_{k+1} + \prod_{j=1}^{k-2} I_j I_k^2 I_{k-1})$, that is, $\mathcal{O}(\prod_{i=1}^{k-1} I_i S^2 C + \prod_{j=1}^{k-2} I_j S C^2)$. The computational complexity linearly increases with the size of the image and the number of channels. Besides, self-attention has some (GPU memory) storage costs. The FCSA storage cost, which depends on S and C, is $\mathcal{O}(\prod_{i=1}^{k-1} I_i S^2 + \prod_{i=1}^{k-2} I_i C^2)$, consistently with the hybrid self-attention considering both spatial and spectral modes.

D. Multi-Source Image Representation

Several networks for MSIF can be separated into two parts: deep and fusion sub-networks. The simplest fusion method relies upon just adding or concatenating features. Instead, in this work, we will introduce two more elaborated fusion strategies: (a) dynamic branch fusion; (b) model-based branch fusion.

Dynamic Branch Fusion: In our previous work [58], we showed that different resolutions have unequal effects on fusion results. Thus, feature maps at different resolutions should be reweighed before being combined by the dynamic branch fusion (DBF) module. The DBF method adds fusion coefficients to features at different resolutions and resizes features at the same resolution before the weighted fusion. The DBF method can be widely applied to different fusion scenarios, thereby we chose it as the baseline for the multi-source image representation (MSIR) module.

Model-based Branch Fusion: The DBF method does not consider physical constraints. Physical constraints are usually introduced by linear observation models [9], [35]. The linear relationships (reflecting prior knowledge) among input and output data of the MSIF problem are computed by solving an optimization problem. By using MSIR with linear observation models, we can draw the following conclusions.

Lemma 1 (Linear Observation Models for MSIF): Assume the MSIF problem with $\mathbf{X} \in \mathbb{R}^{HW \times S}$ denoting the desired results. The linear observed models, having as information sources two cubes $\mathbf{Y} \in \mathbb{R}^{hw \times S}$, and $\mathbf{M} \in \mathbb{R}^{HW \times s}$, where (H, W) and (h, w) denote the different spatial sizes (with a scale ratio $r = \frac{H}{h}$), and S and s indicate the number of spectral bands for the two inputs, are as follows:

$$\mathbf{Y} = f_1(\mathbf{X}), \quad \mathbf{M} = f_2(\mathbf{X}). \tag{7}$$

The functions $f_1(\cdot)$ and $f_2(\cdot)$ represent degradation operators. Thus, the objective function can be formulated as:

$$\mathbf{X} = \arg\min_{\mathbf{X}} ||\mathbf{Y} - f_1(\mathbf{X})||_F + ||\mathbf{M} - f_2(\mathbf{X})||_F + \phi(\mathbf{X}). \tag{8}$$

Remark 3: For the MHIF problem, f_1 and f_2 can be defined as $f_1(\mathbf{X}) = \mathbf{XBS}$ and $f_2(\mathbf{X}) = \mathbf{RX}$, where $\mathbf{B}, \mathbf{S} \in \mathbb{R}^{HW \times hw}$, and $\mathbf{R} \in \mathbb{R}^{S \times s}$ denote the blur operator, the downsampling operator, and the spectral response matrix, respectively. f_1 and f_2 are the spatial and the spectral fidelity terms, respectively. The problem can be solved into the linear least squares framework [15]. Similar linear relationships hold for other related MSIF problems.

It is worth to be pointed out that, to obtain the desired fused image, we adopt the proximal gradient algorithm [86] to solve the problem as follows.

Theorem 3 (Interpretable Model-based Unfolding Representation [86]): Let an observed image, $\mathbf{Y}$, be the corrupted (or degraded) tensor version through a function $f(\cdot)$ of an unknown image, $\mathbf{X}$, having a simplified form as:

$$\mathbf{X} = \arg\min_{\mathbf{X}} \frac{1}{2}||\mathbf{Y} - f(\mathbf{X})||^2 + \phi(\mathbf{X}). \tag{9}$$

There exists, $\hat{\mathbf{X}}$, solution of the algorithm, in the form:

$$\hat{\mathbf{X}}_{(t+1)} = \text{prox}_{(\lambda\eta\phi)}(\mathbf{X}_{(t)} - \eta\nabla g(\mathbf{X}_{(t)})), \tag{10}$$

where $\text{prox}_{(\lambda\eta\phi)}(\cdot)$ denotes a proximal operator, η is a weighting coefficient, $\nabla g(\cdot)$ is the gradient operator, and t is an iteration index.

Problem 9 requires HQS or ADMM frameworks to be solved. For different regularization terms, including deep network priors, we usually can have a closed-form solution. For the MHIF problem, the gradient of $\mathbf{X}$ is: $\nabla g(\mathbf{X}_{(t)}) = (\mathbf{X}_{(t)}\mathbf{D} - \mathbf{Y})\mathbf{D}^T + \mathbf{R}^T(\mathbf{RX}_{(t)} - \mathbf{M})$, where the degradation operator $\mathbf{D} = \mathbf{BS}$.

By separating each sub-problem and transforming it into a specific network form, we can build an optimization-induced deep network through employing the MSIR block. It allows the network to approximate the proximal operator of a regularizer, not just a denoiser [35], [84].

The main difference among several fusion problems is how to formulate a fusion model. In this work, we embed the observed model into the FC-former to realize an interpretable deep network for MSIF. The proposed FC-former network will be introduced in Section III-E.

Remark 4: It is worth to be remarked that the optimization-induced neural network is motivated by the linear observed model and its solution. Under the aforementioned framework, the regularization term can provide a non-linear representation in the objective function. This allows us to estimate the solution by exploiting deep learning and physical constraints.

E. Network Architecture

The overall architecture of the fully-connected transformer (FC-Former) is presented in Fig. 6. It consists of three parallel branches: the main HR feature branch, the MR feature branch, and the LR feature branch. More specifically, the three branches are arranged in parallel, and they are progressively combined to form three stages. The main HR feature branch considers $H \times W$ spatial size images from different domains. The MR feature branch receives the MR input and the feature maps from the HR branch. Similarly, the LR feature branch takes an LR input and the feature maps from the above two branches as input.

For the inputs in each branch, we design MSIR as the head structure of each branch to aggregate feature maps transferred from other branches with source images, as shown in Fig. 6. From an implementation point of view, inspired by HRNet [34], we chose the residual block and bottleneck as building blocks. The convolution kernel of the residual block of each branch is the same. Finally, the stacked residual blocks are arranged behind the MSIR blocks. Therefore, a complete stage is built to extract better features. Finally, we train the proposed model on supervised and unsupervised tasks. Let $\mathbf{I} = \{\mathbf{I_1}, \mathbf{I_2}, \dots\}$ denote input images from different sources. Then, for the supervised task, we use the mean absolution error (i.e., ℓ_1 loss; $\mathcal{L}_1$) and the structural similarity index measure (SSIM) as losses [50] ($\mathcal{L}_{\text{SSIM}}$) to calculate differences between outputs and ground-truths (GTs):

$$\mathbf{\Theta} = \arg\min_{\mathbf{\Theta}} \mathcal{L}_1(f_\Theta(\mathbf{I}), \mathbf{GT}) + \lambda\mathcal{L}_{\text{SSIM}}(f_{\mathbf{\Theta}}(\mathbf{I}), \mathbf{GT}), \tag{11}$$

where λ is set to 0.1 to balance the two losses, $\mathbf{GT}$ is the ground-truth image, and $f_{\mathbf{\Theta}}$ is a non-linear function depending on the learnable parameters $\mathbf{\Theta}$.

For unsupervised tasks, we use the intensity loss ($\mathcal{L}_1$), SSIM loss ($\mathcal{L}_{\text{SSIM}}$), and texture loss ($\mathcal{L}_{\text{text}}$), to compute the loss between output and input images. The details of the loss functions can

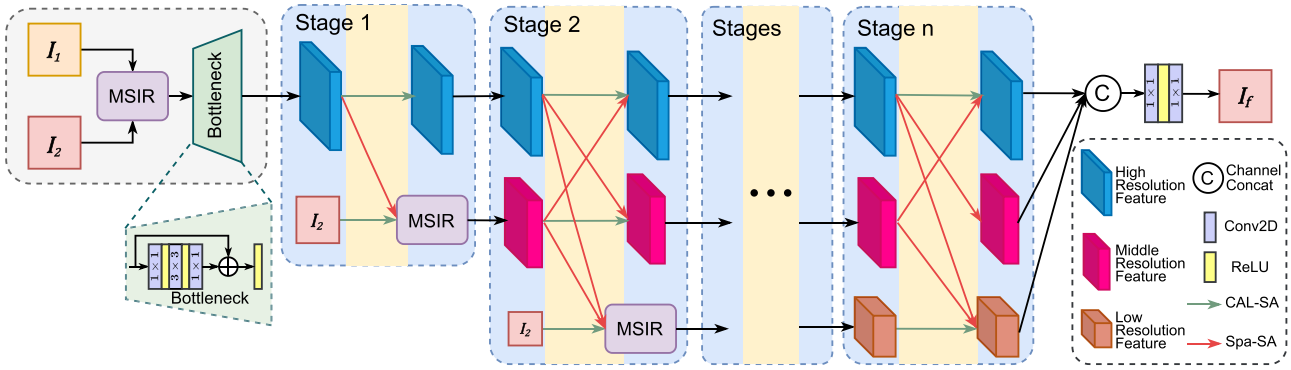


Fig. 6. The overall architecture of FC-Former. The blue boxes represent network stages and the yellow parts denote the FCSA method depicted in Fig. 5.

Algorithm 2: Fully-Connected Transformer Algorithm.

Input: HR, MR, LR feature maps $\mathcal{F}_H$, $\mathcal{F}_M$, and $\mathcal{F}_L$, with $I_1 \times I_2 \times \cdots \times I_N$; HR, MR, and LR source images $\mathcal{I}_H$, $\mathcal{I}_M$, and $\mathcal{I}_L$;

while *until convergence* **do**
- Obtain output tensors $\mathcal{X}_{H,M,L}$ of MSIR by Theorem 3.
- Obtain the feature maps $\{\mathcal{Z}\}$ by the network based on the generalized self-attention in Theorem 2 and the FCSA in Algorithm 1.
- Update parameters Θ of the network via Eq. (11).

end

Result: The fused image.

be found in related works [17], [74], [87].

$$\boldsymbol{\Theta} = \arg\min_{\boldsymbol{\Theta}} \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_{\text{SSIM}} + \lambda_3 \mathcal{L}_{\text{text}}, \tag{12}$$

where each term $\mathcal{L}$ is $\mathcal{L}(f_\Theta(\mathbf{I}), \mathbf{I})$. λ_1, λ_2, and λ_3 are hyper-parameters.

The FC-Former is summarized in Algorithm 2.

IV. Experiments and Results

We assess the performance by comparing the proposed model for different tasks, i.e., the MHIF, the VIS-IR image fusion, and the remote sensing pansharpening. In addition, digital photographic image fusion tasks (i.e., multi-exposure image fusion and multi-focus image fusion) are shown in the supplementary material. MHIF data have different spatial and spectral resolutions. The VIS-IR image fusion task relies upon unsupervised image fusion (combining data at the same resolution). Finally, the remote sensing pansharpening task considers both simulated and real-world data to fully assess the performance of our method and its generalization ability. The proposed approach is implemented in Pytorch and trained on a workstation with 2 NVIDIA GeForce RTX 3090 GPUs and 128 GB memory. For the sake of brevity, we selected results of some representative methods. The interested readers can refer to the supplementary material to have a look at all the outcomes.

A. Multispectral and Hyperspectral Image Fusion

Setup: We test the proposed method on two widely used MHIF datasets (i.e., CAVE [106][1] and Harvard [107][2]) considering 15 state-of-the-art techniques: MTF-GLP-HS [88]$^{\text{JSTARS}'2015}$, CSTF-FUS [108]$^{\text{TIP}'2018}$, BDSD-PC [59]$^{\text{JSTARS}'2015}$, LTTR [109]$^{\text{TNNLS}'2019}$, LTMR [67]$^{\text{TIP}'2019}$, UTV [110]$^{\text{JSTARS}'2020}$, DBIN [77]$^{\text{ICCV}'2019}$, SSRNet [89]$^{\text{TGRS}'2021}$, HSRNet [71]$^{\text{TNNLS}'2021}$, MoG-DCN [35]$^{\text{TIP}'2021}$, Fusformer [90]$^{\text{GRSL}'2022}$, DHIF [91]$^{\text{TCI}'2022}$, 3DTNet [84]$^{\text{IF}'2023}$, MIMO-SST [92]$^{\text{TGRS}'2024}$, DCINN [93]$^{\text{IJCV}'2024}$.

Datasets: We assess the performance on the CAVE and Harvard datasets simulating a scaling factor of 4/8. Details about the simulation procedure are provided in the supplementary material. We randomly chose 20 samples for the simulated training/validation dataset. The remaining 11 samples are used for testing, i.e., *balloons, cd, chart and stuffed toy, clay, fake and real beers, fake and real lemon slices, fake and real tomatoes, feathers, flowers, hairs, and jelly beans*.

Results: The results are reported in Table II. For scaling factor 4, we also showed the true-color images of the fusion results and the corresponding error maps in Fig. 7. It can be noted that both the details and color accuracy of the proposed method are closest to the GT. Besides, the high performance of our technique is also reported using scaling factor 8, see Table II. Table II generally shows that our method achieves competitive results compared to the benchmark.

B. Visible-Infrared Image Fusion

Setup: Since our method is a general model, we can substitute the MHIF fusion task with the visible and infrared (VIS-IR) image fusion problem. The related datasets (i.e., TNO [111][3] and RoadScene [17][4]) are publicly available. To build training and testing data, all red-green-blue (RGB) inputs are converted into the YCbCr color space, and then image fusion is performed between the IR image and the luminance (Y) channel. We compare the proposed FC-former with 11 representative state-of-the-art methods: NSST [94]$^{\text{TIM}'2018}$, DenseFuse [95]$^{\text{TIP}'2018}$, IFCNN

[1] http://www.cs.columbia.edu/CAVE/databases/
[2] http://vision.seas.harvard.edu/hyperspec/
[3] https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029
[4] https://github.com/hanna-xu/RoadScene

TABLE II
QUANTITATIVE RESULTS FOR THE MHIF TASK COMPARING SOME REPRESENTATIVE STATE-OF-THE-ART APPROACHES

	CAVE ×4: Avg±std				Harvard ×4: Avg±std				
	PSNR	SAM	ERGAS	SSIM	PSNR	SAM	ERGAS	SSIM	#params
MTF-GLP-HS [88]	37.74±3.88	4.51±1.60	4.55±2.65	0.974±0.016	41.09±4.37	2.76±0.76	3.41±1.04	0.974±0.013	/
LTMR [67]	37.07±3.60	5.77±2.00	5.06±2.40	0.968±0.021	40.87±3.96	4.00±1.27	4.03±2.17	0.957±0.035	/
DBIN [77]	50.88±4.25	2.15±0.63	1.23±1.03	0.997±0.003	47.91±3.88	2.31±0.46	1.99±0.82	0.988±0.007	0.469M
SSRNet [89]	48.62±3.92	2.54±0.84	1.64±1.22	0.995±0.002	48.00±3.36	2.31±0.60	2.30±1.26	0.987±0.007	0.03M
MoG-DCN [35]	51.69±4.09	1.97±0.61	1.10±0.81	0.997±0.002	47.96±4.10	2.11±0.52	1.91±0.83	0.988±0.007	7.07M
Fusformer [90]	49.98±8.10	2.20±0.85	2.53±5.31	0.994±0.011	47.87±5.13	2.84±2.07	2.04±0.99	0.986±0.010	0.11M
DHIF [91]	51.07±4.16	2.01±0.63	1.22±0.97	0.997±0.002	47.68±3.85	2.32±0.53	1.95±0.92	0.988±0.007	22.46M
3DTNet [84]	51.38±4.18	2.16±0.70	1.15±1.01	0.997±0.003	47.78±4.42	2.04±0.53	2.02±0.93	**0.989±0.006**	3.46M
MIMO-SST [92]	51.01±3.39	2.21±0.66	1.17±0.72	0.997±0.002	47.91±3.30	2.28±0.55	2.02±0.82	0.988±0.006	4.98M
DCINN [93]	52.21±4.25	1.93±0.61	1.04±0.84	0.998±0.002	48.77±3.59	**2.03±0.53**	**1.77±0.73**	0.989±0.007	4.32M
Proposed	**52.48±4.06**	**1.84±0.54**	**0.96±0.71**	**0.998±0.001**	**49.00±3.12**	2.08±0.52	1.85±0.93	0.989±0.007	3.75M
	CAVE ×8: Avg±std				Harvard ×8: Avg±std				
	PSNR	SAM	ERGAS	SSIM	PSNR	SAM	ERGAS	SSIM	#params
MTF-GLP-HS [88]	33.81±3.50	6.25±2.42	3.47±1.82	0.952±0.032	35.69±7.42	3.59±1.20	3.90±3.86	0.940±0.070	/
LTMR [67]	38.41±3.57	5.04±1.70	2.24±0.97	0.974±0.017	42.09±4.56	3.62±1.34	1.80±0.92	0.959±0.060	/
DBIN [77]	48.39±4.83	2.62±0.76	0.82±0.71	0.995±0.004	44.10±6.70	4.02±3.24	1.86±1.66	0.978±0.020	0.47M
SSRNet [89]	46.23±4.19	3.13±0.97	1.05±0.73	0.993±0.004	45.76±3.34	2.99±0.98	1.34±0.74	0.983±0.010	0.03M
MoG-DCN [35]	49.21±4.99	2.44±0.74	0.76±0.63	0.996±0.003	45.14±5.41	3.19±1.45	1.75±1.66	0.980±0.019	7.07M
Fusformer [90]	47.96±7.79	2.75±1.30	1.45±2.69	0.990±0.022	44.93±5.65	3.63±2.40	1.49±0.96	0.979±0.017	0.11M
DHIF [91]	48.46±4.89	2.50±0.79	0.83±0.67	0.996±0.003	45.00±4.13	3.70±1.68	1.32±0.61	0.983±0.011	22.46M
3DTNet [84]	49.41±5.83	2.26±0.66	0.83±1.07	0.996±0.003	44.41±5.38	2.93±0.88	1.55±0.89	0.983±0.010	3.46M
MIMO-SST [92]	48.31±5.04	2.88±0.86	0.89±0.79	0.995±0.004	46.59±3.34	2.91±0.75	2.29±1.03	0.985±0.009	4.98M
DCINN [93]	**49.84±4.83**	**2.39±0.73**	1.40±1.20	0.996±0.003	**46.89±3.77**	**2.74±0.79**	2.28±1.06	0.986±0.009	4.32M
Proposed	49.77±4.85	2.42±0.74	**0.69±0.56**	**0.996±0.002**	46.05±3.74	2.79±0.71	**1.25±0.69**	**0.986±0.009**	3.75M
Ideal value	∞	0	0	1	∞	0	0	1	0

Please, refer to Section IV-a for further details. M stands for million. Bold: best; underline: second best.

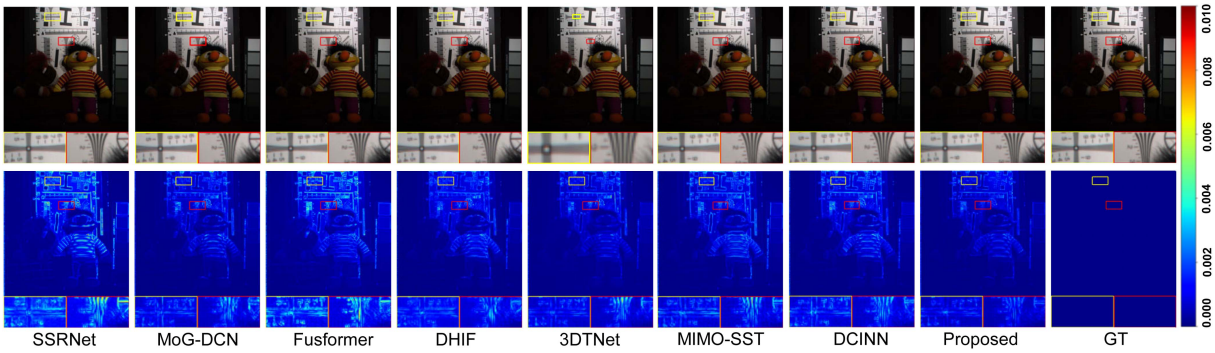


Fig. 7. In the first row, true-color fused images are depicted obtained by the proposed FC-Former and by some representative methods on *chart and stuffed toy* with scaling factor 4 for the CAVE dataset. In the second row, the related error maps (calculated between the fused image and the GT) are represented. Some close-ups are also considered.

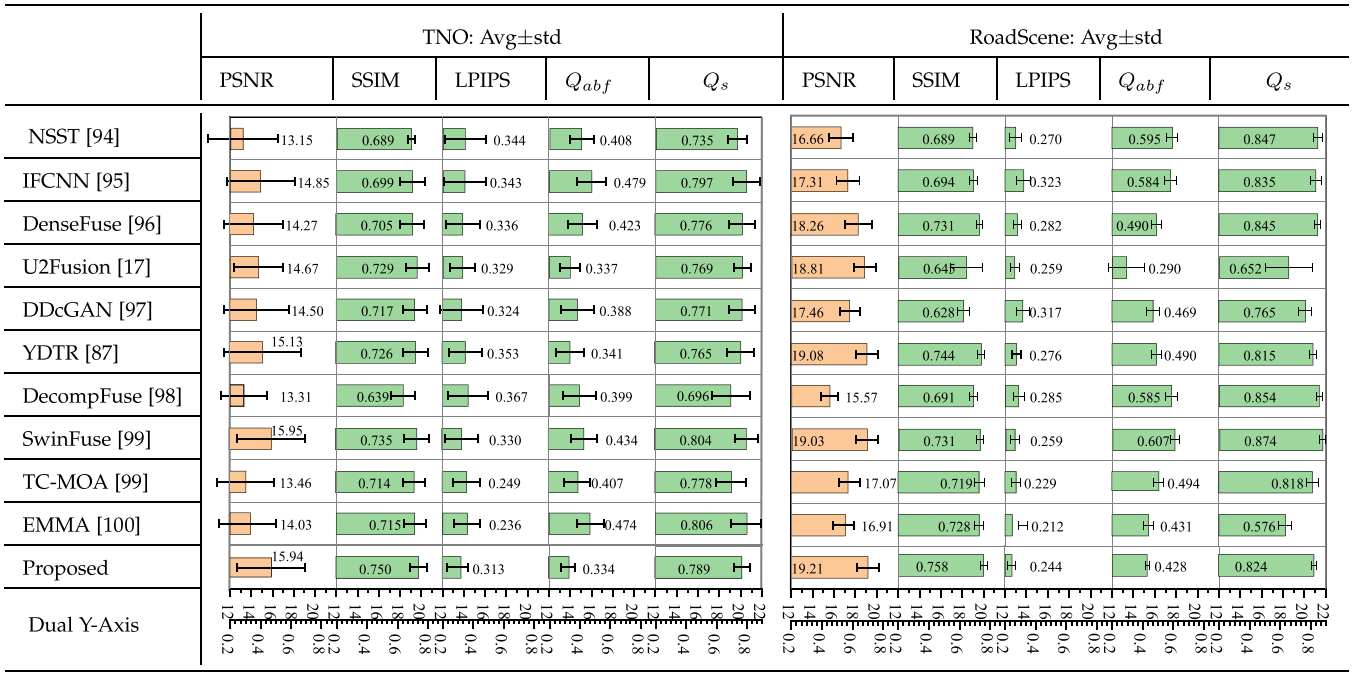
[96][IF′2020], DDcGAN [97][TIP′2020], U2Fusion [17][TPAMI′2020], YDTR [87][TMM′2022], DecompFusion [98][ECCV′2022], SwinFuse [112][TIM′2022], SwinFusion [74][JAS′2022], EMMA [100][CVPR′2024] and TC-MOA [99][CVPR′2024].

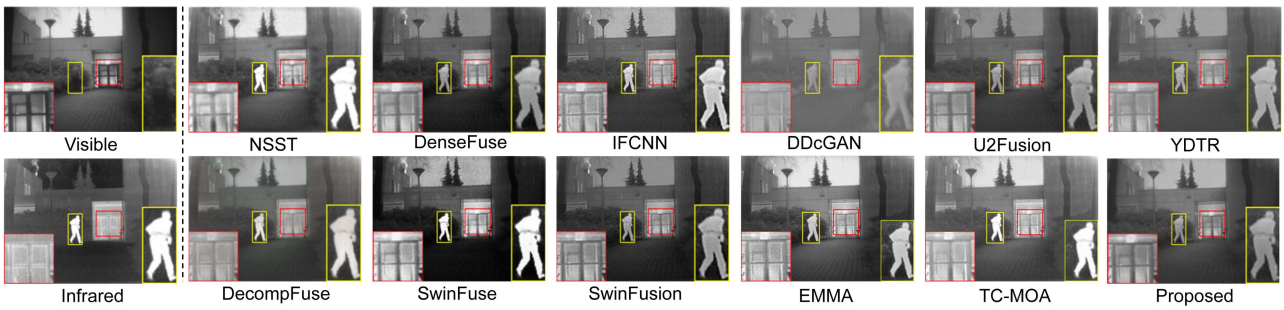
Datasets: According to [113] and the website[5], we have 98/38 training/test images for the TNO dataset. For the RoadScene dataset, we randomly selected 190/10/20 pairs for training/validation/test containing heterogeneous characteristics, such as roads, vehicles, and pedestrians. We use the same data augmentation strategy as U2Fusion [17] (i.e., images are randomly cropped to patches of size of 64×64 with flipping) to enlarge the number of samples.

Results: The quantitative results related to the TNO and RoadScene datasets are shown in Table III. Five quality metrics are used to assess the performance, i.e., the peak signal-to-noise ratio (PSNR) [114], the SSIM [115], the learned perceptual image patch similarity (LPIPS) [116], Q_{abf} [117], and Q_s [118]. It is

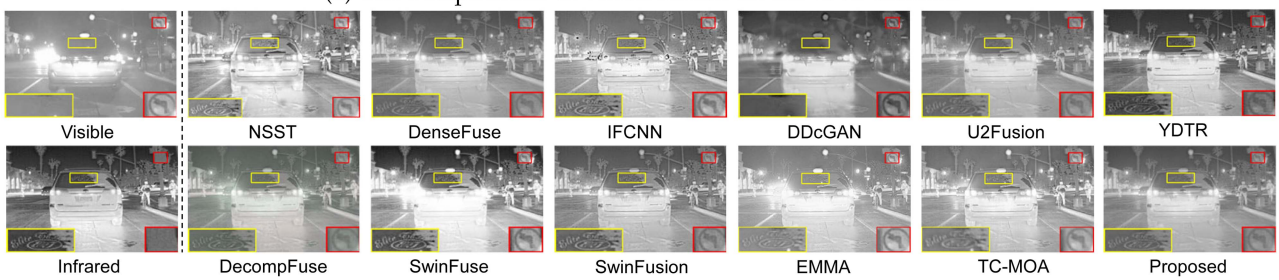
[5]https://github.com/Linfeng-Tang/Image-Fusion

TABLE III
QUANTITATIVE RESULTS FOR THE VIS-IR IMAGE FUSION TASK ON THE TNO AND ROADSCENE DATASETS. PLEASE, REFER TO SECTION IV-B FOR FURTHER DETAILS





(a) The comparisons of visual results on the TNO dataset.



(b) The comparisons of visual results on the RoadScene dataset.

Fig. 8. Comparison among some representative state-of-the-art methods for the VIS-IR image fusion task. Some close-ups are depicted in yellow and red boxes. No error map is depicted because of the absence of a GT.

clear that our FC-Former achieves state-of-the-art performance on both the VIS-IR datasets, getting superior SSIM and LPIPS metrics and top performance (close to the best) on the PSNR, Q_{abf}, and Q_s metrics. Some qualitative results are depicted in Fig. 8(a) and (b). It can be observed that our approach gets high performance, accurately preserving details without introducing issues, such as grayscale biases, artifacts, or noise.

C. Remote Sensing Pansharpening

Setup: We compare our method using a publicly available remote sensing pansharpening dataset [6], namely PanCollection, consisting of data acquired by WorldView-3 (WV3), QuickBird (QB), GaoFen-2 (GF2), and WorldView-2 (WV2) sensors. Reduced resolution data are simulated starting from

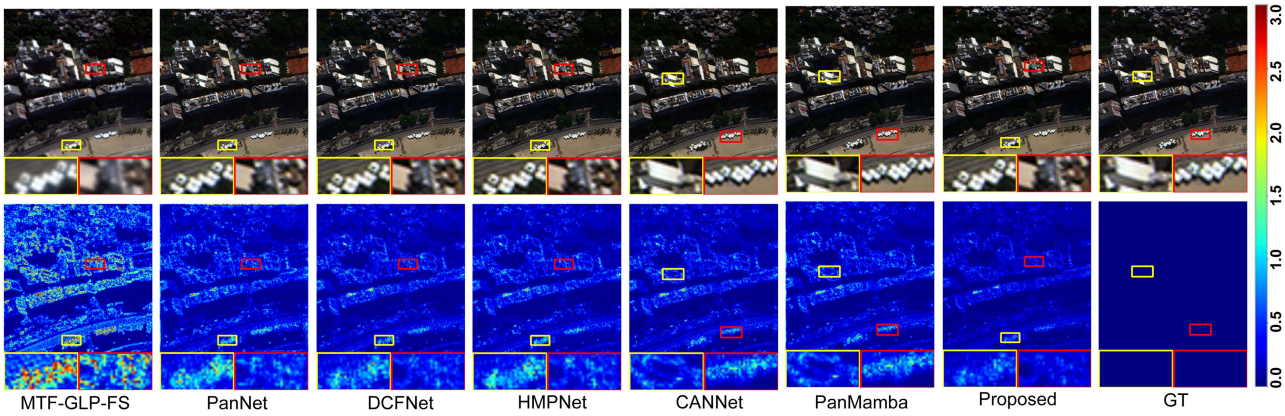


Fig. 9. Visual comparisons involving some representative pansharpening methods on one example of the reduced resolution WV3 dataset. True-color fused images are depicted in the first row. The second row is devoted to error maps between fused images and the GT. Some close-ups are also reported in yellow and red boxes.

real-world images exploiting Wald's protocol [119]. It is very valuable to perform experiments on remote sensing pansharpening because it allows for a more comprehensive evaluation of model performance in real fusion scenarios. In addition, we compare the proposed approach with 15 state-of-the-art methods. They are divided into four classes [120], i.e., CS, MRA, VO, and DL (CNNs and transformers) methods:[6]

1) Component substitution (CS) methods: BT-H [121]$^{\text{GRSL}'2017}$ and BDSD-PC [59]$^{\text{TGRS}'2019}$.
2) Multi-resolution analysis (MRA) approaches: the generalized Laplacian pyramid (GLP) with modulation transfer function (MTF)-matched filters [122] and its full-scale regression version (MTF-GLP-FS) [101]$^{\text{TIP}'2018}$.
3) A variational optimization-based (VO) technique: LRTCFPan [27]$^{\text{TIP}'2023}$.
4) DL methods: 11 CNN-based approaches, such as PNN [69]$^{\text{RS}'2016}$, PanNet [70]$^{\text{ICCV}'2017}$, MSDCNN [123]$^{\text{JSTARS}'2018}$, DiCNN [124]$^{\text{JSTARS}'2019}$, FusionNet [102]$^{\text{TGRS}'2020}$, LAGNet [72]$^{\text{AAAI}'2021}$, DCFNet [58]$^{\text{ICCV}'2021}$, HMPNet [103]$^{\text{TNNLS}'2023}$, CANNet [104]$^{\text{CVPR}'2024}$, and PanMamba [105]$^{\text{Arxiv}'2024}$, and one transformer-based technique, i.e., Invformer [57]$^{\text{AAAI}'2022}$. All the compared DL methods are trained on the same data using default experimental settings (as suggested in the related papers) for fair comparison.

Datasets: The dataset can be downloaded at.[7] We chose WV3, GF2, and QB data for performance assessment, and WV2 data to test network generalization. The number of testing samples for each reduced resolution and full resolution dataset is 20 (i.e., 160 images in total). Please, refer to the supplementary material for further details.

Results: To evaluate the quality of the proposed method, we use reference and no-reference quality metrics.[8] The reduced resolution assessment exploits the following reference-based quality metrics: the spatial correlation coefficient [125] (SCC), the spectral angle mapper [126] (SAM), the erreur relative globale adimensionnelle de synthése [127] (ERGAS), and the Q2n (Q8 for 8-band data and Q4 for 4-band data). From Table IV, the proposed FC-Former outperforms the other methods for almost all metrics, and it is very close to the optimal values for the rest of the cases. Fig. 9 shows the fused results with the related error maps to appreciate the goodness of the outcome of the proposed approach.

To assess the performance on real (full resolution) data, where the reference image is not available, indexes without reference are used, i.e., the spectral distortion (D_λ), the spatial distortion (D_s), and the hybrid quality with no reference (HQNR) indexes [128]. Table IV reports the average performance on the full resolution (FR) examples for the exploited public dataset. Again, FC-Former obtains the best results on average with the lowest standard deviations, showing its superiority and greater stability. Furthermore, in Fig. 10, we show visual results on a full resolution WV3 example. The outcome of the proposed FC-Former shows more details and a better visual quality.

V. Discussions

In this section, we will discuss the components of the proposed FC-Former. Without loss of generality, we consider the MHIF problem using the CAVE $\times 4$ dataset as an example.

A. MSIR

We analyze the FC-Former combined with classical fusion methods. Two model-based fusion methods, model-based branch fusion (MBF)-1 and MBF-2, have been adopted. In MBF-1 [53], the mutual-projected fusion is used to replace the simple addition or concatenation of feature maps. The residual between two features from the in-scale and cross-scale branches is downsampled into the original identity branch, transferring information from both cross-scale and in-scale features. In contrast, MBF-2 [35] adopts model-guided fusion to perform

[6] All the obtained results are reported in the supplementary material.
[7] https://github.com/liangjiandeng/PanCollection
[8] https://github.com/liangjiandeng/DLPan-Toolbox/tree/main/02-Test-toolbox-for-traditional-and-DL(Matlab)

TABLE IV
QUANTITATIVE RESULTS FOR THE REMOTE SENSING PANSHARPENING TASK COMPARING SOME REPRESENTATIVE STATE-OF-THE-ART APPROACHES

		Reduced Resolution (RR): Avg±std				Full Resolution (FR): Avg±std		
		SAM	ERGAS	Q2n	SCC	D_λ	D_s	HQNR
WorldView-3 (WV3, 8 bands)	MTF-GLP-FS [101]	5.32±1.65	4.64±1.44	0.818±0.101	0.899±0.047	0.021±0.008	0.063±0.028	0.918±0.035
	LRTCFPan [27]	4.67±1.28	4.23±1.28	0.836±0.097	0.927±0.023	0.018±0.007	0.053±0.026	0.931±0.031
	PanNet [70]	3.74±0.68	2.82±0.71	0.862±0.106	0.975±0.008	**0.017±0.007**	0.047±0.021	0.937±0.027
	FusionNet [102]	3.31±0.63	2.45±0.59	0.896±0.093	0.980±0.006	0.024±0.009	0.036±0.014	0.941±0.020
	LAGNet [72]	3.10±0.50	2.29±0.56	0.902±0.091	0.983±0.006	0.037±0.015	0.042±0.015	0.923±0.025
	Invformer [57]	3.25±0.64	2.39±0.52	0.906±0.084	0.983±0.005	0.055±0.029	0.068±0.031	0.882±0.049
	DCFNet [58]	3.03±0.74	2.16±0.46	0.905±0.088	0.986±0.004	0.078±0.081	0.051±0.034	0.877±0.101
	HMPNet [103]	3.04±0.52	2.22±0.49	0.913±0.084	0.986±0.004	0.018±0.007	0.053±0.006	0.929±0.011
	CANNet [104]	2.92±0.54	2.20±0.47	0.914±0.083	0.985±0.004	0.020±0.009	0.029±0.008	0.951±0.010
	PanMamba [105]	2.94±0.54	2.20±0.51	0.916±0.090	0.985±0.006	0.020±0.007	0.031±0.003	0.954±0.070
	Proposed	**2.72±0.51**	**1.98±0.43**	**0.919±0.430**	**0.989±0.003**	0.020±0.008	**0.025±0.004**	**0.955±0.012**
GaoFen2 (GF2, 4 bands)	MTF-GLP-FS [101]	1.68±0.35	1.62±0.36	0.890±0.026	0.939±0.016	0.035±0.014	0.143±0.028	0.823±0.035
	LRTCFPan [27]	1.31±0.28	1.30±0.31	0.932±0.033	0.958±0.013	0.033±0.027	0.090±0.014	0.881±0.023
	PanNet [70]	1.04±0.19	1.02±0.15	0.955±0.021	0.980±0.003	0.021±0.011	0.080±0.018	0.901±0.020
	FusionNet [102]	0.98±0.19	1.00±0.20	0.978±0.005	0.978±0.005	0.040±0.013	0.101±0.013	0.863±0.018
	LAGNet [72]	0.80±0.14	0.71±0.11	0.979±0.011	0.989±0.002	0.032±0.013	0.079±0.014	0.891±0.020
	Invformer [57]	0.83±0.14	0.70±0.11	0.977±0.012	0.980 ±0.002	0.059±0.026	0.110±0.015	0.838±0.024
	DCFNet [58]	0.89±0.16	0.81±0.14	0.973±0.010	0.985±0.002	0.023±0.012	0.066±0.010	0.912±0.012
	HMPNet [103]	0.80±0.14	0.56±0.10	0.981±0.030	0.993±0.003	0.080±0.050	0.115±0.012	0.815±0.049
	CANNet [104]	0.72±0.14	0.65±0.12	0.982±0.007	0.991±0.002	0.019±0.010	0.063±0.009	0.919 ±0.011
	PanMamba [105]	0.68±0.12	0.64±0.10	0.982±0.008	0.985±0.006	**0.016±0.008**	0.045±0.009	0.939±0.010
	Proposed	**0.60±0.11**	**0.52±0.09**	**0.987±0.007**	**0.994±0.001**	0.018±0.009	**0.027±0.008**	**0.955±0.011**
QuickBird (QB, 4 bands)	MTF-GLP-FS [101]	7.87±1.67	7.45±0.56	0.834±0.096	0.902±0.025	0.049±0.015	0.138±0.024	0.820±0.034
	LRTCFPan [27]	7.29±1.60	7.03±0.62	0.853±0.095	0.914±0.014	**0.023±0.012**	0.071±0.035	0.909±0.044
	PanNet [70]	5.88±1.07	6.02±0.79	0.881±0.102	0.947±0.018	0.041±0.011	0.114±0.032	0.850±0.039
	FusionNet [102]	4.96±0.84	4.19±0.25	0.923±0.097	0.976±0.010	0.059±0.019	0.052±0.009	0.892±0.022
	LAGNet [72]	4.58±0.77	3.87±0.36	0.932±0.094	0.981±0.009	0.084±0.024	0.068±0.014	0.854±0.018
	Invformer [57]	4.66±0.78	3.70±0.29	0.932±0.007	0.983±0.007	0.174±0.033	0.073±0.024	0.766±0.043
	DCFNet [58]	4.54±0.73	3.85±0.28	0.932±0.093	0.974±0.010	0.045±0.015	0.124±0.027	0.836±0.016
	HMPNet [103]	4.72±0.38	3.66±0.40	0.930±0.110	**0.988±0.009**	0.183±0.054	0.079±0.025	0.754±0.065
	CANNet [104]	4.54±0.79	3.74±0.31	0.935±0.089	0.982±0.007	0.038±0.013	0.047±0.009	**0.917±0.012**
	PanMamba [105]	4.74±0.88	4.38±0.60	0.923±0.092	0.975±0.011	0.049±0.013	0.044±0.016	0.910±0.027
	Proposed	**4.35±0.73**	**3.57±0.27**	**0.938±0.087**	0.984±0.007	0.059±0.019	**0.040±0.018**	0.903±0.018
	Ideal value	**0**	**0**	**1**	**1**	**0**	**0**	**1**

Q8/Q4 is the Q2n index for 8-band/4-band images, respectively. Please, refer to Section IV-c for further details. bold: best; underline: second best.

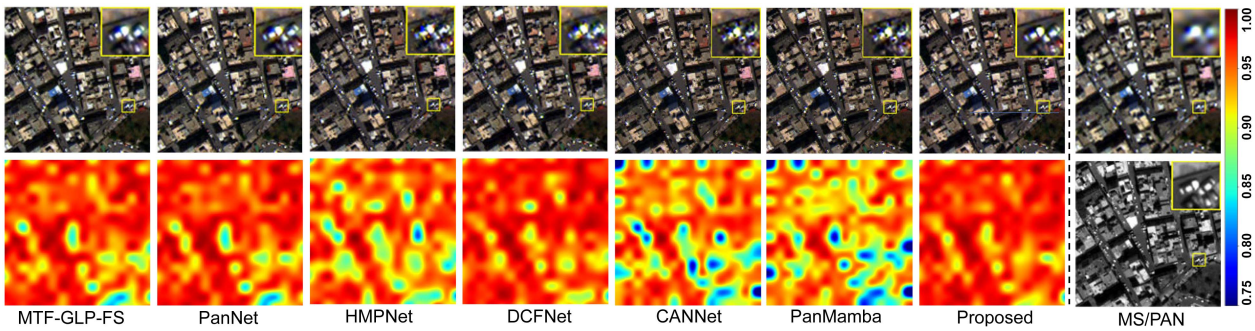


Fig. 10. Visual comparisons involving some representative pansharpening methods on one example of the full resolution WV3 dataset. True-color fused images are depicted in the first row. The second row is devoted to the HQNR maps. Some close-ups are also reported in yellow boxes.

TABLE V
COMPARISON OF BRANCH FUSION APPROACHES ON THE CAVE ×4 DATASET

Method	DBF	MBF-1	MBF-2
PSNR (±std)	52.48±4.06	52.54±4.13	**52.77±4.07**
SAM (±std)	1.84±0.54	**1.77±0.52**	1.78±0.52
ERGAS (±std)	0.95±0.71	**0.94±0.71**	0.95±0.71
SSIM (±std)	0.998±0.0013	0.998±0.0013	0.998±0.0013
FLOPs	4.0G	4.9G	4.0G

G stands for giga (billion). Bold: best.

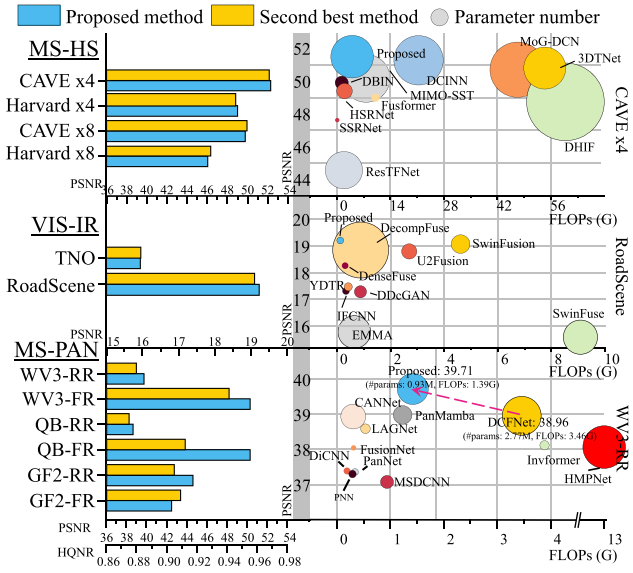


Fig. 11. The PSNR values between the proposed FC-Former and the second best method for the three fusion tasks. The FC-Former achieves robust and superior performance with a small parameter number and FLOPs considering the CAVE ×4, the RoadScene, and the WorldView-3 (WV3) datasets. RR denotes reduced resolution data simulated starting from full resolution (FR) images. The red dotted arrow, in the remote sensing pansharpening case, indicates the performance gain compared to the conference version [58]. The circle radius indicates the parameter number (i.e., the larger the circle, the higher the parameter number).

the MSIR operation. The MBF-2 explicitly incorporates the observation model into the MHIF problem, where a convolution network is used as a denoiser and guidance. In contrast, the deep network based on the FCSA framework can get a nonlinear representation with a non-local self-similarity prior. The DBF method represents the basic implementation for the MSIR module. Table V shows that the FC-Former has a better performance by using MBFs to deal with multi-source inputs. Finally, it indicates that our FC-Former is an interpretable network, which is capable of combining the advantages of the model-driven and data-driven approaches.

B. Complexity Analysis

It is well known that there is a trade-off between the performance of DL methods and the number of parameters (or computational cost). Fig. 11 shows these trade-offs for 23 state-of-the-art approaches belonging to the considered three fusion tasks. It can be concluded that the proposed FC-Former achieves the best trade-off. Moreover, in Fig. 12, we show that the floating point operations per second (FLOPs) of the FC-Former are quite

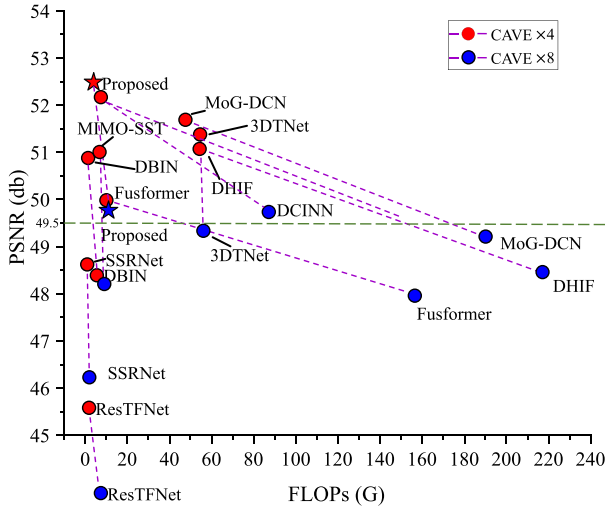


Fig. 12. PSNR vs. FLOPs for all the high performance methods on MSI/HSI images with sizes of $16 \times 16/64 \times 64$ and $16 \times 16/128 \times 128$ related to the CAVE ×4 and ×8 datasets, respectively. The proposed FC-Former (indicated with a star marker) gets the best PSNR values with a small amount of FLOPs.

low, yet it achieves the highest performance in both the CAVE test cases.

VI. ABLATION STUDY

We consider in this section, without loss of generality, the CAVE ×4 dataset to conduct ablation studies.

A. Fully-Connected Self-Attention

The proposed FCSA framework, corresponding to the multilinear algebra in Section III-A, retains several forms of self-attention along different unfolded modes. To explore the effects of the FCSA framework, we perform ablation studies by considering five improved DCFNet methods and using the original DCFNet as a baseline. Table VI reports the average and the corresponding standard deviation results for the proposed FC-Former approaches. It can be observed the beneficial effects of adding self-attention into the baseline. Indeed, fusion results with self-attention get higher performance on all the metrics. When the cross-scale self-attention is the spatial self-attention (Spa-SA) and the in-scale self-attention is the channel self-attention (CAL-SA), the FCSA framework achieves the best results.

B. Spatial Multi-Head Self-Attention

Leveraging the trade-off between global dependency and computational complexity in spatial multi-head self-attention (Spa-MSA), we employ a window-based Spa-MSA to model long-range information along the spatial mode. In Table VII, we compare the effects of different window sizes, reducing it from the HR to the LR branch to have a long-range response with a flexible range. We chose a window size of $16/8/4$ as a good trade-off between computational burden and performance.

TABLE VI
ABLATION STUDIES FOR THE FCSA FRAMEWORK USING CROSS-SCALE AND/OR IN-SCALE ATTENTION

Method		Cross-scale attention		In-scale attention			CAVE $\times 4$: Avg$\pm$std			
		Spa-SA	CAL-SA	Spa-SA	CAL-SA	Patch-SA	PSNR	SAM	ERGAS	SSIM
DCFNet	I						49.50±3.84	2.35±0.74	1.50±0.98	0.996±0.0018
	II		✓				51.35±3.56	2.15±0.69	1.07±0.73	0.997±0.0012
FC-Former	III	✓					51.50±3.56	2.23±0.75	1.05±0.72	0.997±0.0011
	IV	✓		✓			52.08±3.90	1.92±0.59	0.98±0.71	**0.998±0.0013**
	V					✓	52.10±3.79	1.97±0.57	1.01±0.70	0.997±0.0011
	VI	✓			✓		**52.48±4.06**	**1.84±0.54**	**0.96±0.71**	**0.998±0.0013**

The baseline is represented by the DCFNet without self-attention blocks. Bold: best.

TABLE VII
A COMPARISON OF DIFFERENT WINDOW SIZES FOR THE WINDOW-BASED SPA-MSA FOR THE HR, THE MR, AND THE LR BRANCHES

Window Size	32/16/8	16/8/4	8/4/2
PSNR (±std)	52.45±4.03	**52.48±4.06**	52.46±3.99
SAM (±std)	**1.82±0.53**	1.84±0.54	1.82±0.54
ERGAS (±std)	0.96±0.71	0.96±0.71	0.96±0.71
SSIM (±std)	0.998±0.001	0.998±0.001	0.998±0.001

BOLD: BEST.

VII. CONCLUSION

In this paper, inspired by multilinear algebra, we proposed the mathematical idea of the generalized self-attention to unify and generalize existing self-attention mechanisms. Based on this generalized mechanism, we developed the first fully-connected self-attention framework that captures intra- and cross-scale patterns, as well as local and non-local similarities. Through theoretical analysis and broad experiments, the proposed FCSA framework addresses the representation issue at different dimensions (modes) and scales while achieving better detail reconstruction and lower computational costs. Afterwards, we built a fully-connected transformer network using the FCSA framework, called FC-Former. In this case, the multi-source image representation module provides support to improve the physical interpretation of the network and to guide the FCSA regularization. FC-Former demonstrated superior performance with high efficiency and low parameters (and computational costs) for MHIF, VIS-IR image fusion, remote sensing pansharpening, and digital photographic image fusion. Thanks to the positive impact of strong feature representations for different fusion tasks, the proposed method can outperform some state-of-the-art methods, specially designed for the above-mentioned problems, demonstrating its usefulness for a wide range of image processing tasks.

REFERENCES

[1] T. Zhou, S. Ruan, and S. Canu, "A review: Deep learning for medical image segmentation using multi-modality fusion," *Array*, vol. 3, 2019, Art. no. 100004.

[2] Y. Guo, D. Zhou, X. Ruan, and J. Cao, "Variational gated autoencoder-based feature extraction model for inferring disease-MIRNA associations based on multiview features," *Neural Netw.*, vol. 165, pp. 491–505, 2023.

[3] W. Li, Y. Guo, B. Wang, and B. Yang, "Learning spatiotemporal embedding with gated convolutional recurrent networks for translation initiation site prediction," *Pattern Recognit.*, vol. 136, 2023, Art. no. 109234.

[4] X. Zhang and Y. Demiris, "Visible and infrared image fusion using deep learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 10535–10554, Aug. 2023.

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

[6] L. J. Deng et al., "Machine learning in pansharpening: A benchmark, from shallow to deep networks," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 3, pp. 279–315, Sep. 2022.

[7] W. He et al., "Non-local meets global: An iterative paradigm for hyperspectral image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 2089–2107, Apr. 2022.

[8] Y. Guo, D. Zhou, P. Li, C. Li, and J. Cao, "Context-aware poly(A) signal prediction model via deep spatial–temporal neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 8241–8253 2022.

[9] X. Deng and P. L. Dragotti, "Deep convolutional neural network for multi-modal image restoration and fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3333–3348, Oct. 2020.

[10] S. Yang, D. Zhou, J. Cao, and Y. Guo, "LightingNet: An integrated learning method for low-light image enhancement," *IEEE Trans. Comput. Imag.*, vol. 9, pp. 29–42, 2023.

[11] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019.

[12] L. Tang, J. Yuan, and J. Ma, "Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network," *Inf. Fusion*, vol. 82, pp. 28–42, 2022.

[13] H. Li, X.-J. Wu, and J. Kittler, "RFN-Nest: An end-to-end residual fusion network for infrared and visible images," *Inf. Fusion*, vol. 73, pp. 72–86, 2021.

[14] D. Liu et al., "Transfusion: Multi-view divergent fusion for medical image segmentation with transformers," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Interv.*, 2022, pp. 485–495.

[15] Q. Xie, M. Zhou, Q. Zhao, Z. Xu, and D. Meng, "MHF-Net: An interpretable deep network for multispectral and hyperspectral image fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1457–1473, Mar. 2022.

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

[17] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 502–518, Jan. 2022.

[18] Z. Liu, E. Blasch, Z. Xue, J. Zhao, R. Laganiere, and W. Wu, "Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: A comparative study," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 94–109, Jan. 2012.

[19] G. Vivone, M. Dalla Mura, A. Garzelli, and F. Pacifici, "A benchmarking protocol for pansharpening: Dataset, preprocessing, and quality assessment," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6102–6118, 2021.

[20] Y. Yan, J. Liu, S. Xu, Y. Wang, and X. Cao, "MD^3 Net: Integrating model-driven and data-driven approaches for pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5411116.

[21] Y. Liang, P. Zhang, Y. Mei, and T. Wang, "PMACNet: Parallel multiscale attention constraint network for pan-sharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 5512805.

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

[23] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[24] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10012–10022.

[25] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 568–578.

[26] D. Liu, B. Wen, Y. Fan, C. C. Loy, and T. S. Huang, "Non-local recurrent network for image restoration," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 1680–1689.

[27] Z. C. Wu, T. Z. Huang, L. J. Deng, J. Huang, J. Chanussot, and G. Vivone, "LRTCFPan: Low-rank tensor completion based framework for pansharpening," *IEEE Trans. Image Process.*, vol. 32, pp. 1640–1655, 2023.

[28] S. Karim, G. Tong, J. Li, A. Qadir, U. Farooq, and Y. Yu, "Current advances and future perspectives of image fusion: A comprehensive review," *Inf. Fusion*, vol. 90, pp. 185–217, 2023.

[29] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 136–144.

[30] Y. Wang, L.-J. Deng, T.-J. Zhang, and X. Wu, "SSconv: Explicit spectral-to-spatial convolution for pansharpening," in *Proc. ACM Int. Conf. Multimedia*, 2021, pp. 4472–4480.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[32] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assisted Interv.*, 2015, pp. 234–241.

[34] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5693–5703.

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

[36] Q. Ma, J. Jiang, X. Liu, and J. Ma, "Deep unfolding network for spatiospectral image super-resolution," *IEEE Trans. Comput. Imag.*, vol. 8, pp. 28–40, 2021.

[37] D. Geman and G. Reynolds, "Constrained restoration and the recovery of discontinuities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 3, pp. 367–383, Mar. 1992.

[38] D. Geman and C. Yang, "Nonlinear image recovery with half-quadratic regularization," *IEEE Trans. Image Process.*, vol. 4, no. 7, pp. 932–946, Jul. 1995.

[39] S. Boyd et al., "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. TrendsR PLX Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[40] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep CNN denoiser prior for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3929–3938.

[41] K. Zhang, L. V. Gool, and R. Timofte, "Deep unfolding network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3217–3226.

[42] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9446–9454.

[43] K. Zhang, Y. Li, W. Zuo, L. Zhang, L. Van Gool, and R. Timofte, "Plug-and-play image restoration with deep denoiser prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6360–6376, Oct. 2022.

[44] L. Wang, C. Sun, M. Zhang, Y. Fu, and H. Huang, "DNU: Deep non-local unrolling for computational spectral imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1661–1671.

[45] Y. Han, G. Huang, S. Song, L. Yang, H. Wang, and Y. Wang, "Dynamic neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7436–7456, Nov. 2022.

[46] S. Peng, L.-J. Deng, J.-F. Hu, and Y. Zhuo, "Source-adaptive discriminative kernels based network for remote sensing pansharpening," in *Proc. Int. Joint Conf. Artif. Intell.*, 2022, pp. 1283–1289.

[47] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using swin transformer," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 1833–1844.

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

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

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

[51] M. Li, Y. Fu, and Y. Zhang, "Spatial-spectral transformer for hyperspectral image denoising," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 1368–1376.

[52] Y. Peng, Y. Zhang, B. Tu, Q. Li, and W. Li, "Spatial–spectral transformer with cross-attention for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5537415.

[53] Y. Mei, Y. Fan, Y. Zhou, L. Huang, T. S. Huang, and H. Shi, "Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5690–5699.

[54] S. Zhou, J. Zhang, W. Zuo, and C. C. Loy, "Cross-scale internal graph neural network for image super-resolution," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 3499–3509.

[55] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear analysis of image ensembles: Tensorfaces," in *Proc. Eur. Conf. Comput. Vis.*, 2002, pp. 447–460.

[56] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, 2009.

[57] M. Zhou, X. Fu, J. Huang, F. Zhao, A. Liu, and R. Wang, "Effective pan-sharpening with transformer and invertible neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5406815.

[58] X. Wu, T.-Z. Huang, L.-J. Deng, and T.-J. Zhang, "Dynamic cross feature fusion for remote sensing pansharpening," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 14687–14696.

[59] G. Vivone, "Robust band-dependent spatial-detail approaches for panchromatic sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6421–6433, Sep. 2019.

[60] G. Vivone et al., "Pansharpening based on semiblind deconvolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 1997–2010, Apr. 2015.

[61] G. Vivone, S. Marano, and J. Chanussot, "Pansharpening: Context-based generalized Laplacian pyramids by robust regression," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6152–6167, Sep. 2020.

[62] N. Akhtar and A. Mian, "Hyperspectral recovery from RGB images using Gaussian processes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 100–113, Jan. 2020.

[63] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[64] Q. Zhang, Y. Liu, R. S. Blum, J. Han, and D. Tao, "Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review," *Inf. Fusion*, vol. 40, pp. 57–75, 2018.

[65] X. Fu, Z. Lin, Y. Huang, and X. Ding, "A variational pan-sharpening with local gradient constraints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10265–10274.

[66] L.-J. Deng, G. Vivone, W. Guo, M. Dalla Mura, and J. Chanussot, "A variational pansharpening approach based on reproducible kernel Hilbert space and heaviside function," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4330–4344, Sep. 2018.

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

[68] T. Xu, T.-Z. Huang, L.-J. Deng, and N. Yokoya, "An iterative regularization method based on tensor subspace representation for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5529316.

[69] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sens.*, vol. 8, no. 7, 2016, Art. no. 594.

[70] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, "PanNet: A deep network architecture for pan-sharpening," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 5449–5457.

[71] J.-F. Hu, T.-Z. Huang, L.-J. Deng, T.-X. Jiang, G. Vivone, and J. Chanussot, "Hyperspectral image super-resolution via deep spatiospectral attention convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 7251–7265, Dec. 2022.

[72] Z.-R. Jin, T.-J. Zhang, T.-X. Jiang, G. Vivone, and L.-J. Deng, "LAGConv: Local-context adaptive convolution kernels with global harmonic bias for pansharpening," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 1113–1121.

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

[74] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer," *IEEE-CAA J. Automatica Sinica*, vol. 9, no. 7, pp. 1200–1217, Jul. 2022.

[75] L. Tang, Y. Deng, Y. Ma, J. Huang, and J. Ma, "SuperFusion: A versatile image registration and fusion network with semantic awareness," *IEEE-CAA J. Automatica Sinica*, vol. 9, no. 12, pp. 2121–2137, Dec. 2022.

[76] H. Liu, C. Feng, R. Dian, and S. Li, "SSTF-Unet: Spatial–spectral transformer-based U-Net for high-resolution hyperspectral image acquisition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 12, pp. 18222–18236, Dec. 2023.

[77] W. Wang, W. Zeng, Y. Huang, X. Ding, and J. W. Paisley, "Deep blind hyperspectral image fusion," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 4149–4158.

[78] B. Lecouat, J. Ponce, and J. Mairal, "Fully trainable and interpretable non-local sparse models for image restoration," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 238–254.

[79] N. Park and S. Kim, "How do vision transformers work?," 2021, *arXiv:2202.06709*.

[80] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[81] A. Kolesnikov et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[82] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," 2019, *arXiv: 1912.01703*.

[83] H. Chen et al., "Pre-trained image process.ing transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12299–12310.

[84] Q. Ma, J. Jiang, X. Liu, and J. Ma, "Learning a 3D-CNN and transformer prior for hyperspectral image super-resolution," *Inf. Fusion*, vol. 100, 2023, Art. no. 101907.

[85] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–41, 2022.

[86] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 399–406.

[87] W. Tang, F. He, and Y. Liu, "YDTR: Infrared and visible image fusion via Y-shape dynamic transformer," *IEEE Trans. Multimedia*, vol. 25, pp. 5413–5428, 2023.

[88] M. Selva, B. Aiazzi, F. Butera, L. Chiarantini, and S. Baronti, "Hyper-sharpening: A first approach on SIM-GA data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 3008–3024, Jun. 2015.

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

[90] J.-F. Hu, T.-Z. Huang, L.-J. Deng, H.-X. Dou, D. Hong, and G. Vivone, "Fusformer: A transformer-based fusion network for hyperspectral image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6012305.

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

[92] J. Fang, J. Yang, A. Khader, and L. Xiao, "MIMO-SST: Multi-input multi-output spatial-spectral transformer for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5510020.

[93] W. Wang, L.-J. Deng, R. Ran, and G. Vivone, "A general paradigm with detail-preserving conditional invertible network for image fusion," *Int. J. Comput. Vis.*, vol. 132, no. 4, pp. 1029–1054, 2024.

[94] M. Yin, X. Liu, Y. Liu, and X. Chen, "Medical image fusion with parameter-adaptive pulse coupled neural network in nonsubsampled Shearlet transform domain," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 1, pp. 49–64, Jan. 2019.

[95] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, May 2019.

[96] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "IFCNN: A general image fusion framework based on convolutional neural network," *Inf. Fusion*, vol. 54, pp. 99–118, 2020.

[97] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4980–4995, 2020.

[98] P. Liang, J. Jiang, X. Liu, and J. Ma, "Fusion from decomposition: A self-supervised decomposition approach for image fusion," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 719–735.

[99] P. Zhu, Y. Sun, B. Cao, and Q. Hu, "Task-customized mixture of adapters for general image fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 7099–7108.

[100] Z. Zhao et al., "Equivariant multi-modality image fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 25912–25921.

[101] G. Vivone, R. Restaino, and J. Chanussot, "Full scale regression-based injection coefficients for panchromatic sharpening," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3418–3431, Jul. 2018.

[102] L.-J. Deng, G. Vivone, C. Jin, and J. Chanussot, "Detail injection-based deep convolutional neural networks for pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6995–7010, Aug. 2021.

[103] X. Tian, K. Li, W. Zhang, Z. Wang, and J. Ma, "Interpretable model-driven deep network for hyperspectral, multispectral, and panchromatic image fusion," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 10, pp. 14382–14395, Oct. 2024.

[104] Y. Duan, X. Wu, H. Deng, and L.-J. Deng, "Content-adaptive non-local convolution for remote sensing pansharpening," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 27738–27747.

[105] X. He et al., "Pan-mamba: Effective pan-sharpening with state space model," 2024, *arXiv:2402.12192*.

[106] F. Yasuma, T. Mitsunaga, D. Iso, and S. K. Nayar, "Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2241–2253, Sep. 2010.

[107] A. Chakrabarti and T. Zickler, "Statistics of real-world hyperspectral images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 193–200.

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

[109] R. Dian, S. Li, and L. Fang, "Learning a low tensor-train rank representation for hyperspectral image super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2672–2683, Sep. 2019.

[110] T. Xu, T.-Z. Huang, L.-J. Deng, X.-L. Zhao, and J. Huang, "Hyperspectral image superresolution using unidirectional total variation with tucker decomposition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4381–4398, 2020.

[111] A. Toet, "The TNO multiband image data collection," *Data Brief*, vol. 15, pp. 249–251, 2017.

[112] Z. Wang, Y. Chen, W. Shao, H. Li, and L. Zhang, "SwinFuse: A residual swin transformer fusion network for infrared and visible images," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 5016412.

[113] L. Tang, H. Zhang, H. Xu, and J. Ma, "Deep learning-based image fusion: A survey," *J. Image Graph.*, vol. 28, no. 1, pp. 3–36, 2023.

[114] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electron. Lett.*, vol. 44, no. 13, pp. 800–801, 2008.

[115] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[116] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.
[117] C. S. Xydeas et al., "Objective image fusion performance measure," *Electron. Lett.*, vol. 36, no. 4, pp. 308–309, 2000.
[118] G. Piella and H. Heijmans, "A new quality metric for image fusion," in *Proc. IEEE 2003 Int. Conf. Image Process.*, 2003, pp. III– 173.
[119] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogrammetric Eng. Remote Sens.*, vol. 63, no. 6, pp. 691–699, 1997.
[120] G. Vivone et al., "A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 1, pp. 53–81, Mar. 2021.
[121] S. Lolli, L. Alparone, A. Garzelli, and G. Vivone, "Haze correction for contrast-based multispectral pansharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2255–2259, Dec. 2017.
[122] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, "Mtf-tailored multiscale fusion of high-resolution ms and pan imagery," *Photogrammetric Eng. Remote Sens.*, vol. 72, no. 5, pp. 591–596, 2006.
[123] Q. Yuan, Y. Wei, X. Meng, H. Shen, and L. Zhang, "A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 3, pp. 978–989, Mar. 2018.
[124] L. He et al., "Pansharpening via detail injection based convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 4, pp. 1188–1204, Apr. 2019.
[125] J. Zhou, D. L. Civco, and J. A. Silander, "A wavelet transform method to merge Landsat TM and spot panchromatic data," *Int. J. Remote Sens.*, vol. 19, no. 4, pp. 743–757, 1998.
[126] R. H. Yuhas, A. F. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm," in *Proc. Summaries Third Annu. JPL Airborne Geosci. Workshop*, 1992, vol. 1, pp. 147–149.
[127] L. Wald, *Data Fusion: Definitions and Architectures: Fusion of Images of Different Spatial Resolutions*. Vouillé, France: Presses des MINES, 2002.
[128] A. Arienzo, G. Vivone, A. Garzelli, L. Alparone, and J. Chanussot, "Full-resolution quality assessment of pansharpening: Theoretical and hands-on approaches," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 3, pp. 168–201, Sep. 2022.

Xiao Wu received the MSc degree from the School of Mathematical Sciences, University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2023. He is currently admitted to the UESTC and studies for the PhD degree under prof. Ting-Zhu Huang. His research interests include theories and applications of machine learning and deep learning in image processing.

Zi-Han Cao received the BS degree from the School of Information and Communication Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2023. He is currently working toward the MS degree under prof. Liang-Jian Deng in the School of Mathematics, University of Electronic Science and Technology of China. His research interests include computer vision, machine learning, and applications on low-level vision tasks including super-resolution, image fusion, and inverse problems.

Ting-Zhu Huang (Member, IEEE) received the BS, MS, and PhD degrees in computational mathematics from the Department of Mathematics, Xi'an Jiaotong University, Xi'an, China, in 1986, 1992, and 2001, respectively. He is currently a professor with the School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu, China. His research interests include scientific computation and applications, numerical algorithms for image processing, numerical linear algebra, preconditioning technologies, and matrix analysis with applications. Dr. Huang is an editor of the *Scientific World Journal*, *Advances in Numerical Analysis*, the *Journal of Applied Mathematics*, the *Journal of Pure and Applied Mathematics: Advances in Applied Mathematics*, and the *Journal of Electronic Science and Technology, China*.

Liang-Jian Deng (Senior Member, IEEE) received the BS and PhD degrees in applied mathematics from the School of Mathematical Sciences, University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2010 and 2016, respectively. He is currently a research fellow with the School of Mathematical Sciences, UESTC. From 2013 to 2014, he was a Joint-Training PhD student with the Case Western Reserve University, Cleveland, OH, USA. In 2017, he was a postdoc with Hong Kong Baptist University (HKBU). In addition, he also stayed with Isaac Newton Institute for Mathematical Sciences, Cambridge University and HKBU for short visits. His research interests include the use of partial differential equations (PDE), optimization modeling, and deep learning to address several tasks in image processing, and computer vision, e.g., resolution enhancement and restoration.

Jocelyn Chanussot (Fellow, IEEE) received the MSc degree in electrical engineering from the Grenoble Institute of Technology (Grenoble INP), Grenoble, France, in 1995, and the PhD degree from the Université de Savoie, Annecy, France, in 1998. Since 1999, he has been with Grenoble INP, where he is currently a professor of signal and image processing. His research interests include image analysis, hyperspectral remote sensing, data fusion, machine learning, and artificial intelligence. Dr. Chanussot was a member of the Institut Universitaire de France from 2012 to 2017. He was the vice-president of the *IEEE Geoscience and Remote Sensing Society (GRSS)*, in charge of meetings and symposia from 2017 to 2019. He was the general chair of the first IEEE GRSS Workshop on Hyperspectral Image and Signal Processing, Evolution in Remote sensing. He is an associate editor for the *IEEE Transactions on Geoscience and Remote Sensing*, *IEEE Transactions on Image Processing* and the *Proceedings of the IEEE*. He was the editor-in-chief of the *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* from 2011 to 2015. In 2014, he served as a guest editor for the *IEEE Signal Processing Magazine*. He has been a Highly Cited Researcher (Clarivate Analytics/Thomson Reuters), since 2018.

Gemine Vivone (Senior Member, IEEE) received the BSc and MSc degrees (summa cum laude), and the PhD degree in information engineering from the University of Salerno, Fisciano, Italy, in 2008, 2011, and 2014, respectively. He is a senior researcher with the National Research Council (Italy). His main research interests focus on image fusion, statistical signal processing, deep learning, and classification and tracking of remotely sensed images. Dr. Vivone is a co-chair of the IEEE GRSS Image Analysis and Data Fusion Technical Committee, a member of the IEEE Task Force on "Deep Vision in Space", and he was the Leader of the Image and Signal Processing Working Group of the IEEE Image Analysis and Data Fusion Technical Committee (2020-2021). Dr. Vivone is currently an area editor for Elsevier Information Fusion, and associate editor for *IEEE Transactions on Geoscience and Remote Sensing (TGRS)* and *IEEE Geoscience and Remote Sensing Letters (GRSL)*. Moreover, he is an Editorial Board Member for Nature Scientific Reports and MDPI Remote Sensing. Dr. Vivone received the IEEE GRSS Early Career Award in 2021, the Symposium Best Paper Award at IEEE International Geoscience and Remote Sensing Symposium (IGARSS), in 2015 and the Best Reviewer Award of the IEEE Transactions on Geoscience and Remote Sensing, in 2017. Moreover, he is listed in the World's Top 2% Scientists by Stanford University.