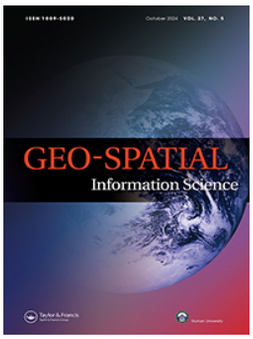

Taylor & Francis
Taylor & Francis Group

Geo-spatial Information Science

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tgsi20

A novel cross fusion model with fine-grained detail reconstruction for remote sensing image pan-sharpening

Chuang Liu, Zhiqi Zhang, Mi Wang, Shao Xiang & Guangqi Xie

To cite this article: Chuang Liu, Zhiqi Zhang, Mi Wang, Shao Xiang & Guangqi Xie (07 Nov 2024): A novel cross fusion model with fine-grained detail reconstruction for remote sensing image pan-sharpening, Geo-spatial Information Science, DOI: 10.1080/10095020.2024.2416899

To link to this article: https://doi.org/10.1080/10095020.2024.2416899

Published online: 07 Nov 2024.

GEO-SPATIAL INFORMATION SCIENCE
https://doi.org/10.1080/10095020.2024.2416899

A novel cross fusion model with fine-grained detail reconstruction for remote sensing image pan-sharpening

Chuang Liu [a], Zhiqi Zhang [a,b], Mi Wang [b], Shao Xiang [b] and Guangqi Xie [a,b]

[a]School of Computer Science, Hubei University of Technology, Wuhan, China; [b]State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China

ABSTRACT

Pan-sharpening aims to obtain high resolution multispectral (HRMS) images by integrating the information in the panchromatic and multispectral images. Existing pan-sharpening methods have demonstrated impressive sharpening performance. However, these methods inherently overlook the complementary characteristics and interaction between diverse source images, resulting in sharpened outcomes accompanied by distortion. To solve the above problems, we construct a novel cross fusion model with fine-grained detail reconstruction from the perspective of frequency-domain. The motivation of the model is twofold: (1) to reconstruct spatial detail representations from diverse source images, laying the foundation for the generation of fine details in the subsequent fused images; and (2) to enhance the interaction between diverse source features during the fusion process in order to attain high-fidelity fusion outcomes. Based on the theoretical model, we develop a frequency-spectral dual domain cross fusion network (CF2N) utilizing the deep learning technique. Consequently, the CF2N consist of two main stages, namely frequency-domain dominated detail reconstruction (FD2R) and frequency-spectral cross fusion (FSCF). Specifically, a more reasonable reconstruction of fine frequency details in HRMS can be achieved by performing adaptive weighted fusion of frequency details in the FD2R stage. Furthermore, the FSCF module, which seamlessly integrates frequency- and spectral-domain details in a highly interactive cross fusion manner. As a result, the CF2N possesses the capability to attain high frequency-spectral fidelity results with excellent interpretability. Extensive experiments show the superior performance of ours over state of the art, while maintaining high efficiency. All implementations of this work will be published at our website.

ARTICLE HISTORY
Received 19 June 2024
Accepted 10 October 2024

KEYWORDS
Pan-sharpening; remote sensing; deep learning; image fusion; model construction; detail reconstruction; cross fusion

1. Introduction

The rapid advancement of satellite and sensor technologies has led to the widespread utilization of remote sensing (RS) images. High-resolution multispectral (HRMS) images are a prerequisite for the successful execution of downstream vision tasks such as RS classification (Tariq et al. 2022; Xu et al. 2023; Zhang et al. 2023) and detection (Li et al. 2022; Wang et al. 2023; Zuo et al. 2024). Further, the accuracy of these tasks can provide strong technical support for disaster warning (Anno et al. 2023; Masoumi and Genderen 2023) and geological survey (Zhang et al. 2022b; Liu et al. 2023b; Cheng and Li 2023). However, the limitations of physical conditions make it impossible to acquire HRMS directly from a single satellite sensor. Consequently, only paired panchromatic (PAN) and multispectral (MS) images can be obtained from different sensors. In order to cater to the requirement of downstream applications, the process of PAN and MS image fusion, commonly referred to as pan-sharpening, has been initiated. The objective of pan-sharpening is to comprehensively exploit the information present in diverse source images, ultimately leading to HRMS images with fine-grained spatial texture and high spectral fidelity.

So far, pan-sharpening methods can be broadly categorized into traditional methods and deep learning (DL) methods. Most traditional methods employ the linear transformation to manually extract features from diverse source images. These methods are favored by some RS software, such as ENVI and PCI, owing to their high efficacy and independence from hardware devices. However, the features obtained from these methods exhibit limited representation capabilities and are incapable of adapting to intricate sharpening scenarios. Recently, with the increase in the means of data acquisition, data-driven DL methods have been extensively studied (Bouasria et al. 2022; Li et al. 2024). Various novel techniques have been introduced into the field of pan-sharpening, such as residual learning (Yang et al. 2017a), attention mechanism (Zhang et al. 2022a) and multi-branch structure (Jian et al. 2023). These techniques

CONTACT Zhiqi Zhang zzq540@hbut.edu.cn

significantly enhance the representation capability across diverse source data and enable DL-based methods to surpass traditional methods in terms of sharpening performance. According to the fusion pipelines, existing DL-based methods can be categorized into super-resolution (SR)-based image dimension concatenation methods (Cai and Huang 2021; Jin et al. 2022), dual branch-based feature dimension concatenation methods (Zhang et al. 2019; Zhang et al. 2022a) and detail injection-based methods (Deng et al. 2021; Lu et al. 2022) The pan-sharpening task is treated as a SR problem in SR-based methods, as shown in Figure 1(a). The SR-based fusion pipeline, akin to the common single-input SR architecture, takes the PAN and MS as the input to the network after channel-wise concatenation. Subsequently, the mapping of the input to HRMS is learned through the designed SR network. The dual branch-based feature dimension concatenation methods take into account different modal discrepancies between original image pairs, as illustrated in Figure 1(b). They employ two sub-networks to parallelly extract the shallow features of the PAN and MS, followed by feature dimension concatenation, and finally integration of the feature through a feature extraction sub-network. Some methods (Diao et al. 2022) employ the same sub-networks to extract shallow features from diverse source images, which makes them similar to the pipeline in Figure 1(a) with exception of the feature dimension concatenating. As depicted in Figure 1(c), some studies (Chen, Liu, and Fang 2024; Ke et al. 2023) have utilized the nonlinear fitting capability of deep neural networks (DNNs) to obtain a nonlinear injection gain. This can alleviate the distortion caused by the linear injection gain obtained from the traditional detail injection-based methods.

From the perspective of the fusion pipelines, existing DL methods can be roughly classified into three classes. In general, the first type of methods implicitly constructs relationships through DNNs without considering the unique characteristics of PAN and MS images. This renders DNNs incapable of effectively utilizing the characteristics of diverse source images, resulting in the loss of crucial information. The second type of methods extracts features from diverse source images separately. In most cases, the second class of methods is superior to the first in terms of fusion performance due to its consideration of the uniqueness of diverse source data. However, experiments have revealed that in the event of severe image degradation, the fusion outcomes exhibit apparent spatial and spectral distortions (Yang et al. 2017a; Zhang et al. 2022a). This is attributed to inadequate interaction between diverse source features during the fusion stage, resulting in the loss of significant complementary information. In addition, interpreting and explicating the behavior of the above two types of methodologies proves to be arduous, as they lack interpretability due to their predominant empirical design frameworks (Mai, Lam, and Lee 2022; Truong, Lam, and Lee 2024). The third category of methods has the potential to enhance the interpretability of DL models by combining the classical fusion pipelines with DL techniques. They take the differences between PAN and up-sampled MS (UPMS) images as the representation of the spatial information that subsequent DNNs need to learn. However, the

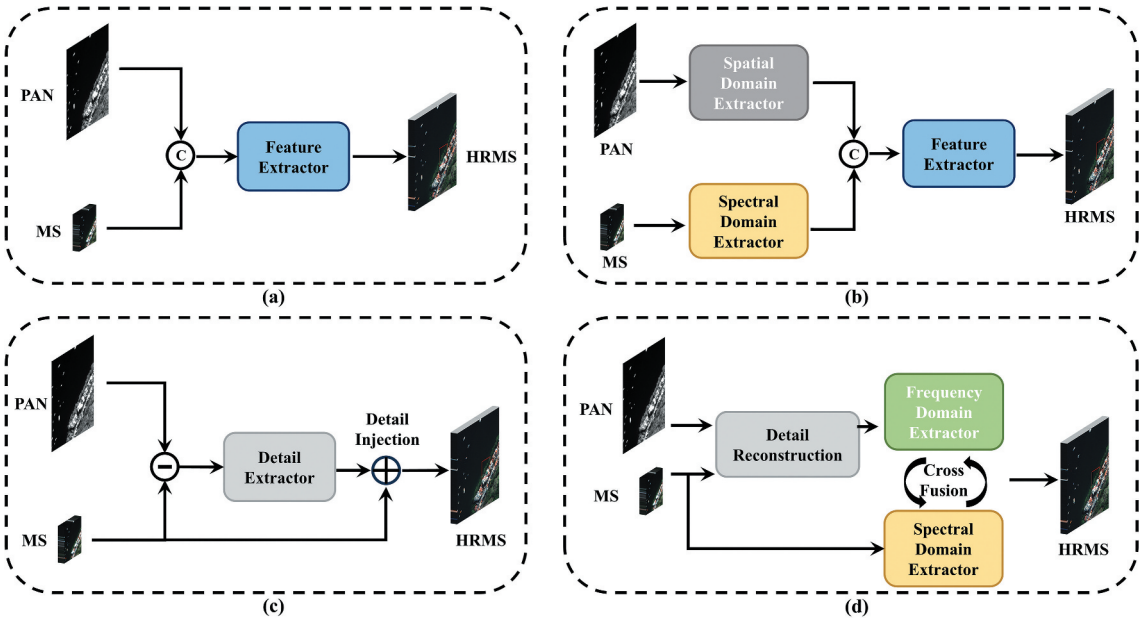

Figure 1. Fusion pipelines used by existing dl-based pan-sharpening methods. (a) SR-based image dimension concatenation methods. (b) Dual branch-based feature dimension concatenation methods. (c) Detail injection-based methods. (d) A novel cross fusion model with fine-grained detail reconstruction we proposed, which reconstructs the unique and complementary detail representation in diverse source images and integrates frequency-spectral details utilizing a cross fusion manner.

unique and complementary spatial information present in MS is overlooked in this representation of detail. It is well-known that PAN and MS images are imaged by different sensors, and that radiometric discrepancies exist between them. This means that the HRMS images contain spatial details that are absent in PAN but present in MS (Choi, Yu, and Kim 2011). In other words, it is possible to obtain fused images with finer textures by considering the spatial information present in both PAN and MS images. As a result, the limitations of the above three categories of methods are mainly attributed to two aspects: (1) insufficient exploration and integration of spatial information from diverse source images, resulting in a limited representation of the spatial information fed to the subsequent DNNs; and (2) during the fusion phase, there is inadequate interaction between the diverse domains, resulting in the apparent distortion of fusion outcomes appearing in one domain, spatial or spectral.

In order to eliminate the limitations, we construct a novel cross fusion model with fine-grained detail reconstruction, as shown in Figure 1(d). The model attempts to reconstruct the detail injection model from the perspective of the frequency domain in two aspects. This investigation explores the uniqueness and complementarity between the frequency information in diverse source images, taking into account the existence of details in the MS that are complementary to those in the PAN. This will serve as guidelines for the generation of frequency details in the fused image. On the other hand, realizing the importance of the interaction between the information in diverse domains, we introduce a spectral term based on the above. The spectral term aims to deeply cross-fuse the frequency details with corresponding spectral details, leading to a more reasonable spectral distribution in the fused image. Guided by the theoretical model, we develop a frequency-spectral dual domain cross fusion network utilizing DL techniques, commonly referred to as CF2N. The CF2N consists of two main stages: frequency-domain dominated detail reconstruction (FD2R) and frequency-spectral cross fusion (FSCF). In the FD2R stage, the discrete wavelet transform (DWT) is introduced to investigate the representation of details between characteristics from diverse source images. The DWT is capable of preserving the frequency details in the original image pairs due to its reversibility. In the meantime, we perform adaptive weighted fusion on the frequency details and the corresponding position information, which comprehensively considers the unique and complementary information in diverse source images, thereby facilitating the reconstruction of fine-grained frequency details. Moreover, the FSCF module (FSCFM) provides a concrete implementation of the spectral terms in the constructed model. It has been specifically engineered to integrate both frequency and spectral details in a highly interactive cross fusion manner, resulting in fused images with exceptional frequency-spectral fidelity. Our proposed method outperforms existing methods in terms of both performance and effectiveness, showing the advantages of the proposed methodology. Experiments in real-world scenarios further demonstrate that our method exhibits better generalization capability. The following contribution can be drawn.

(1) Considering the unique and complementary characteristics and the interaction between diverse source images, a novel cross fusion model with fine-grained detail reconstruction is constructed, which aims to preserve the frequency-spectral details in diverse source images. Guided by this theoretical model, we propose an efficient network known as CF2N, comprising of two main stages, namely FD2R and FSCF. The CF2N is capable of achieving high frequency-spectral fidelity outcomes with excellent interpretability.
(2) The FD2R investigates a comprehensive representation of spatial information from a frequency-domain perspective, which is capable of adaptively weighting frequency information from various directions in diverse source images to obtain fine-grained details. The reconstructed details can provide the foundation for the generation of fine details in the subsequent fused images.
(3) Realizing the significance of the interaction between information across diverse domains, we have constructed the FSCFM, which provides a concrete implementation of the spectral terms in the constructed model. The FSCFM is capable of seamlessly integrating frequency and spectral details in a highly interactive manner, thereby achieving fusion outcomes with high frequency-spectral fidelity.
(4) Extensive experiments and comparative analyses on five datasets demonstrate that the proposed method is capable of achieving high frequency-spectral fidelity sharpening outcomes in homogeneous fusion tasks, such as MS image pan-sharpening and hyperspectral (HS) image pan-sharpening, while maintaining a high level of efficiency. The robust generalization capability and scalability of the proposed methodology are further confirmed by extended experiments conducted on heterogeneous fusion task, i.e. synthetic aperture radar (SAR)-optical fusion.

2. Related work

2.1. Traditional pan-sharpening methods

Several studies have classified traditional pan-sharpening methods into three categories, namely component substitution (CS), multi-resolution analysis (MRA) and variational optimization (VO) (Vivone et al. 2021; Misra et al. 2023a). The spatial components of the UPMS are substituted by PAN in CS-based methods. Due to the complete retention of information within the PAN, they exhibit high spatial fidelity. However, these methods are susceptible to severe spectral distortion due to rigid transformation operations. This category includes methods such as the adaptive Gram Schmidt (GSA) (Aiazzi, Baronti, and Selva 2007), the partial replacement adaptive CS (Choi, Yu, and Kim 2011) and BDSDPC (Vivone 2019). Different from CS, the MRA-based methods focus on improving spectral fidelity. They use MRA tools to extract spatial information from PAN without disrupting the spectral distribution of the original MS. The additive wavelet luminance proportional (AWLP) (Otazu et al. 2005) and morphological filters (MF) (Restaino et al. 2016) are the predominant methods based on MRA. However, perfection remains elusive. MRA-based methods are difficult to extract sufficient spatial details under the premise of maintaining high spectral fidelity, which makes them vulnerable to spatial blurring. A method called BFLP was proposed to enhance the spatial fidelity, but its effectiveness depends on the precision of the parameter settings (Kaplan and Erer 2014). The VO-based methods comprise two fundamental stages, i.e. constructing the energy function and computing the optimal solution. Early VO-based methods mainly relied on constructed models through compressive sensing and dictionary learning. In recent times, numerous improved methods have been proposed, such as ADMM (Kaplan et al. 2019), MMMT (Yang et al. 2017b), and FSRIC (Ghahremani et al. 2019). These methods focus on enhancing the spatial details, but neglect spectral constraints, which can result in spectral distortion.

2.2. DI-based pan-sharpening methods

Traditional methods rely on linear transformations, which have limited fitting capability. In recent years, DNNs have garnered significant attention owing to their potent nonlinear fitting capabilities. Inspired by the first DL-based super-resolution network SRCNN, Masi et al. (2016) proposed a DNN-based pan-sharpening network. Despite the fact that PNN contains only three convolutional layers, it demonstrated a notable sharpening effect, laying the foundation for subsequent DL-based pan-sharpening methods. Yang et al. (2017a) utilized high-pass filtering to extract high-frequency (HF) feature from PAN, and utilized the residual structure to retain the extracted HF details. A gradient projection pan-sharpening network, influenced by VO theory and DNNs, was proposed (Xu et al. 2021). Cai and Huang (2021) regarded the pan-sharpening task as a SR task, and proposed a SR-based progressive pan-sharpening network. In order to establish the global spatial-spectral dependence, several methods employed improved attention modules to construct the network, leading to satisfactory sharpened results, such as TRRNet (Zhang et al. 2022a), LAGConv (Jin et al. 2022) and RSANet (Liu et al. 2023a). Although these methods can achieve satisfactory results, they lack interpretation due to the block-box nature of DNNs. To address this issue, researchers have improved the interpretability of DL-based methods by utilizing the combination of traditional fusion pipelines and DL techniques. For example, Deng et al. (2021) amalgamated the concepts of CS and MRA to construct an interpretable DNN called FusionNet. Motivated by FusionNet, Chen, Liu, and Fang (2024) employed guided filtering to enhance original image pairs, resulting in more granular detail representations. However, these methods only perform well on single datasets. It is imperative to enhance the generalization capability, as the missing precision may result in inadequate sharpened effects in some scenes. Truong, Lam, and Lee (2024) first attempted to treat the pan-sharpening task as a tensor rank minimization problem, and employed the concept of detail injection to extract spatial details in PAN. This method is flexible and can be applied to images of different quality, improving the interpretability of DL-based methods. Misra et al. (2023b) proposed SPRINT in order to maintain the spectral characteristics of MS in the process of fusion. As the first method to utilize both Digital Elevation Model and solar angles simultaneously to determine the weight, SPRINT has the potential to not only enhance the spatial fidelity of the sharpened image, but also mitigate the impact of radiation distortion to the greatest extent. In addition, there are several studies dedicated to achieving superior sharpening performance in the frequency domain. Xing et al. (2024) combined frequency transform with deformable self-attention to integrate local and non-local features in diverse source images. Diao et al. (2022) employed Gaussian filtering to extract HF and low-frequency (LF) information from diverse source images, then employed two different branches to learn HF and LF information, respectively, and finally integrated the information of the two different frequencies. In order to mine more potential spatial details, Zhuo et al. (2022) used five different high-pass filters to extract the HF information in PAN image. Zhao et al. (2022) employed high-pass operators to fully explore the low-, medium-, and high-frequency information in the original image pairs. Unlike these

methods, our methodology is devoted to reconstructing the detail injection model from a frequency domain perspective. Subsequently, we implement the constructed fusion model utilizing DL techniques in order to achieve high frequency-spectral fidelity results with good interpretability and generalization capability.

2.3. Detail injection models

Both CS and MRA, in fact, can be summarized as detail injection-based methods. Detail injection models aim to extract details from PAN in a specific manner, and then incorporate the extracted details into the UPMS. The process could be summarized as follows.

$$\tilde{\mathcal{M}} = \bar{\mathcal{M}} + g \cdot \mathcal{D}_P, \tag{1}$$

where $\tilde{\mathcal{M}} \in R^{H\times W\times B}$ and $\bar{\mathcal{M}} \in R^{H\times W\times B}$ represent the HRMS and UPMS with height, width, and number of bands H, W, and B, respectively; g and $\mathcal{D}_\mathcal{P}$ represent the ratio of injection gain and the spatial details extracted from the PAN. The primary discrepancy between the CS and MRA pertains to manner in which the details are extracted.

For the CS-based method (Aiazzi, Baronti, and Selva 2007; Chavez and Kwarteng 1989; Choi, Yu, and Kim 2011; Mallat 1989; Tu et al. 2004; Vivone 2019), $\mathcal{D}_\mathcal{P}$ can be quantified as the discrepancy between $\mathcal{P}$ and $\mathcal{I}_{\bar{\mathcal{M}}}$.

$$\mathcal{D}_\mathcal{P} = P - \mathcal{I}_{\bar{\mathcal{M}}}, \tag{2}$$

where $\mathcal{P} \in R^{H\times W\times 1}$ represents the PAN image, and $\mathcal{I}_{\bar{\mathcal{M}}}$ represents the intensity component in $\bar{\mathcal{M}}$. Whereas MRA-based methods (Kaplan and Erer 2014; Kaplan et al. 2019; Otazu et al. 2005; Restaino et al. 2016) denote $\mathcal{D}_\mathcal{P}$ by the discrepancy between $\mathcal{P}$ and $\mathcal{P}_L$, which can be quantified as:

$$\mathcal{D}_P = P - P_L, \tag{3}$$

where $\mathcal{P}_L$ represents the low-pass version of $\mathcal{P}$.

The above two methods employ a linear injection model, which is typically inapplicable to the relative spectral response of sensors because the overlap between spectral responses of $\bar{\mathcal{M}}$. In recent years, several researchers have taken advantage of the nonlinear mapping capability of DNNs to develop nonlinear injection models to replace the linear detail injection in traditional methods, such as DiCNN (Vivone et al. 2015), FusionNet (Deng et al. 2021), and GF-CSTNet (Chen, Liu, and Fang 2024). These methods obtain injected details by tuning the parameters in the training phase, thus obtaining the fine details in fused images. The process could be summarized as follows.

$$\tilde{\mathcal{M}} = \bar{\mathcal{M}} + DNN(\mathcal{D}_P, \theta), \tag{4}$$

where θ represents the parameters that can be learned in *DNN*. The development of DL-based models that incorporate the concept of detail injection is a promising method, and the key is to effectively extract the precise details.

3. Methodology

In this section, we introduce a novel cross fusion model and optimize it by analyzing the unique and complementary characteristics in diverse source images. Furthermore, we develop the frequency-spectral dual domain cross fusion network, CF2N, which implements the constructed model by utilizing DL techniques. The CF2N achieves high frequency-spectral fidelity sharpened results by sequentially executing frequency-domain dominated detail reconstruction and frequency-spectral cross fusion of diverse source images.

3.1. Construction of the fusion model

The CS- and MRA-based methods implement the detail injection model by using traditional analysis tools, which have good interpretability. Preliminary assumptions about the spectral model that determines the projection of the MS image into the PAN domain (for CS-based methods) or the spherical of the spatial filter (for MRA-based methods) are crucial for traditional methods. Errors at this stage may have a significant impact on the results, which could negatively impact the performance of pan-sharpening. Deng et al. (2021) combined the advantages of CS and MRA, which represent the spatial details that the DNNs needs to learn by subtracting UPMS from the duplicated version of PAN. The energy function for this extraction can be quantified as:

$$\mathcal{E} = \| \mathcal{C}\tilde{\mathcal{M}}_b - \kappa_b \mathcal{C}\left(\bar{P}_b - \bar{\mathcal{M}}_b\right) \|, \tag{5}$$

where $\mathcal{E}$ represents the energy function associated with details; $\mathcal{C}$ represents a series of convolution operation; $\bar{\mathcal{P}} \in R^{H\times W\times B}$ represents the PAN duplication in the channel dimension such that it coincides with the number of bands in $\bar{\mathcal{M}}$; κ_b is the injection gain factor of the b-th band, which is learned by DNNs. Such a detail extraction takes the PAN-specific spatial information as the learning target of the DNNs. The detail representation of FusionNet overlooks the unique and complementary information in $\bar{\mathcal{M}}$, which results in distortions in the fused images. Inspired by FusionNet, Chen, Liu, and Fang (2024) designed GF-CSTNet, which utilizes guided filtering to extract the spatial details in the original pairs before subtracting them from diverse source images. The energy function for the extraction of GF-CSTNet can be quantified as:

$$\bar{\mathcal{E}} = \| \mathcal{C}\tilde{\mathcal{M}}_b - \kappa_b \mathcal{C}\left(GF(\bar{P}_b) - GF\left(\bar{\mathcal{M}}_b\right)\right) \|, \quad (6)$$

where GF represents the guided filtering. The GF-CSTNet exhibits superior sharpening performance in comparison to FusionNet, but still overlooks the unique and complementary information in $\bar{\mathcal{M}}$. As a result, these representations render the fusion performance highly dependent on the extraction and generalization capabilities of the subsequent network. When the sharpening scene is intricate, it can be challenging for the model to attain satisfactory outcomes.

It is widely acknowledged that paired $\mathcal{P}$ and $\mathcal{M}$ are captured by different sensors within a single scene, and there are radiometric disparities between them. As a result, $\mathcal{M}$ possess spatial details that complement the $\mathcal{P}$, which are omitted by Equation (5) and Equation (6). Furthermore, objects in RS images often have great discrepancies, which means that targeted gain coefficients are applied to different local areas to represent local information more accurately. In other words, the acquisition of spatial details end-to-end in terms of bands does not extract specific local details. In light of the above issues, we propose three improvements to Equation (5) and Equation (6). First, we comprehensively explore the unique and complementary detail representations in diverse source images from a frequency domain perspective. These reconstructed details are capable of providing rich detail support for the fused image. Second, we substitute the band-by-band scale factor with a local scale factor for each pixel, taking into account the disparities in diverse regions. Third, we construct the relationship from reconstructed details to HRMS pixel by pixel. As a result, a new energy function is quantified as follows.

$$\tilde{\mathcal{E}} = \| \mathcal{C}\tilde{\mathcal{M}}^{\varsigma}_{(i,j)} - \kappa_{(i,j)} \mathcal{C}\mathcal{D}^{\varsigma}_{(i,j)} \|, \quad (7)$$

where ς represents the frequency information in a certain direction and $\kappa_{(i,j)}$ denotes the weight coefficient located at (i,j); Besides, $\mathcal{D}^{\varsigma} \in R^{H\times W\times B}$ represents reconstructed details in the frequency domain, which lays the foundation for the construction of reasonable detail in the fused image; $\kappa_{(i,j)}$ can be optimized by bringing $\mathcal{D}^{\varsigma}_{(i,j)}$ close to the details in $\tilde{\mathcal{M}}$. The preciseness of details can lead to the formation of reasonable spectral distribution within the sharpened images, which in turn can lead to the formation of finer textures. Thus, precise spatial details and spectral information can be mutually supportive. Considering the aforementioned factors, the details required in HRMS can be divided into frequency-domain details and spectral-domain details. The frequency-domain details refer to the spatial detail information reconstructed from various frequency directions, whereas the spectral-domain details denote the spectral information corresponding to the frequency details. Consequentially, we determine the value of κ by both frequency-domain and spectral-domain details. The optimized solution can be quantified as:

$$\kappa_{(i,j)} = \underset{\kappa_{(i,j)}}{argmin} \| \tilde{\mathcal{D}}_{(i,j)} - \kappa_{(i,j)} \mathcal{C}\left(\mathcal{D}^{\varsigma}_{(i,j)}, \mathcal{S}_{(i,j)}\right) \|, \quad (8)$$

$$\kappa_{(i,j)} = f\left(\varsigma_{(i,j)}, s_{(i,j)}\right), \quad (9)$$

where $\tilde{\mathcal{D}}_{(i,j)} = \tilde{\mathcal{M}}_{(i,j)}$ represents the details (including frequency details and spectral details) in $\tilde{\mathcal{M}}$, and $(\varsigma_{(i,j)}, s_{(i,j)})$ represent the frequency-spectral details located at (i,j). $f(\cdot)$ represents the function that combines the details in frequency- and spectral-domain. On this basis, we can get the frequency details and corresponding spectral details located at (i,j), which can be described as follows.

$$\hat{\mathcal{D}}_{(i,j)} = \kappa_{(i,j)} \mathcal{C}\left(\mathcal{D}^{\varsigma}_{(i,j)}, \mathcal{S}_{(i,j)}\right), \quad (10)$$

$$\tilde{\mathcal{M}} = \bar{\mathcal{M}} + \hat{\mathcal{D}}, \quad (11)$$

where $\hat{\mathcal{D}}$ represents the estimated frequency-spectral details.

The objective of Equation (8) is to investigate the correlation between diverse source images and HRMS in the frequency and spectral domains. The process involves two pivotal steps. First, taking into consideration the uniqueness and complementarity of spatial details in diverse source images, the representation of details in diverse source images is comprehensively explored in the frequency domain. This attempt aims to reconstruct the spatial details required for the fused image. Second, since information in the frequency domain and information in the spectral domain are mutually supportive, a spectral term is introduced into the proposed energy function. This attempt aims to obtain the spectral details corresponding to the reconstructed frequency details. Guided by the theoretical model, we develop CF2N to solve the required κ and $\mathcal{D}^{\varsigma}$ in Equation (10). Consequently, the proposed CF2N comprises two primary modules, namely FD2R and FSCFM. In FD2R, the frequency information of various directions in diverse source images is meticulously considered in the frequency domain, resulting in the estimated frequency details $\mathcal{D}^{\varsigma}$. Based on this, the FSCFM provides a concrete implementation of the spectral terms in the constructed model, which combines $\mathcal{D}^{\varsigma}$ and corresponding spectral details to obtain the determined κ.

3.2. Frequency-domain dominated detail reconstruction

Despite the existence of several pan-sharpening methods based on the frequency domain, an efficient

technique for reconstructing unique and complementary detail representations in diverse source images is still lacking. Given the existence of complementary details in diverse source images, we endeavor to investigate the unique and complementary detail representations from a frequency domain perspective. In particular, we focus on the interaction between the frequency components in various directions, instead of directly subtracting or splicing the diverse source images or features. The frequency information of various directions in diverse source images is fused adaptively in order to reconstruct the unique and complementary detail representations. As shown in Figure 2, the frequency information from various directions in $\mathcal{P}$ and $\bar{\mathcal{M}}$ is hierarchically integrated to obtain the fine-grained details $\mathcal{D}^{\varsigma}$. It is apparent that $\mathcal{D}^{\varsigma}$ contains a wealth of spatial detail information derived from the original image pairs, laying the groundwork for the generation of spatial details in the fused image.

Initially, $\mathcal{P}$ and $\bar{\mathcal{M}}$ are fed into the Head Residual Blocks (HRB), which can be quantified as:

$$\mathcal{F}^{p}, \mathcal{F}^{m} = HRB(P), HRB\left(\bar{\mathcal{M}}\right), \tag{12}$$

where $\mathcal{F}^{p} \in R^{H\times W\times C}$ and $\mathcal{F}^{m} \in R^{H\times W\times C}$ denote the output features after encoding by *HRB*. The structure of *HRB* is shown in the upper left corner of Figure 2. It adopts a simplistic residual structure, wherein the initial convolution block is used to increase the dimensionality of the input image, whereas the subsequent residual blocks are used to extract the shallow coded features. Subsequently, the coded features are fed into the FD2R, which comprehensively considers the frequency components of various directions in diverse source images. These frequency components are integrated hierarchically to obtain fine-grained frequency details. Specifically, we employ 2D Haar wavelet transform to get frequency information in various directions. The filters for various directions are quantified as follows.

$$f_{LL} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, f_{LH} = \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix},$$
$$f_{HL} = \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}, f_{LL} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}. \tag{13}$$

After performing the Haar transform on the above encoded features, the frequency information in various directions can be quantified as:

$$\mathcal{F}^{x}_{LL}, \mathcal{F}^{x}_{LH}, \mathcal{F}^{x}_{HL}, \mathcal{F}^{x}_{HH} = DWT(\mathcal{F}^{x}), s.t.x \in \{p.m\}, \tag{14}$$

$$\begin{cases} \mathcal{F}^{x}_{LL}(i,j) = \mathcal{F}^{x}(2i-1,2j-1) + \mathcal{F}^{x}(2i-1,2j) + \mathcal{F}^{x}(2i,2j-1) + \mathcal{F}^{x}(2i,2j) \\ \mathcal{F}^{x}_{LH}(i,j) = -\mathcal{F}^{x}(2i-1,2j-1) - \mathcal{F}^{x}(2i-1,2j) + \mathcal{F}^{x}(2i,2j-1) + \mathcal{F}^{x}(2i,2j) \\ \mathcal{F}^{x}_{HL}(i,j) = -\mathcal{F}^{x}(2i-1,2j-1) + \mathcal{F}^{x}(2i-1,2j) - \mathcal{F}^{x}(2i,2j-1) + \mathcal{F}^{x}(2i,2j) \\ \mathcal{F}^{x}_{HH}(i,j) = \mathcal{F}^{x}(2i-1,2j-1) - \mathcal{F}^{x}(2i-1,2j) - \mathcal{F}^{x}(2i,2j-1) + \mathcal{F}^{x}(2i,2j) \end{cases} \tag{15}$$

As mentioned above, both SR-based methods and dual-branch-based methods use concatenation operations to combine images or features from diverse sources. It is tough to disseminate complementary information as the concatenation process is incapable of fully exploring the interactions and correlations between diverse source data. Furthermore, the

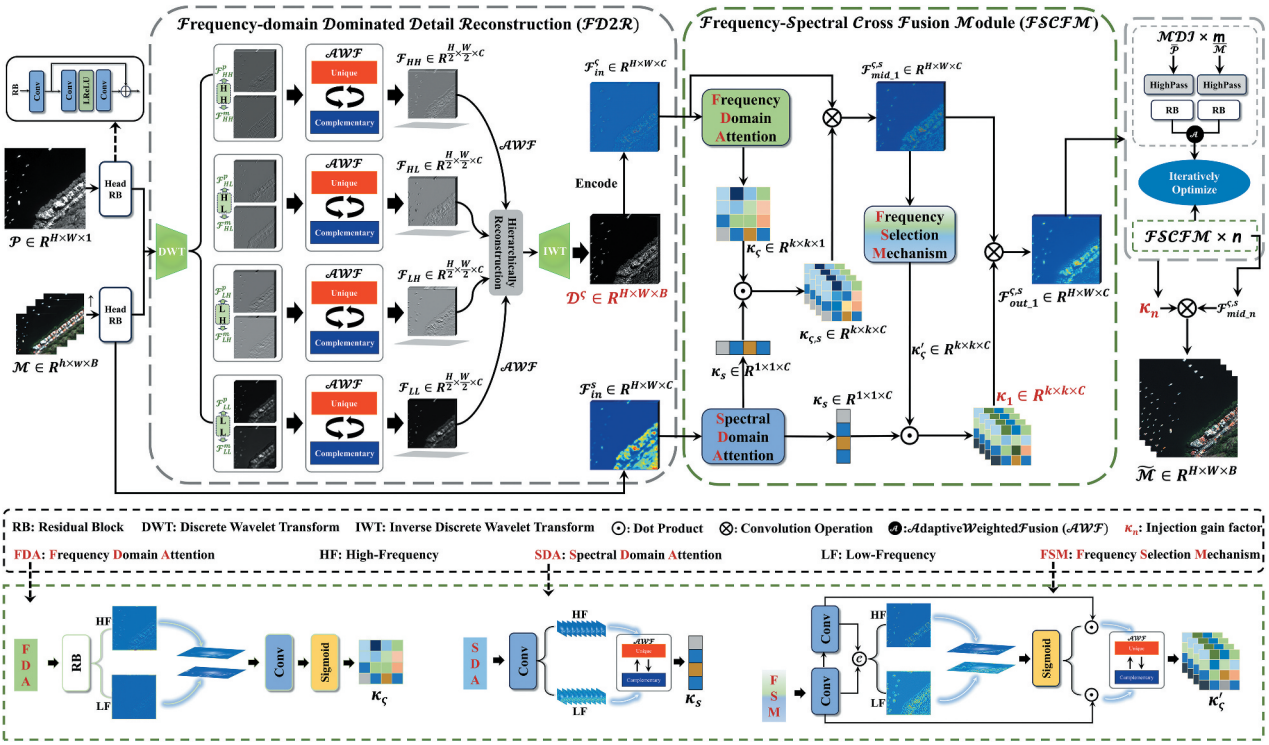

Figure 2. The flowchart depicting the proposed CF2N, which is guided by the constructed cross fusion model. The CF2N consists of two main stages: FD2R and FSCF.

concatenation operation can result in a significant increase in the channel dimension, thereby elevating both the space complexity and computational complexity of the model. The pixel- wise addition can deal with this issue. Despite the fact that this rigid operation preserves the uniqueness of diverse source data, it may result in the accumulation of redundant information, rendering the model unable to explicitly acquire the crucial complementary information. In order to reconstruct the unique and complementary information between diverse source images while avoiding a large amount of redundant information, a learnable parameter is introduced to adaptively fuse the frequency information in diverse source images. Specifically, the unique and complementary frequency characteristics of diverse sources in various directions are obtained through adaptive weighted fusion (AWF) of the frequency components in the corresponding directions of diverse source features. This process can be quantified as:

$$\mathcal{F}_{\varsigma} = \mathcal{A}\left(\mathcal{F}^{p}_{\varsigma}, \mathcal{F}^{m}_{\varsigma}\right), s.t. \varsigma \in \{LL, LH, HL, HH\}, \quad (16)$$

where $\mathcal{F}_{\varsigma} \in R^{\frac{H}{2}\times\frac{W}{2}\times C}$ represents the outcome of adaptive weighted fusion of frequency components from diverse sources in various directions. Besides, $\mathcal{A}$ represents adaptive weighted fusion. The AWF adjusts the magnitude of the weights in accordance with the specific details contributed in the frequency features of $\mathcal{P}$ and $\bar{\mathcal{M}}$, which can be quantified as:

$$\mathcal{F}_{\varsigma} = \beta \cdot \mathcal{F}^{p}_{\varsigma} + (1-\beta) \cdot \mathcal{F}^{m}_{\varsigma}. \quad (17)$$

The learnable parameter β in the AWF process can be manipulated to adjust the contribution degree of diverse source frequency components, thereby enabling the acquisition of fine-grained details with unique and complementary representations. As an illustration, it can be observed from Figure 2 that $\mathcal{F}_{HL}$ contains unique and complementary details in $\mathcal{F}^{p}_{HL}$ and $\mathcal{F}^{m}_{HL}$, which facilitate the formation of finer frequency details. Similarly, we integrate unique and complementary feature representations of the various directions hierarchically. It is worth noting that the LF component of the MS is omitted during the AWF process, i.e. $\mathcal{F}^{m}_{LL}$. This is due to the fact that we primarily reconstruct the HF information, which are incorporated into the subsequent FSCF stage to preserve more details. Finally, we obtain the reconstructed frequency details via Inverse DWT (IWT). This can be quantified as:

$$\mathcal{D}^{\varsigma} = IWT\left(HRB\left(\mathcal{A}\left(\sum_{\varsigma}\mathcal{F}_{\varsigma}\right)\right)\right), \quad (18)$$

where $\mathcal{D}^{\varsigma} \in R^{H\times W\times B}$ represents the reconstructed frequency details. By utilizing this methodology, all HF components are taken into account, thereby facilitating the reconstruction of fine-grained frequency details. In addition, the distinction between the AWF, channel-wise concatenation operation, and pixel-wise addition operation will be further explored in the first part of Section 4.5.

3.3. Frequency-spectral cross fusion

Most of the methods ignore the importance of the interaction between the information in diverse domains. They solely combine PAN and MS in the image or feature dimensions, resulting in the apparent distortion of fusion outcomes appearing in one domain, spatial or spectral. The Equation (8) and Equation (9) suggest that the estimated details ought to preserve the frequency-spectral fidelity of HRMS, i.e. the fine-grained frequency details and spectral details. Therefore, we construct a network representing $\kappa_{(i,j)}\mathcal{C}\left(\mathcal{D}^{\varsigma}_{(i,j)}, \mathcal{S}_{(i,j)}\right)$ and train it to approximate the details in $\mathcal{M}$. In order to attain this objective, we have developed the FSCFM to comprehensively amalgamate the reconstructed frequency details with the corresponding spectral details. As depicted in Figure 3, the goal of FSCFM is to calculate the weight coefficient of each pixel $\kappa_{(i,j)}$ integrating $\mathcal{D}^{\varsigma}$ obtained in FD2R stage and the encoded spectral information. The proposed FSCFM establishes the representation of frequency details and the representation of spectral details. Details in diverse domains are, in the meantime, deeply cross-fused.

As shown in the right of Figure 2, the FSCFM comprises three elaborate components, namely frequency-domain attention (FDA), spectral-domain attention (SDA), and frequency selection mechanism (FSM). The FDA primarily concentrates on the frequency domain information in the same spectral band, and generates adaptive weights κ_{ς} for each pixel in frequency domain. The SDA primarily concentrates on the spectral information to get the spectral adaptive weights κ_{s} in different spectral bands. The above process can be quantified as:

$$\kappa_{\varsigma}(i,j) = \sigma\left(Conv\left(HF\left(HRB\left(\varsigma_{(i,j)}\right)\right) \circ LF\left(HRB\left(\varsigma_{(i,j)}\right)\right)\right)\right), \quad (19)$$

$$\kappa_{s}(i,j) = A\left(HF\left(Conv\left(s_{(i,j)}\right)\right), LF\left(Conv\left(s_{(i,j)}\right)\right)\right), \quad (20)$$

where $\circ$ represents the concatenation operation. The input features of FDA ($\varsigma_{(i,j)}$) and SDA ($s_{(i,j)}$) are obtained by shallow encoding on $\mathcal{D}^{\varsigma}$ and $\bar{\mathcal{M}}$, as illustrated in Figure 3. Given that each spectral band possesses unique frequency-spectral characteristics, we execute the dot product operation on the adaptive weights across diverse domains. This operation can enhance the correlation of information in frequency domain and spectral domain,

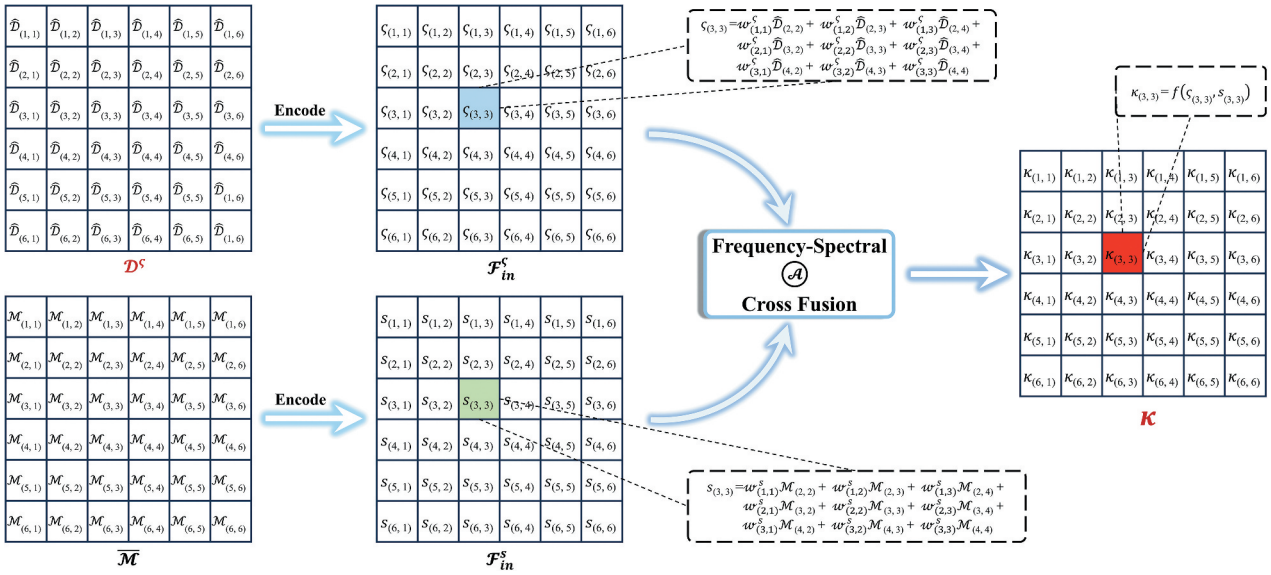

Figure 3. Illustration of the proposed FSCFM. We give two example matrices $\mathcal{D}^{\varsigma} \in R^{6\times6}$ and $\bar{\mathcal{M}} \in R^{6\times6}$, and the goal of FSCFM is to calculate the weight coefficient of each pixel $\kappa_{(i,j)}$ integrating frequency details obtained in FD2R and the corresponding spectral details. The encoder employs only one convolutional layer, and the convolution kernel size, step size and padding of this convolutional layer are 3, 1, 1 respectively.

which can elucidate the frequency-spectral importance of distinct regions. The above process can be quantified as:

$$\kappa_{\varsigma,s}(i,j) = \kappa_{\varsigma}(i,j) \odot \kappa_{s}(i,j). \tag{21}$$

$\kappa_{\varsigma,s}$ stands for the frequency-spectral adaptive weights. Subsequently, we execute a convolution operation on the frequency-spectral adaptive weights with the input features of the FDA. This operation aims to perform frequency-spectral cross fusion in the frequency domain, thereby enabling the adaptive addition of spectral detail information for each detail in the frequency domain.

$$\mathcal{F}^{\varsigma,s}_{mid} = \kappa_{\varsigma,s} \otimes \mathcal{F}^{\varsigma}_{in}. \tag{22}$$

The representation of $\mathcal{F}^{\varsigma,s}_{mid}$ is visualized in Figure 2. It is evident that, following the execution of the frequency-spectral cross-fusion, the frequency details in $\mathcal{D}^{\varsigma}$ are accompanied by the corresponding spectral information.

Furthermore, we consider that the scales of feature in RS images can be significantly different, such as cars and flower beds. The selection mechanism is introduced to extract the precise frequency-spectral details with varying scales. The receptive field of the second convolution is expanded by combining two convolutional kernels in series, and these two convolutions in parallel to extract features with varying scales. The adaptive weights $\kappa^{'}_{\varsigma}$ for details with varying scales are obtained, which enable the generation of adaptive scaled receptive fields for details with varying scales. In this way, more precise frequency-spectral details can be obtained. The above process can be quantified as follows.

$$\mathcal{F}_{sma} = Conv\left(\mathcal{F}^{\varsigma,s}_{mid}\right), \tag{23}$$

$$\mathcal{F}_{lar} = Conv(\mathcal{F}_{sma}), \tag{24}$$

$$\mathcal{F}_{var} = HF(\mathcal{F}_{sma} \circ \mathcal{F}_{lar}) \circ LF(\mathcal{F}_{sma} \circ \mathcal{F}_{lar}), \tag{25}$$

$$\kappa^{'}_{\varsigma} = A(\sigma(\mathcal{F}_{var}) \odot \mathcal{F}_{sma}, \sigma(\mathcal{F}_{var}) \odot \mathcal{F}_{lar}), \tag{26}$$

where $\mathcal{F}_{sma}$ and $\mathcal{F}_{lar}$ represent the features of varying scales derived from the small receptive field and the extended receptive field, respectively. $\mathcal{F}_{var}$ contains feature representations for various scales and different frequency directions. By establishing the correlation between $\mathcal{F}_{var}$ and features of varying scales ($\mathcal{F}_{sma}$ and $\mathcal{F}_{lar}$), and subsequently performing AWF, $\kappa^{'}_{\varsigma}$ is capable of adapting to frequency features of various scales.

On this foundation, we further enhance the correlation between frequency-domain features and spectral-domain features to obtain κ. Furthermore, the precise frequency-domain information obtained is fully cross-fused with the spectral-domain information once more. The above process can be quantified as follows.

$$\kappa = \kappa^{'}_{\varsigma} \odot \kappa_{s}, \tag{27}$$

$$\mathcal{F}^{\varsigma,s}_{out} = \kappa \otimes \mathcal{F}^{\varsigma,s}_{mid}. \tag{28}$$

Identically, the representation of $\mathcal{F}^{\varsigma,s}_{out}$ is visualized in Figure 2. It is apparent that subsequent to the selective cross-fusion of frequency information at various

scales, targets at various scales are accompanied by specific spectral information, thereby promoting the generation of finer frequency-spectral details.

The first half of the FSCF stage employs the $\mathcal{F}_{in}^{c}$ and $\mathcal{F}_{in}^{s}$ as inputs to obtain fine-grained frequency-spectral details. Meanwhile, we introduce MDI (see Section 3.4) in order to prevent the loss of crucial details. The subsequent FSCFMs take the output features of the upper-level FSCFM and features from MDI as inputs in order to iteratively refine the extracted frequency-spectral details. The specific structure of MDI will be described in the following. In addition, the number of FSCFMs and MDIs will be discussed in Section 4.5.

3.4. Multi-level detail injection

The MDI is introduced in the FSCF stage to avoid the loss of crucial frequency-domain and spectral-domain details. Specifically, we extract the HF information from the $\bar{\mathcal{P}}$ and $\bar{\mathcal{M}}$ in parallel by employing the high-pass filtering operator. Subsequently, the HF features are extracted by utilizing the simplistic residual block as HRB. Finally, an AWF operation is performed on both. The obtained detailed features are fed into the FSCFM, which is highly cross-fused with the output of the previous level. In this way, it is feasible to replenish detailed information lost during the FSCF stage, facilitate the fusion of frequency-spectral details, and achieve high-fidelity results.

Ultimately, by utilizing the constructed FSCFM and MDI, it is possible to iteratively optimize κ and $\mathcal{D}^{c}$ during the training process. Furthermore, the Mean Absolute Error (MAE) is utilized to instruct CF2N to generate sharpened images that exhibits fine-grained frequency-spectral details in accordance with the ground truth (GT).

4. Experiments

4.1. Datasets and settings

We conduct extensive experiments on three publicly available datasets, which are available at https://github.com/liangjiandeng/PanCollection (Deng et al. 2022). These three datasets come from three satellites, namely, GaoFen-2 (GF2), QuickBird (QB), and WorldView-3 (WV3). Table 1 shows the basic information of each dataset in detail. As there are no authentic HRMS images to serve as GT for model training, low-resolution versions of the original image pairs are obtained by following Wald's protocol in these publicly available datasets. Therefore, the original MS can be regarded as GT. It is worth noting that the experiment mainly consists of reduced-scale test (R-Test) and full-scale test (F-Test), wherein the R-Test mainly evaluates the fitting capability of the model, while the F-Test mainly evaluates the generalization capability in real-world scenarios.

In the R-Test, we employ six well-known metrics to evaluate the fusion performance at reduced-scale, namely the Structure Similarity Index Measure (SSIM) (Wang et al. 2004), the Erreur Relative Globale Adimensionnelle de Synthèse (ERGAS) (Vivone et al. 2015), the Spectral Angle Mapper (SAM) (Yuhas, Goetz, and Boardman 1992), the spatial Correlation Coefficients (CC), the Relative Average Spectral Error (RASE) (Zhou, Civco, and Silander 1998), and the Q2n index (Garzelli and Nencini 2009). Additionally, we employ five non-reference metrics to evaluate the fusion performance at full-scale in the F-Test, namely the spectral distortion index (D_λ), the spectral distortion index from Khan's protocol (D_λ^F), the spatial distortion index (D_s), the quality with no reference (QNR) (Alparone et al. 2008), and the hybrid quality with no reference (HQNR) (Aiazzi et al. 2014).

We choose the proposed method with fourteen most advanced methods, include five traditional methods and nine DL-based methods. Among the traditional methods, EXP (Aiazzi et al. 2002) is an up-sampling method, also called 23-tap polynomial interpolation. BT-H (Lolli et al. 2017) and BDSDPC (Vivone 2019) are CS-based methods, while MF (Restaino et al. 2016) and MTF-GLP-FS (FS) (Vivone, Restaino, and Chanussot 2018) are MRA-based methods. The remaining nine DL-based methods comprise PanNet (Yang et al. 2017a), GPPNN (Xu et al. 2021), FusionNet (Deng et al. 2021), SRPPNN (Cai and Huang 2021), TRRNet (Zhang et al. 2022a), LAGConv (Jin et al. 2022), AWFLN (Lu et al. 2023), RSANet (Liu et al. 2023a), and MMFN (Jian et al. 2023). To ensure a fair comparison, we use the authors' officially released code and the setup described in the original paper. All of the codes are executed on a computer that is equipped with an i5 -11,600 CPU and two GTX-3060

Table 1. Basic information of each dataset.

		Bit depth	Band	Resolution (m)	Train set		Valid set		R-Test set		F-Test set		Scene
					Size	Number	Size	Number	Size	Number	Size	Number	
GF2	PAN	10	1	0.8	64 × 64	19,809	64 × 64	2201	256 × 256	20	512 × 512	20	coasts,
	MS		4	3.2	16 × 16		16 × 16		64 × 64		128 × 128		vegetation,
QB	PAN	11	1	0.61	64 × 64	17,139	64 × 64	1905	256 × 256	20	512 × 512	20	buildings,
	MS		4	2.44	16 × 16		16 × 16		64 × 64		128 × 128		urban
WV3	PAN	11	1	0.3	64 × 64	9714	64 × 64	1080	256 × 256	20	512 × 512	20	
	MS		8	1.2	16 × 16		16 × 16		64 × 64		128 × 128		

GPUs. In addition, for the sake of reproducibility, all implementations of this work will be published at https://github.com/JUSTM0VE0N/CF2N.

4.2. Reduced-scale test

(1) Results on GF2 dataset (4-Band): The dataset primarily contains the environment such as streets, buildings, and croplands. The qualitative results of each advanced method are shown in Figure 4, which includes five typical scenarios, such as roads, rivers, croplands, small buildings, and large buildings. In addition, the scores obtained by each scenario on the two metrics ERGAS and Q2n are above the subjective results, which can measure the spatial fidelity and spectral fidelity of the fusion results. Traditional methods appear to exhibit varying degrees of distortion in most cases. The CS-based methods BT-H and BDSDPC demonstrate high spatial fidelity, such as clear bridge reflections in the river and fine building edges. This phenomenon can be attributed to the complete retention of information within the PAN. However, the redundant spatial details in PAN disturb the spectral distribution in the original MS, resulting in obvious spectral distortion in the fusion results. Comparative to CS, MRA-based methods are known for their high spectral fidelity. Nevertheless, owing to the inaccurate filter shape, the MRA-based methods MF and FS encounter obstacles in achieving high spatial fidelity while maintaining a reasonable spectral distribution, such as blurred bridge edges and cropland boundaries. SR-based methods such as SRPPPNN, LAGConv, AWFLN, and RSANet exhibit superior sharpening outcomes when compared with traditional methods. The details that are not present in the GT are sharpened by these methods, for instance, the road area in Figure 4(i) and the cropland area in Figure 4(iii). In addition, these methods exhibit excessively smooth texture in certain local areas, such as the white building edge area in Figure 4(iv) and the building center texture area in Figure 4(v). Some crucial information is lost during the fusion process as the SR-based methods fail to account for the uniqueness of diverse source data, which results in evident distortions in certain local areas. The dual-branch-based method, PanNet, exhibits severe spatial and spectral distortions that surpass those of traditional methods in some scenarios, such as blurred croplands in Figure 4(iii) and distorted building tones in Figure 4(v). The TRRNet and MMFN demonstrate excellent spatial and spectral retention capabilities. However, it is found from the corresponding AEMs in Figure 4(i),(iv),(v) that when the target scale is small, they are unable to preserve edge details well. This is attributed to the fact that these dual-branch structures ignore the interaction between diverse source features during the fusion process. Contrary to this, GPPNN and FusionNet based on the detail injection model can preserve finer details, such as the building areas in Figure 4(i),(iv), (v).

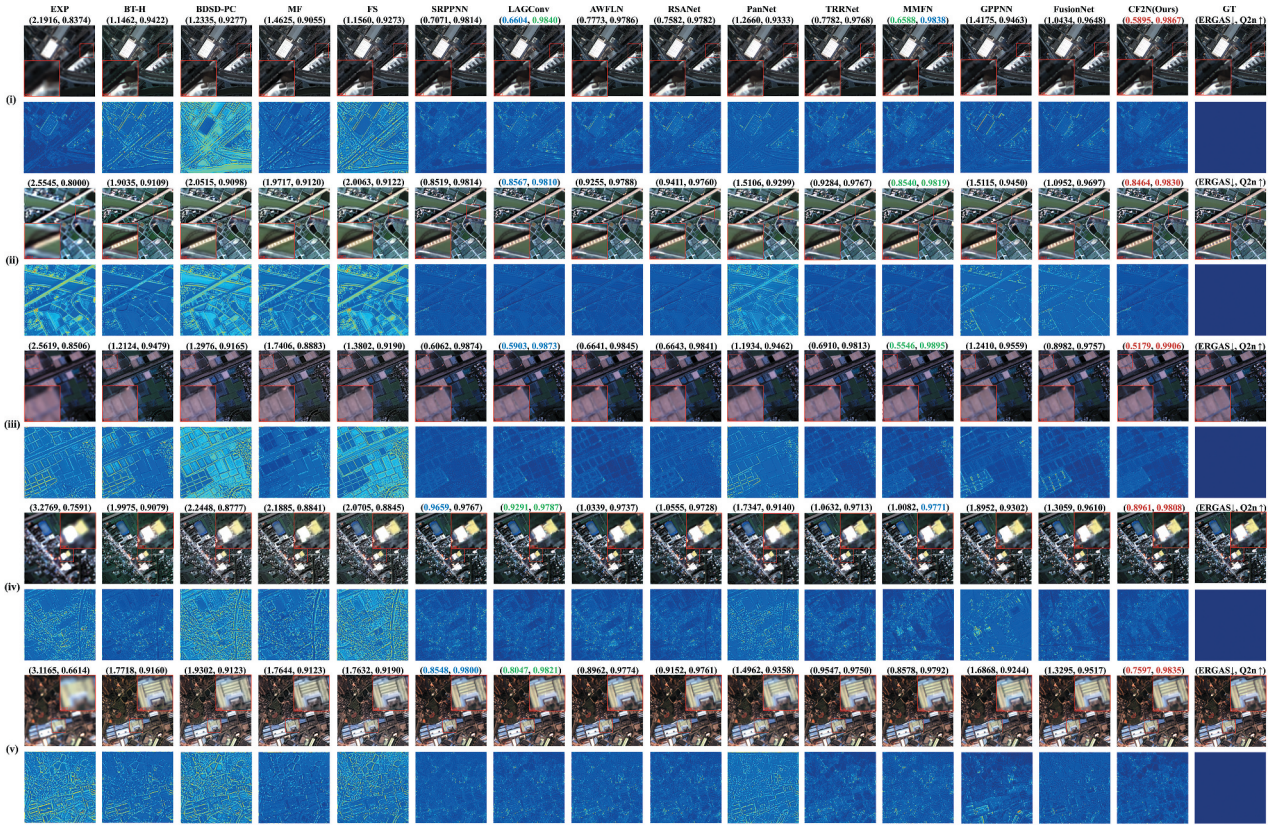

Figure 4. Qualitative evaluation results on GF2 R-Test set. The fusion results are presented in odd rows, while the corresponding AEMs are listed in even rows.

When the target area grows in scale, GPPNN and FusionNet tend to lose a lot of edge details. This is due to the fact that they solely concentrate on PAN-specific spatial information, neglecting the unique information in MS and the complementary spatial information with PAN, which have an impact on the spatial texture and spectral distribution. In contrast, as depicted in Figure 4(i)-(v), the proposed CF2N possesses the capability to preserve more spatial and spectral details, irrespective of whether the targets are small-scale or large-scale, owing to its emphasis on unique and complementary detail representations in diverse source images and the interaction between diverse source features.

Table 2 presents the quantitative results along with the mean and standard achieved for all samples in GF2 R-Test set. The top three performers on each metric are labeled red, green and blue respectively. The results show that DL-based methods perform better than traditional methods on all metrics. The achievement of top scores on most metrics with significant advantages is achieved by CF2N due to the concurrent focus on frequency details and spectral details as well as the deep interaction between the two. The last row of Table 2 shows an increase in percentage points compared to the second- and third-best methods. In particular, our performance on ERGAS and SCC is 8.02% and 0.06% superior to the second place, respectively. This indicates that the fusion results obtained by our method exhibit a lower global error and a higher spatial correlation, thereby leading to a more convenient spatial fidelity. Moreover, our performance on SAM, RASE and Q2n is 5.30%, 8.37% and 0.28% superior to the second place, respectively, indicating that the proposed CF2N preserves the most authentic and reasonable spectral distribution.

(2) Results on QB dataset (4-Band): The environment in the dataset includes oceans, cities, forests, etc. The qualitative outcomes of each method on the QB R-Test set are depicted in Figure 5, which encompasses five typical scenarios, including ports, urban buildings, sparse buildings, forests, and dense buildings. In general, each method exhibits a discernible decline in performance owing to the improvement of image quality and the complexity of the sharpening scene. Traditional methods exhibit evident spatial and spectral distortions, such as distorted water tone in Figure 5(i), rough texture in Figure 5(ii), (iv), and blurred artifacts in building areas in Figure 5(ii), (iv). The SR-based methods LAGConv, AWFLN and RSANet show more or less spectral distortions, such as the distorted water tone in Figure 5(i) and distorted spectral information in the playground part in Figure 5(ii); SRPPNN loses some crucial information, as shown in the area around the wave trajectory in Figure 5(i) and the building area in Figure 5(ii), (v). To learn the mapping to HRMS, these SR-based methods directly stitch two images from diverse sources in the spectral dimension; however, they neglect the uniqueness of diverse source data, resulting in the loss of frequency or spectral information. The PanNet shows spatial fidelity close to GT in the zoomed-in region of Figure 5(i), but it shows worse spectral retention compared to traditional methods in all examples. TRRNet and MMFN show slight spectral and spatial distortions, which are manifested in the distorted water color and building tone in Figure 5 (i)-(ii), as well as blurred paths in Figure 5(iii). These dual-branch-based methods employ two independent branches to extract the features of PAN and MS and then directly stitch them in the spectral dimension. They ignore the necessity of interaction between diverse source features, which results in poor performance in balancing spatial and spectral retention. GPPNN and FusionNet exhibit obvious spatial and spectral distortions due to their neglect of the unique information in MS and the complementary spatial information with PAN, which is crucial for fusion. The proposed CF2N shows slight spectral distortions as illustrated in Figure 5(i). In general, our method shows the fusion quality closest to GT in different scenarios, such as richer spectral information of

Table 2. Quantitative comparison on GF2 R-Test set.

Category	Method	SSIM↑	ERGAS↓	SAM↓	sCC↑	RASE↓	Q2n↑
Up-sample	EXP	0.9116 ± 0.0291	2.4094 ± 0.4647	1.8531 ± 0.3459	0.9542 ± 0.0203	8.5984 ± 1.7342	0.7971 ± 0.0430
CS-based	BT-H	0.9630 ± 0.0119	1.5526 ± 0.3548	1.6819 ± 0.3084	0.9690 ± 0.0125	5.5194 ± 1.2519	0.9089 ± 0.0284
	BDSDPC	0.9549 ± 0.0163	1.6954 ± 0.3896	1.7243 ± 0.3118	0.9689 ± 0.0119	6.1165 ± 1.4205	0.8847 ± 0.0300
MRA-based	MF	0.9516 ± 0.0149	1.7763 ± 0.3009	1.6842 ± 0.3115	0.9670 ± 0.0104	6.4951 ± 1.2576	0.8784 ± 0.0240
	FS	0.9533 ± 0.0160	1.6201 ± 0.3526	1.6807 ± 0.3394	0.9685 ± 0.0118	5.8813 ± 1.3128	0.8904 ± 0.0250
SR-based	SRPPNN	**0.9905** ± 0.0021	0.7299 ± 0.1088	0.8250 ± 0.1387	0.9933 ± 0.0028	2.6163 ±0.4054	0.9778 ± 0.0085
	LAGConv	**0.9909** ± 0.0018	**0.7223** ± 0.0963	**0.8099**±0.1282	**0.9933** ± 0.0027	**2.5852** ± 0.3656	**0.9785** ± 0.0096
	AWFLN	0.9888 ± 0.0025	0.8030 ± 0.1192	0.9208 ± 0.1554	0.9916 ± 0.0033	2.8562 ± 0.4421	0.9745 ± 0.0091
	RSANet	0.9887 ± 0.0026	0.8011 ± 0.1248	0.8846 ± 0.1514	0.9918 ± 0.0034	2.8742 ± 0.4696	0.9734 ± 0.0103
Dual-branch-based	PanNet	0.9720 ± 0.0065	1.3438 ± 0.1944	1.5439 ± 0.2339	0.9751 ± 0.0097	4.8840 ± 0.7304	0.9154 ± 0.0352
	TRRNet	0.9896 ± 0.0025	0.8246 ± 0.1153	0.8858 ± 0.1240	0.9927 ± 0.0030	2.9839 ± 0.4240	0.9685 ± 0.0138
	MMFN	**0.9909** ± 0.0026	**0.7167** ± 0.1331	**0.7710** ± 0.1486	**0.9940** ± 0.0028	**2.5714** ± 0.4952	**0.9796** ± 0.0070
Detail Injection-based	GPPNN	0.9714 ± 0.0080	1.4138 ± 0.2505	1.3435 ± 0.2331	0.9728 ± 0.0103	5.1413 ± 0.9417	0.9374 ± 0.0125
	FusionNet	0.9818 ± 0.0055	1.0438 ± 0.2102	1.0135 ± 0.1981	0.9894 ± 0.0046	3.7563 ± 0.7580	0.9603 ± 0.0101
	Ours	0.9830 ± 0.0039	**0.6592** ± 0.1140	**0.7301** ± 0.1323	**0.9946** ± 0.0025	**2.3563** ± 0.4222	**0.9823** ± 0.0070
		–	(−8.02/8.74%)	(−5.30/9.85%)	(+0.06/0.13%)	(−8.37/8.85%)	(+0.28/0.39%)

Figure 5. Qualitative evaluation results on QB R-Test set. The fusion results are presented in odd rows, while the corresponding AEMs are listed in even rows.

building areas in Figure 5(ii), clearer road texture in Figure 5(iii), more consistent spectral distribution in Figure 5(iv) and finer building edges in Figure 5(v). These are inextricably linked to our methodology's emphasis on the unique and complementary detail representation of diverse source data, as well as the interaction between diverse domain characteristics.

Table 3 shows that the proposed CF2N achieve the best score on all metrics. When compared to the second-best method, our scores on SSIM, ERGAS, and SCC are superior by 0.44%, 5.74%, and 0.26% better, respectively. This indicates that our methodology is capable of retaining more comprehensive details, owing to our emphasis on frequency and spectral details. Moreover, our scores on SAM, RASE, and Q2n are at least 3.66%, 5.59%, and 0.41% better, respectively, which further verifies the high spectral fidelity of the fusion results obtained by our methodology. These demonstrate that, despite the more intricate nature of the sharpening scenes, our method still possesses the strongest capacity for achieving the desired results.

(3) Results on WV3 dataset (8-Band): The proposed CF2N achieves commendable outcomes on two 4-band datasets. To further assess the generalizability, we conduct qualitative and quantitative experiments on the 8-band WV3 dataset. The environment in the dataset includes mountains, rivers,

Table 3. Quantitative comparison on QB R-Test set.

Category	Method	SSIM↑	ERGAS↓	SAM↓	sCC↑	RASE↓	Q2n↑
Up-sample	EXP	0.6879 ± 0.0840	12.0189 ± 1.5432	8.5575 ± 1.7025	0.8701 ± 0.0264	46.6474 ± 8.3320	0.5782 ± 0.0774
CS-based	BT-H	0.8684 ± 0.0288	7.4939 ± 0.5895	7.2981 ± 1.3674	0.9416 ± 0.0104	29.1453 ± 3.8422	0.8295 ± 0.0963
	BDSDPC	0.8624 ± 0.0293	7.6084 ± 0.5744	8.1813 ± 1.7780	0.9324 ± 0.0131	29.6034 ± 3.9402	0.8287 ± 0.0966
MRA-based	MF	0.8473 ± 0.0353	8.9013 ± 3.0752	8.0350 ± 1.6997	0.9294 ± 0.0135	33.1591 ± 6.2940	0.8120 ± 0.0950
	FS	0.8935 ± 0.0281	7.4454 ± 0.5512	7.8662 ± 1.6282	0.9378 ± 0.0124	29.0174 ± 3.7189	0.8336 ± 0.0932
SR-based	SRPPNN	0.9498 ± 0.0059	4.3333 ± 0.2882	5.1869 ± 0.8108	0.9789 ± 0.0040	16.5415 ± 1.3385	0.9140 ± 0.1125
	LAGConv	**0.9593** ± 0.0046	**3.8610** ± 0.3068	**4.7172** ± 0.7511	**0.9837** ± 0.0035	**14.9851** ± 1.5373	**0.9315** ± 0.0895
	AWFLN	**0.9592** ± 0.0056	3.9246 ± 0.3209	**4.6689** ± 0.7859	**0.9844** ± 0.0037	15.2497 ± 1.8084	**0.9320** ± 0.0851
	RSANet	0.9586 ± 0.0052	**3.8686** ± 0.2722	4.7659 ± 0.7969	0.9834 ± 0.0037	**15.0549** ± 1.6040	0.9301 ± 0.0916
Dual-branch-based	PanNet	0.8804 ± 0.0204	6.9987 ± 0.5552	7.9971 ± 1.4499	0.9320 ± 0.0155	27.3428 ± 3.4054	0.8413 ± 0.1040
	TRRNet	0.9445 ± 0.0164	4.8495 ± 0.9964	5.0151 ± 0.9622	0.9786 ± 0.0078	18.8421 ± 4.3051	0.9146 ± 0.0866
	MMFN	0.9332 ± 0.0207	5.7966 ± 1.1541	5.5140 ± 1.0623	0.9720 ± 0.0107	22.5006 ± 5.1659	0.8995 ± 0.0858
Detail Injection-based	GPPNN	0.9049 ± 0.0193	7.0470 ± 0.6431	5.8599 ± 1.0636	0.9587 ± 0.0085	27.1333 ± 3.9996	0.8717 ± 0.0799
	FusionNet	0.9521 ± 0.0068	4.2581 ± 0.2493	4.9853 ± 0.8163	0.9807 ± 0.0038	16.4903 ± 1.5234	0.9224 ± 0.0940
	Ours	**0.9635** ± 0.0045	**3.6392** ± 0.2836	**4.4978** ± 0.7600	**0.9870** ± 0.0031	**14.1474** ± 1.5612	**0.9358** ± 0.0887
		(+0.44/0.45%)	(−5.74/5.93%)	(−3.66/4.66%)	(+0.26/0.34%)	(−5.59/6.03%)	(+0.41/0.46%)

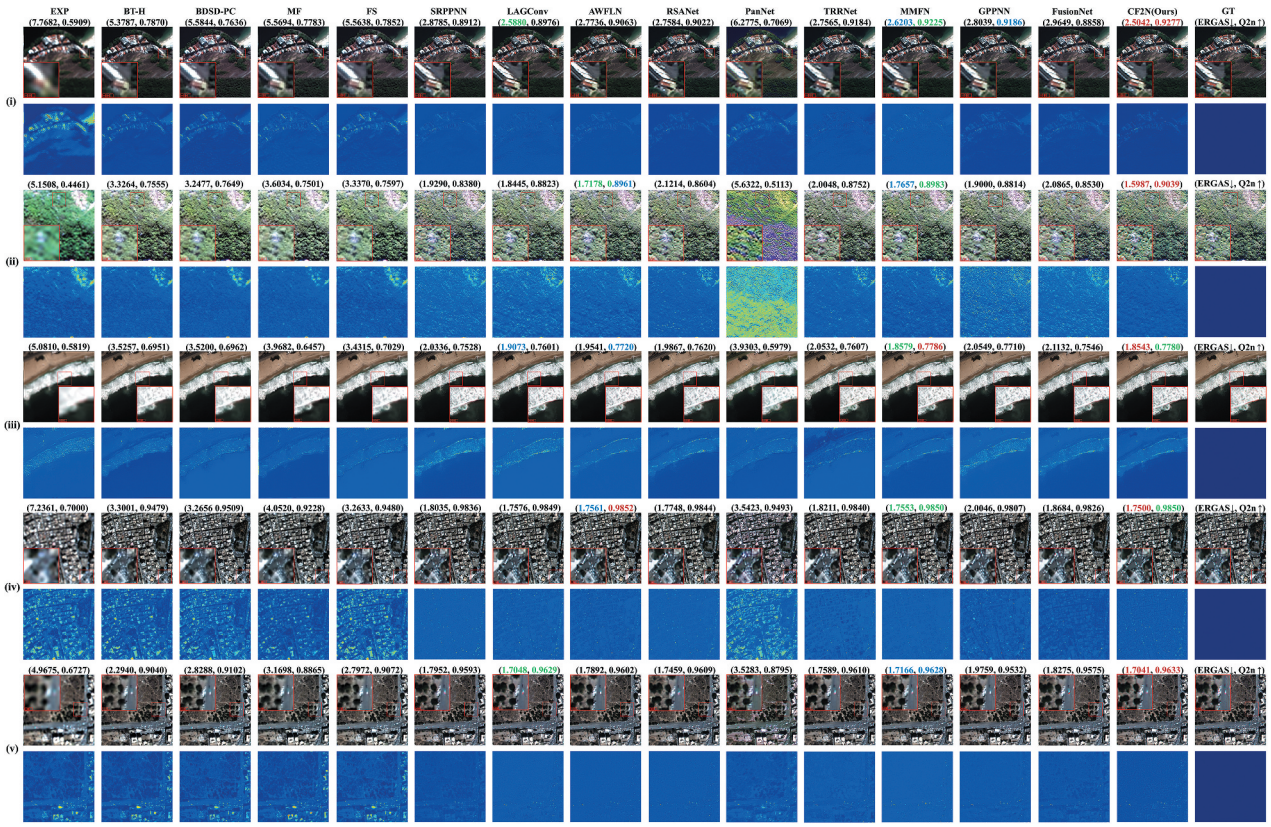

Figure 6. Qualitative evaluation results on WV3 R-Test set. The fusion results are presented in odd rows, while the corresponding AEMs are listed in even rows.

cities, etc. The qualitative outcomes of each method on the QB R-Test set are depicted in Figure 6, which encompasses five typical scenarios, Mountains, jungles, waves, dense buildings, and wide roads. The performance of the traditional method is further degraded upon enhancement of image quality. Specifically, BT-H and BDSDPC, which are known for their high spatial fidelity, lost a lot of texture details in the shrubbery in Figure 6(i)-(ii). The MRA methods MF and FS, which are known for spectral fidelity, sharpened the emerald green shrubbery to white green, accompanied by blurring. These occurrences can be attributed to the limited linear representation ability, which proves challenging to apply to intricate scenes of sharpening. The SR-based methods SRPPNN, LAGConv, AWFLN, and RSANet show comparable sharpening results. Compared with GT, these methods exhibit a certain gap in spatial and spectral fidelity. The discrepancy is attributed to the fact that the SR method directly learns the mapping of input image pairs to HRMS, disregarding the uniqueness of diverse source data, resulting in the inevitable loss of spatial and spectral details. PanNet reveals unacceptable results. It employs two distinct branches to extract HF information from diverse source images through Gaussian filtering, and then directly splices the HF information from diverse sources. Large amounts of redundant information from diverse sources cause the network to crash because they fail to fully interact with the frequencies of diverse sources. It can be seen from Figure 6(i)-(v) that PanNet actually retains a good spatial texture, but its spectral distribution is extremely poor, which further proves the necessity of interaction. The other two dual-branch-based methods TRRNet and MMFN demonstrate considerable sharpening results. Similar to SR-based methods, they also lost crucial spatial and spectral information, such as the whitened jungle in Figure 6(ii). These methods take into account the uniqueness of diverse source data during the feature extraction stage, which helps them retain more spatial texture in some scenes, such as the wave area in Figure 6(iii) and the building area in Figure 6(iv). However, they ignore the interaction between diverse source features during the fusion process, rendering it challenging to achieve global spatial-spectral retention. Due to the lack of the unique information in MS and the complementary spatial information with PAN, GPPNN and FusionNet based on detail injection show evident spatial and spectral distortions, such as the white jungle in Figure 6(ii), and the blurred edge texture in Figure 6(iii)-(iv). In contrast, the CF2N exhibits the most superior spatial and spectral

Table 4. Quantitative comparison on WV3 R-Test set.

Category	Method	SSIM↑	ERGAS↓	SAM↓	sCC↑	RASE↓	Q2n↑
Up-sample	EXP	0.7246 ± 0.0933	7.1354 ± 1.5641	5.8351 ± 1.6720	0.9226 ± 0.0290	22.2931 ± 6.3631	0.6027 ± 0.0889
CS-based	BT-H	0.9018 ± 0.0221	4.5107 ± 1.2973	4.8984 ± 1.2695	0.9586 ± 0.0123	13.9308 ± 4.7883	0.8182 ± 0.0993
	BDSDPC	0.8947 ± 0.0285	4.6499 ± 1.4270	5.4643 ± 1.6708	0.9552 ± 0.0117	13.8559 ± 4.7681	0.8117 ± 0.1036
MRA-based	MF	0.8846 ± 0.0244	4.9190 ± 1.2875	5.3162 ± 1.4722	0.9527 ± 0.0136	15.0103 ± 4.7843	0.7957 ± 0.1005
	FS	0.8905 ± 0.0303	4.6450 ± 1.4062	5.3228 ± 1.6112	0.9560 ± 0.0114	13.7713 ± 4.8366	0.8177 ± 0.0989
SR-based	SRPPNN	0.9708 ± 0.0078	2.3587 ± 0.5462	3.1867 ± 0.5588	0.9859 ± 0.0056	7.3885 ± 2.0307	0.8927 ± 0.0889
	LAGConv	**0.9731** ± 0.0073	**2.2435** ± 0.5154	3.1007 ± 0.5211	0.9870 ± 0.0053	**7.0068** ± 1.8672	0.9039 ± 0.0864
	AWFLN	0.9723 ± 0.0086	2.2969 ± 0.5278	**3.0712** ± 0.5684	**0.9878** ± 0.0052	7.2374 ± 1.9565	**0.9086** ± 0.0832
	RSANet	0.9707 ± 0.0079	2.2904 ± 0.4945	3.1252 ± 0.5458	0.9864 ± 0.0056	7.1681 ± 1.8556	0.8995 ± 0.0882
Dual-branch-based	PanNet	0.8944 ± 00183	5.2335 ± 1.5127	6.9395 ± 1.3863	0.9321 ± 0.0355	15.5878 ± 4.1562	0.7491 ± 0.1464
	TRRNet	0.9718 ± 0.0078	2.3638 ± 0.5133	3.2627 ± 0.5607	0.9847 ± 0.0058	7.3628 ± 1.9141	0.9045 ± 0.0858
	MMFN	**0.9737** ± 0.0078	**2.2488** ± 0.4954	**2.9772** ± 0.5516	**0.9883** ± 0.0052	**7.0415** ± 1.8871	**0.9138** ± 0.0813
Detail Injection-based	GPPNN	0.9690 ± 0.0091	2.4828 ± 0.5251	3.3093 ± 0.6258	0.9800 ± 0.0078	7.4752 ± 1.8249	0.9045 ± 0.0821
	FusionNet	0.9680 ± 0.0085	2.4778 ± 0.5775	3.3821 ± 0.6256	0.9836 ± 0.0064	7.6414 ± 2.0502	0.8939 ± 0.0876
	Ours	**0.9751** ± 0.0075	**2.1591** ± 0.4628	**2.9341** ± 0.5288	**0.9892** ± 0.0049	**6.8298** ± 1.8060	**0.9145** ± 0.0805
		(+0.14/0.21%)	(−3.76/3.99%)	(−1.45/4.46%)	(+0.10/0.14%)	(−2.53/3.01%)	(+0.08/0.65%)

retention, particularly in the fine texture of shrubs and consistent flower hues in Figure 6(i)-(ii), the finer wave texture in Figure 6(iii), and the more accurate road details in Figure 6(iv)-(v). These results can be attributed to the proposed methodology's emphasis on frequency detail reconstruction and spectral retention, as well as the consideration of frequency-spectral targets at various scales in different scenarios.

Similar quantitative outcomes can be observed in Table 4. CF2N achieves top scores on all metrics due to its focus on frequency details and spectral details, as well as the deep interaction between the two. The proposed CF2N achieves scores of at least 0.14%, 3.76% and 0.10% better on SSIM, ERGAS and SCC compared to other methods, indicating that our method can retain fine-grained details, owing to our emphasis on frequency details and spectral details. Furthermore, our scores on SAM, RASE, and Q2n are at least 3.76%, 2.53%, and 0.08% better than other methods, respectively, which further confirms the high spectral fidelity of the proposed method.

4.3. Full-scale test

(1) Results on GF2 dataset (4-Band): The generalization capability of each method is evaluated in real-world sharpening scenarios by F-Test. We show all types of ground objects in the GF2 dataset, including roads, rivers, bridges, croplands, small and large buildings, to evaluate the sharpening performance of the model under different environmental conditions. The corresponding QNR and HQNR scores are presented above each image, indicating the overall quality of the fused image. As shown in Figure 7(i), when the spatial resolution is 0.8 m and the sharpening environment is relatively single, each method can achieve good sharpening quality. However, most DL-based methods such as SRPPNN, LAGConv, AWFLN, RSANet, TRRNet, MMFN and FusionNet still have some flaws, which is manifested in the gray building edge being blue as shown in the zoomed-in area. This is attributable to the inaccurate detail representation and inadequate interaction of these methods. GPPNN, which is based on the detail injection model, possesses a robust spatial retention capability. However, it exhibits the loss of spectral information in small-scale building areas, as depicted in Figure 7(i)-(ii). Each method shows significant performance degradation when the sharpening scene becomes intricate, as shown in Figure 7(iii). All SR-based methods suffer from the loss of certain spatial or spectral details, as exemplified by blurred bridge edges, car roof colors that are inconsistent with MS, and distorted croplands. The reason for this is that SR-based methods ignore the uniqueness of diverse source data, resulting

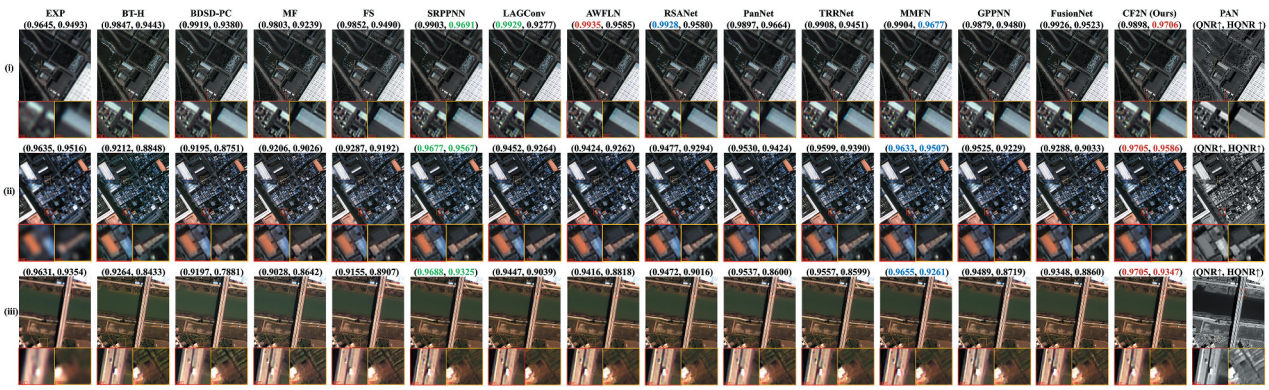

Figure 7. Qualitative evaluation results on GF2 F-Test set.

in the loss of crucial information, such as edge details and spectral details. PanNet exhibits severe spectral distortion, as evidenced by the hue of the river. TRRNet and MMFN show slight spatial and spectral distortions, which is manifested in the blurring of the boundary area of the cropland center. GPPNN and FusionNet show notable spatial and spectral distortions, such as blurred bridges and obvious chromatic aberration. Their omission of the unique information in MS and the complementary spatial information with PAN is crucial for fusion. The proposed CF2N demonstrates satisfactory sharpening quality in different sharpening environments, such as the spatial details consistent with PAN and retaining a reasonable spectral distribution in Figure 7(i)-(iii), which benefits from the model's emphasis on the uniqueness and complementarity of diverse source images and the concrete implementation of the deep interaction of diverse domain features.

Table 5 presents the quantitative results achieved for all samples in GF2 F-Test set. CF2N achieves the best mean and the lowest standard deviation on all non-reference metrics, with the exception of D_λ, demonstrating its generalization ability and stability of the proposed CF2N at full-scale. Specifically, the proposed CF2N's score on the spatial distortion index D_s is at least 4.71% superior to any other method, indicating that our fusion exhibits the highest level of spatial fidelity, owing to our comprehensive consideration of uniqueness and complementarity in diverse source data. The score obtained on the spectral distortion index D_λ^F is at least 2.53% better than any other method, which means that our fusion has the best spectral fidelity. This is attributable to the concrete implementation of spectral terms within the constructed model, thereby enabling the complete integration of frequency details and corresponding spectral details. Furthermore, our scores on the two comprehensive quality assessment matrices QNR and HQNR are at least 0.20 and 0.54% better, respectively, indicating that our fusion outcomes exhibit high frequency-spectral fidelity.

(2) Results on QB dataset (4-Band): As illustrated in Figure 8, when the spatial resolution is increased to 0.61 m and the sharpening scene becomes intricate, the sharpening performance of each method is significantly degraded. Despite the high spatial fidelity of CS-based methods BT-H and BDSDPC, they exhibit incredible spectral distortion, especially BT-H. In contract to CS, the MRA-based method exhibits superior spectral fidelity, however, it fails to account for spatial information, such as smooth sea surface and blurred building edges. The SR-based methods SRPPNN, LAGConv and RSANet show obvious spectral distortion, such as the unreasonable spectral distribution in Figure 8(i)-(iii). PanNet and TRRNet also show

Table 5. Quantitative comparison on GF2 F-Test set.

Category	Method	$D_\lambda\downarrow$	$D_\lambda^F\downarrow$	$D_s\downarrow$	QNR↑	HQNR↑
Up-sample	EXP	0.0000 ± 0.0000	0.0140 ± 0.0048	0.0263 ± 0.0176	0.9737 ± 0.0176	0.9601 ± 0.0187
CS-based	BT-H	0.0216 ± 0.0114	0.0639 ± 0.0232	0.0555 ± 0.0179	0.9242 ± 0.0262	0.8841 ± 0.0296
	BDSDPC	0.0130 ± 0.0095	0.0810 ± 0.0286	0.0559 ± 0.0196	0.9319 ± 0.0260	0.8678 ± 0.0368
MRA-based	MF	0.0326 ± 0.0188	0.0584 ± 0.0200	0.0593 ± 0.0208	0.9104 ± 0.0361	0.8858 ± 0.0297
	FS	0.0236 ± 0.0134	0.0375 ± 0.0130	0.0524 ± 0.0173	0.9254 ± 0.0273	0.9121 ± 0.0210
SR-based	SRPPNN	**0.0058** ± 0.0056	**0.0198** ± 0.0079	**0.0276** ± 0.0111	**0.9668** ± 0.0149	**0.9531** ± 0.0135
	LAGConv	0.0113 ± 0.0085	0.0296 ± 0.0089	0.0397 ± 0.0138	0.9495 ± 0.0192	0.9319 ± 0.0161
	AWFLN	0.0120 ± 0.0085	0.0314 ± 0.0145	0.0405 ± 0.0143	0.9480 ± 0.0196	0.9294 ± 0.0192
	RSANet	0.0099 ± 0.0083	0.0297 ± 0.0100	0.0378 ± 0.0133	0.9528 ± 0.0184	0.9337 ± 0.0160
Dual-branch-based	PanNet	0.0121 ± 0.0095	0.0336 ± 0.0262	0.0356 ± 0.0127	0.9528 ± 0.0193	0.9320 ± 0.0296
	TRRNet	0.0074 ± 0.0067	0.0394 ± 0.0229	0.0310 ± 0.0118	0.9619 ± 0.0152	0.9308 ± 0.0241
	MMFN	**0.0066** ± 0.0061	**0.0211** ± 0.0085	**0.0295** ± 0.0111	**0.9641** ± 0.0151	**0.9499** ± 0.0124
Detail Injection-based	GPPNN	0.0100 ± 0.0067	0.0432 ± 0.0159	0.0371 ± 0.0126	0.9533 ± 0.0174	0.9213 ± 0.0201
	FusionNet	0.0157 ± 0.0097	0.0407 ± 0.0110	0.0483 ± 0.0168	0.9369 ± 0.0232	0.9130 ± 0.0203
	Ours	**0.0059** ± 0.0056	**0.0193** ± 0.0078	**0.0263** ± 0.0106	**0.9680** ± 0.0143	**0.9550** ± 0.0117
		–	(−2.53/8.53%)	(−4.71/10.85%)	(+0.12/0.40%)	(+0.20/0.54%)

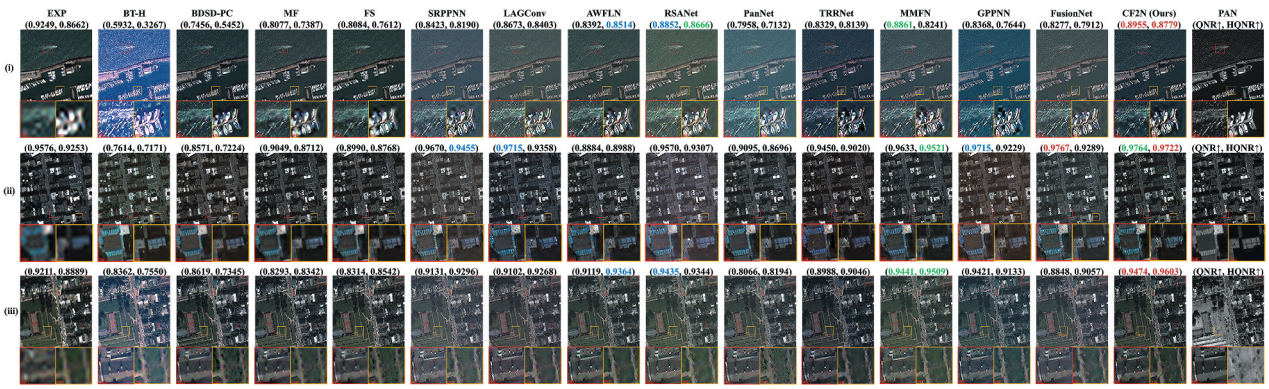

Figure 8. Qualitative evaluation results on QB F-Test set.

spectral distortion, such that PanNet sharpens the sea tone in Figure 8(i) to blue-green, while TRRNet sharpens the vegetation area in Figure 8(iii) to black. Similarly, the model-based methods GPPNN and FusionNet show obvious spectral distortion. GPPNN sharpens the playground tone to dark black, whereas FusionNet sharpens an additional path that does not exist in the original image pair. This is due to the emphasis placed on spatial details in PAN, while ignoring the guiding role of complementary spatial details in MS for sharpening. In contrast, the fusion results of AWFLN and MMFN are more acceptable. These distortions can be attributed to the fact that these methods do not fully exploit the detailed representation in diverse source images and the insufficient information interaction between diverse source features. When compared to EXP and PAN, it can be observed that the proposed CF2N exhibits spatial the details that are consistent with PAN and the spectral distribution that is closest to EXP. For instance, in the port area, only our method and traditional methods preserve the spatial texture in MS. This is due to the fact that the CF2N takes into consideration the uniqueness and complementarity of diverse source data, which is overlooked by other DL-based methodologies.

The quantitative results on the QB F-Test set are depicted in Table 6. The proposed CF2N achieves a clear advantage in two comprehensive quality assessment metrics, QNR and HQNR, indicating that it is capable of obtaining fusion results with high frequency-spectral fidelity. Specifically, the proposed CF2N's score on the spectral distortion index D_λ^F is at least 8.11% superior to any other method, indicating that our fusion exhibits the highest degree of spectral fidelity, which is attributable to the complete integration of frequency details and corresponding spectral details. Although the CF2N did not achieve the best score on D_λ and D_s, but our scores on the two comprehensive quality assessment matrices QNR and HQNR are at least 0.45 and 1.53% better, respectively. This indicates that CF2N achieves an optimal balance between spatial and spectral preservation.

(3) Results on WV3 dataset (8-Band): The spatial resolution of PAN in this dataset is 0.3 m, which includes dense buildings, vegetation and vehicles. As shown in Figure 9, when the data is fine-grained, the detail texture retention of the DL-based method is questioned, which contradicts the scores of quantitative metrics. In the dense vehicle area shown in Figure 9(i),(iii), the CS-based method is very close to PAN in terms of spatial fidelity, but there is a slight spectral distortion. In contrast, the MRA-based method is slightly inferior to CS-based in terms of spatial fidelity, albeit it exhibits advantages in spectral fidelity. SR-based and dual-branch-based methods

Table 6. Quantitative comparison on QB F-Test set.

Category	Method	$D_\lambda\downarrow$	$D_\lambda^F\downarrow$	$D_s\downarrow$	QNR↑	HQNR↑
Up-sample	EXP	0.0001 ± 0.0000	0.0348 ± 0.0071	0.0529 ± 0.0156	0.9470 ± 0.0156	0.9142 ± 0.0190
CS-based	BT-H	0.1457 ± 0.0458	0.2344 ± 0.0738	0.1244 ± 0.0506	0.7489 ± 0.0690	0.6731 ± 0.0863
	BDSDPC	**0.0261** ± 0.0142	0.1948 ± 0.0334	0.1415 ± 0.0306	0.8362 ± 0.0346	0.6919 ± 0.0464
MRA-based	MF	0.0350 ± 0.0211	0.0609 ± 0.0178	0.1105 ± 0.0246	0.8589 ± 0.0403	0.8356 ± 0.0342
	FS	0.0342 ± 0.0199	0.0454 ± 0.0149	0.1146 ± 0.0225	0.8554 ± 0.0356	0.8454 ± 0.0303
SR-based	SRPPNN	0.0384 ± 0.0327	0.0601 ± 0.0204	0.0279 ± 0.0193	0.9353 ± 0.0465	0.9140 ± 0.0339
	LAGConv	0.0406 ± 0.0312	0.0648 ± 0.0173	0.0201 ± 0.0098	0.9402 ± 0.0362	**0.9164** ± 0.0224
	AWFLN	0.0376 ± 0.0236	**0.0333** ± 0.0148	0.0490 ± 0.0233	0.9154 ± 0.0349	0.9194 ± 0.0276
	RSANet	0.0348 ± 0.0235	0.0620 ± 0.0129	**0.0240** ± 0.0141	**0.9422** ± 0.0355	0.9155 ± 0.0289
Dual-branch-based	PanNet	0.0563 ± 0.0659	0.0763 ± 0.0263	0.1262 ± 0.1265	0.8321 ± 0.1465	0.8073 ± 0.1213
	TRRNet	0.0458 ± 0.0389	0.0788 ± 0.0174	0.0304 ± 0.0103	0.9253 ± 0.0406	0.8933 ± 0.0214
	MMFN	0.0333 ± 0.0250	**0.0466** ± 0.0250	**0.0156** ± 0.0081	**0.9517** ± 0.0293	**0.9386** ± 0.0285
Detail Injection-based	GPPNN	**0.0306** ± 0.0266	0.0827 ± 0.0223	0.0288 ± 0.0204	0.9419 ± 0.0421	0.8912 ± 0.0363
	FusionNet	0.0418 ± 0.0345	0.0741 ± 0.0231	0.0284 ± 0.0128	0.9315 ± 0.0532	0.8998 ± 0.0317
	Ours	**0.0277** ± 0.0239	**0.0306** ± 0.0172	**0.0170** ± 0.0099	**0.9560** ± 0.0308	**0.9530 ±** 0.0218
		–	(−8.11/34.33%)	–	(+0.45/1.46%)	(+1.53/3.65%)

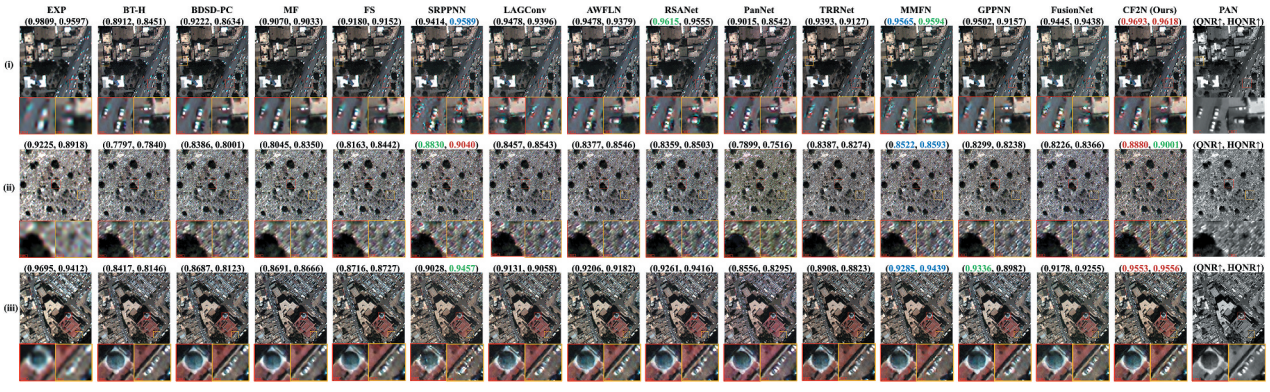

Figure 9. Qualitative evaluation results on WV3 F-Test set.

result in the loss of key information during the process of feature extraction or fusion due to their imprecise detail representation and inadequate interaction. Among them, SRPPNN, LAGConv, RSANet, TRRNet, and MMFN demonstrate poor sharpening performance, especially for vehicles with severe deformation. The vehicle edge contour is preserved by PanNet, but there is an apparent spectral distortion. The fusion performance of AWFLN, GPPNN, and CF2N is comparable, with AWFLN exhibiting superior spatial fidelity, whereas our method exhibits superior spectral fidelity. In the dense vegetation area shown in Figure 9(ii), only the SRPPNN and the proposed CF2N show similar spectral distribution to EXP, and other methods show obvious distortion. Compared with SRPPNN, the fusion results of our method have finer spatial texture. In general, the proposed CF2N can also achieve high frequency-spectral fidelity in fine-grained scenarios, which is due to the uniqueness and complementarity of diverse source data and the necessity of interaction between diverse domain characteristics.

Table 7 shows that our method yields the best scores for two spectral distortion metrics (D_λ and D_λ^F) and a comprehensive quality assessment matric (QNR). Specifically, the scores obtained by CF2N on D_λ and D_λ^F are at least 10.08% and 16.75% superior to other methods, respectively, indicating that our fusion exhibits the highest degree of spectral fidelity. The concrete implementation of spectral terms within the constructed model enables the complete integration of frequency details and corresponding spectral details. The score obtained on QNR is at least 1.45% better than any other method, which means that CF2N can obtain high-fidelity fusion results. Even though our method is slightly inferior to SRPPNN in HQNR, coupled with the poor fusion quality of SRPPNN in qualitative assessment, we hold the belief that the proposed CF2N holds the greatest promise, even in the face of high-fine-grained sharpening scenarios. In addition, as the image quality increases, i.e. the spatial resolution increases from 0.8 m to 0.3 m, the score obtained by the proposed method is gradually improved compared to the second-best method. For example, the QNR scores obtained by our methodology on GF2, QB, and WV3 are 0.12%, 0.45%, and 1.55% higher than the second place, respectively. This further underscores the high level of adaptability to different scenes and robust generalization in intricate scenes.

4.4. Efficiency analysis

We compare the proposed CF2N with the existing state-of-the-art methods in terms of four aspects, including the floating-point operations (FLOPs), number of parameters (NoPs), multiply-accumulate operations (MACs), and the test runtime (Time). The NoPs is utilized to quantify the spatial complexity of the model, while the FLOPs and MACs are utilized to quantify the computational complexity, and the Time is utilized to quantify the temporal complexity. As depicted in Table 8, CF2N exhibits significant advantages in the NoPs, FLOPs, and MACs, albeit with a slight decline in Time. Furthermore, taking into account the significance of model sharpening performance, we affix model performance with four efficiency aspects to provide a comprehensive assessment. Our method exhibits the optimal performance in both R-Test and F-Test, as shown in Figure 10. In

Table 7. Quantitative comparison on WV3 F-Test set.

Category	Method	$D_\lambda\downarrow$	$D_\lambda^F\downarrow$	$D_s\downarrow$	QNR↑	HQNR↑
Up-sample	EXP	0.0000 ± 0.0000	0.0232 ± 0.0064	0.0340 ± 0.0133	0.9660 ± 0.0166	0.9437 ± 0.0166
CS-based	BT-H	0.0268 ± 0.0198	0.0574 ± 0.0226	0.1001 ± 0.0375	0.8764 ± 0.0516	0.8489 ± 0.0511
	BDSDPC	**0.0129** ± 0.0096	0.0624 ± 0.0228	0.0912 ± 0.0362	0.8973 ± 0.0432	0.8528 ± 0.0509
MRA-based	MF	0.0266 ± 0.0194	0.0275 ± 0.0100	0.0853 ± 0.0293	0.8909 ± 0.0438	0.8898 ± 0.0355
	FS	0.0197 ± 0.0168	**0.0197** ± 0.0076	0.0851 ± 0.0307	0.8973 ± 0.0432	0.8971 ± 0.0352
SR-based	SRPPNN	0.0338 ± 0.0188	**0.0164** ± 0.0072	**0.0355** ± 0.0145	**0.9320** ± 0.0246	**0.9487** ± 0.0165
	LAGConv	0.0178 ± 0.0126	0.0258 ± 0.0108	0.0572 ± 0.0192	0.9261 ± 0.0256	0.9185 ± 0.0230
	AWFLN	0.0164 ± 0.0136	0.0203 ± 0.0097	0.0633 ± 0.0233	0.9216 ± 0.0334	0.9179 ± 0.0292
	RSANet	0.0203 ± 0.0148	0.0232 ± 0.0096	0.0498 ± 0.0232	0.9309 ± 0.0270	0.9281 ± 0.0244
Dual-branch-based	PanNet	0.0251 ± 0.0192	0.0627 ± 0.0237	0.0933 ± 0.0348	0.8845 ± 0.0485	0.8503 ± 0.0482
	TRRNet	0.0171 ± 0.0150	0.0364 ± 0.0139	0.0735 ± 0.0236	0.9110 ± 0.0353	0.8931 ± 0.0326
	MMFN	0.0224 ± 0.0141	**0.0200** ± 0.0077	**0.0446** ± 0.0207	**0.9340** ± 0.0242	**0.9363** ± 0.0221
Detail Injection-based	GPPNN	**0.0130** ± 0.0135	0.0415 ± 0.0160	0.0622 ± 0.0239	0.9259 ± 0.0344	0.8991 ± 0.0331
	FusionNet	0.0230 ± 0.0179	0.0259 ± 0.0098	0.0548 ± 0.0222	0.9236 ± 0.0293	0.9208 ± 0.0247
	Ours	**0.0116 ±** 0.0078	**0.0164 ±** 0.0069	**0.0404** ± 0.0148	**0.9485 ±** 0.0181	**0.9439** ± 0.0154
		(−10.0 8/10.77%)	(−16.75/18.00%)	–	(+1.55/1.77%)	–

Table 8. Quantitative comparison of efficiency on WV3 dataset.

	PanNet	GPPNN	FusionNet	SRPPNN	TRRNet	LAGConv	AWFLN	RSANet	MMFN	Ours
NoPs (M)	0.31	0.48	0.31	3.44	5.23	0.30	0.34	**0.20**	4.74	0.25
FLOPs (G)	2.50	1.39	2.45	10.59	3.75	**0.26**	0.70	0.71	10.85	1.50
MACs (G)	1.25	0.69	1.23	5.30	1.87	**0.13**	0.35	0.35	5.42	0.75
Time (s)	**0.003**	0.012	0.005	0.014	0.085	0.011	0.007	0.059	0.014	0.024

The spatial scale of the image in the efficiency test is 64×64.

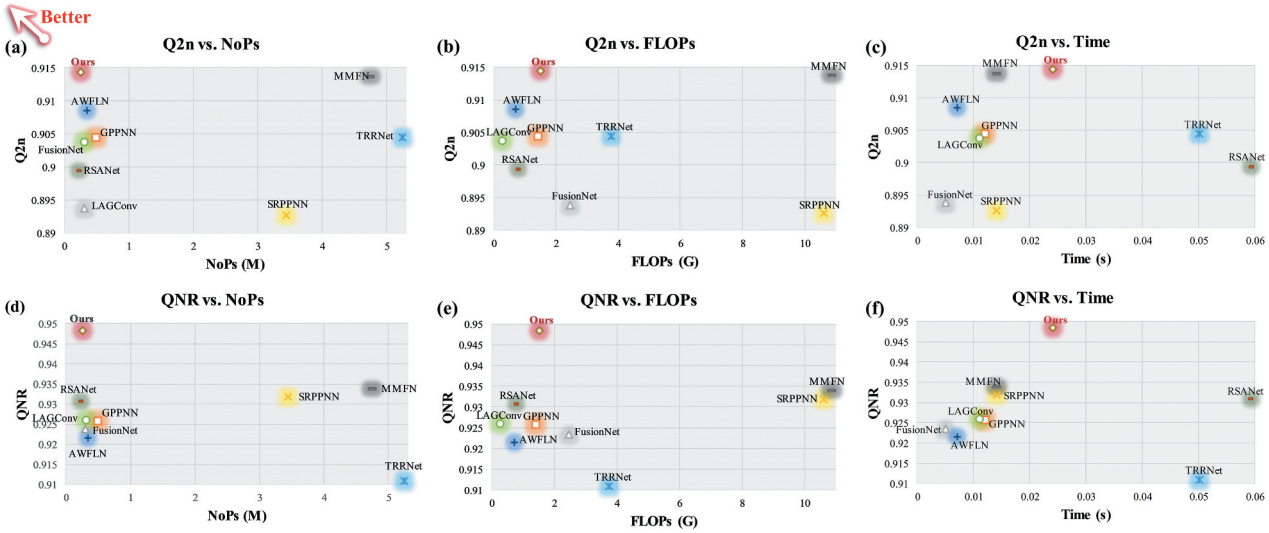

Figure 10. Comparison of the efficiency of existing state-of-the-art dl-based methods. The first line shows the relationship between reduced-scale performance and efficiency, and the second line shows the relationship between full-scale performance and efficiency. The first column shows the relationship between performance and NoPs, the second column shows the relationship between performance and the computational complexity, and the second column shows the relationship between performance and the time complexity.

terms of performance vs. NoPs, our method achieves the best performance while NoPs is second only to RSANet. In terms of performance vs. computational complexity, our method achieves a significant enhancement in performance, accompanied by a rise in computational complexity. Our method achieves the best balance between performance and Time at full-scale, but is inferior to MMFN at reduced-scale. In general, the proposed CF2N attains the optimal balance between fusion performance and efficiency.

4.5. Ablation experiments

We conduct ablation experiments on various components of CF2N, primarily comprising of 1) the components in FD2R, 2) the components in FSCFM, and 3) the number of FSCFMs and MDIs in the FSCF stage. All ablation experiments are conducted on the more challenging WV3 dataset.

(1) FD2R. we conduct ablation experiments on FD2R in order to test the impact of fine-grained details on the performance of pan-sharpening. Specifically, we proceed with sequential training without FD2R (w/o FD2R), replacing AWF in FD2R with Concatenation (FD2R_Cat), replacing AWF in FD2R with Addition (FD2R_Add), and implementing our proposed FD2R. The qualitative results are shown in Figure 11(a). It can be seen that without FD2R and FD2R_Cat temporarily out obvious spatial distortion and spectral distortion. It is

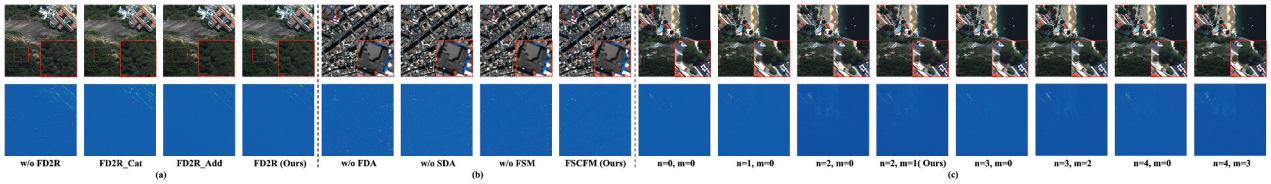

Figure 11. Qualitative results of ablation experiments conducted for various components. (a) FD2R. (b) FSCFM. (c) The number of FSCFMs and MDIs.

Table 9. Ablation experiment on FD2R.

	Quantitative Results					Efficiency Analysis			
Method	ERGAS↓	SAM↓	sCC↑	RASE↓	Q2n↑	NoPs (M)	FLOPs (G)	MACs (G)	Time (s)
w/o FD2R	2.5737 ± 0.5481	3.8875 ± 0.7351	0.9813 ± 0.0090	7.9055 ± 1.6727	0.8612 ± 0.1093	**0.12**	**0.99**	**0.50**	**0.016**
FD2R_Cat	2.3297 ± 0.4899	3.2264 ± 0.5348	0.9878 ± 0.0055	7.4501 ± 1.9048	0.9000 ± 0.0927	1.35	3.83	1.91	0.028
FD2R_Add	2.2619 ± 0.4973	3.0381 ± 0.5507	0.9877 ± 0.0054	7.1030 ± 1.8347	0.9105 ± 0.0812	0.25	1.50	0.75	0.022
FD2R	**2.1591** ± 0.4628	**2.9341** ± 0.5288	**0.9892** ± 0.0049	**6.8298** ± 1.8060	**0.9145** ± 0.0805	0.25	1.50	0.75	0.024

worthwhile mentioning that our method can be regarded as a generalized version of FD2R_Add. Our method achieves the minimal residuals in AEMs. Table 9 shows the results of quantitative experiments, as well as the efficiency metrics. The proposed method achieves the optimal sharpening performance, albeit with a slight efficiency degradation. It is noteworthy that the concatenation of information from various directions by FD2R_Cat results in a significant increase of channels, which is accompanied by a significant increase in the NoPs and computation. Overall, we believe that the proposed FD2R is an integral part that guides the model to reconstruct finer details.

(2) FSCFM. The FSCFM is an additional important component of the proposed CF2N, which provides a concrete implementation of the spectral terms in the proposed theoretical model. we sequentially train the module without FDA, without SDA, without FSM, and the proposed FSCFM. Figure 11(b) shows the sharpened results and AEMs, while Table 10 shows the quantitative experimental outcomes. The results in Table 10 support the conclusion that our method exhibits the mini mal residual, as demonstrated by AEMs. We observe significant spatial and spectral distortion in both without FDA and without SDA. Furthermore, the inclusion of the FSM significantly enhances the granularity of the sharpening results.

(3) Number of FSCFMs and MDIs. To examine the impact of the number of FSCFMs and MDIs in the FSCF stage, we conduct the corresponding ablation experiments. The results are shown in Table 11 and Figure 11(c). It is evident that the sharpening performance is significantly improved with the increase in number. Furthermore, it has been observed that the incorporation of MDI results in a slight enhancement in sharpening performance, albeit the efficiency of the model is significantly diminished. As a result, we select two FSCFMs and one MDI as our baseline methodology for the FSCF stage.

4.6. Hyperspectral sharpening experiments

In contrast to MS data, HS data possesses a greater amount of information, typically comprising hundreds of spectral bands, albeit with a further decrease in spatial resolution. In order to cater to a wider range of interpretation scenarios, a substantial number of pan-sharpening techniques are naturally applied to HS sharpening tasks. To verify the universality of the proposed CF2N, we conduct additional experiments on two HS datasets, the Pavia Center and Botswana datasets (Zhuo et al. 2022). These two datasets are available at https://github.com/liangjiandeng/HyperPanCollection. Six advanced HS sharpening methods were selected for comparative experiments, comprising EXP (Aiazzi et al. 2002), CNMF (Yokoya, Yairi, and Iwasaki 2012), GFPCA (Liao et al. 2015), PSRT (Deng et al. 2023), DBDENet (Qu et al. 2022) and HyperDSNet (Zhuo et al. 2022). Among these, CNMF and GFPCA are traditional methods; while the PSRT is based on the SR, DBDENet is based on dual-branch, and HyperDSNet is based on the detail injection model. Additionally, we select five matrics commonly employed in HS sharpening tasks for quantitative evaluation, including SSIM (Yuhas, Goetz, and

Table 10. Ablation experiment on FSCFM.

	Quantitative Results					Efficiency Analysis			
Method	ERGAS↓	SAM↓	sCC↑	RASE↓	Q2n↑	NoPs (M)	FLOPs (G)	MACs (G)	Time (s)
w/o FDA	2.2348 ± 0.5121	2.9878 ± 0.5436	0.9886 ± 0.0049	7.0742 ± 1.9260	0.9138 ± 0.0808	0.25	1.50	0.75	0.020
w/o SDA	2.2878 ± 0.5265	3.0216 ± 0.5588	0.9883 ± 0.0049	7.1599 ± 2.0074	0.9120 ± 0.0809	**0.24**	**1.35**	**0.68**	**0.018**
w/o FSM	2.1848 ± 0.4682	2.9601 ± 0.5341	0.9886 ± 0.0050	6.8881 ± 1.8553	0.9137 ± 0.0817	0.24	1.40	0.70	0.021
FSCFM	**2.1591** ± 0.4628	**2.9341** ± 0.5288	**0.9892** ± 0.0049	**6.8298** ± 1.8060	**0.9145** ± 0.0805	0.25	1.50	0.75	0.024

Table 11. Ablation experiment about the number of FSCFMs and MDIs.

	Quantitative Results					Efficiency Analysis			
Method	ERGAS↓	SAM↓	sCC↑	RASE↓	Q2n↑	NoPs (M)	FLOPs (G)	MACs (G)	Time (s)
$n = 0, m = 0$	2.3880 ± 0.5541	3.1528 ± 0.5755	0.9861 ± 0.0055	7.5480 ± 2.1512	0.9052 ± 0.0851	**0.14**	**0.59**	**0.29**	**0.005**
n=1, m=0	2.3394 ± 0.4900	3.1452 ± 0.5673	0.9868 ± 0.0055	7.2730 ± 1.8995	0.9022 ± 0.0872	0.18	0.94	0.47	0.008
$n = 2, m$=0	2.2462 ± 0.4915	3.0313 ± 0.5509	0.9882 ± 0.0051	7.0641 ± 1.8840	0.9097 ± 0.0832	0.23	1.31	0.65	0.011
$n = 2, m$=1	**2.1591** ± 0.4628	**2.9341** ± 0.5288	0.9892 ± 0.0049	**6.8298** ± 1.8060	0.9145 ± 0.0805	0.25	1.50	0.75	0.024
$n = 3, m = 0$	2.2192 ± 0.4760	3.0279 ± 0.5596	0.9887 ± 0.0048	6.9544 ± 1.8246	0.9114 ± 0.0819	0.28	1.67	0.84	0.014
n=3, $m = 2$	2.2112 ± 0.5045	2.9350 ± 0.5323	0.9893 ± 0.0048	6.9594 ± 1.9002	0.9141 ± 0.0807	0.34	2.22	1.11	0.056
$n = 4, m = 0$	2.2025 ± 0.4914	2.9749 ± 0.5367	0.9887 ± 0.0049	6.9463 ± 1.8686	0.9122 ± 0.0817	0.32	2.04	1.02	0.019
$n = 4, m = 3$	2.1963 ± 0.4823	2.9515 ± 0.5596	**0.9895** ± 0.0046	6.8786 ± 1.8579	**0.9147** ± 0.0803	0.43	3.00	1.50	0.088

Boardman 1992), PSNR, SAM (Wang et al. 2004), CC, and ERGAS (Vivone et al. 2015).

(1) Results on Pavia Center dataset (102-Band): The images in this dataset have been captured by the Reflective Optical System Imaging Sensor, which provides HS with 102 bands within the spectral range of 0.4-0.9 μm. As shown in Figure 12(a), the dataset comprises sharpening environments for urban buildings and streets with a spatial resolution of 1.3 m. The traditional methods CNMF and GFPCA exhibit evident spatial and spectral distortions, particularly GFPCA. The SR-based method PSRT and the dual-branch-based method DBDENet exhibit slight spectral distortions, such as the distortion of the black building in the enlarged area into black gray, which can be attributed to inaccurate detail representation during the feature extraction stage or inadequate feature interaction during the feature fusion stage. The detail injection-based method HyperDSNet shows slight spatial distortion, which is manifested in the pixels that do not exist in GT in the edge area of black buildings. The reason for this is that it employs five distinct HF operators to extract the details in PAN, but ignores the complementary details in HS data. The proposed CF2N, on the other hand, exhibits the highest subjective performance, owing to the fact that it fully considers the uniqueness and complementarity of diverse source data, as well as the necessity of interaction between diverse domain features. In addition, the spectral vectors for a spatial location in the test sample are depicted in the left of Figure 13. It can be observed that the spectral vectors of CF2N are closest to GT, further proving that our method possesses the most superior spectral preservation capability. In Table 12, the proposed CF2N achieve the best score on all metrics. Specifically, the scores obtained on SSIM and SAM are at least 0.40% and 4.46% better than any other methods, respectively, which indicates that CF2N does the best in spectral fidelity. The scores obtained on PSNR, CC, and ERGAS are at least 0.43%, 0.10%, and 1.90% superior to other methods, respectively, indicating that CF2N exhibits the highest level of spatial fidelity. These are inextricably linked to our methodology's emphasis on the unique and complementary detail representation of diverse source data, as well as the interaction between diverse domain characteristics.

(2) Results on Botswana dataset (145-Band): The images in this dataset have been captured by the Earth Observing-1 (EO-1) Hyperion satellite in Botswana, which provides HSI with 145 bands in the spectral band range of 0.4-2.5 μm. Similarly, the traditional methods still show obvious spatial distortion and spectral distortion. The DBDENet shows slight spectral distortion, such as the loss of some crucial spectral information in the lower left corner of the amplification region. HyperDSNet exhibits slight spectral distortion, such as the distortion of the middle portion of the enlarged region

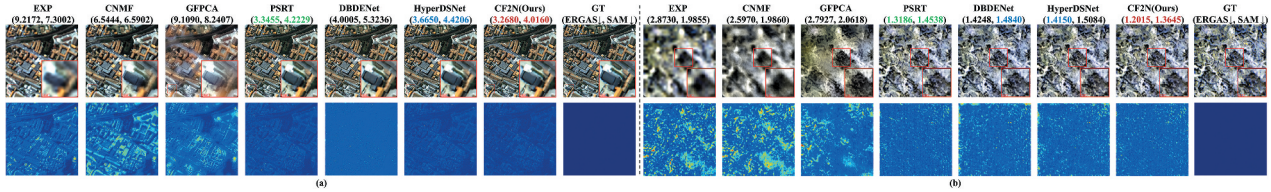

Figure 12. Qualitative evaluation results on two HS dataset. (a) Pavia Center dataset. (b) Botswana dataset.

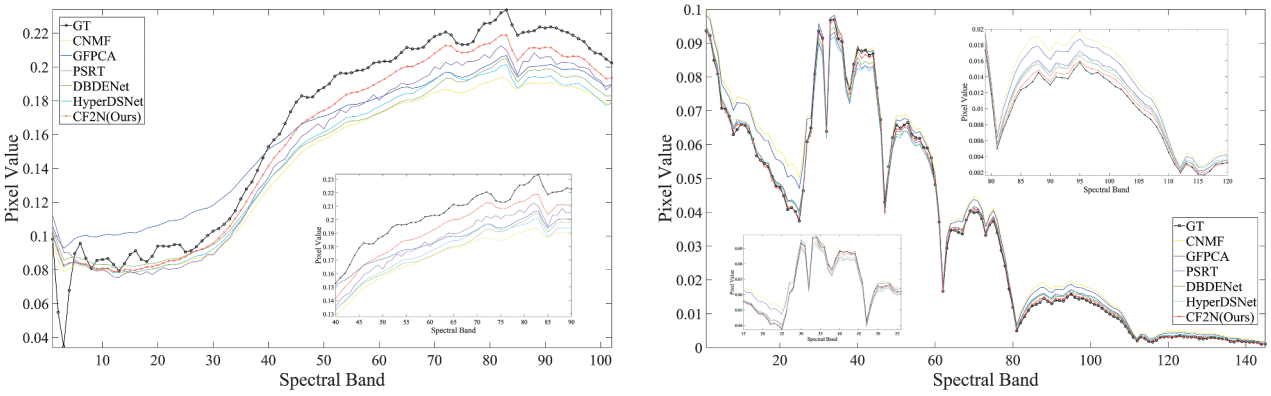

Figure 13. The comparisons of spectral vectors from different spatial locations in Figure 12.

Table 12. Quantitative comparison on the Pavia Center and Botswana dataset.

	Pavia Center dataset (102-Band)					Botswana dataset (145-Band)				
Method	SSIM↑	PSNR↑	SAM↓	CC↑	ERGAS↓	SSIM↑	PSNR↑	SAM↓	CC↑	ERGAS↓
EXP	0.5996 ± 0.0054	25.9337 ± 0.0487	7.1899 ± 0.1103	0.7848 ± 0.0030	9.1698 ± 0.0474	0.6753 ± 0.0216	31.3785 ± 0.7912	2.0029 ± 0.2670	0.7135 ± 0.0414	3.0295 ± 0.4455
CNMF	0.7819 ± 0.0149	29.4520 ± 0.3558	6.6936 ± 0.1034	0.8949 ± 0.0029	6.3537 ± 0.1907	0.7339 ± 0.0253	31.9540 ± 1.0506	2.0324 ± 0.2435	0.7505 ± 0.0490	2.8813 ± 0.4748
GFPCA	0.6161 ± 0.0002	26.0518 ± 0.0206	8.3355 ± 0.0949	0.8040 ± 0.0022	9.0318 ± 0.0772	0.7673 ± 0.0231	31.9927 ± 0.8340	2.0289 ± 0.2472	0.7979 ± 0.0516	2.8646 ± 0.4096
PSRT	**0.9281** ± 0.0012	**35.1552** ± 0.3082	**4.3610** ± 0.1381	**0.9684** ± 0.0018	**3.4371** ± 0.0915	**0.9521** ± 0.0066	**37.5537** ± 2.6206	**1.5498** ± 0.1986	**0.9023** ± 0.0621	**1.8140** ± 0.5508
DBDENet	0.9098 ± 0.0020	33.8079 ± 0.1241	5.4532 ± 0.1296	0.9591 ± 0.0014	4.0457 ± 0.0453	0.9438 ± 0.0084	36.8447 ± 2.4657	**1.6256** ± 0.2346	0.8981 ± 0.0554	1.8693 ± 0.5218
HyperDSNet	**0.9145** ± 0.0013	**34.4274** ± 0.2144	**4.5454** ± 0.1248	**0.9632** ± 0.0017	**3.7359** ± 0.0708	**0.9464** ± 0.0077	**37.0110** ± 2.4849	1.6464 ± 0.2302	**0.9004** ± 0.0558	**1.8512** ± 0.5023
Ours	**0.9318** ± 0.0010 (+0.40/1.89%)	**35.3055** ± 0.3616 (+0.43/2.55%)	**4.1665** ± 0.1505 (−4.46/8.34%)	**0.9694** ± 0.0020 (+0.10/0.64%)	**3.3718** ± 0.1038 (−1.90/9.75%)	**0.9562** ± 0.0064 (+0.43/1.04%)	**38.3753** ± 2.6736 (+2.19/3.69%)	**1.4577** ± 0.1872 (−5.94/10.33%)	**0.9150** ± 0.0537 (+1.41/1.62%)	**1.6760** ± 0.5306 (−7.61/9.46%)

into a light-yellow hue. This can be attributed to the inaccurate detail reconstruction and inadequate interaction of these methods, resulting in the loss of crucial information. PSRT and the proposed CF2N show considerable sharpening performance. In contrast, CF2N has texture details and spectral distribution closer to GT. In addition, we show the spectral vectors for one spatial location in the test sample shown in Figure 13. It can be observed that the spectral vectors of CF2N are closest to GT, further indicating that our method possesses the most superior spectral preservation capability. In the corresponding quantitative assessment, the scores obtained by CF2N on SSIM and SAM are at least 0.43% and 5.94% better than other methods, respectively. This indicates that CF2N exhibits the highest level of spectral fidelity. The scores obtained on PSNR, CC, and ERGAS are at least 2.19%, 1.41%, and 7.61% superior to other methods, indicating that CF2N exhibits the highest level of spatial fidelity.

In general, the proposed CF2N is capable of adapting to these environments and exhibiting superior sharpening performance in the HS pan-sharpening task, regardless of whether it is a high-fine-grained urban environment or a coarse-grained swamp environment. Benefiting from the uniqueness and complementarity of diverse source data and the necessity of interaction between diverse domain features.

4.7. Extended experiments on heterogeneous fusion

The proposed CF2N achieves superior fusion performance when applied to homogeneous fusion tasks, such as MS image pan-sharpening and HS image pan-sharpening. We conduct extended experiments in a heterogeneous fusion task, specifically the SAR-optical fusion (Hong et al. 2024; Yu et al. 2023), to further validate the generalization capability and scalability of CF2N. Specifically, we conduct SAR-Optical fusion cloud removal experiments on the global dataset SEN2MS-CR (Ebel et al. 2020), which contains 122,218 pairs of SAR data from Sentinel-1 and MS cloudy and cloud-free data from Sentinel-2 with 13 bands. The height and width of each patch are 256. Six non-overlapping Regions of Interest in winter are randomly chosen for model training, validation, and testing. The number of patches in the training set, validation set, and test set is 3166, 718, and 780, respectively. This dataset is available at https://mediatum.ub.tum.de/1554803. In addition, three advanced cloud removal methods were used for comparative analysis, including U-Net3D (Rose et al. 2019), DSen2-CR (Meraner et al. 2020) and GLF-CR (Xu et al. 2022).

Figure 14 depicts the fusion scenarios under two different cloud coverages. When the cloud is thin, the proposed CF2N shows a considerable cloud removal effect. However, the fusion result exhibits obvious distortion when the cloud is thick and covers a large area. It is apparent that the proposed model retains the edge texture of the cloud part to a certain extent, owing to its excessive focus on reconstructing the details in diverse source images. Furthermore, it is evident from Table 13 that CF2N exhibits superior performance in PSNR, MAE and RMSE, thereby enabling it to achieve higher reconstruction quality. However, CF2N fails to excel in SSIM and SAM, further indicating that the spatial structure and spectral correlation of the fused image are affected by the interference of clouds. In addition, the proposed CF2N clearly outperforms both DSen2-CR and GLF-CR in terms of effectiveness. In general, our methodology demonstrates the potential to be extended to other multimodal fusion tasks.

5. Discussion

Given the importance of fine detail representation and the interaction between diverse domains for the pan-sharpening task, we have developed a novel cross fusion model. Guided by this theoretical model, we propose an efficient network known as CF2N. CF2N possesses both the interpretability of traditional methods and the adaptability of DL-based methods. In both R-Test and F-Test, we evaluate the

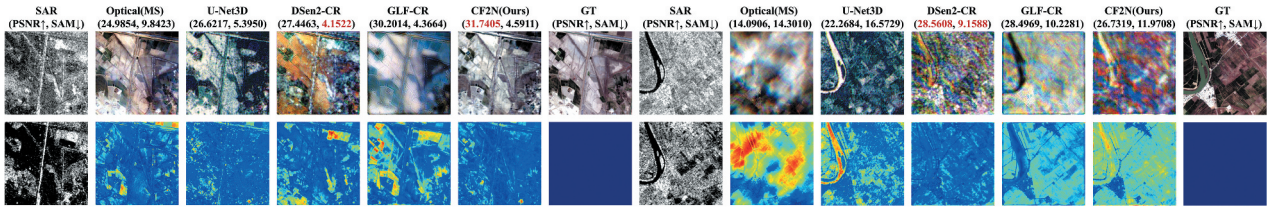

Figure 14. Qualitative evaluation results on heterogeneous sar-optical fusion.

Table 13. Quantitative comparison on heterogeneous sar-optical fusion.

	Quantitative Results					Efficiency Analysis*			
Method	SSIM↑	PSNR↑	MAE↓	SAM↓	RMSE↓	NoPs(M)	FLOPs(G)	MACs(G)	Time(s)
U-Net3D	0.7683 ± 0.0349	24.6062 ± 1.5872	0.0428 ± 0.0071	12.2276 ± 2.7467	0.0598 ± 0.0109	0.95	**24.95**	**12.47**	0.036
DSen2-CR	0.8655 ± 0.0477	28.1201 ± 2.3621	0.0315 ± 0.0106	**6.6350** ± 1.9170	0.0408 ± 0.0122	18.95	2482.36	1241.18	**0.029**
GLF-CR	**0.8759** ± 0.0397	27.1886 ± 2.2097	0.0341 ± 0.0107	7.5923 ± 2.4394	0.0453 ± 0.0136	10.06	334.07	167.04	0.132
Ours	0.8561 ± 0.0433	**29.0263** ± 1.7701	**0.0276** ± 0.0061	6.8566 ± 2.0214	**0.0361** ± 0.0074	**0.63**	58.09	29.05	0.037

The spatial scale of the image in the efficiency test is 256×256.

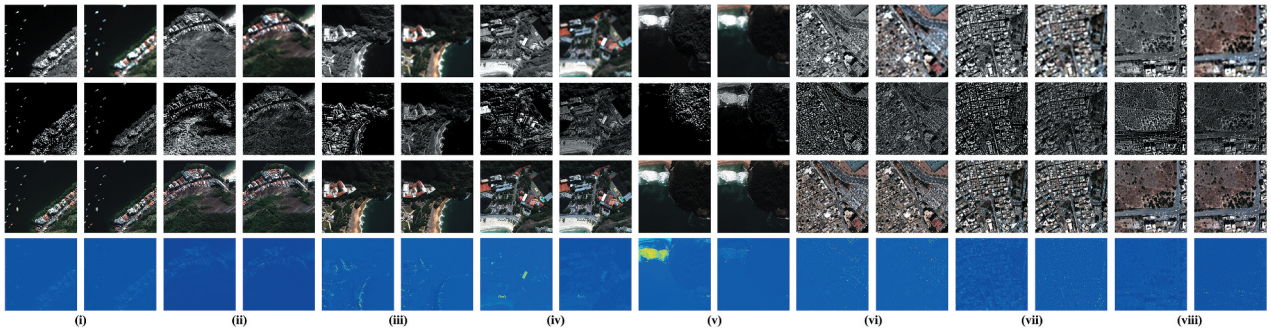

Figure 15. Eight examples of detail representation in different methods. The first line shows the original image pairs with the left side being PAN and the right side being MS; the second, third and fourth lines show the detail representations, the fusion results and the corresponding AEMs, respectively, with the left side being FusionNet and the right side being the proposed CF2N.

sharpening performance of each method under varying experimental conditions, including image quality, sensor type, and sharpening scene. Experiments have demonstrated that with the enhancement of image quality and the increasing complexity of sharpening scenes, the majority of methods exhibit a discernible decline in performance. In contrast, our method shows strong robustness and generalization ability in different scenarios, owing to its emphasis on unique and complementary detail representations in diverse source images and the interaction between diverse source features. Furthermore, experiments on two HS datasets further validate that the proposed method is capable of achieving superior sharpening performance in both coarse-grained and fine-grained environments. As depicted in Figure 15, the visualization of the detail representations in different models serves to more intuitively validate the necessity of detail reconstruction. The detail representations that require extraction for the subsequent networks constructed by FusionNet do not exhibit the spatial details in diverse source images. This is due to the fact that they solely consider the spatial information unique to the PAN, omitting the unique spatial information present in MS and the complementary information present in diverse source images. In contrast, the reconstructed detail representation of our method is richer than the details in any of the original image pairs, due to the fact that we consider both the unique and complementary information in diverse source images. As a result, our methodology yields more reasonable frequency details, thereby providing a solid foundation for the generation of HRMS. However, as shown in Figure 9 of the experimental findings on the WV3 dataset, it is evident that excessive attention to spatial details in both PAN and MS leads to spectral distortion in certain local regions of the fused image, particularly when the sharpening target is a vehicle. In addition, the efficiency analysis indicates that the proposed method still has room for improvement in terms of timeliness.

The proposed method also shows excellent generalization ability and scalability in SAR-optical fusion task, owing to the effective exploration of unique and complementary details in diverse source data. However, in instances where the cloud in the optical image is thick and the coverage area is large, the fusion outcome of the proposed method exhibits evident spectral distortion. The reason for this is that the cloud interference causes the model to focus on the edge texture of the cloud. Furthermore, the introduction of task-specific prior knowledge is absent, such as the extraction and utilization of cloud mask information, as well as the consideration of the impact of speckle noise in SAR and its polarization characteristics. In general, our methodology is capable of preserving finer details, including frequency details and spectral details, thereby achieving high frequency-spectral fidelity in homogeneous pan-sharpening tasks. Simultaneously, it demonstrates that it can be extended to a heterogeneous fusion task.

6. Conclusions

With regards to the issues of inaccurate detail representation, inadequate feature interaction, lack of interpretability, high model complexity, and high computational complexity in the pan-sharpening task, we propose a novel cross fusion model with fine-grained detail reconstruction and develop the CF2N guided by this model. The CF2N consists of two main stages: FD2R and FSCF. Specifically, in order to achieve a more reasonable reconstruction of the fine spatial information, we design the FD2R by analyzing the diverse source images to HRMS detail-to-detail relationship in the frequency-domain. The reconstructed details can provide the foundation for the

generation of fine details in the subsequent fused images. The FSCF stage, integrates the localization between frequency details and corresponding spectral details through a highly cross fusion manner, taking into account the interaction of diverse domain details in the fusion process. The CF2N can balance the sharpening performance and efficiency through the analysis and ablation experiments of each module. High frequency-spectral fidelity fusion results can be achieved on five datasets in two sharpening tasks, demonstrating the adaptability and robustness of the proposed method under diverse experimental conditions. In R-Test, our method achieves ERGAS scores that are at least 8.02%, 5.74%, and 3.76% superior to other advanced methods on GF2, QB, and WV3, respectively. In F-Test, the QNR scores obtained by our method on GF2, QB, and WV3 are at least 0.12%, 0.45%, and 1.55% higher than other methods, respectively. These outcomes indicate that the proposed method possesses robust fitting capability at reduced-scale and generalization capability in real-world sharpening scenarios. In the HS sharpening experiment, the ERGAS scores obtained by our method on the Pavia Center with a resolution of 1.3 m and Botswana datasets with a resolution of 30 m are at least 1.90% and 7.61% superior to other methods, respectively. This further confirms that our method is suitable for both coarse- and fine-grained scenes. Furthermore, the performance of the proposed method on heterogeneous SAR-Optical fusion proves its potential for other RS image fusion tasks. In the forthcoming research, we will further explore the theory model with a particular emphasis on enhancing the balance between spatial and spectral preservation and the issue of low timeliness will be avoided. Furthermore, we will endeavor to introduce task-specific prior knowledge on the basis of the existing model, thereby enabling its application to a wider range of multimodal fusion scenarios.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Science Fund for Distinguished Young Scholars [grant number 62425102], National Key R&D Program of China [grant number 2022YFB3902800], and National Natural Science Foundation of China [grant number 62301214].

Notes on contributors

Chuang Liu is currently working toward the M.Eng. degree in computer technology with the School of Computer Science, Hubei University of Technology, Wuhan, China. His research interests include remote sensing image processing, multimodal image fusion, low-level vision, machine learning, and deep learning.

Zhiqi Zhang received the B.Sc. degree in geographic information systems from Huazhong Agricultural University, Wuhan, China, in 2006, the B.Eng. degree in computer science and technology from the Huazhong University of Science and Technology, Wuhan, in 2006, and the M.Eng. degree in computer technology and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, in 2015 and 2018, respectively. He is currently an Associate Professor with the School of Computer Science, Hubei University of Technology, Wuhan. His research interests include system architecture, algorithm optimization, AI, and high-performance processing of remote sensing.

Mi Wang received the B.Sc., M.Sc., and Ph.D. degrees in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in1997, 1999, and 2001, respectively. Since 2008, he has been a Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University. His research interests include measurable seamless stereo orthoimage databases, geographic information systems (GIS), and the integration of global navigation satellite systems, remote sensing, and GIS.

Shao Xiang received the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Surveying, Mapping and Remote Sensing Information Engineering, Wuhan University, Wuhan, China, in 2024. His research interests include artificial intelligence, pattern recognition, and remote sensing image processing.

Guangqi Xie received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, in 2022. He is currently a lecturer with the School of Computer Science, Hubei University of Technology, Wuhan, China. His research interest includes image matching and registration, panchromatic sharpening, and image super-resolution.

ORCID

Chuang Liu http://orcid.org/0009-0001-8246-3417
Zhiqi Zhang http://orcid.org/0000-0003-1914-9430
Mi Wang http://orcid.org/0000-0003-2799-5987
Shao Xiang http://orcid.org/0009-0005-1426-7931
Guangqi Xie http://orcid.org/0000-0002-0292-1626

Data availability statement

The MS data supporting the findings of this study are available in PanCollection at https://doi.org/10.1109/mgrs.2022.3187652. The HS data supporting the findings of this study are available in HyperPanCollection at https://doi.org/10.

1109/jstars.2022.3202866. The SAR data supporting the findings of this study are available in SEN2MS-CR at https://doi.org/10.1109/TGRS.2020.3024744. All implementations of this paper will be published at https://github.com/JUSTM0VE0N/CF2N.

References

Aiazzi, B., L. Alparone, S. Baronti, R. Carlà, A. Garzelli, and L. Santurri. 2014. "Full Scale Assessment of Pansharpening Methods and Data Products." *Proceedings of SPIE*, Amsterdam, Netherlands, October. https://doi.org/10.1117/12.2067770.

Aiazzi, B., L. Alparone, S. Baronti, and A. Garzelli. 2002. "Context-Driven Fusion of High Spatial and Spectral Resolution Images Based on Oversampled Multiresolution Analysis." *IEEE Transactions on Geoscience & Remote Sensing* 40 (10): 2300–2312. https://doi.org/10.1109/tgrs.2002.803623.

Aiazzi, B., S. Baronti, and M. Selva. 2007. "Improving Component Substitution Pansharpening Through Multivariate Regression of MS + Pan Data." *IEEE Transactions on Geoscience & Remote Sensing* 45 (10): 3230–3239. https://doi.org/10.1109/tgrs.2007.901007.

Alparone, L., B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva. 2008. "Multispectral and Panchromatic Data Fusion Assessment without Reference." *Photogrammetric Engineering & Remote Sensing* 74 (2): 193–200. https://doi.org/10.14358/pers.74.2.193.

Anno, S., H. Tsubasa, S. Sugita, S. Yasumoto, M. Lee, Y. Sasaki, and K. Oyoshi. 2023. "Challenges and Implications of Predicting the Spatiotemporal Distribution of Dengue Fever Outbreak in Chinese Taiwan Using Remote Sensing Data and Deep Learning." *Geo-Spatial Information Science* 27 (4): 1155–1161. https://doi.org/10.1080/10095020.2022.2144770.

Bouasria, A., K. I. Namr, A. Rahimi, E. M. Ettachfini, and B. Rerhou. 2022. "Evaluation of Landsat 8 Image Pansharpening in Estimating Soil Organic Matter Using Multiple Linear Regression and Artificial Neural Networks." *Geo-Spatial Information Science* 25 (3): 353–364. https://doi.org/10.1080/10095020.2022.2026743.

Cai, J., and B. Huang. 2021. "Super-Resolution-Guided Progressive Pansharpening Based on a Deep Convolutional Neural Network." *IEEE Transactions on Geoscience & Remote Sensing* 59 (6): 5206–5220. https://doi.org/10.1109/tgrs.2020.3015878.

Chavez, P. S., and A. Y. Kwarteng. 1989. "Extracting Spectral Contrast in Landsat Thematic Mapper Image Data Using Selective Principal Component Analysis." *Photogrammetric Engineering & Remote Sensing* 55 (3): 339–348.

Chen, Y. X., H. Q. Liu, and F. M. Fang. 2024. "A Novel Pansharpening Method Based on Cross Stage Partial Network and Transformer." *Scientific Reports* 14 (1). https://doi.org/10.1038/s41598-024-63336-w.

Cheng, X. H., and Z. L. Li. 2023. "Modeling Information Flow from Multispectral Remote Sensing Images to Land Use and Land Cover Maps for Understanding Classification Mechanism." *Geo-Spatial Information Science*: 1–17. November. https://doi.org/10.1080/10095020.2023.2275625.

Choi, J., K. Yu, and Y. Kim. 2011. "A New Adaptive Component-Substitution-Based Satellite Image Fusion by Using Partial Replacement." *IEEE Transactions on Geoscience & Remote Sensing* 49 (1): 295–309. https://doi.org/10.1109/tgrs.2010.2051674.

Deng, L. J., G. Vivone, C. Jin, and J. Chanussot. 2021. "Detail Injection-Based Deep Convolutional Neural Networks for Pansharpening." *IEEE Transactions on Geoscience & Remote Sensing* 59 (8): 6995–7010. https://doi.org/10.1109/tgrs.2020.3031366.

Deng, L. J., G. Vivone, M. Paoletti, G. Scarpa, J. He, Y. Zhang, J. Chanussot, and A. J. Plaza. 2022. "Machine Learning in Pansharpening: A Benchmark, from Shallow to Deep Networks." *IEEE Geoscience and Remote Sensing Magazine* 10 (3): 279–315. https://doi.org/10.1109/mgrs.2022.3187652.

Deng, S. Q., L. J. Deng, X. Wu, R. Ran, D. F. Hong, and G. Vivone. 2023. "PSRT: Pyramid Shuffle-And-Reshuffle Transformer for Multispectral and Hyperspectral Image Fusion." *IEEE Transactions on Geoscience & Remote Sensing* 61:1–15. https://doi.org/10.1109/tgrs.2023.3244750.

Diao, W. X., F. Zhang, H. T. Wang, W. B. Wan, J. D. Sun, and K. Zhang. 2022. "HLF-Net: Pansharpening Based on High- and Low-Frequency Fusion Networks." *IEEE Geoscience & Remote Sensing Letters* 19:1–5. https://doi.org/10.1109/lgrs.2022.3225974.

Ebel, P., A. Meraner, M. Schmitt, and X. X. Zhu. 2020. "Multisensor Data Fusion for Cloud Removal in Global and All-Season Sentinel-2 Imagery." *IEEE Transactions on Geoscience and Remote Sensing*, 1–13. https://doi.org/10.1109/TGRS.2020.3024744.

Garzelli, A., and F. Nencini. 2009. "Hypercomplex Quality Assessment of Multi/Hyperspectral Images." *IEEE Geoscience & Remote Sensing Letters* 6 (4): 662–665. https://doi.org/10.1109/lgrs.2009.2022650.

Ghahremani, M., Y. H. Liu, P. S. T. Yuen, and A. Behera. 2019. "Remote Sensing Image Fusion via Compressive Sensing." *Isprs Journal of Photogrammetry & Remote Sensing* 152:34–48. https://doi.org/10.1016/j.isprsjprs.2019.04.001.

Hong, Y., T. J. Xie, L. K. Luo, M. Wang, D. R. Li, Q. Zhang, and T. Xu. 2024. "Area Extraction and Growth Monitoring of Sugarcane from Multi-Source Remote Sensing Images Under a Polarimetric SAR Data Compensation Based on Buildings." *Geo-Spatial Information Science*: 1–18. July. https://doi.org/10.1080/10095020.2024.2381607.

Jian, L. H., S. W. Wu, L. H. Chen, G. Vivone, R. Rayhana, and D. Zhang. 2023. "Multi-Scale and Multi-Stream Fusion Network for Pansharpening." *Remote Sensing* 15 (6): 1666. https://doi.org/10.3390/rs15061666.

Jin, Z. R., T. J. Zhang, T. X. Jiang, G. Vivone, and L. J. Deng. 2022. "LAGConv: Local-Context Adaptive Convolution Kernels with Global Harmonic Bias for Pansharpening." *Proceedings of the AAAI Conference on Artificial Intelligence*, Vancouver, Canada. 36 (1): 1113–1121. https://doi.org/10.1609/aaai.v36i1.19996.

Kaplan, N. H., and İ. Erer. 2014. "Bilateral Filtering-Based Enhanced Pansharpening of Multispectral Satellite Images." *IEEE Geoscience & Remote Sensing Letters* 11 (11): 1941–1945. https://doi.org/10.1109/lgrs.2014.2314389.

Kaplan, N. H., I. Erer, O. Ozcan, and N. Musaoglu. 2019. "MTF Driven Adaptive Multiscale Bilateral Filtering for Pansharpening." *International Journal of Remote Sensing* 40 (16): 6262–6282. https://doi.org/10.1080/01431161.2019.1590874.

Ke, C. J., W. Zhang, Z. Y. Wang, J. Y. Ma, and X. Tian. 2023. "Coarse-To-Fine Cross-Domain Learning Fusion

Network for Pansharpening." *IEEE Transactions on Geoscience & Remote Sensing* 61:1–14. https://doi.org/10.1109/tgrs.2023.3299336.

Li, X., Y. H. Xu, L. Ma, Z. Yang, Z. H. Huang, H. Y. Hong, and J. W. Tian. 2022. "Multi-Source Weakly Supervised Salient Object Detection via Boosting Weak-Annotation Source and Constraining Object Structure." *Digital Signal Processing* 126:103461. https://doi.org/10.1016/j.dsp.2022.103461.

Li, Y. J., S. P. Chen, K. Hwang, X. Q. Ji, Z. Lei, Y. Zhu, F. Ye, and M. J. Liu. 2024. "Spatio-Temporal Data Fusion Techniques for Modeling Digital Twin City." *Geo-Spatial Information Science*: 1–24. May. https://doi.org/10.1080/10095020.2024.2350175.

Liao, W. Z., X. Huang, F. V. Coillie, S. Gautama, A. Pizurica, W. Philips, H. Liu, et al. 2015. "Processing of Multiresolution Thermal Hyperspectral and Digital Color Data: Outcome of the 2014 IEEE GRSS Data Fusion Contest." *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing* 8 (6): 2984–2996. https://doi.org/10.1109/jstars.2015.2420582.

Liu, C., L. Wei, Z. Zhang, X. Feng, and S. Xiang. 2023a. "Recursive Self-Attention Modules Based Network for Panchromatic and Multispectral Image Fusion." *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing* 16:10067–10083. https://doi.org/10.1109/jstars.2023.3327167.

Liu, G. Y., R. Li, J. Xia, Z. H. Liu, J. Cai, H. Y. Wu, and M. J. Peng. 2023b. "Dual-Environment Feature Fusion-Based Method for Estimating Building-Scale Population Distributions." *Geo-Spatial Information Science*: 1–16. December. https://doi.org/10.1080/10095020.2023.2281571.

Lolli, S., L. Alparone, A. Garzelli, and G. Vivone. 2017. "Haze Correction for Contrast-Based Multispectral Pansharpening." *IEEE Geoscience & Remote Sensing Letters* 14 (12): 2255–2259. https://doi.org/10.1109/lgrs.2017.2761021.

Lu, H., Y. Yang, S. Huang, X. Chen, B. Chi, A. Liu, and W. Tu. 2023. "AWFLN: An Adaptive Weighted Feature Learning Network for Pansharpening." *IEEE Transactions on Geoscience & Remote Sensing* 61:1–15. https://doi.org/10.1109/tgrs.2023.3241643.

Lu, H., Y. Yang, S. Huang, W. Tu, and W. G. Wan. 2022. "A Unified Pansharpening Model Based on Band-Adaptive Gradient and Detail Correction." *IEEE Transactions on Image Processing* 31:918–933. https://doi.org/10.1109/tip.2021.3137020.

Mai, T. T. N., E. Y. Lam, and C. Lee. 2022. "Deep Unrolled Low-Rank Tensor Completion for High Dynamic Range Imaging." *IEEE Transactions on Image Processing* 31:5774–5787. https://doi.org/10.1109/tip.2022.3201708.

Mallat, S. G. 1989. "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation." *IEEE Transactions on Pattern Analysis & Machine Intelligence* 11 (7): 674–693. https://doi.org/10.1109/34.192463.

Masi, G., D. Cozzolino, L. Verdoliva, and G. Scarpa. 2016. "Pansharpening by Convolutional Neural Networks." *Remote Sensing* 8 (7): 594. https://doi.org/10.3390/rs8070594.

Masoumi, Z., and J. V. Genderen. 2023. "Artificial Intelligence for Sustainable Development of Smart Cities and Urban Land-Use Management." *Geo-Spatial Information Science* 27 (4): 1212–1236. https://doi.org/10.1080/10095020.2023.2184729.

Meraner, A., P. Ebel, X. X. Zhu, and M. Schmitt. 2020. "Cloud Removal in Sentinel-2 Imagery Using a Deep Residual Neural Network and SAR-Optical Data Fusion." *Isprs Journal of Photogrammetry & Remote Sensing* 166:333–346. https://doi.org/10.1016/j.isprsjprs.2020.05.013.

Misra, I., M. K. Rohil, S. M. Moorthi, and D. Dhar. 2023a. "CLIM: Co-Occurrence with Laplacian Intensity Modulation and Enhanced Color Space Transform for Infrared-Visible Image Fusion." *Infrared Physics & Technology* 135:104951–104951. https://doi.org/10.1016/j.infrared.2023.104951.

Misra, I., M. K. Rohil, S. M. Moorthi, and D. Dhar. 2023b. "SPRINT: Spectra Preserving Radiance Image Fusion Technique Using Holistic Deep Edge Spatial Attention and Minnaert Guided Bayesian Probabilistic Model." *Signal Processing: Image Communication* 113:116920–116920. https://doi.org/10.1016/j.image.2023.116920.

Otazu, X., M. Gonzalez-Audicana, O. Fors, and J. Nunez. 2005. "Introduction of Sensor Spectral Response into Image Fusion Methods. Application to Wavelet-Based Methods." *IEEE Transactions on Geoscience & Remote Sensing* 43 (10): 2376–2385. https://doi.org/10.1109/tgrs.2005.856106.

Qu, J., S. Hou, W. Dong, S. Xiao, Q. Du, and Y. Li. 2022. "A Dual-Branch Detail Extraction Network for Hyperspectral Pansharpening." *IEEE Transactions on Geoscience & Remote Sensing* 60:1–13. https://doi.org/10.1109/tgrs.2021.3130420.

Restaino, R., G. Vivone, M. D. Mura, and J. Chanussot. 2016. "Fusion of Multispectral and Panchromatic Images Based on Morphological Operators." *IEEE Transactions on Image Processing* 25 (6): 2882–2895. https://doi.org/10.1109/tip.2016.2556944.

Rose, R., R. Cheong, L. Wang, S. Ermon, M. Burke, and D. Lobell. 2019. "Semantic Segmentation of Crop Type in Africa: A Novel Dataset and Analysis of Deep Learning Methods." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 75–82.

Tariq, A., J. Yan, A. S. Gagnon, M. R. Khan, and F. Mumtaz. 2022. "Mapping of Cropland, Cropping Patterns and Crop Types by Combining Optical Remote Sensing Images with Decision Tree Classifier and Random Forest." *Geo-Spatial Information Science* 26 (3): 302–320. https://doi.org/10.1080/10095020.2022.2100287.

Truong, T. N. M., E. Y. Lam, and C. Lee. 2024. "Deep Unfolding Tensor Rank Minimization with Generalized Detail Injection for Pansharpening." *IEEE Transactions on Geoscience & Remote Sensing* 62:1–18. https://doi.org/10.1109/tgrs.2024.3392215.

Tu, T. M., P. S. Huang, C. L. Hung, and C. P. Chang. 2004. "A Fast Intensity–Hue–Saturation Fusion Technique with Spectral Adjustment for IKONOS Imagery." *IEEE Geoscience & Remote Sensing Letters* 1 (4): 309–312. https://doi.org/10.1109/lgrs.2004.834804.

Vivone, G. 2019. "Robust Band-Dependent Spatial-Detail Approaches for Panchromatic Sharpening." *IEEE Transactions on Geoscience & Remote Sensing* 57 (9): 6421–6433. https://doi.org/10.1109/tgrs.2019.2906073.

Vivone, G., L. Alparone, J. Chanussot, M. D. Mura, A. Garzelli, G. A. Licciardi, R. Restaino, and L. Wald. 2015. "A Critical Comparison Among Pansharpening Algorithms." *IEEE Transactions on Geoscience & Remote Sensing* 53 (5): 2565–2586. https://doi.org/10.1109/tgrs.2014.2361734.

Vivone, G., M. D. Mura, A. Garzelli, R. Restaino, G. Scarpa, M. O. Ulfarsson, L. Alparone, and J. Chanussot. 2021. "A New Benchmark Based on Recent Advances in

Multispectral Pansharpening: Revisiting Pansharpening with Classical and Emerging Pansharpening Methods." *IEEE Geoscience and Remote Sensing Magazine* 9 (1): 53–81. https://doi.org/10.1109/mgrs.2020.3019315.

Vivone, G., R. Restaino, and J. Chanussot. 2018. "Full Scale Regression-Based Injection Coefficients for Panchromatic Sharpening." *IEEE Transactions on Image Processing* 27 (7): 3418–3431. https://doi.org/10.1109/tip.2018.2819501.

Wang, S., X. Jiang, E. Spyrakos, J. Li, C. McGlinchey, A. Maria Constantinescu, and A. N. Tyler. 2023. "Water Color from Sentinel-2 MSI Data for Monitoring Large Rivers: Yangtze and Danube." *Geo-Spatial Information Science* 27 (3): 854–869. https://doi.org/10.1080/10095020.2023.2258950.

Wang, Z., A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. 2004. "Image Quality Assessment: From Error Visibility to Structural Similarity." *IEEE Transactions on Image Processing* 13 (4): 600–612. https://doi.org/10.1109/tip.2003.819861.

Xing, Y., L. Qu, K. Zhang, Y. Zhang, X. Zhang, and Y. Zhang. 2024. "Complementary Fusion Network Based on Frequency Hybrid Attention for Pansharpening." *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2650–2654. Seoul, Korea, Republic of, https://doi.org/10.1109/ICASSP48485.2024.10446416.

Xu, F., Y. Shi, P. Ebel, L. Yu, G. S. Xia, W. Yang, and X. X. Zhu. 2022. "GLF-CR: SAR-Enhanced Cloud Removal with Global–Local Fusion." *Isprs Journal of Photogrammetry & Remote Sensing* 192:268–278. https://doi.org/10.1016/j.isprsjprs.2022.08.002.

Xu, R., C. Wang, J. G. Zhang, S. Xu, W. Meng, and X. P. Zhang. 2023. "RSSFormer: Foreground Saliency Enhancement for Remote Sensing Land-Cover Segmentation." *IEEE Transactions on Image Processing* 32:1052–1064. https://doi.org/10.1109/tip.2023.3238648.

Xu, S., J. Zhang, Z. Zhao, K. Sun, and J. Liu. 2021. "Deep Gradient Projection Networks for Pan-Sharpening." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 1366–1375. June. https://doi.org/10.1109/cvpr46437.2021.00142.

Yang, J., X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley. 2017a. "PanNet: A Deep Network Architecture for Pan-Sharpening." *2017 IEEE International Conference on Computer Vision (ICCV)*, 1753–1761. Venice, Italy, https://doi.org/10.1109/ICCV.2017.193.

Yang, Y., W. Wan, S. Huang, P. Lin, and Y. Que. 2017b. "A Novel Pan-Sharpening Framework Based on Matting Model and Multiscale Transform." *Remote Sensing* 9 (4): 391. https://doi.org/10.3390/rs9040391.

Yokoya, N., T. Yairi, and A. Iwasaki. 2012. "Coupled Nonnegative Matrix Factorization Unmixing for Hyperspectral and Multispectral Data Fusion." *IEEE Transactions on Geoscience & Remote Sensing* 50 (2): 528–537. https://doi.org/10.1109/tgrs.2011.2161320.

Yu, X., J. Pan, M. Wang, and J. Xu. 2023. "A Curvature-Driven Cloud Removal Method for Remote Sensing Images." *Geo-Spatial Information Science* 27 (4): 1326–1347. https://doi.org/10.1080/10095020.2023.2189462.

Yuhas, R. H., A. F. H. Goetz, and J. W. Boardman. 1992. "Discrimination Among Semi-Arid Landscape Endmembers Using the Spectral Angle Mapper (SAM) Algorithm." *Summaries 3rd Annual JPL Airborne Geoscience Workshop*, Washington, DC, USA, 1:147–149. https://api.semanticscholar.org/CorpusID:126879175.

Zhang, K., Z. Li, F. Zhang, W. Wan, and J. Sun. 2022a. "Pan-Sharpening Based on Transformer with Redundancy Reduction." *IEEE Geoscience & Remote Sensing Letters* 19:1–5. https://doi.org/10.1109/lgrs.2022.3186985.

Zhang, W., B. Hu, G. Brown, and S. Meyer. 2023. "Beaver Pond Identification from Multi-Temporal and Multi-Sourced Remote Sensing Data." *Geo-Spatial Information Science* 27 (4): 953–967. https://doi.org/10.1080/10095020.2023.2183144.

Zhang, Y., C. Liu, M. Sun, and Y. Ou. 2019. "Pan-Sharpening Using an Efficient Bidirectional Pyramid Network." *IEEE Transactions on Geoscience & Remote Sensing* 57 (8): 5549–5563. https://doi.org/10.1109/tgrs.2019.2900419.

Zhang, Z., Z. Qu, S. Liu, D. Li, J. Cao, and G. Xie. 2022b. "Expandable On-Board Real-Time Edge Computing Architecture for Luojia3 Intelligent Remote Sensing Satellite." *Remote Sensing* 14 (15): 3596. https://doi.org/10.3390/rs14153596.

Zhao, X. N., C. Y. Zhao, T. J. Zhang, and L. J. Deng. 2022. "Cross-Frequency Detail Compensation Network for Pansharpening." *Proceedings of the 2022 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Kuala Lumpur, Malaysia, July. https://doi.org/10.1109/igarss46834.2022.9883203.

Zhou, J., D. L. Civco, and J. A. Silander. 1998. "A Wavelet Transform Method to Merge Landsat TM and SPOT Panchromatic Data." *International Journal of Remote Sensing* 19 (4): 743–757. https://doi.org/10.1080/014311698215973.

Zhuo, Y. W., T. J. Zhang, J. F. Hu, H. X. Dou, T. Z. Huang, and L. J. Deng. 2022. "A Deep-Shallow Fusion Network with Multidetail Extractor and Spectral Attention for Hyperspectral Pansharpening." *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing* 15:7539–7555. https://doi.org/10.1109/jstars.2022.3202866.

Zuo, X., Z. Shao, J. Wang, X. Huang, and Y. Wang. 2024. "A Cross-Stage Features Fusion Network for Building Extraction from Remote Sensing Images." *Geo-Spatial Information Science*: 1–15. https://doi.org/10.1080/10095020.2024.2307922.