

Remote sensing image pansharpening based on a nested multi-scale fusion network

Zhao Jing-Chao, Ji Ping, Wang Qun-Ming*

(College of Surveying and Geo-Informatics, Tongji University, Shanghai 200092, China)

Abstract: Pansharpening aims to fuse a coarse spatial resolution multispectral (MS) image with a fine spatial resolution panchromatic (PAN) image to generate a fused image with fine spatial and spectral resolution. Existing deep learning-based pansharpening methods still face difficulties in simultaneously extracting local high-frequency details and preserving global spectral consistency. Consequently, fused images are prone to spectral distortion when spatial details are enhanced. Moreover, the spatial resolution gap between PAN and MS images often causes feature misalignment during the fusion process, resulting in visual artifacts such as local edge ghosting and color distortion. To address these issues, this paper proposes a Nested Multi-Scale Fusion Network (NMSFusion), which performs synergistic modeling at both local and global levels. At the local level, a Multi-Scale Gated Block (MSGB) is utilized to extract high-frequency details from the PAN image. At the global level, an Artifact-Free Residual Fusion Module (ARFM) is designed to coordinate global semantics, ensuring spectral consistency between the fused images and the original MS images. Additionally, an Adaptive Coordinate Encoding Module (ACEM) is introduced to alleviate the feature misalignment caused by the resolution gap. Finally, conventional deep networks with random weight initialization tend to disrupt the original color distribution, leading to convergence difficulties and spectral distortion in the early training stages. To tackle this, we propose a plug-and-play Identity Initialization Mechanism (IIM). By constraining the initial weight of network layers, IIM forces the initial network output to equal the interpolated MS image, thereby providing a reasonable initial state for optimization and promoting stable convergence. Experimental results on three remote sensing datasets demonstrate that NMSFusion outperforms nine representative pansharpening methods in both quantitative metrics and visual assessments. Furthermore, ablation studies demonstrate that the proposed modules not only synergistically enhance the fusion performance but also guarantee an efficient and lightweight network architecture.

Key words: remote sensing image fusion, pansharpening, deep learning, multi-scale

PACS:

Introduction

Pansharpening, also known as spatio-spectral fusion, aims to fuse multispectral (MS) and panchromatic (PAN) images acquired by the same satellite platform to generate an MS image with both fine spatial and spectral resolutions. Early pansharpening methods were mainly based on physically interpretable algorithms. According to the strategies used for spatial-detail injection, these methods can be classified into: component substitution (CS), multiresolution analysis (MRA), variational optimization (VO), and geostatistical methods. The fundamental principle of CS methods is to project the MS image into a transformed feature space, where the spatial component is subsequently replaced by the PAN image. This category encompasses not only early techniques such as the Intensity-Hue-Saturation (IHS) transform^[1], Principal Component Analysis (PCA)^[2], and the Brovey

transform^[3], but also sensor-adaptable methods, including Gram-Schmidt (GS) smoothing^[4], Gram-Schmidt Adaptive (GSA)^[5], and Partial Replacement Adaptive Component Substitution (PRACS)^[6]. MRA methods utilize multi-scale spatial filters to extract high-frequency details from the PAN image, and inject them into the upsampled MS image. Representative algorithms include the Discrete Wavelet Transform (DWT)^[7], High-Pass Filtering (HPF), Smoothing Filter-based Intensity Modulation (SFIM)^[8], the À Trous Wavelet Transform (ATWT)^[9], and the Modulation Transfer Function-Generalized Laplacian Pyramid (MTF-GLP)^[10]. VO methods^[11-13] formulate pansharpening as an ill-posed inverse problem, and solve it via energy functional minimization. Beyond classical P+XS and total variation-based models, recent advancements feature complex optimization frameworks that utilize joint sparse representation

Foundation items: Supported by the Fundamental Research Funds for the Central Universities (22120260026); the Excellent Young Scientists Fund of the National Natural Science Foundation of China (42222108); the General Program of the National Natural Science Foundation of China (42171345)

*Corresponding author: E-mail: wqm11111@126.com

(e. g., J-SparseFI), low-rank tensor decomposition, and Bayesian estimation. Geostatistical methods primarily construct variograms to quantitatively characterize the spatial correlation between different bands (including MS and PAN bands). Representative methods in this category include Area-to-Point Regression Kriging (AT-PRK)^[14] and Downscaling Cokriging (DSCK)^[15].

In recent years, deep learning has been widely applied to pansharpening, shifting the focus from mathematical optimization to end-to-end feature learning. According to architectural evolution and fusion paradigms, these existing methods fall into four major categories. The first category encompasses single-stream models driven by residual learning. As early applications of deep learning to pansharpening, these models focus on directly fitting non-linear mappings within a unified network backbone. A pioneering work in this category is the Pansharpening Neural Network (PNN)^[16], which demonstrated the efficacy of a simple three-layer convolution for the joint spectral-spatial features extraction. To alleviate the optimization difficulties brought by network deepening, the Pansharpening Network (PanNet)^[17] innovatively trained the network in the high-frequency domain, enabling the model to focus on learning residual details. Subsequently, to address the loss of high-frequency details caused by single-step direct upsampling, the Super-Resolution-guided Progressive Pansharpening Neural Network (SRPNN)^[18] incorporated super-resolution concepts and adopted a progressive upsampling strategy to reconstruct images. Furthermore, to overcome the limitations of a single receptive field in representing complex land cover structures, the Multi-Scale Deep Convolutional Neural Network (MSDCNN)^[19] introduced multiscale convolutional filters and dense connections. Building upon this, the Detail Injection Convolutional Neural Network (DiCNN)^[20] mitigates direct mapping distortion by introducing a component-substitution-inspired dual-stream architecture (detail extraction and injection). Although single-stream networks have been continuously refined, the differences in imaging principles between PAN and MS images make a single backbone prone to early confusion of spatial and spectral features during the initial stages of feature extraction. To effectively decouple multi-source heterogeneous information in early network stages, the second category comprises multi-branch architectures driven by cross-modal interaction, which has become a key paradigm in recent studies. The Two-stream Fusion Network (TFNet) method^[21] employed a dual-branch encoder to independently extract features from both modalities, followed by fusion and decoding. However, this simple late fusion strategy overlooks deep cross-modal interactions within the intermediate layers. To address this, the Bidirectional Pyramid Network (BDPN) method^[22] designed a bidirectional pyramid network, achieving deep alignment of the two modalities across different spatial scales. To address the limitations of static feature alignment in complex remote sensing scenes, the Dynamic Cross-feature Fusion Network (DCFNet) method^[23] proposed a dynamic cross-feature

fusion mechanism, allowing features to conduct adaptive information exchange within the network. In addition, to eliminate the reliance on massive amounts of registered label data, the FusionNet^[24] explored unsupervised paradigms to maintain data fidelity without explicit ground-truth constraints. As network architectures grow increasingly complex, purely data-driven "black-box" models suffer from limited physical interpretability and great susceptibility to overfitting. Consequently, to address these limitations, a third category—model-driven deep unfolding networks—has emerged. These methods aim to integrate traditional mathematical optimization mechanisms with deep learning. For example, the Gradient Projection Pansharpening Neural Network (GPPNN)^[25] enhances conventional network stacking by unfolding the traditional gradient projection optimization algorithm into the forward propagation process. Furthermore, GPPNN explicitly incorporates the sensor point spread function and the spatial downsampling operator into the network as physical priors. This dual-driven paradigm of data and model effectively endows deep networks with mathematical interpretability and higher spectral fidelity. Meanwhile, to overcome the restricted receptive fields of traditional Convolutional Neural Networks (CNNs) and capture global context, a fourth category—rooted in global modeling and generative approaches—has gained significant attention. To mitigate the over-smoothing tendency inherent in traditional CNNs, early works such as the Pansharpening Generative Adversarial Network (PSGAN)^[26] and Pan-GAN^[27] introduced generative adversarial mechanisms, synthesizing more realistic visual texture details through adversarial training. However, the effectiveness of adversarial training is often hindered by inherent mode collapse and training instability. To stably expand the receptive field during feature extraction, attention mechanisms were introduced into pansharpening. For example, the Adaptive Discriminative Kernel Network (ADKNet)^[28] employs an adaptive discriminative kernel to achieve spatially aware detail extraction. In recent years, driven by the evolution of vision foundation models, the PanFormer^[29] utilizes self-attention mechanisms to establish long-range dependencies based on the Transformer^[30].

Despite these advances, when confronted with the complex distribution of ground features in fine spatial resolution remote sensing imagery, existing models still struggle to optimally balance spatial details and spectral fidelity, while failing to fully mitigate cross-resolution feature misalignment. First, reconciling receptive field expansion with spatial detail preservation remains inherently difficult, posing a major challenge for models to simultaneously retain spectral consistency and high-frequency features. While many methods have introduced multi-scale architectures to expand the receptive field, their reliance on pooling or strided convolutions inevitably degrades the high-frequency edge details of the PAN image during spatial downsampling. Conversely, by strictly maintaining fine spatial resolution feature maps, the network is confined to local receptive fields, severely

limiting its ability to model large-scale spectral contextual dependencies. Consequently, this architectural dilemma renders the fusion results highly susceptible to either over-smoothing or spectral distortion. Second, spatial feature misalignment remains a critical challenge during cross-resolution fusion. Pansharpening requires precise pixel-level mapping between coarse spatial resolution color information and fine spatial resolution geometric structures. However, existing architectures rely primarily on standard convolutions, whose inherent translation equivariance renders the network spatially blind to the absolute positions of pixels (as revealed by CoordConv^[31]). This deficiency in spatial coordinate awareness causes the network to be highly susceptible to feature misalignment and ghosting artifacts, particularly when propagating multi-source features across layers in regions with complex textures. Finally, deep multi-branch architectures introduce optimization challenges, particularly regarding training convergence and gradient flow. In pursuit of higher representational capacity, existing pansharpening network architectures have become increasingly deep and complex. However, during the initial training stages, multi-branch networks governed by standard random initialization schemes (e. g. , Kaiming^[32] or Xavier initialization) often inadvertently disrupt the intrinsic spectral distribution of the original MS data. This makes them prone to feature variance explosion and high-frequency noise, which subsequently leads to gradient shock and irreversible early color distortion.

To address the challenges of insufficient high-frequency detail extraction, cross-scale feature misalignment, reconstruction artifacts, and unstable training of deep networks in pansharpening, this paper proposes a Nested Multi-Scale Fusion Network (NMSFusion) based on U-Net^[33]. This method constructs a framework comprising four sequential stages: feature extraction, spatial alignment, residual fusion, and stable optimization. Specifically, in the extraction stage, the Multi-Scale Gating Block (MSGB) integrates multi-scale convolution with channel gating to enhance the capability of extracting local high-frequency details. In the alignment stage, the Adaptive Coordinate Encoding Module (ACEM) introduces continuous coordinate encoding to provide a spatial reference for cross-resolution feature interactions, thereby mitigating feature misalignment. In the fusion stage, the Artifact-free Residual Fusion Module (ARFM) integrates multi-scale features through a residual mechanism to suppress the artifacts generated during cross-layer fusion. In the optimization stage, the Identity Initialization Mechanism (IIM) utilizes physical priors to constrain the initial state of the network, facilitating the stable convergence of deep models. The synergy of the aforementioned modules effectively balances the enhancement of spatial details and the high-fidelity reconstruction of the original data.

In summary, the main contributions of this paper are summarized as follows:

1) Construction of a nested multi-scale fusion framework: An architecture encompassing feature extraction,

spatial alignment, and residual fusion is established, achieving a deep synergy between global semantics and local details.

2) Development of a spatial detail extraction and alignment mechanism: The MSGB and ACEM modules are designed to handle multi-scale local feature capture and cross-resolution spatial localization, respectively, which enhances the accuracy in modeling complex ground features.

3) Design of an artifact-free feature reconstruction module: A residual fusion scheme based on smooth stepping is proposed, which effectively alleviates the distribution discrepancy during cross-layer feature fusion and suppresses reconstruction artifacts.

4) Introduction of a stabilized training strategy: The IIM is proposed to constrain the initial state of the neural network via physical priors, alleviating the gradient fluctuation problem in the early stages of deep network training and increasing convergence efficiency.

1 Methods

This section details the proposed NMSFusion network. To address the difficulty in balancing spatial detail preservation and global spectral-context extraction, NMSFusion adopts a nested multi-scale architecture based on a U-Net backbone (shown in Fig. 1). At the global level, a U-Net-style backbone with additive skip connections is employed to extract multi-level global features. At the local level, the MSGB is utilized to effectively capture local detail information within each resolution level. Furthermore, the network incorporates the ACEM and the IIM to respectively enhance the spatial alignment capability of cross-resolution features and improve the stability of network training.

1.1 Overall Architecture

The forward propagation process of NMSFusion primarily consists of four stages: shallow feature extraction, multi-scale encoding, cross-level fusion, and fine spatial resolution reconstruction. Let the coarse spatial resolution MS image and the fine spatial resolution PAN image be denoted as $\mathbf{MS} \in \mathbb{R}^{B \times H \times W}$ and $\mathbf{PAN} \in \mathbb{R}^{1 \times rH \times rW}$, respectively, where B represents the number of MS bands, H and W denote the spatial height and width, 1 represents a single channel, and r is the spatial resolution ratio between the PAN and MS images. First, the network utilizes bicubic interpolation to upsample \mathbf{MS} to the spatial resolution of \mathbf{PAN} , thereby generating the MS base $\mathbf{MS}_{\text{up}} \in \mathbb{R}^{1 \times rH \times rW}$. Subsequently, the network utilizes independent convolutional layers to perform shallow feature extraction on \mathbf{MS}_{up} and \mathbf{PAN} , generating the initial high-dimensional fusion feature $\mathbf{F}_0 \in \mathbb{R}^{C \times rH \times rW}$, where C denotes the number of channels after initial feature extraction. Simultaneously, the spatial positional priors provided by the ACEM are injected to establish a unified global coordinate constraint for subsequent feature alignment. Building upon this, the features enter a pyramid encoder containing three resolution levels. Within the same level, MSGB is consecutively stacked to mine multi-scale local features. For cross-level interaction, strided convolu-

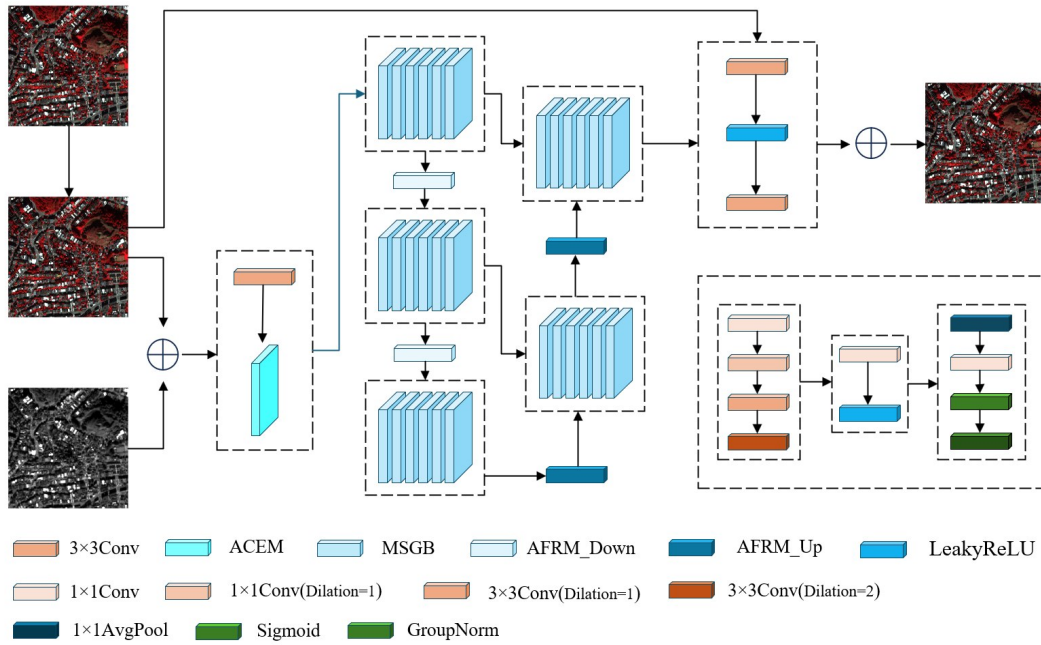


Fig. 1 The Pansharpening Framework Based on the Nested Multi-Scale Fusion Network
图1 基于嵌套多尺度融合网络的全色融合框架

tions are utilized for spatial downsampling and channel dimension expansion. As the receptive field progressively expands, the network effectively extracts large-scale spectral-contexts and macroscopic structural information within the deepest bottleneck layer.

To recover the target spatial resolution, the decoder utilizes the ARFM for smooth feature upsampling. Through element-wise additive skip connections, high-frequency spatial details from the corresponding encoder levels are guided into the decoding stage to alleviate information loss caused by feature resampling. The progressively decoded features are finally mapped back to the image space by the reconstruction module, generating a high-frequency residual map $\mathbf{R}_{out} \in \mathbb{R}^{B \times rH \times rW}$ consistent with the spatial dimensions of the MS base. Under the constraint of IIM, the final fine spatial resolution MS image $\mathbf{SR} \in \mathbb{R}^{B \times rH \times rW}$ is produced by integrating the MS base and the reconstructed residual, namely:

$$\mathbf{SR} = \mathbf{MS}_{up} + \alpha \cdot \mathbf{R}_{out}, \quad (1)$$

where α is a learnable weight parameter. During the initial stage of training, α can drive the network output to approximate \mathbf{MS}_{up} , thereby effectively enhancing the stability of early-stage optimization.

1.2 Multi-Scale Gating Block (MSGB)

The traditional U-Net architecture may attenuate high-frequency spatial details during successive downsampling, which ultimately hinders the fine recovery of local textures and edge information. To facilitate sufficient interaction between local details and multi-scale contextual features within a single resolution level, this paper develops the MSGB. The module adopts a "split-transform-merge" paradigm, the core objective of which is to concurrently model local details and regional contextual information across varying receptive fields, while main-

taining a manageable computational overhead. Specifically, the input features are first divided equally into four parallel branches along the channel dimension to construct a contextual feature pyramid at the local scale. The functions of the four branches are as follows:

- 1) Pixel-wise mapping branch (1×1 convolution): It focuses on cross-channel feature aggregation and linear recombination, preserving the original pixel-level positional information without altering the receptive field.
- 2) Local detail branch (standard 3×3 convolution; $d = 1$): It focuses on extracting compact texture edges and high-frequency spatial structures.
- 3) Regional context branch (dilated 3×3 convolution; $d = 2$): It aims to perceive the neighborhood dependencies of medium-scale ground features, such as buildings and roads.
- 4) Wide-area perception branch (dilated 3×3 convolution; $d = 3$): It is used to extract a broader range of spectrally homogeneous regions (e. g., water bodies and vegetation) and assists in mitigating the interference of local noise.

Subsequently, the multi-scale features extracted from the four parallel branches are concatenated along the channel dimension. A fusion convolution is then applied to aggregate these diverse representations and project them back to the original channel dimensionality. To adaptively modulate the relative contributions of various receptive field branches across different spatial regions, a Simplified Channel Attention (SCA) mechanism is introduced. In contrast to the conventional SE module, SCA replaces the multi-layer perceptron with a single 1×1 convolution to reduce the parameter overhead. Specifically, the global average pooling is used to compress the spatial dimensions, a 1×1 convolution is used to learn

cross-channel correlations, and a Sigmoid activation is applied to generate the final attention weights. These weights are element-wise multiplied with the fused features, thereby achieving adaptive feature gating. Finally, the recalibrated features are added to the original input via a residual connection, and group normalization is employed to complete feature normalization, aiming to enhance the optimization stability of the model under small-batch training environments.

In the design of the MSGB, feature gating only employs the simplified channel attention mechanism, without introducing spatial attention. This design is primarily based on a trade-off between feature modeling requirements and computational overhead. On the one hand, the overall U-Net architecture and the multi-scale convolutions within the MSGB already possess the function of extracting high-frequency spatial features. Thus, the introduction of spatial attention is prone to generating redundancy in feature modeling. On the other hand, the pixel-level operations of spatial attention would increase the floating-point operations (FLOPs) of the model. Therefore, this paper adopts channel attention as the gating mechanism. By dynamically recalibrating the channel weights of feature branches at different scales, it achieves the filtering and injection of high-frequency spatial features under the condition of controlled computational complexity.

1.3 Adaptive Coordinate Encoding Module (ACEM)

The key to pansharpening lies in accurately aligning and injecting the spatial structural information from the PAN image into the MS features. However, conventional convolutions are inherently translation-equivariant, rendering them insensitive to absolute spatial coordinates. This deficiency makes the network highly susceptible to feature misalignment. To address the issue, this paper designs the ACEM to supplement the network with stable global positional priors. The core mechanism of this module can be summarized into two parts: "two-dimensional (2D) continuous coordinate construction" and "zero-gated adaptive injection".

First, the ACEM utilizes sine and cosine functions of different frequencies to generate a base dictionary \mathbf{P}_{base} containing rich positional mapping information along both the horizontal and vertical spatial dimensions. Unlike traditional positional encoding that relies on fixed resolutions, the ACEM introduces a continuous domain resampling mechanism. Let the current visual feature tensor be $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$. After dynamically acquiring its spatial dimensions (H, W) , the module utilizes a bilinear interpolation function $\mathcal{B}(\cdot)$ to perform smooth resampling on the base dictionary \mathbf{P}_{base} , generating the target coordinate tensor $\mathbf{P}_{\text{target}}$ that adapts to the current resolution, namely:

$$\mathbf{P}_{\text{target}} = \mathcal{B}(\mathbf{P}_{\text{base}}(H, W)) \quad . \quad (2)$$

This mechanism breaks the network's reliance on fixed spatial grids, enabling it to flexibly adapt to the training and inference of remote sensing image patches of arbitrary sizes. Specifically, when the inference size is larger than the training size (e. g. , expanded from $64 \times$

64 to 256×256), the ACEM does not extrapolate into an unknown absolute coordinate domain. Instead, it consistently normalizes the spatial positions of the input image to the same continuous reference interval. Therefore, the coordinate tensor mapping process for large-scale images is essentially performing a higher-density spatial sampling (i. e. , interpolation) within this fixed normalized interval. This continuous domain resampling mechanism helps reduce the risk of coordinate distribution shift caused by scale variations, thereby enhancing the stability of the model's cross-scale inference.

In the feature fusion stage, to prevent the positional priors from causing excessively strong interference with the visual feature representation during the early training stages, this paper introduces a learnable global scalar Y as a gating parameter, injecting the coordinate encoding into the feature stream in the form of a residual. The fusion process can be expressed as:

$$\mathbf{Y} = \mathbf{X} + Y \cdot \mathbf{P}_{\text{target}} \quad , \quad (3)$$

where \mathbf{Y} represents the output feature after fusing the absolute spatial priors. During model initialization, Y is set to 0. This zero-gating strategy prompts the network to primarily rely on visual texture content for feature alignment in the early training stage, preventing spatial coordinate signals from dominating prematurely. As the optimization deepens, the network can progressively and adaptively adjust the weight of Y based on gradient back-propagation, thereby effectively improving the convergence stability of the model during the initial optimization phase.

1.4 Artifact-free Residual Fusion Module (ARFM)

In low-level vision tasks, conventional U-Net architectures typically rely on max pooling, transposed convolutions, and channel-wise concatenation to facilitate multi-scale feature integration. However, in pansharpening scenarios, such operations are prone to introducing issues such as spatial information loss, checkerboard artifacts, and inconsistent cross-layer feature distributions. To address these limitations, this paper designs the ARFM (as shown in Fig. 2), which improves the decoding process from two perspectives: spatial reconstruction sampling and the cross-level feature interaction. Specifically:

1) Feature spatial resampling: In the downsampling stage, the ARFM employs strided convolutions (stride = 2) instead of max pooling, avoiding the direct discarding of non-extremum pixels. This design transforms downsampling into an adaptive information compression process, which helps maintain the coherence of local high-frequency structures. In the upsampling stage, to alleviate the periodic checkerboard artifacts that transposed convolutions might introduce, the ARFM adopts bilinear interpolation coupled with standard convolution to increase the spatial resolution, thereby generating a smoother feature representation.

2) Additive residual interaction: Conventional cross-layer channel concatenation not only significantly increases the feature dimensionality of the decoder, but is also prone to introducing distribution discrepancies be-

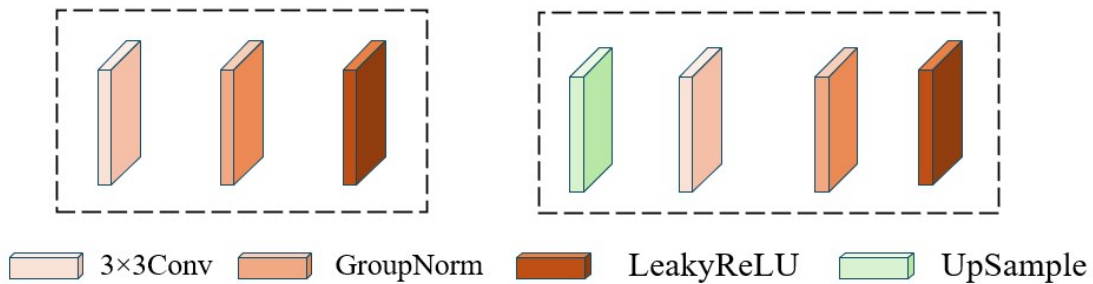


Fig. 2 Schematic diagram of the Artifact-free Residual Fusion Module (ARFM)
图2 无伪影残差融合模块(ARFM)结构示意图

tween shallow textures and deep semantics. The ARFM adopts an element-wise addition approach to replace channel concatenation. In this mechanism, the low-frequency semantic features of the decoder serve as the base, while the shallow features transmitted from the encoder serve as high-frequency spatial residuals for supplementary fusion. This parameter-free interaction method transforms the decoding process into a residual learning paradigm, effectively reducing the optimization difficulty of the model.

1.5 Identity Initialization Mechanism (IIM)

NMSFusion encompasses a complex multi-branch and multi-level feature interaction structure. If conventional random initialization strategies (such as Kaiming or Xavier initialization) are directly adopted, high-frequency noise accumulation is prone to occur during the early stages of training, which subsequently triggers gradient instability and spectral feature distortion. To this end, this paper introduces the IIM to constrain the network parameters, enabling the model to closely approximate identity mapping during the initial training phase. The IIM primarily acts synergistically at both the local and global levels.

At the local level, the IIM initializes the weights and biases of the feature fusion convolutions within each MSGB to zero. This ensures that during the initial forward propagation of the model, the feature output of the multi-scale branches is zero, and the module is equivalent to a residual flow of skip connections, thereby maintaining the stable transmission of visual signals within the deep network.

At the global level, the IIM applies zero initialization to both the terminal projection layer and the feature scaling parameter within the final reconstruction module. This strategy biases the network toward directly outputting the upsampled MS base MS_{up} during the initial phases of training, thereby circumventing the premature fitting of high-frequency edge information. Consequently, the pansharpening process can be transformed into a smoother residual optimization process, enhancing the convergence stability and reconstruction fidelity of the model.

The IIM applies zero initialization exclusively to the weights and biases of the terminal convolutional layers within the residual branches (i. e. , immediately prior to

feature fusion and output), while the preceding feature extraction layers retain Kaiming initialization. During the first backpropagation, because the inputs to the zero-convolution layers are non-zero and asymmetric feature maps generated by the front-end network, the parameter gradients of the terminal convolutional layers exhibit asymmetry according to the chain rule, which breaks the symmetrical state of the network parameters after the first iteration. Furthermore, based on the residual structure, gradients can be backpropagated along the backbone path, avoiding the issues of vanishing gradients and dead neurons that branch networks might cause, maintaining the numerical stability of the network during the initial training stage.

2 Experiments

2.1 Experimental Data and Design

To objectively evaluate the performance of the NMS-Fusion network, the PanCollection dataset constructed by the Wuhan University was adopted as the baseline dataset. This dataset generates standardized paired training samples through spatial downsampling of real MS images, facilitating model training and accuracy evaluation. Three fine spatial resolution satellite datasets-WorldView-3, QuickBird, and GaoFen-2 were selected for this study. These datasets exhibit significant physical differences in spatial resolution, modulation transfer function, and spectral response, which is conducive to assessing the model's generalization capability across multiple sensors. Specifically, the WorldView-3 (WV3) dataset comprises PAN images with a spatial resolution of 0.31 m and eight-band MS images at 1.24 m. It covers dense building clusters and complex urban details such as road markings, which was primarily used to test the capability for high-frequency detail reconstruction under complex spatial textures and effectiveness in suppressing spatial aliasing. Meanwhile, the spatial resolution of the PAN and four-band MS of the QuickBird (QB) dataset is 0.61 m and 2.44 m, respectively. The scenes include numerous vegetation and water boundary, which is suitable for examining the network's spectral consistency in naturally transitioning areas. Furthermore, the spatial resolution of the PAN and four-band MS of the domestic GaoFen-2 (GF2) satellite are 0.8 m and 3.2 m, respectively. Independent training and evaluation were conducted on

three datasets. Regarding dataset partitioning, 9, 714, 17, 139, and 19, 809 training sample pairs were extracted from the WV3, QB, and GF2 datasets, respectively. By applying a standard 9:1 training-to-validation split, 1, 080, 1, 905, and 2, 201 pairs were subsequently reserved for validation. The sizes of the PAN and MS images for the aforementioned training and validation data are 64×64 pixels and 16×16 pixels, respectively. Furthermore, to better align with real-world remote sensing interpretation requirements, 20 pairs of PAN and MS images with dimensions of 256×256 pixels and 64×64 pixels were constructed for each dataset for quantitative evaluation. During the training phase, the network utilized the L1 loss function and the AdamW optimizer for parameter updates. The initial learning rate of the model was set to 1×10^{-4} , and the weight decay was set to 0. A step decay strategy was adopted for dynamic learning rate adjustment, where the learning rate was reduced by a factor of 0.5 every 50 epochs. The model was trained for a total of 300 epochs, with the batch size uniformly set to 24.

NMSFusion was compared with nine recently developed pansharpening networks, including: PNN^[16], PanNet^[17], DiCNN^[20], MSDCNN^[19], FusionNet^[24], DRPNN^[34], BDPN^[22], DCFNet^[23], and ADKNet^[28]. For quantitative evaluation, six metrics were selected: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Spectral Angle Mapper (SAM), Relative Dimensionless Global Error in Synthesis (ERGAS), Correlation Coefficient (CC), and Root Mean Square Error (RMSE).

2.2 Results for the WV3 Dataset

Fig. 3 illustrates the fusion results for three test regions in the WV3 dataset, covering different land cover characteristics. As observed, shallower CNN models (e. g. , DiCNN and PNN), constrained by their simple interpolation architectures, exhibit pronounced spatial blurring and edge aliasing. Furthermore, lacking fine-grained cross-modal alignment mechanisms, several deep networks (such as ADKNet and BDPN) suffer from severe spectral overflow and halo artifacts in boundary areas with sharp color contrast (e. g. , transitions between red roofs and dark shadows). Conversely, the proposed NMSFusion not only reliably reconstructs high-frequency edges but also accurately preserves spectral fidelity in both dark shadows and large homogeneous regions, yielding visual results closer to the reference images.

Table 1 presents the quantitative evaluation results for these three scenes. As noted, NMSFusion produces the greatest accuracy across all six metrics, striking an optimal balance between spatial detail enhancement and spectral fidelity. For example, in Region 3, the PSNR of NMSFusion reaches 39.511 1 dB, which is 1.15 dB larger than the second-best method DCFNet (38.356 6 dB). Furthermore, in Region 1, when enhancing high-frequency edge injection, traditional deep networks (e. g. , BDPN and DiCNN) incur relatively obvious spectral degradation, with their SAM values rising to 2.924 8 and 2.610 8, respectively. In contrast, NMSFusion maintains the SAM at a lower level of 1.897 0.

2.3 Results for the QB Dataset

Fig. 4 shows the fusion results of three regions in the QB dataset. Different from the structural blurring or edge bleeding frequently encountered in other methods, the fusion results of NMSFusion exhibit fewer artifacts and higher fidelity, demonstrating the advantages of its multi-scale architecture in adaptively extracting and fusing cross-modal features. Table 2 presents the corresponding quantitative evaluation results. Despite the significant differences in spatial resolution and spectral response curves between the QB and WV3 datasets, NMSFusion still demonstrates great robustness, achieving the greatest accuracy across all six metrics for the three test regions. Specifically, while maintaining high spatial clarity, NMSFusion exhibits a strong spectral preservation capability. For example, in Region 3, whereas the SAM values of some traditional deep networks exceed 6.0 due to insufficient cross-modal feature alignment, NMSFusion achieves a SAM of 4.8086, representing superior spectral fidelity.

2.4 Results for the GF2 Dataset

Fig. 5 illustrates the fusion results for three regions in the GF2 dataset. The results demonstrate that NMSFusion effectively suppresses color misalignment and shadow overflow, while reconstructing sharper boundaries. The quantitative evaluation results in Table 3 indicate that NMSFusion maintains stable fusion performance, achieving the greatest accuracy among all methods. For example, in Regions 2 and 3, the PSNR of NMSFusion reaches 44.0953 dB and 47.9541 dB, respectively, which are at least 2.1 dB larger than the second-best method.

2.5 Comprehensive Analysis

Table 4 summarizes the average quantitative metrics of all 10 methods across 60 independent test samples (20 samples per dataset) from the WV3, QB, and GF2 datasets. As observed, NMSFusion achieves the optimal performance across all six metrics on the three datasets. Specifically, regarding PSNR, NMSFusion surpasses the second-best method by approximately 0.6 dB, 0.2 dB and 1.52 dB on the WV3, QB and GF2 datasets, respectively.

Meanwhile, NMSFusion effectively mitigates the spectral distortion induced by conventional deep networks (e. g. , ADKNet and BDPN). Despite varying sensor-specific physical characteristics—such as the unique signal-to-noise ratio (SNR) and modulation transfer function of GF2—the proposed method consistently maintains stable performance, demonstrating great cross-sensor robustness. This indicates that the proposed architecture, which synergistically integrates a macroscopic U-Net topology with microscopic multi-scale gated residual fusion, effectively circumvents the risk of overfitting to specific data distributions.

2.6 Ablation Study

To evaluate the contributions of the MSGB, ARFM, and ACEM in the proposed NMSFusion network, this section conducts ablation experiments on the WV3 datas-

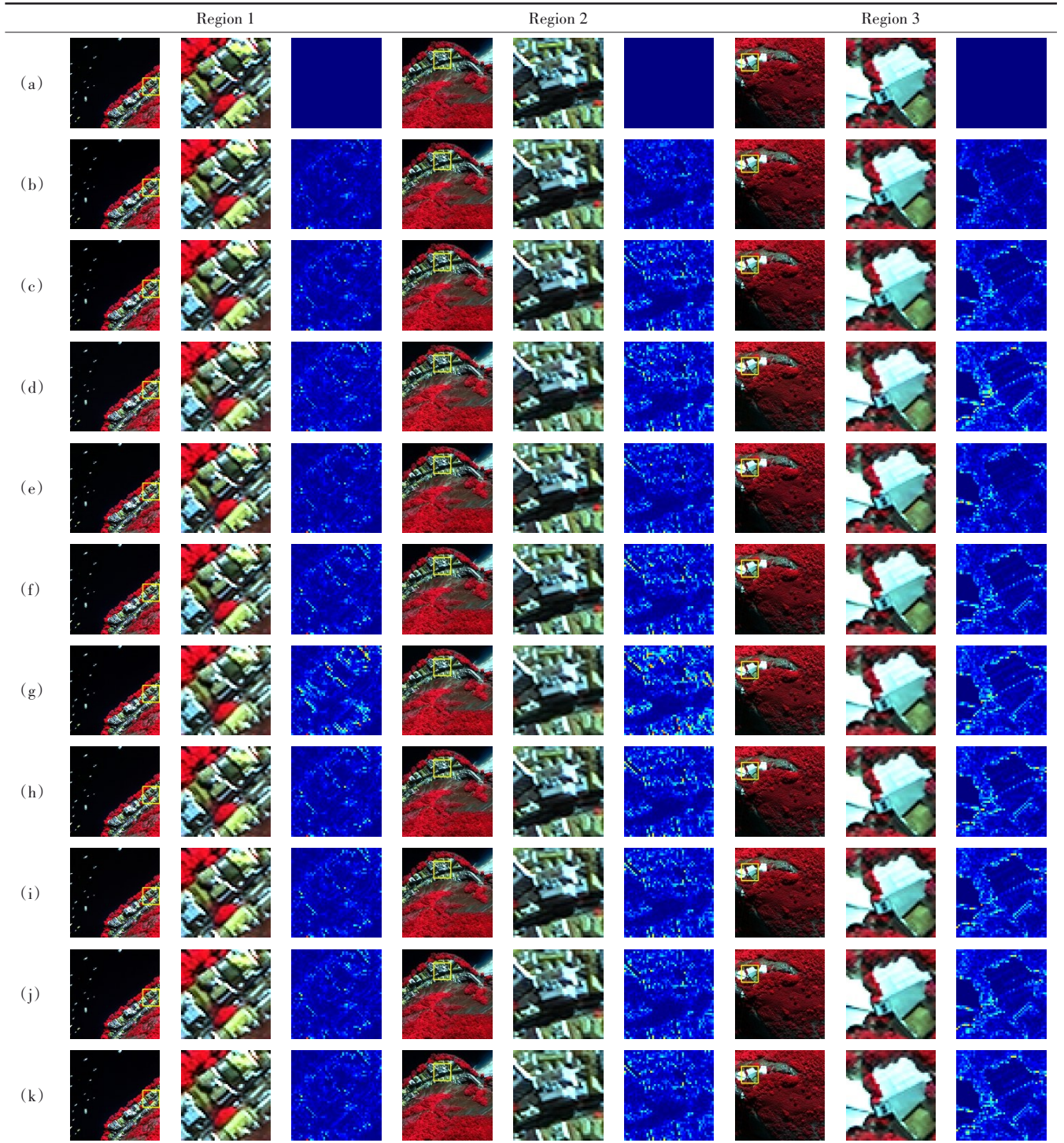


Fig. 3 Visual comparison of different pansharpening methods for three specific regions from the WV3 dataset: (a) Reference image; (b) NMSFusion; (c) ADKNet; (d) BDPN; (e) DCFNet; (f) DRPNN; (g) DiCNN; (h) FusionNet; (i) MSDCNN; (j) PNN; (k) PanNet (For each region, the first column displays the full view, the second column displays the locally magnified view, and the third column displays the corresponding error map)

图3 不同全色融合方法在WV3数据集三个特定区域上的视觉对比:(a)参考图像;(b)NMSFusion;(c)ADKNet;(d)BDPN;(e)DCFNet;(f)DRPNN;(g)DiCNN;(h)FusionNet;(i)MSDCNN;(j)PNN;(k)PanNet(每个区域的第一列为完整展示,第二列为局部缩放,第三列为相应误差图)

et. The experiment adopts the base U-Net excluding the aforementioned modules as the baseline model. By progressively adding each module, the accuracy (as shown in Table 5) and computational complexity (number of pa-

rameters and FLOPs, as shown in Table 6) of different model variants are comparatively analyzed. When the MSGB is removed, the feature extraction stage of the network employs a basic two-layer residual convolutional

Table 1 Quantitative evaluation results of different pansharpening methods for the three specific regions from the WV3 dataset**表 1 不同全色融合方法在 WV3 数据集三个特定区域上的定量评估结果**

		PSNR	SSIM	SAM	ERGAS	CC	RMSE
Region 1	NMSFusion	39.583 8	0.987 7	1.897 0	9.644 1	0.984 0	0.010 5
	ADKNet	37.882 5	0.982 9	2.246 4	11.719 8	0.976 5	0.012 8
	BDPN	36.543 2	0.975 8	2.924 8	13.733 5	0.966 2	0.014 9
	DCFNet	38.783 5	0.985 7	2.148 9	10.529 1	0.981 2	0.011 5
	DRPNN	37.630 2	0.981 5	2.426 4	12.023 2	0.974 8	0.013 1
	DiCNN	35.638 0	0.973 6	2.610 8	14.800 1	0.960 7	0.016 5
	FusionNet	37.416 2	0.981 1	2.417 9	12.356 7	0.973 4	0.013 5
	MSDCNN	37.453 8	0.980 8	2.496 7	12.323 9	0.973 3	0.013 4
	PNN	37.104 6	0.978 8	2.629 4	12.765 1	0.970 7	0.014 0
	PanNet	37.176 0	0.980 3	2.498 4	12.665 6	0.971 0	0.013 8
Region 2	NMSFusion	36.768 8	0.969 6	3.027 0	6.908 2	0.981 5	0.014 5
	ADKNet	35.076 3	0.958 6	3.623 0	8.545 9	0.972 2	0.017 6
	BDPN	33.811 8	0.941 3	4.570 5	9.959 8	0.960 7	0.020 4
	DCFNet	36.049 5	0.964 6	3.288 7	7.563 0	0.977 9	0.015 8
	DRPNN	35.066 4	0.957 2	3.766 8	8.611 9	0.971 6	0.017 6
	DiCNN	33.272 4	0.940 4	4.377 7	10.597 6	0.958 5	0.021 7
	FusionNet	34.842 0	0.955 5	3.845 3	8.879 2	0.969 8	0.018 1
	MSDCNN	34.692 4	0.954 7	3.909 0	9.011 4	0.968 9	0.018 4
	PNN	34.344 2	0.950 9	4.100 8	9.356 6	0.965 9	0.019 2
	PanNet	34.617 9	0.954 1	3.887 6	9.172 1	0.967 3	0.018 6
Region 3	NMSFusion	39.511 1	0.979 9	3.146 5	5.861 6	0.992 6	0.010 6
	ADKNet	37.791 0	0.973 6	3.697 2	7.085 7	0.989 6	0.012 9
	BDPN	35.808 4	0.955 9	4.953 8	9.002 3	0.982 6	0.016 2
	DCFNet	38.356 6	0.975 1	3.482 9	6.975 5	0.989 4	0.012 1
	DRPNN	38.000 4	0.973 2	3.694 7	7.162 0	0.989 1	0.012 6
	DiCNN	35.648 6	0.958 1	4.930 3	8.866 1	0.983 6	0.016 5
	FusionNet	37.631 8	0.972 7	3.806 3	7.380 1	0.988 8	0.013 1
	MSDCNN	36.877 4	0.969 2	4.181 9	7.841 2	0.987 2	0.014 3
	PNN	36.515 0	0.967 3	4.338 4	8.156 2	0.986 0	0.014 9
	PanNet	36.757 5	0.968 1	4.218 7	7.973 3	0.986 9	0.014 5

block (comprising standard 3×3 convolutions and activation functions) for structural padding. When the ARFM is removed, the upsampling and downsampling operations of the network revert to standard transposed convolutions and max pooling, and the feature fusion method in the skip connection stage is altered from element-wise addition to channel concatenation followed by a 1×1 convolution for dimensionality reduction. Since the ACEM does not alter the feature dimensions of the backbone network, omitting it simply entails bypassing the spatial coordinate fusion step. The aforementioned alternative schemes ensure that all model variants maintain proper forward and backward propagation.

As presented in Table 5, the integration of each proposed module yields targeted enhancement. For example, the addition of the ARFM optimizes the SAM to the largest value of 5.5507, demonstrating that the residual feature fusion mechanism effectively mitigates spectral

distortion. Furthermore, the subsequent incorporation of the ACEM, culminating in the complete NMSFusion model, brings consistent gains in PSNR and ERGAS. This validates that the injection of spatial positional priors guides the network to align high-frequency spatial structures more precisely.

Table 6 details the influence of each proposed module on model complexity. Notably, the MSGB significantly reduces both the parameter count and computational cost, demonstrating its high efficiency in feature extraction. While the ARFM incurs a marginal increase in parameters, it maintains a compact structural profile, ensuring that the overall computational overhead remains low. Furthermore, the integration of the ACEM and the IIM strategy imposes negligible additional computational burden. Overall, compared to the baseline network, the complete NMSFusion architecture delivers superior performance while remarkably reducing the parameter count

Table 2 Quantitative evaluation results of different pansharpening methods for the three specific regions from the QB dataset**表 2 不同全色融合方法在 QB 数据集三个特定区域上的定量评估结果**

		PSNR	SSIM	SAM	ERGAS	CC	RMSE
Region 1	NMSFusion	39.1068	0.9697	4.5129	7.5412	0.9819	0.0111
	ADKNet	35.3983	0.9433	5.6373	11.5369	0.9578	0.0170
	BDPN	35.8562	0.9394	6.4551	10.9445	0.9624	0.0161
	DCFNet	38.8141	0.9683	4.6681	7.7900	0.9806	0.0115
	DRPNN	37.9132	0.9612	5.0855	8.6196	0.9764	0.0127
	DiCNN	35.0164	0.9344	6.2303	12.0529	0.9542	0.0177
	FusionNet	35.8840	0.9443	5.8103	10.8553	0.9629	0.0161
	MSDCNN	35.1831	0.9359	6.0399	11.7708	0.9563	0.0174
	PNN	35.1597	0.9366	5.9409	11.8365	0.9557	0.0175
	PanNet	35.1746	0.9379	5.9832	11.8098	0.9559	0.0174
Region 2	NMSFusion	37.0901	0.9588	4.4214	6.6902	0.9820	0.0140
	ADKNet	32.0291	0.9258	5.2081	11.9666	0.9415	0.0250
	BDPN	32.8693	0.9222	6.1338	10.8620	0.9525	0.0227
	DCFNet	36.7455	0.9567	4.5587	6.9576	0.9805	0.0145
	DRPNN	35.9401	0.9499	4.7499	7.6268	0.9765	0.0160
	DiCNN	32.6799	0.9100	5.7488	11.0960	0.9503	0.0232
	FusionNet	33.7968	0.9307	5.2125	9.7601	0.9616	0.0204
	MSDCNN	32.6628	0.9186	5.4273	11.1010	0.9500	0.0233
	PNN	32.8232	0.9102	5.5332	10.9142	0.9520	0.0228
	PanNet	32.9803	0.9189	5.3951	10.7069	0.9536	0.0224
Region 3	NMSFusion	36.5318	0.9572	4.8086	7.3341	0.9835	0.0149
	ADKNet	31.6374	0.9203	5.6934	12.8927	0.9480	0.0262
	BDPN	33.4798	0.9202	6.6409	10.4417	0.9665	0.0212
	DCFNet	36.1632	0.9551	4.9451	7.6557	0.9819	0.0156
	DRPNN	34.9185	0.9462	5.1921	8.8351	0.9758	0.0180
	DiCNN	32.1535	0.9066	6.3681	12.1612	0.9544	0.0247
	FusionNet	33.3226	0.9262	5.7394	10.6165	0.9652	0.0216
	MSDCNN	32.7015	0.9144	5.9676	11.3940	0.9599	0.0232
	PNN	32.4567	0.9102	6.0222	11.7254	0.9576	0.0238
	PanNet	32.6528	0.9143	5.9356	11.4633	0.9594	0.0233

by 38.8% and the computational overhead by 35.8%.

Additionally, Table 7 presents an ablation study of removing individual modules. The results indicate that the performance of the complete NMSFusion model is superior to all its ablation variants. Specifically, the MSGB contributes the most to the spatial reconstruction accuracy, while the ARFM and ACEM primarily enhance spectral fidelity.

It is worth noting that the incremental gains in global quantitative metrics (e.g., PSNR) following the integration of each module appear relatively limited. This is primarily because local high-frequency structural improvements are often diluted by large numbers of smooth background pixels during global statistical averaging. Specifically, the spatial misalignment corrections provided by the ACEM and the artifact suppression achieved by the ARFM are highly concentrated in these localized high-frequency regions. Thus, such nuanced variations

are difficult to capture fully through holistic error metrics. Furthermore, the contributions of each module extend across distinct dimensions. For example, the MSGB enhances accuracy without significantly inflating computational overhead, while the IIM leverages its initialization mechanism to reduce early-stage spectral distortion, thereby bolstering convergence stability. Ultimately, NMSFusion strikes an optimal balance among accuracy, local feature preservation, computational efficiency, and training stability.

2.7 Comparison of Computational Efficiency

To further evaluate the computational efficiency of NMSFusion, this paper compares the number of parameters, FLOPs, and average inference time of various methods, with the results presented in Table 8. Compared to shallow networks such as PNN and PanNet, the computational overhead of NMSFusion is larger, primarily due to the introduction of the cross-resolution feature alignment

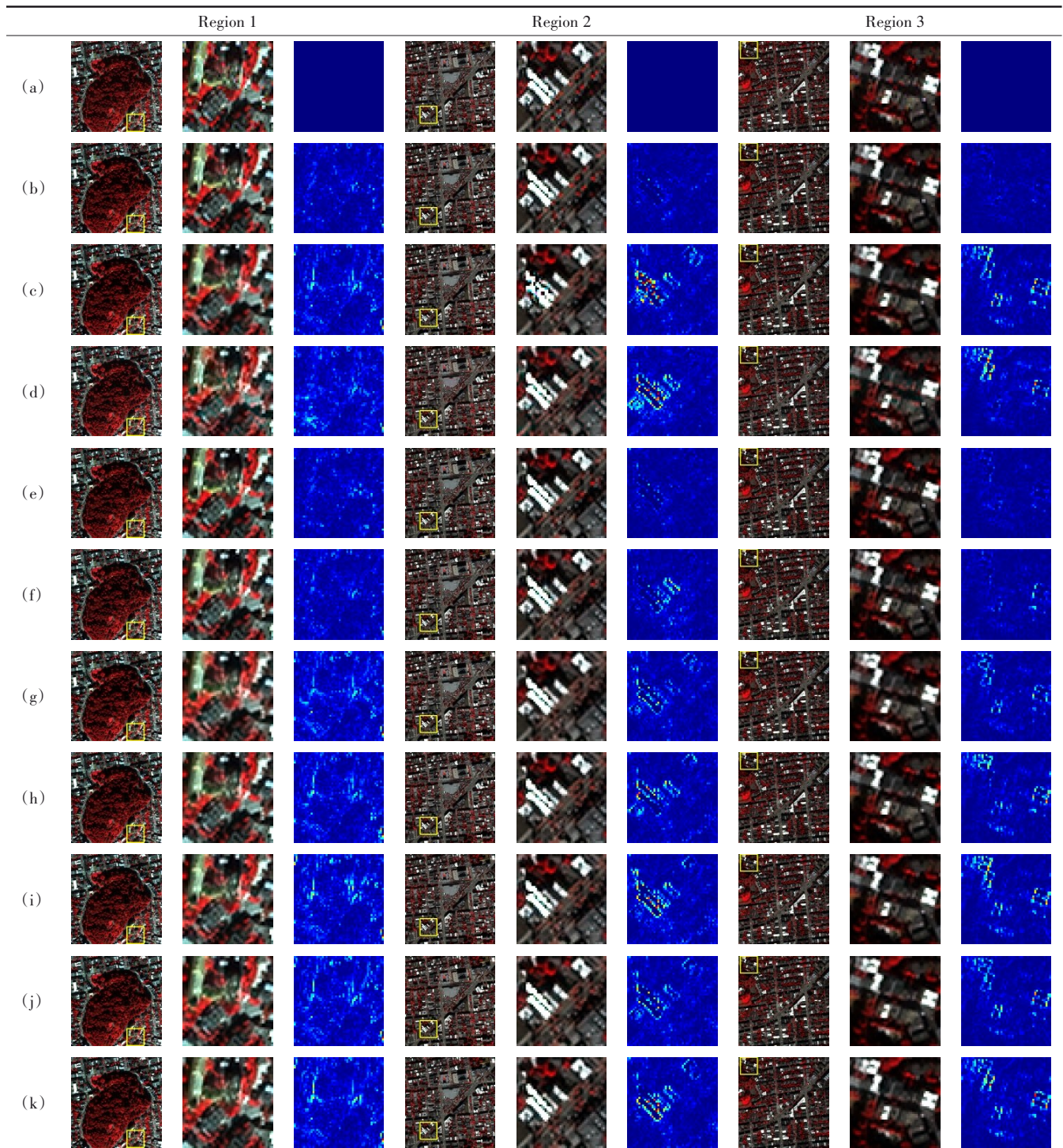


Fig. 4 Visual comparison of different pansharpening methods for three specific regions from the QB dataset: (a) Reference image; (b) NMSFusion; (c) ADKNet; (d) BDPN; (e) DCFNet; (f) DRPNN; (g) DiCNN; (h) FusionNet; (i) MSDCNN; (j) PNN; (k) PanNet (For each region, the first column displays the full view, the second column displays the locally magnified view, and the third column displays the corresponding error map)

图4 不同全色融合方法在QB数据集三个特定区域上的视觉对比:(a) 参考图像;(b) NMSFusion;(c) ADKNet;(d) BDPN;(e) DCFNet;(f) DRPNN;(g) DiCNN;(h) FusionNet;(i) MSDCNN;(j) PNN;(k) PanNet(每个区域的第一列为完整展示,第二列为局部缩放,第三列为相应误差图)

and multi-scale gating mechanisms. However, compared with some high-complexity models developed in recent years, NMSFusion exhibits a relatively lower computational cost. For example, the FLOPs of NMSFusion are

12.726 G, significantly lower than the 135.001 G of ADKNet. The inference time of NMSFusion is 11.82 ms, which is also lower than those of DCFNet and BDPN. Overall, NMSFusion achieves a favorable balance be-

Table 3 Quantitative evaluation results of different pansharpening methods for the three specific regions from the GF2 dataset**表3 不同全色融合方法在GF2数据集三个特定区域上的定量评估结果**

		PSNR	SSIM	SAM	ERGAS	CC	RMSE
Region 1	NMSFusion	42.846 9	0.981 9	0.662 5	1.146 3	0.997 4	0.007 2
	ADKNet	38.790 7	0.960 9	0.886 4	1.789 0	0.993 7	0.011 5
	BDPN	35.078 6	0.928 4	1.228 5	2.773 9	0.985 0	0.017 6
	DCFNet	40.971 9	0.976 7	0.818 8	1.421 6	0.996 1	0.008 9
	DRPNN	39.939 2	0.968 2	0.872 3	1.587 2	0.995 1	0.010 1
	DiCNN	35.239 0	0.929 0	1.145 8	2.705 6	0.985 8	0.017 3
	FusionNet	36.406 4	0.944 9	0.973 5	2.367 0	0.989 0	0.015 1
	MSDCNN	37.316 9	0.948 6	0.941 9	2.105 8	0.991 2	0.013 6
	PNN	36.405 2	0.941 5	1.006 3	2.346 2	0.989 3	0.015 1
	PanNet	37.199 0	0.948 1	0.953 2	2.161 2	0.990 9	0.013 8
Region 2	NMSFusion	44.095 3	0.985 8	0.594 0	1.075 0	0.997 5	0.006 2
	ADKNet	40.707 8	0.972 6	0.765 4	1.579 9	0.994 7	0.009 2
	BDPN	37.232 6	0.946 5	1.152 2	2.348 5	0.988 0	0.013 8
	DCFNet	41.975 6	0.980 5	0.746 7	1.365 1	0.996 0	0.008 0
	DRPNN	41.589 6	0.976 6	0.751 6	1.425 1	0.995 6	0.008 3
	DiCNN	37.644 1	0.952 7	1.008 8	2.250 1	0.989 2	0.013 1
	FusionNet	38.921 1	0.963 4	0.836 4	1.938 5	0.991 9	0.011 3
	MSDCNN	39.583 0	0.964 7	0.838 3	1.793 4	0.993 1	0.010 5
	PNN	38.807 1	0.960 7	0.884 1	1.960 0	0.991 8	0.011 5
	PanNet	39.521 0	0.965 1	0.839 8	1.815 9	0.993 0	0.010 6
Region 3	NMSFusion	47.954 1	0.991 6	0.398 2	0.851 0	0.994 8	0.004 0
	ADKNet	45.781 9	0.986 6	0.491 3	1.088 3	0.991 3	0.005 1
	BDPN	42.992 0	0.975 9	0.713 8	1.479 1	0.982 9	0.007 1
	DCFNet	45.660 1	0.987 9	0.519 9	1.109 4	0.991 4	0.005 2
	DRPNN	45.950 4	0.986 8	0.493 1	1.065 7	0.991 6	0.005 0
	DiCNN	42.454 0	0.975 8	0.686 3	1.584 9	0.981 1	0.007 5
	FusionNet	44.782 9	0.983 9	0.533 6	1.211 9	0.988 8	0.005 8
	MSDCNN	44.682 2	0.983 2	0.551 9	1.235 3	0.988 8	0.005 8
	PNN	44.094 6	0.981 7	0.574 0	1.315 3	0.987 1	0.006 2
	PanNet	44.660 4	0.983 2	0.547 6	1.238 4	0.988 7	0.005 8

tween fusion performance and computational overhead.

3 Conclusions

To address the inherent dilemma between enhancing spatial details and preserving spectral fidelity in remote sensing image pansharpening, as well as the limitations of traditional deep convolutional networks in leading to spatial misalignment and checkerboard artifacts, this paper proposes the NMSFusion method. It achieves breakthroughs from the perspectives of network architecture, feature interaction, and model initialization strategies. By introducing the ACEM, global spatial anchors are injected into the network, thereby reducing spatial misalignment during feature transmission across different resolutions. Simultaneously, a plug-and-play IIM is proposed to avoid the gradient shock and early spectral distortion that frequently occur in deep networks. The pro-

posed method was experimentally validated on three mainstream remote sensing datasets. The results show that the proposed method achieves superior accuracy compared to nine mainstream pansharpening networks. While enhancing the visual clarity of the fused images, the NMSFusion network effectively mitigates spectral distortion. Notably, it maintains high fusion accuracy even while achieving a substantial reduction in parameter count and computational overhead of FLOPs.

The fusion results generated by NMSFusion have potential value for downstream applications. For example, they can provide finer texture information for feature extraction in highly heterogeneous regions, such as urban scenes, and for detection of small-sized targets of interest. Furthermore, the architecture of NMSFusion is not restricted to the pansharpening task, but also exhibits significant expansion potential in cross-resolution spatial-spectral fusion tasks, such as MS and hyperspectral im-

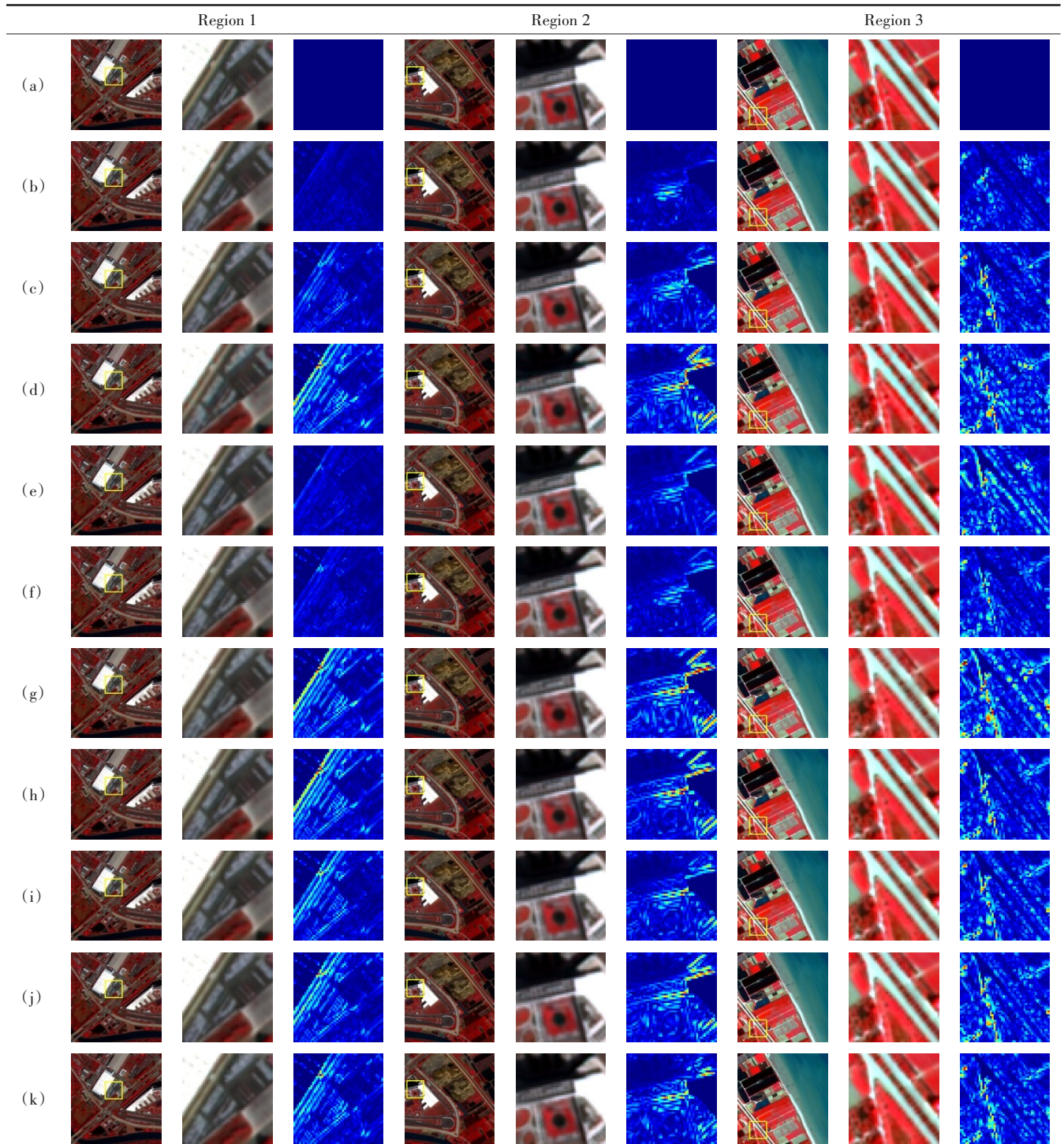


Fig. 5 Visual comparison of different pansharpening methods for three specific regions from the GF2 dataset; (a) Reference image; (b) NMSFusion; (c) ADKNet; (d) BDPN; (e) DCFNet; (f) DRPNN; (g) DiCNN; (h) FusionNet; (i) MSDCNN; (j) PNN; (k) PanNet (For each region, the first column displays the full view, the second column displays the locally magnified view, and the third column displays the corresponding error map)

图5 不同全色融合方法在GF2数据集三个特定区域上的视觉对比:(a) 参考图像;(b) NMSFusion;(c) ADKNet;(d) BDPN;(e) DCFNet;(f) DRPNN;(g) DiCNN;(h) FusionNet;(i) MSDCNN;(j) PNN;(k) PanNet(每个区域的第一列为完整展示,第二列为局部缩放,第三列为相应误差图)

age fusion. For example, by adjusting the input channels of the network and utilizing the ACEM to handle the corresponding resolution differences, this framework holds

the promise of injecting the spatial structures of MS images while preserving the spectral features of hyperspectral images.

Table 4 Average quantitative evaluation results of different methods on 20 test samples from the WV3, QB and GF2 datasets**表4 不同方法在WV3、QB和GF2数据集的20个测试样本上的平均定量评估结果**

		PSNR	SSIM	SAM	ERGAS	CC	RMSE
WV3	NMSFusion	37.566 6	0.974 2	2.891 0	6.034 8	0.983 2	0.013 5
	ADKNet	36.331 5	0.967 2	3.291 2	6.979 4	0.978 7	0.015 5
	BDPN	34.578 0	0.950 1	4.225 4	8.592 5	0.964 2	0.019 0
	DCFNet	36.948 6	0.970 8	3.126 4	6.545 5	0.978 9	0.014 5
	DRPNN	36.384 2	0.966 3	3.376 2	7.014 1	0.976 1	0.015 4
	DiCNN	34.631 3	0.953 9	3.961 5	8.482 6	0.968 2	0.018 8
	FusionNet	35.842 1	0.963 4	3.556 4	7.403 1	0.975 0	0.016 5
	MSDCNN	35.713 3	0.962 5	3.633 2	7.488 8	0.974 0	0.016 7
	PNN	35.458 6	0.960 7	3.735 5	7.709 0	0.972 1	0.017 2
	PanNet	35.549 8	0.961 5	3.665 1	7.651 8	0.973 0	0.017 0
QB	NMSFusion	37.912 5	0.960 6	4.550 8	7.391 4	0.976 1	0.012 9
	ADKNet	33.917 3	0.926 3	5.445 7	11.786 4	0.941 0	0.020 8
	BDPN	34.461 3	0.922 4	6.354 2	11.067 2	0.946 9	0.019 4
	DCFNet	37.708 2	0.959 4	4.667 6	7.552 5	0.975 0	0.013 2
	DRPNN	35.760 6	0.944 9	5.057 8	9.610 5	0.956 8	0.016 7
	DiCNN	33.634 6	0.913 7	6.044 0	12.118 8	0.937 7	0.021 3
	FusionNet	34.487 3	0.927 6	5.559 8	11.073 4	0.946 3	0.019 4
	MSDCNN	33.910 4	0.919 1	5.724 3	11.786 4	0.940 0	0.020 7
	PNN	33.803 9	0.916 5	5.755 9	11.887 8	0.939 8	0.020 9
	PanNet	33.911 5	0.919 7	5.716 5	11.788 8	0.939 8	0.020 7
GF2	NMSFusion	43.276 3	0.982 7	0.714 9	1.282 4	0.994 3	0.007 0
	ADKNet	40.290 4	0.968 0	0.959 2	1.821 9	0.988 8	0.009 9
	BDPN	37.272 9	0.943 1	1.300 0	2.566 8	0.977 9	0.014 0
	DCFNet	41.754 6	0.978 7	0.855 7	1.519 1	0.992 2	0.0083
	DRPNN	40.981 1	0.972 5	0.913 9	1.681 0	0.990 4	0.009 1
	DiCNN	37.364 9	0.946 2	1.213 4	2.544 4	0.978 4	0.013 8
	FusionNet	38.771 1	0.958 9	1.038 5	2.180 4	0.984 3	0.011 8
	MSDCNN	39.145 2	0.960 4	1.042 0	2.083 0	0.985 6	0.011 3
	PNN	38.472 3	0.954 8	1.106 7	2.253 3	0.983 1	0.012 2
	PanNet	39.173 3	0.960 1	1.039 4	2.087 4	0.985 7	0.011 2

Table 5 Comparison of accuracy metrics among different model variants on the WV3 dataset**表5 不同模型变体的精度指标对比(以WV3数据集为例)**

Model variant	PSNR	SSIM	SAM	ERGAS
Baseline	30.995 9	0.893 1	5.552 1	12.970 1
+MSGB	30.996 0	0.893 1	5.551 2	12.970 4
+ARFM	30.995 9	0.893 1	5.550 7	12.970 6
+ACEM	30.996 4	0.893 1	5.553 7	12.969 0
NMSFusion	30.996 3	0.893 1	5.555 6	12.968 4

Table 6 Analysis of the computational complexity and parameter count of different model variants on the WV3 dataset**表6 各个模型变体的计算复杂度与参数量变化分析(以WV3数据集为例)**

Model variant	Added key module	Parameters	Computation	Parameter Change
Baseline		0.282 4	0.238 6	
+MSGB	MSGB	0.162 2	0.134 2	-0.120 2
+ARFM	ARFM	0.172 7	0.153 1	+0.010 5
+ACEM	ACEM	0.172 7	0.153 1	+0.000 0
NMSFusion	IIM	0.172 7	0.153 1	+0.000 0

Table 7 Accuracy comparison of the versions by removing individual modules on the WV3 dataset**表 7 剥离单一模块的消融实验精度对比(以 WV3 数据集为例)**

Model variant	PSNR	SSIM	SAM	ERGAS
-MSGB	30.995 9	0.892 9	5.556 1	12.968 6
-ARFM	30.996 2	0.893 0	5.558 0	12.969 1
-ACEM	30.996 1	0.893 1	5.557 5	12.968 5
NMSFusion	30.996 3	0.893 1	5.555 6	12.968 4

Table 8 Comparison of computational complexity and inference time among different pansharpening methods**表 8 各种全色融合方法的计算复杂度与推理时间对比**

	Parameters (M)	FLOPs (G)	Average inference time (ms)
PNN	0.104	6.833	4.07
PanNet	0.083	5.169	1.80
DRPNN	0.433	28.388	11.71
MSDCNN	0.229	14.960	10.77
BDPN	1.487	60.860	17.18
DiCNN	0.047	3.058	1.18
DCFNet	3.249	61.350	31.50
ADKNet	2.074	135.001	80.43
FusionNet	0.079	5.134	1.71
NMSFusion	1.071	12.726	11.82

References

- [1] Anand S, Sharma R. Pansharpening and spatiotemporal image fusion method for remote sensing[J]. *Engineering Research Express*, 2024, 6(2): 022201.
- [2] Shah V P, Younan N H, King R L. An efficient pan-sharpening method via a combined adaptive PCA approach and contourlets [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2008, 46(5): 1323-1335.
- [3] Dong W, Yang Y, Qu J, et al. Hyperspectral pansharpening via local intensity component and local injection gain estimation [J]. *IEEE Geoscience and Remote Sensing Letters*, 2022, 19: 1-5.
- [4] Xu Y, Smith S E, Grunwald S, et al. Effects of image pansharpening on soil total nitrogen prediction models in south India [J]. *Geoderma*, 2018, 320: 52-66.
- [5] Choi J, Yu K, Kim Y. A new adaptive component-substitution-based satellite image fusion by using partial replacement [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2011, 49(1): 295-309.
- [6] Fang Y, Cai Y, Fan L. SDRCNN: A single-scale dense residual connected convolutional neural network for pansharpening[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023, 16: 6325-6338.
- [7] Ding J, Xu H, Zhou S. CDFAN: Cross-domain fusion attention network for pansharpening[J]. *Entropy*, 2025, 27(6): 567.
- [8] Lian Z, Zhan Y, Zhang W, et al. Recent advances in deep learning-based spatiotemporal fusion methods for remote sensing images[J]. *Sensors*, 2025, 25(4): 1093.
- [9] Anand S, Sharma R. Pansharpening and spatiotemporal image fusion method for remote sensing[J]. *Engineering Research Express*, 2024, 6(2): 022201.
- [10] Aiuzzi B, Alparone L, Baronti S, et al. MTF-tailored multi-scale fusion of high-resolution MS and pan imagery [J]. *Photogrammetric Engineering & Remote Sensing*, 2006, 72 (5) : 591-596.
- [11] Palsson F, Sveinsson J R, Ulfarsson M O. A new pansharpening algorithm based on total variation [J]. *IEEE Geoscience and Remote Sensing Letters*, 2014, 11(1): 318-322.
- [12] Ballester C, Caselles V, Igual L, et al. A variational model for P+XS image fusion [J]. *International Journal of Computer Vision*, 2006, 69(1): 43-58.
- [13] Garzelli A, Nencini F, Capobianco L. Optimal MMSE pan sharpening of very high resolution multispectral images [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2008, 46(1): 228-236.
- [14] Wang Q, Shi W, Atkinson P M. Area-to-point regression kriging for pan-sharpening [J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2016, 114: 151-165.
- [15] Pardo-Igúzquiza E, Chica-Olmo M, Atkinson P M. Downscaling cokriging for image sharpening [J]. *Remote Sensing of Environment*, 2006, 102(1-2): 86-98.
- [16] Masi G, Cozzolino D, Verdoliva L, et al. Pansharpening by convolutional neural networks [J]. *Remote Sensing*, 2016, 8 (7): 594.
- [17] Yang J, Fu X, Hu Y, et al. PanNet: A deep network architecture for pan-sharpening [C]//2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017: 1753-1761.
- [18] Cai J, Huang B. Super-resolution-guided progressive pansharpening based on a deep convolutional neural network [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2021, 59 (6): 5206-5220.
- [19] Yuan Q, Wei Y, Meng X, et al. A multiscale and multidepth convolutional neural network for remote sensing imagery pansharpening [J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2018, 11 (3) : 978-989.
- [20] He L, Rao Y, Li J, et al. Pansharpening via detail injection based convolutional neural networks [J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019, 12(4): 1188-1204.
- [21] Liu X, Liu Q, Wang Y. Remote sensing image fusion based on two-stream fusion network [J]. *Information Fusion*, 2020, 55: 1-15.
- [22] Zhang Y, Liu C, Sun M, et al. Pan-sharpening using an efficient bidirectional pyramid network [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2019, 57(8): 5549-5563.
- [23] Wu X, Huang T Z, Deng L J, et al. Dynamic cross feature fusion for remote sensing pansharpening [C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, 2021: 14667-14676.
- [24] Jung C, Zhou K, Feng J. Fusionnet: Multispectral fusion of RGB and NIR images using two stage convolutional neural networks [J]. *IEEE Access*, 2020, 8: 23912-23919.
- [25] Xu S, Zhang J, Zhao Z, et al. Deep gradient projection networks for pan-sharpening [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, 2021: 1366-1375.
- [26] Liu Q, Zhou H, Xu Q, et al. PSGAN: A generative adversarial network for remote sensing image pan-sharpening [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2021, 59 (12): 10227-10242.
- [27] Ma J, Yu W, Chen C, et al. Pan-GAN: An unsupervised pan-

- sharpening method for remote sensing image fusion[J]. Information Fusion, 2020, 62: 110-120.
- [28] Peng S, Deng L J, Hu J F, et al. Source-adaptive discriminative kernels based network for remote sensing pansharpening [C]//Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. Vienna, Austria: International Joint Conferences on Artificial Intelligence Organization, 2022: 1283-1289.
- [29] Zhou H, Liu Q, Wang Y. PanFormer: A transformer based model for pan-sharpening[C]//2022 IEEE International Conference on Multimedia and Expo (ICME). Taipei, Taiwan: IEEE, 2022: 1-6.
- [30] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. 2017: 5998-6008.
- [31] Liu R, Lehman J, Molino P, et al. An intriguing failing of convolutional neural networks and the CoordConv solution [C]//Advances in Neural Information Processing Systems. 2018: 9605-9616.
- [32] He K, Zhang X, Ren S, et al. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification [C]//2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015: 1026-1034.
- [33] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation [M]//Navab N, Hornegger J, Wells W M, et al. Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015; Vol. 9351. Cham: Springer International Publishing, 2015: 234-241.
- [34] Wei Y, Yuan Q, Shen H, et al. Boosting the accuracy of multi-spectral image pansharpening by learning a deep residual network[J]. IEEE Geoscience and Remote Sensing Letters, 2017, 14(10): 1795-1799.

基于嵌套多尺度融合网络的遥感图像全色融合

赵景超, 纪萍, 王群明*

(同济大学 测绘与地理信息学院, 上海 200092)

摘要:全色融合通过融合多光谱数据与全色图像,生成全色图像空间分辨率下的多光谱图像。现有基于深度学习的全色融合方法难以兼顾局部高频细节提取与全局光谱一致性保真,进而使得融合后的影像在增强空间分辨率时易伴随光谱畸变,而在维持光谱特征时又容易导致空间细节的平滑。同时,全色与多光谱影像的跨分辨率差异会导致网络在融合过程中发生特征错位,进而引发局部边缘重影、色彩溢出等视觉失真现象。为此,本文提出了一种嵌套多尺度融合网络(NMSFusion)。该方法从宏观与微观层面进行协同建模。在微观层级,利用多尺度门控块(MSGB)提取全色图像的局部高频细节;在宏观层级,通过无伪影残差融合模块(ARFM)统筹全局语义,确保融合结果与原始多光谱影像的光谱一致性。此外,该方法引入了自适应坐标编码模块(ACEM),拟缓解不同分辨率特征错位问题。最后,针对传统深层网络在随机权重初始化下易破坏原始色彩分布,导致训练初期面临收敛困难与光谱畸变的问题,本文提出了一种即插即用的恒等初始化机制(IIM)。该机制通过约束网络各层的初始权重参数,强制模型在训练起点的输出为插值后的多光谱影像,从而为模型优化提供了合理的初始状态,有效促进了网络的稳定收敛。在三个遥感数据集上的实验表明,NMSFusion在量化指标与视觉评价上均优于9种主流的全色融合网络。同时,消融实验分析表明,各模块不仅协同提升了模型的融合性能,还保证了网络结构的高效与轻量化。

关键词:遥感图像融合;全色融合;深度学习;多尺度