

Article

Conditional Skipping Mamba Network for Pan-Sharpening

Yunxuan Tang ¹, Huaguang Li ¹ , Peng Liu ¹ and Tong Li ^{2,3,*}

¹ School of Information Science and Engineering, Yunnan University, Kunming 650500, China; yxtang@ynu.edu.cn (Y.T.); huagl@ynu.edu.cn (H.L.)

² College of Big Data, Yunnan Agricultural University, Kunming 650201, China

³ The Key Laboratory for Crop Production and Smart Agriculture of Yunnan Province, Yunnan Agricultural University, Kunming 650201, China

* Correspondence: tli@ynu.edu.cn

Abstract: Pan-sharpening aims to generate high-resolution multispectral (HRMS) images by combining high-resolution panchromatic (PAN) images with low-resolution multispectral (LRMS) data, while maintaining the symmetry of spatial and spectral characteristics. Traditional convolutional neural networks (CNNs) struggle with global dependency modeling due to local receptive fields, and Transformer-based models are computationally expensive. Recent Mamba models offer linear complexity and effective global modeling. However, existing Mamba-based methods lack sensitivity to local feature variations, leading to suboptimal fine-detail preservation. To address this, we propose a Conditional Skipping Mamba Network (CSMN), which enhances global-local feature fusion symmetrically through two modules: (1) the Adaptive Mamba Module (AMM), which improves global perception using adaptive spatial-frequency integration; and (2) the Cross-domain Mamba Module (CDMM), optimizing cross-domain spectral-spatial representation. Experimental results on the IKONOS and WorldView-2 datasets demonstrate that CSMN surpasses existing state-of-the-art methods in achieving superior spectral consistency and preserving spatial details, with performance that is more symmetric in fine-detail preservation.

Keywords: pan-sharpening; Mamba network; adaptive fusion; multi-frequency fusion



Citation: Tang, Y.; Li, H.; Liu, P.; Li, T. Conditional Skipping Mamba Network for Pan-Sharpening. *Symmetry* **2024**, *16*, 1681. <https://doi.org/10.3390/sym16121681>

Academic Editor: Antonio Palacios

Received: 13 November 2024

Revised: 7 December 2024

Accepted: 13 December 2024

Published: 19 December 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The inherent limitations of satellite imaging sensors make it difficult to obtain images that combine high spatial and spectral resolution simultaneously [1]. Nevertheless, the complementary properties of multispectral (MS) and PAN images provide a solution to the challenges in producing HRMS images from satellite systems. Specifically, low spatial resolution MS images offer extensive spectral information across multiple bands, whereas high spatial resolution PAN images exhibit limited spectral variability. Consequently, various pan-sharpening techniques have been developed to fuse MS and PAN images, aiming to generate HRMS images that effectively preserve the spatial resolution of PAN images alongside the spectral characteristics of MS images [2,3].

Traditional pan-sharpening techniques are generally classified into multi-resolution analysis methods [4], model-based optimization approaches [5], and component substitution methods [6]. Multi-resolution analysis methods involve extracting high-frequency spatial details from PAN images to merge with MS images. Model-based optimization approaches conceptualize the pan-sharpening process as a mathematical optimization problem, employing prior knowledge to achieve an optimal fusion outcome. Component substitution methods identify key features from MS and PAN images through reversible projections, subsequently merging them via targeted fusion strategies. Despite their utility, these methods often struggle to capture the global semantic information inherent in MS and PAN images and are typically limited by restrictive, often imprecise, linear assumptions regarding the relationship between spectral and spatial features. Recently, researchers have

leveraged the strong nonlinear fitting capabilities of CNNs to improve pan-sharpening performance [7,8]. However, CNN-based methods are limited in modeling global dependencies [9,10], and include approaches based on frequency-domain features [11], multi-scale feature models [12], deep models using the Transformer framework [9], and prior feature-driven methods [8]. INNformer [10], which utilizes a Transformer module to maximize global information extraction, suffers from increased computational complexity and limited global semantic integrity due to its sliding window-based feature extraction approach. Although MSDDN [13] introduces Fourier transforms to enhance global feature richness, its use of fixed convolution parameters reduces model adaptability. Moreover, approaches using transposed self-attention [14] achieve computational efficiency at the expense of global information and input generalization.

Recently, the Mamba model, based on state-space modeling (SSM), has been proposed as a solution to these challenges, offering nonlinear fitting capabilities with linear computational complexity and robust global perception [15–17]. S2DBPN [18] enhances spatial and spectral details through spatial and spectral back-projection mechanisms; however, it demonstrates instability in scenarios involving nonlinear degradation relationships or insufficient data. Although MPEFNet [19] proposes some innovations in its architecture, it fails to fully consider the influence of MS and PAN features, leading to significant loss of spatial and spectral information. Zhang et al. [20] proposed improving the spatial and spectral quality of remote sensing image fusion through local dissimilarity suppression, multiscale fusion, and feature-guided enhancement. However, the method suffers from high computational complexity and may not fully eliminate the impact of local dissimilarity (LD) in certain specific scenarios.

To address these limitations, we introduce a novel skipping Mamba network (MSMN), specifically tailored for pan-sharpening applications. This model harnesses the distinct characteristics of MS and PAN images by employing multi-scale features and frequency-domain information, facilitating a more robust integration of global and local information. As a result, the CSMN achieves superior fusion of spatial and spectral information across various feature hierarchies. The network introduces Mamba SSM for efficient global information modeling with linear complexity, while the Skipping Skipping Mamba connections preserves multi-level feature information, ensuring the complete transmission of high spatial resolution and spectral details. The AMM and CDMM further enhance multi-modal feature integration and global consistency by incorporating channel features, spatial information, and frequency-domain data. Furthermore, incorporating ADA enables the model to dynamically tailor the feature extraction and fusion processes, effectively addressing the complexity of diverse input features. Experimental evaluations indicate that the proposed network consistently outperforms current state-of-the-art methods on several benchmark datasets, including IKONOS and WorldView-2, particularly excelling in detail preservation and spectral consistency.

Additionally, the model achieves these results with a notable reduction in computational complexity. The complete architecture is depicted in Figure 1.

Our contributions are as follows:

- We propose the skipping Mamba network with hierarchical Mamba connections to preserve original features while integrating complementary PAN and MS information, enhancing spatial detail, spectral consistency, and sensitivity to local variations.
- We introduce the AMM, combining the Mamba state-space model with channel features for adaptive multi-modal feature extraction and improved global perception.
- We present the CDMM, enabling efficient spatial-spectral feature fusion using ADA, boosting fusion robustness.
- We evaluate our method on IKONOS and WorldView-2 datasets, demonstrating significant improvements in both quantitative and qualitative performance.

The structure of this paper is as follows: Section 2 reviews related work in pan-sharpening; Section 3 details our proposed methodology; Section 4 presents experimental validation and results; and Section 5 concludes the study.

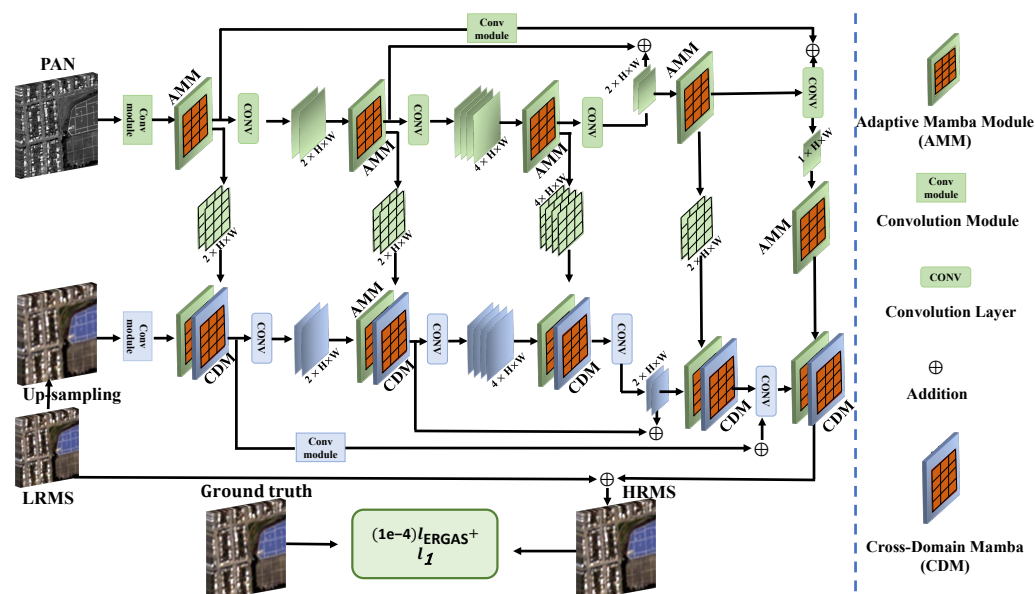


Figure 1. The proposed MSMN architecture features multiple iterative blocks, each containing key sub-blocks: the AMM and the CDM. These components collectively enhance local adaptivity and cross-domain feature integration, improving the overall model’s capability for pan-sharpening tasks.

2. Relate Work

Pan-sharpening techniques are generally divided into traditional approaches and deep learning-based methods. Traditional approaches depend on predefined prior models, primarily encompassing component substitution, variational optimization, and multi-resolution analysis techniques. Component substitution methods, for example, enhance the spatial resolution of the LRMS image by incorporating spatial details from the PAN image [21]. Multi-resolution analysis techniques utilize multi-scale decomposition to handle images of different resolutions and improve detail representation by performing feature fusion in the multi-scale domain [22,23]. Variational optimization methods formulate image fusion as an energy minimization problem and achieve optimal fusion through iterative solutions [24]. While these methods have demonstrated promising results in specific scenarios, they frequently encounter performance limitations with complex image structures, largely due to their heavy reliance on handcrafted feature representations [25].

In recent years, deep learning has made substantial advances in pan-sharpening. The pioneering PNN model, which utilized a straightforward three-layer neural network, achieved notable improvements in fusion performance [26]. Subsequent works adopted more sophisticated network architectures, such as PanNET, which used residual network modules to extract high-frequency details [27], and MSDCNN, which leveraged multi-scale convolutions to capture features from remote sensing images at varying scales [16]. The SRPPNN model further enhanced image resolution and fusion quality through a progressive upsampling strategy [28]. With advancements in self-attention mechanisms, Transformers have been applied to the pan-sharpening field, where models such as INNformer and Panformer have greatly enhanced multi-scale feature extraction and fusion capabilities [29,30]. Moreover, models such as SFINet and MSDDN employed Fourier transforms to capture global feature information, achieving remarkable progress in high-frequency detail learning [18,31]. SSM has recently emerged as a powerful neural architecture for global information modeling across various tasks.

The initial S4 model incorporated state-space structures to effectively capture global dependencies in long sequences [32]. Building on S4, the S5 model optimized computational complexity, enabling more efficient global modeling [33]. The H3 model further refined the architecture, making it competitive with mainstream Transformer models in language modeling tasks [34]. The Mamba model introduced an input-adaptive mechanism, enhancing inference speed and computational efficiency while maintaining performance [33].

Additionally, SSM has shown promise in vision tasks, as demonstrated by Vision Mamba and VMamba, which have exhibited strong performance in image classification and segmentation tasks [16,35]. However, the application of state-space models in multi-modal image fusion, particularly in the pan-sharpening domain, remains largely unexplored.

In response to these challenges, we introduce a novel Conditional spiking Mamba network that fully exploits multi-scale features and frequency-domain information to enhance the fusion of global and local features. This approach achieves effective spatial and spectral complementarity across various feature levels, maintaining low computational complexity while substantially improving accuracy, thus highlighting its potential for advanced pan-sharpening tasks.

3. Method

This section presents a detailed explanation of the proposed CSMN. First, we introduce the theoretical foundation of SSM. Then, we provide an overview of the architecture of the conditional spiking Mamba network, followed by an in-depth discussion on the AMM and the CDMM.

3.1. Preliminaries

3.1.1. State-Space Models

(SSMs) are part of a mathematical framework widely used for sequence-to-sequence modeling [32,36]. These models effectively capture the internal state and dynamic evolution of a system, characterized by a constant temporal behavior known as linear flexibility and expressiveness. An SSM is defined as follows:

$$\begin{aligned} O(s) &= I_{i(s)} + H_{h(s)}, \\ l(\dot{s}) &= \lambda i(s) + \theta h(s). \end{aligned} \quad (1)$$

where $h(s) \in R$ denotes the input, $l(s) \in R$ represents the hidden state. Here, $l(\dot{s})$ indicates the time derivative of $i(s)$, N is the state dimension, and $\lambda \in R^{N \times N}$, $\theta \in R^{N \times 1}$, $I \in R^{1 \times N}$ and $H \in R$ are system matrices.

For discrete sequence tasks, SSMs utilize a zero-order hold (ZOH) discretization method to convert the continuous system parameters λ and θ into discrete parameters $\tilde{\lambda}$ and $\tilde{\theta}$, using a time scale parameter $\delta \in R$. The discretization process is expressed as follows:

$$\tilde{\lambda} = \exp(\delta\lambda) \quad (2)$$

$$\tilde{\theta} = (\delta\lambda)^{-1}(\exp(\delta\lambda) - I) \cdot \delta\theta \approx \delta\theta \quad (3)$$

Based on these parameters, the discrete form of the SSM equation can be represented as follows:

$$\begin{aligned} O(n) &= \tilde{I}i(n) + \tilde{H}h(n), \\ i(n) &= \tilde{\lambda}i(n-1) + \tilde{\theta}h(n). \end{aligned} \quad (4)$$

Typically, the residual connection $\tilde{H}$ is omitted, leading to the following simplified form:

$$O(n) = \tilde{I}i(n) \quad (5)$$

In practice, $i(n)$ is usually a feature vector of size I , and Equation (5) is applied independently to each feature.

3.1.2. Selective Scan Mechanism

Although SSMs are efficient in discrete sequence modeling, their inherent linear time-invariant (LTI) nature limits the model's flexibility to adapt to varying input conditions. Specifically, the LTI property implies that the system parameters remain unchanged regardless of input variations, hindering the model's ability to adjust its dynamics to different in-

puts. This static behavior restricts the model's performance in scenarios involving complex and dynamic sequences, particularly in long sequences and intricate interaction contexts.

To address these limitations, researchers proposed the selective state-space model (S6), also known as the Mamba model [33]. The Mamba model introduces a selective mechanism that enables state-space parameters to be dynamically adjusted based on input characteristics, thereby overcoming the limitations of traditional LTI models. In the Mamba model, matrices $I \in R^{L \times N}$, $\theta \in R^{L \times N}$, and $O \in R^{L \times N}$ are generated from input data $h \in R^{L \times H}$, endowing the model with contextual awareness. This context-aware mechanism allows the Mamba model to respond flexibly to diverse input features, thereby capturing complex interactions more effectively. Compared to traditional SSMs, the Mamba model significantly enhances modeling capabilities in handling complex dynamic systems and diverse sequence data.

3.2. Overall Architecture

As illustrated in Figure 1, the proposed framework consists of an AMM (detailed in Section 3.2.1) and a CDMM (discussed in Section 3.2.2), forming the architecture of the CSMN for pan-sharpening. The AMM combines the Mamba SSM with channel features to achieve adaptive multi-modal feature extraction and fusion, enhancing the network's global perception and generalization capabilities. The CDMM performs a deep fusion of spatial and spectral features and uses ADA to improve the network's responsiveness to different input features, thereby enhancing cross-domain information fusion efficiency.

3.2.1. AMM

To improve the performance of SSMs in feature extraction tasks, we propose an AMM that integrates the dynamic Mamba mechanism with a channel attention mechanism to increase the network's responsiveness to complex inputs. The core design objective of the AMM is to enhance the modeling capability for diverse information in input data through adaptive parameter adjustment and multi-scale feature extraction, as shown in Figure 2.

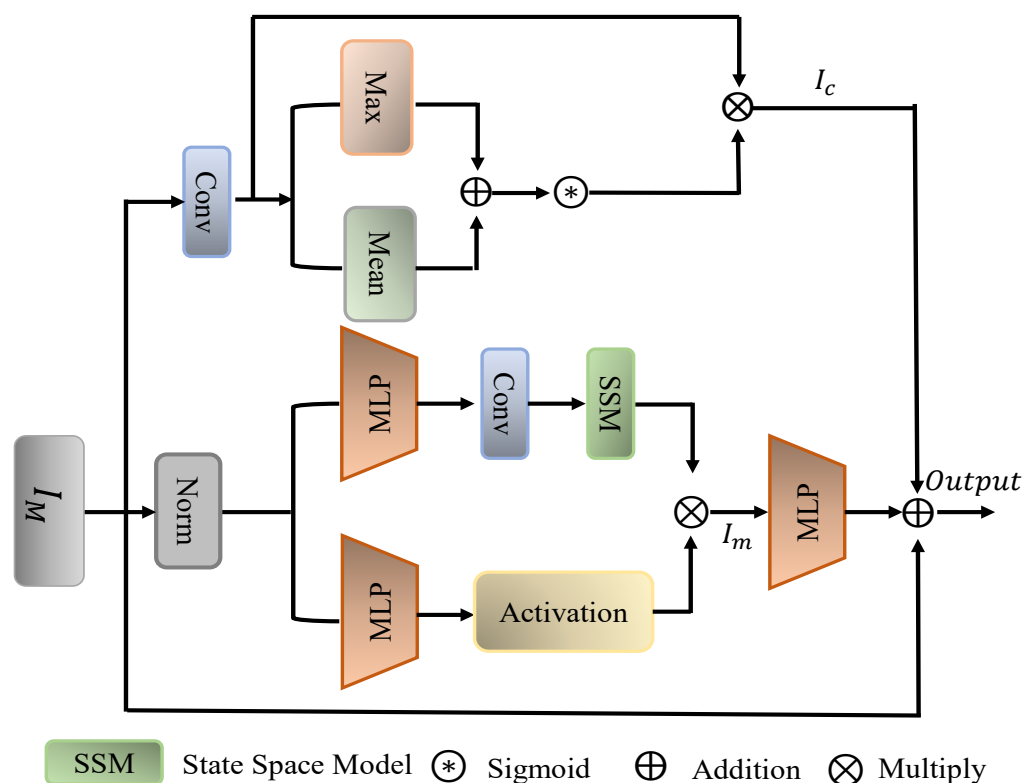


Figure 2. Illustrative breakdown of the components in the AMM.

In this module, the input $I_M \in R^{H \times W \times C}$ is processed using convolutional layers with adaptive max pooling and average pooling to aggregate features, followed by a Sigmoid activation function to generate weighting coefficients, which are used to weight the pooled features. This mechanism automatically enhances the model's focus on key regions, reducing the risk of feature loss due to redundant information. The process can be expressed as follows:

$$I_c = \text{Max}(\text{Conv}(I_M)) + \text{Mean}(\text{Conv}(I_M)) \times \text{Conv}(I_M) \quad (6)$$

where $I_c \in R^{H \times W \times C}$, $\text{Conv}(\cdot)$ denotes convolutional layers, $\text{Max}(\cdot)$ denotes max pooling, and $\text{Mean}(\cdot)$ denotes average pooling. To fully utilize multi-scale ensure stable transmission between different layers.

3.2.2. Cross-Domain Mamba

As shown in Figure 3, the integration of temporal and frequency-domain information within the cross-Mamba block is crucial for enhancing the model's representational capacity, which is achieved through multi-dimensional feature fusion. First, the input is normalized using LayerNorm to reduce distribution bias and improve convergence speed. The normalized input is then processed by the Mamba Module, which operates within the state-space framework, effectively capturing temporal dependencies similar to traditional SSM. According to the reasoning in the AMM, the Mamba output can be expressed as follows:

$$\text{Mamba}(I_{\text{fusion}}) = \text{MLP}(I_{\text{PAN}} + I_{\text{MS}}) \quad (7)$$

here, MLP represents the Multi-head attention mechanism, $I_{\text{PAN}} \in R^{H \times W \times C}$ represents the PAN image, and $I_{\text{MS}} \in R^{H \times W \times C}$ denotes the MS image.

Additionally, the cross-domain Mamba innovatively incorporates frequency-domain processing. Using a 1×1 convolution, the feature maps are projected into the frequency domain to effectively capture the spectral representations of the input features. This enhancement improves the model's ability to identify critical features and mitigates the risk of missing important information that might occur in purely temporal or spatial analyses. To accommodate different scales, the cross-domain Mamba employs adaptive convolutional kernels and performs 3×3 convolution operations to further integrate features at varying resolutions. This multi-scale fusion technique ensures effective capture of both fine-grained and coarse-grained features, enabling a more comprehensive representation of the input data. This can be expressed as follows:

$$\text{ADA}(\text{Conv}(I_S)) = \text{Avg}(\text{Max}(\text{Conv}(I_{\text{fusion}}))) + \text{Fourier}(\text{Conv}(I_{\text{fusion}})) \quad (8)$$

where Fourier represents the Fourier frequency-domain enhancement, Avg denotes average pooling, and Max signifies max pooling. Therefore, based on Equations (9) and (10), the final output of the cross-domain Mamba can be represented as follows:

$$O_{\text{CDM}} = I_{\text{fusion}} + \text{Conv}(I_{\text{PAN}} + I_{\text{MS}}) + I_S \quad (9)$$

where $I_S \in R^{H \times W \times C}$ denotes the frequency-domain enhanced output. Through residual connections, the features from temporal, spatial, and frequency domains are fused with the original input. This design not only preserves the critical information from the initial input but also enhances the representational capability through repeated processing. Thus, the cross-domain Mamba adheres to the principles of state-space model fusion and leverages the richness of multi-dimensional information to achieve superior performance in complex tasks.

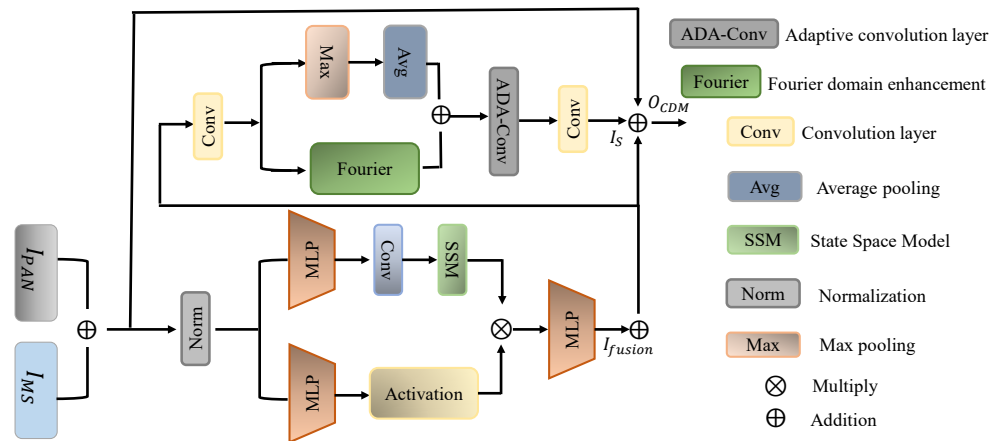


Figure 3. Illustrative breakdown of the components in the CDM.

3.3. Optimization Function

In super-resolution reconstruction tasks, the objective is to generate high-resolution images from low-resolution inputs. To measure the discrepancy between the generated images and the ground-truth high-resolution images, we propose a hybrid loss function that combines the L1 loss and the Erreur Relative Globale Adimensionnelle de Synthèse (ERGAS) loss to enhance performance in super-resolution tasks. The L1 loss measures the absolute difference between the predicted and true values, providing robustness against outliers and effectively promoting the recovery of image details. It is defined as follows:

$$L_{l1} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (10)$$

where y_i denotes the pixel value of the ground-truth high-resolution image, $\hat{y}_i$ represents the pixel value generated by the model, and N is the total number of pixels. To further improve the model's performance, we incorporate the ERGAS loss. The ERGAS metric is specifically designed to evaluate the reconstruction quality of multispectral images, effectively accounting for both spatial and spectral information. The formula is given as follows:

$$L_{ERGAS} = \frac{100}{R} \sqrt{\frac{1}{N} \sum_{i=1}^N \left(\frac{y_i - \hat{y}_i}{y_i} \right)^2} \quad (11)$$

where R is the resolution of the reconstructed image, and y_i and $\hat{y}_i$ are defined as above. During training, if the ERGAS loss is enabled, the final loss function is defined as follows:

$$L_{total} = L_{l1} + \rho L_{ERGAS} \quad (12)$$

where ρ is a hyperparameter for the L_{ERGAS} loss. By adjusting ρ , an optimal balance between the L_{l1} loss and L_{ERGAS} loss can be achieved, promoting comprehensive learning in different aspects of the model.

4. Experiments

4.1. Datasets

To evaluate the effectiveness and robustness of the proposed method, we utilized two widely available datasets, IKONOS and WorldView-2 (WV-2). WV-2 captures eight LRMS bands (red, green, blue, near-infrared 1, coastal, yellow, red edge, and near-infrared 2), along with a high-resolution PAN channel. The spatial resolution ratio between the PAN and LRMS images is 4, with the PAN image having a resolution of 0.5 m, and the LRMS image having a resolution of 2 m. The radiometric resolution is 11 bits. In contrast, the LRMS images from IKONOS consist of four spectral bands: near-infrared, red, blue, and

green. The IKONOS dataset contains 120 training samples and 80 test samples, while the WorldView-2 dataset includes 400 training samples and 100 test samples, allowing for a more comprehensive analysis of spectral information.

Due to the lack of GT data for pan-sharpening fusion [37], we employed the Wald protocol [38] to generate a reduced-resolution dataset. In this approach, MS and PAN images are downsampled by a factor of four to create low-resolution samples suitable for model training, with the original MS images serving as the ground truth. This results in 64×64 low-resolution MS images and 256×256 high-resolution PAN images. Additionally, all full-resolution MS and PAN images are standardized to a resolution of 256×256 . These synthesized datasets provide a robust foundation for model training and performance evaluation, ensuring the reliability and systematic integrity of the experiments.

4.2. Benchmarks and Evaluation Metrics

We conducted a comprehensive evaluation of the proposed method using various benchmarks. To assess the fusion results on the reduced-resolution dataset, we adopted multiple performance metrics, including the global ERGAS [39], the spectral angle mapper (SAM) [40], the spatial correlation coefficient (SCC) [41], and the image quality index (Q) [42]. Specifically, ERGAS quantifies the global relative error of the reconstructed image, while SAM and SCC evaluate the spectral similarity and spatial correlation, respectively, between the fused image and the reference image. The Q index measures image quality by comparing brightness, contrast, and structure between the test and reference images. Additionally, we use the no-reference quality evaluation metric (QNR) [43], expressed as follows:

$$QNR = (1 - D_s)^\alpha (1 - D_\lambda)^\beta \quad (13)$$

where D_λ (spectral distortion) and D_s (spatial distortion) are used to comprehensively evaluate the spectral and spatial fidelity of the full-scale fused image [19]. Some of the quantitative analysis results were based on previous work [44]. For comparative analysis, we evaluate our proposed CSMN against eight advanced techniques. These include traditional methods like Brovey [45] from CS and ATWT-M2 [46] from MRA, as well as deep learning-based methods such as MSDCNN [12], BDPN [47], MUCNN [48], S2DPBN [18], DMLD [20], and MPEFNet [19]. Notably, the first two are traditional image processing techniques, while the remaining six are deep learning-based. All deep learning methods were re-trained on the same training and testing datasets to ensure a fair comparison.

4.3. Implementation Details

The training was conducted on a high-performance computing platform equipped with three NVIDIA A100 40 GB GPUs, an Intel Xeon Gold 5318Y CPU, and 256 GB RAM, running a 64-bit Linux operating system. This platform provides strong computational support for the training and evaluation of deep learning models, making it suitable for handling large-scale datasets and complex models. We used PyTorch 1.11.0 as the deep learning framework, while evaluation metrics were computed using MATLAB R2023a. The optimization process adopted the Adam optimizer with a learning rate of 0.0001, and the model was trained for a total of 400 epochs, with model checkpoints saved every 10 epochs.

4.4. Comparative Analysis

4.4.1. Reduced-Resolution Datasets

Qualitative comparison: To thoroughly validate the fusion performance of the proposed method, Figures 4 and 5 show qualitative analysis results on the IKONOS and WV-2 datasets, respectively. The top rows of the figures display the final outputs of different fusion methods, while the bottom rows illustrate the error distribution maps between these fused results and the reference images. From the comparison, it is evident that traditional fusion methods struggle to balance spectral and spatial details effectively, resulting in either pronounced spectral distortions or excessive enhancement of texture details. This imbalance is particularly evident in complex scenarios, highlighting the limitations of tradi-

tional approaches in handling multi-source data. In contrast, deep learning-based fusion strategies achieve better unification of spectral and spatial information, thereby significantly improving detail preservation and spectral consistency. Compared to existing methods, the proposed approach excels in maintaining spectral and spatial fidelity, preserving the integrity of the original spectral distribution while accurately representing image textures. Further analysis of the error maps reveals that our method produces darker residual images, indicating a smaller discrepancy between the fused and reference images, demonstrating higher similarity and consistency. These results confirm the superior performance of our method in fusion research.

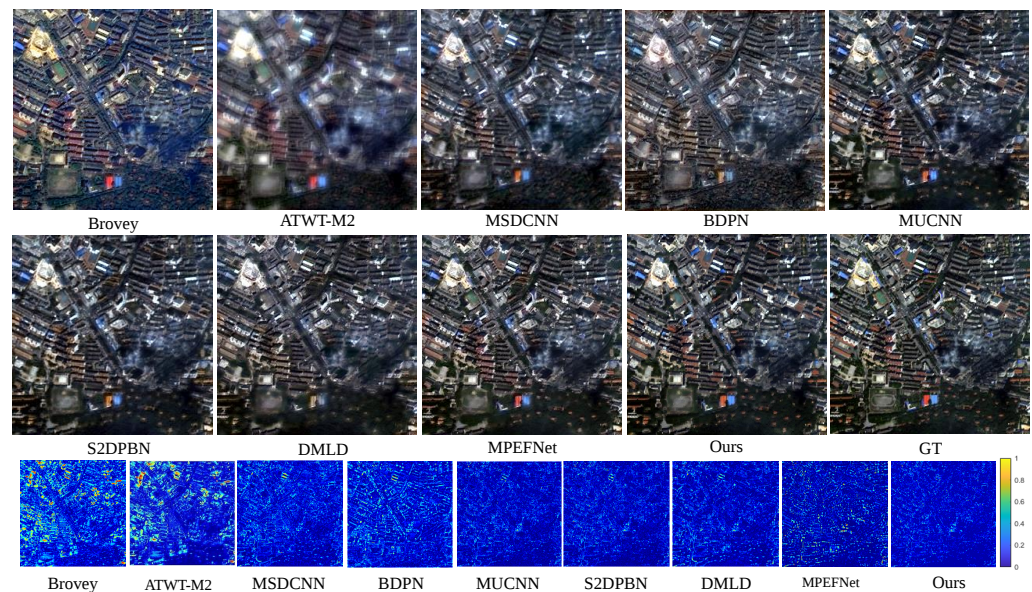


Figure 4. Qualitative results on reduced-resolution IKONOS datasets. The top row shows the fused outputs, while the bottom row depicts the error maps between the fused results and reference images.

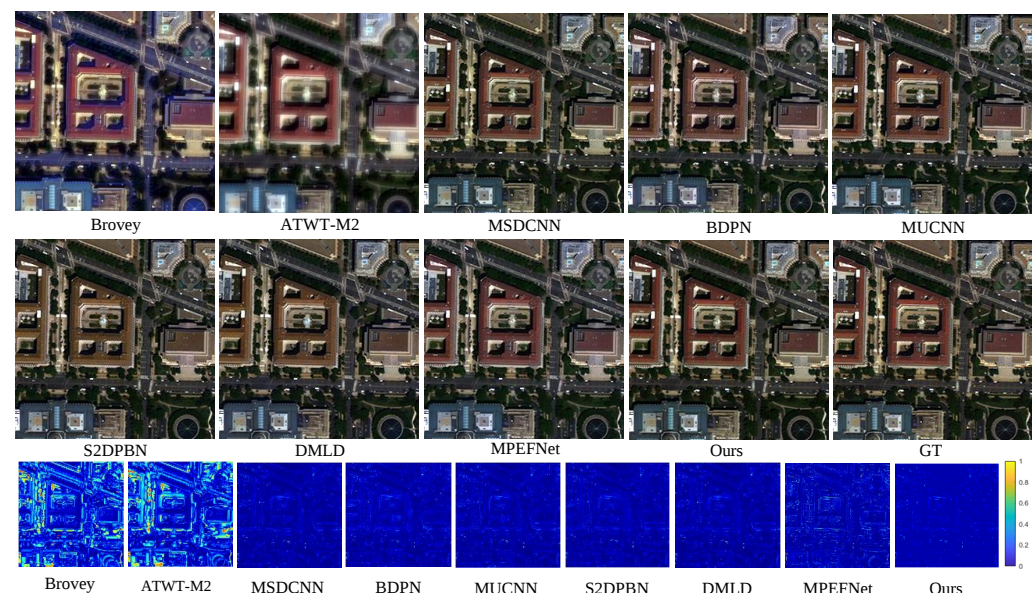


Figure 5. Qualitative results on reduced-resolution WV-2 datasets. The top row shows the fused outputs, while the bottom row depicts the error maps between the fused results and reference images.

Quantitative evaluation: The quantitative results on the two reduced-resolution datasets are presented in Tables 1 and 2. It is clear that the proposed method achieves the best overall performance across all evaluation metrics, showcasing exceptional fusion

capabilities. Specifically, our method demonstrates significant advantages in both spectral fidelity and spatial resolution, surpassing other traditional and deep learning-based methods. This indicates that our approach effectively balances spectral consistency and spatial precision, addressing the complex relationship between these two factors in multi-source image fusion. Additionally, the consistent improvements across various quantitative metrics further validate the adaptability and robustness of the proposed method across different datasets, suggesting strong cross-dataset generalization capabilities. Therefore, compared to other methods, our fusion strategy leads in flexibility, efficiency, and general applicability, making it a promising solution for future multi-source image fusion research.

Table 1. Quantitative comparison of all methods on the IKONOS simulation dataset. arrows indicate preference: $\uparrow$ for larger values, $\downarrow$ for smaller values. The best results are shown in red font.

Methods	Reduced-Resolution				Full-Resolution		
	Q4 $\uparrow$	ERGAS $\downarrow$	SAM $\downarrow$	SCC $\uparrow$	QNR $\uparrow$	$D_s\downarrow$	$D_\lambda\downarrow$
Brovey	0.7347	2.5267	3.4047	0.888	0.7084	0.2143	0.1097
ATWT-M2	0.6919	2.869	3.4583	0.8323	0.7605	0.1559	0.1089
MSDCNN	0.8766	1.6187	2.3738	0.9474	0.8563	0.1071	0.0468
BDPN	0.8434	1.9006	3.0374	0.9277	0.7802	0.1545	0.0802
MUCNN	0.8822	1.5532	2.2227	0.9476	0.8333	0.1026	0.0812
S2DPBN	0.8655	1.6726	2.4063	0.9469	0.8409	0.0924	0.0788
DMLD	0.8560	1.8216	2.6823	0.9397	0.8387	0.1081	0.0694
MPEFNet	0.8737	1.6605	2.3495	0.9427	0.8560	0.1147	0.0370
OURS	0.9507	1.0895	1.5867	0.9624	0.8268	0.0916	0.0928

Table 2. Quantitative comparison of all methods on the WV-2 simulation dataset. Arrows indicate preference: $\uparrow$ for larger values, $\downarrow$ for smaller values. The best results are shown in RED font.

Methods	Reduced-Resolution				Full-Resolution		
	Q8 $\uparrow$	ERGAS $\downarrow$	SAM $\downarrow$	SCC $\uparrow$	QNR $\uparrow$	$D_s\downarrow$	$D_\lambda\downarrow$
Brovey	0.8212	6.3161	7.9286	0.9007	0.8688	0.1088	0.0251
ATWT-M2	0.7234	7.3883	7.9224	0.8382	0.8389	0.1088	0.0587
MSDCNN	0.9605	3.2738	5.1168	0.9632	0.8731	0.094	0.0363
BDPN	0.9483	3.7056	5.8499	0.947	0.8732	0.1005	0.0293
MUCNN	0.9543	3.4941	5.3528	0.9558	0.8709	0.0966	0.036
S2DPBN	0.9587	3.3087	5.1763	0.9619	0.8614	0.0885	0.055
DMLD	0.9552	3.4982	5.3348	0.9581	0.8660	0.1076	0.0296
MPEFNet	0.9527	3.5880	5.4751	0.9538	0.8907	0.0928	0.0181
OURS	0.9637	3.1135	4.9105	0.9643	0.8750	0.0964	0.0312

4.4.2. Full-Scale Datasets

Qualitative comparison: To further validate the fusion performance of the proposed method, Figures 6 and 7 present the qualitative experimental results on two different datasets at full-scale resolution. The comparative analysis reveals that traditional fusion methods show significant limitations in balancing spectral and spatial information, often enhancing one at the expense of the other. Specifically, these methods tend to over-enhance spatial details or cause noticeable spectral distortions, resulting in unnatural image effects. Similarly, some deep learning-based fusion methods fail to address this issue, exhibiting severe spectral artifacts that lead to local distortions in spectral information. Additionally, these deep learning methods lack robustness in spatial detail restoration, resulting in images that are deficient in texture information and lack spatial coherence. In contrast, our proposed model excels by introducing finer spectral and spatial feature extraction and fusion mechanisms, achieving superior performance in maintaining spectral distribution integrity and consistency. Particularly, our fusion results accurately restore the texture structure of the source images and maintain high consistency with the high-resolution PAN image in terms of spatial features. Therefore, the qualitative analysis further confirms the

advantage of our approach in complex scenarios by achieving a better balance between spectral and spatial information, significantly improving the overall quality and visual appeal of the fused images.

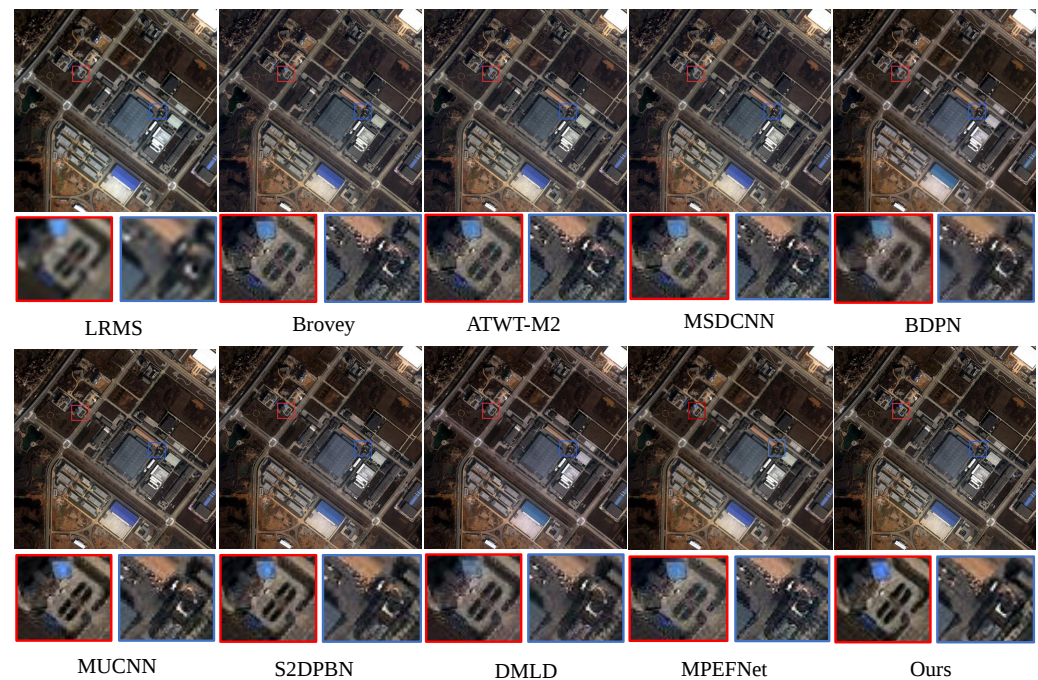


Figure 6. Qualitative analysis for the full-scale evaluation on the IKONOS datasets. The red and blue frames highlight details at different positions within the image.

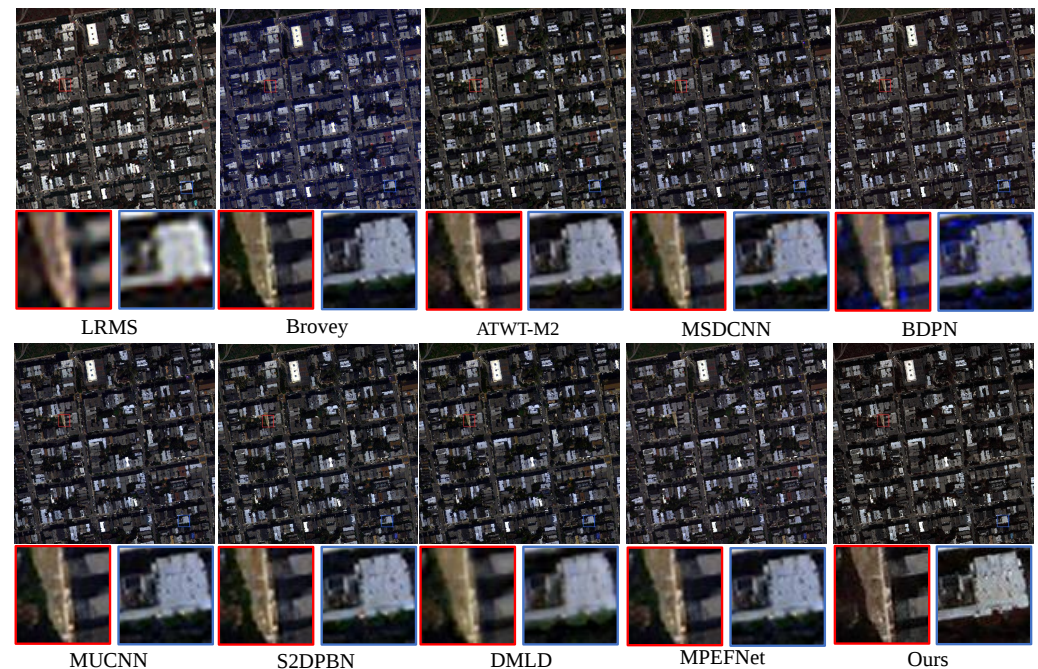


Figure 7. Qualitative analysis for the full-scale evaluation on the WV-2 datasets. The red and blue frames highlight details at different positions within the image.

Quantitative evaluation: To comprehensively evaluate the fusion performance of the proposed method under full-scale conditions, Tables 1 and 2 show the quantitative evaluation results using three no-reference metrics on two datasets. Analysis of the data clearly shows that our method achieves superior performance on the overall QNR (quality with

no reference) metric, significantly outperforming other comparative methods. Moreover, for the spectral distortion metric D_λ and spatial distortion metric D_s , which are crucial parameters for measuring spectral and spatial fidelity, our method effectively achieves a balance between the two, avoiding the biases or trade-offs seen in traditional methods. This indicates that the proposed fusion strategy not only achieves high spectral and spatial consistency in complex scenarios but also exhibits strong stability and adaptability across different datasets. The superior performance on the QNR metric, which simultaneously measures both spectral and spatial fidelity, further validates the superiority of the proposed method in overall fusion quality.

4.4.3. Ablation Study

The ablation study focuses on three key modules to analyze the impact of different strategies on overall model performance. Figure 8 presents the structure comparison of different strategy modules: (a) integrates the AMM with the Mamba framework, (b) combines the CMD with the Mamba framework, and (c) shows the complete architecture with both AMM and CMD modules integrated. Through comparative experiments, we systematically evaluate the contributions of each strategy to model performance. Specifically, Table 3 presents the quantitative results of the AMM module, and Figure 9 shows the visual results of different module combinations. It can be observed that simply combining CMD with the Mamba module yields only minor improvements in certain metrics, and the overall performance remains limited in balancing spectral and spatial fusion. However, when the AMM module is incorporated into the Mamba framework, various metrics show fluctuations. Despite a general trend of improvement, the model fails to achieve optimal accuracy in local feature extraction and maintaining global consistency.

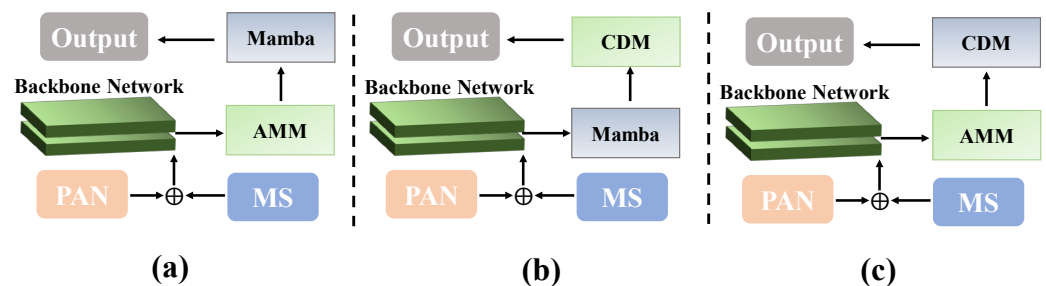


Figure 8. Ablation study on different framework combinations. (a) AMM combined with Mamba, (b) CMD combined with Mamba, (c) our complete model architecture.

Table 3. Mean objective evaluation of different model combinations in the ablation study on the IKONOS (top) and WV-2 (bottom) datasets. $\uparrow$ indicates that higher values are desirable, while $\downarrow$ indicates that lower values are preferred. The best results are shown in RED font.

	Versions	AMM	CMD	Q4 $\uparrow$	ERGAS $\downarrow$	SAM $\downarrow$	SCC $\uparrow$
IKONOS Dataset	(I)	✓	×	0.8785	1.5491	2.2442	0.9536
	(II)	×	✓	0.8759	1.5633	2.2758	0.9492
	Ours	✓	✓	0.9507	1.0895	1.5867	0.9624
	Versions	AMM	CMD	Q8 $\uparrow$	ERGAS $\downarrow$	SAM $\downarrow$	SCC $\uparrow$
WV-2 Dataset	(I)	✓	×	0.963	3.1258	4.889	0.9665
	(II)	×	✓	0.9621	3.1695	4.9491	0.9656
	Ours	✓	✓	0.9637	3.1135	4.9105	0.9643

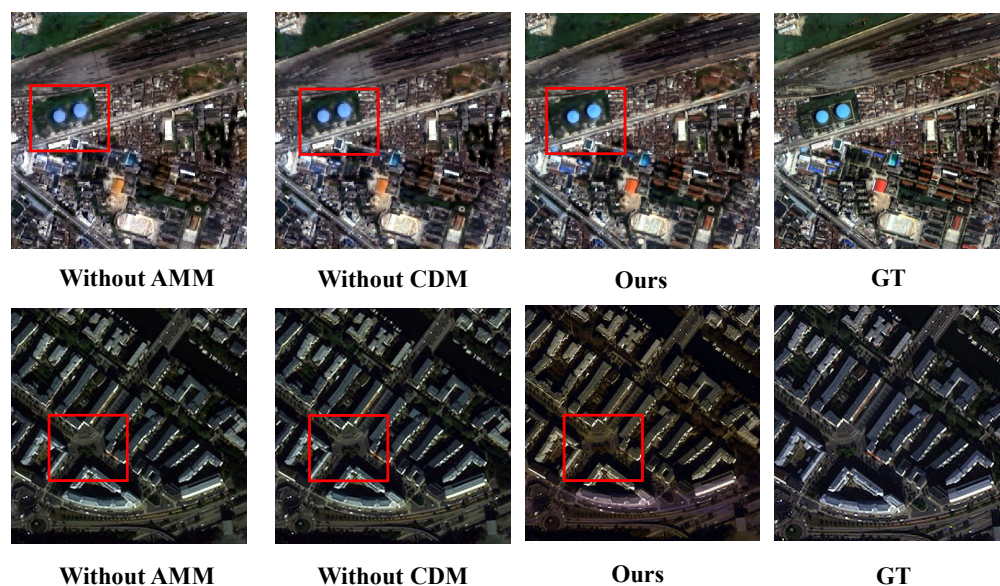


Figure 9. Visual comparison from the ablation study across two datasets. The content within the red box exhibits significant differences.

Further analysis of the combined AMM and CMD modules reveals that this structure effectively enhances feature representation across different scales, significantly optimizing the model's generalization on multi-modal data and achieving a higher balance in spectral and spatial fusion. The effectiveness of this combined strategy is also reflected in the noticeable improvements across various evaluation metrics, demonstrating the strong complementarity of the AMM and CMD modules in fusion strategy design. Thus, in practical applications, we adopted the joint combination of AMM and CMD and achieved the best performance in complex scenarios through further parameter tuning, highlighting the proposed method's effectiveness and superiority in multi-modal data fusion. Overall, the combination of AMM and CMD modules outperforms other strategies in both model accuracy and robustness, making it a critical component of our final solution design.

5. Limitations

Although the CSMN achieves competitive results in pan-sharpening, there remain opportunities for further enhancement. Future research should consider evaluating the model on larger and more diverse datasets to improve its generalization and robustness. Additionally, incorporating other advanced feature extraction methods, such as object detection and semantic segmentation, could extend its applicability to broader multi-modal fusion tasks. Current work mainly focuses on spectral and spatial feature fusion, while temporal and multi-source sensor data have not been fully explored. Integrating these modalities could establish a more comprehensive and robust fusion framework. Moreover, addressing environmental noise and complex backgrounds in remote sensing images by introducing noise suppression and background enhancement mechanisms may further optimize the model's real-world performance.

6. Conclusions

We propose a novel CSMN designed for pan-sharpening, overcoming traditional limitations in global context modeling and local detail preservation. By leveraging the Mamba SSM, CSMN efficiently captures complex long-range dependencies while maintaining linear computational complexity. Two specialized modules, AMM and CDM, are introduced to enhance local detail extraction and cross-domain feature integration. Experimental results on IKONOS and WorldView-2 datasets show that CSMN outperforms state-of-the-art methods across multiple metrics, achieving superior spectral and spatial

fidelity. Qualitative and quantitative evaluations confirm its efficacy in complex scenarios. Ablation studies further reveal the complementary strengths of AMM and CDM, indicating that their combination significantly improves global and local feature fusion. CSMN thus offers a robust and efficient solution for multi-source image fusion with strong potential for real-world applications.

Author Contributions: Conceptualization, Y.T. and H.L.; methodology, T.L.; writing—original draft preparation, P.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Major Special Science and Technology Project of Yunnan Province, no. 202202AE09002105.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhang, B.; Wu, Y.; Zhao, B.; Chanussot, J.; Hong, D.; Yao, J.; Gao, L. Progress and challenges in intelligent remote sensing satellite systems. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 1814–1822. [\[CrossRef\]](#)
2. Casagli, N.; Intrieri, E.; Tofani, V.; Gigli, G.; Raspini, F. Landslide detection, monitoring and prediction with remote-sensing techniques. *Nat. Rev. Earth Environ.* **2023**, *4*, 51–64. [\[CrossRef\]](#)
3. Wald, L.; Ranchin, T.; Mangolini, M. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. *Photogramm. Eng. Remote Sens.* **1997**, *63*, 691–699.
4. Ghahremani, M.; Ghassemian, H. Nonlinear IHS: A promising method for pan-sharpening. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1606–1610. [\[CrossRef\]](#)
5. Sebastian, R. An overview of gradient descent optimization algorithms. *arXiv* **2016**, arXiv:1609.04747.
6. Lu, H.; Yang, Y.; Huang, S.; Tu, W.; Wan, W. A unified pansharpening model based on band-adaptive gradient and detail correction. *IEEE Trans. Image Process.* **2021**, *31*, 918–933. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Zhang, H.; Wang, H.; Tian, X.; Ma, J. P2Sharpen: A progressive pansharpening network with deep spectral transformation. *Inf. Fusion* **2023**, *91*, 103–122. [\[CrossRef\]](#)
8. Xu, S.; Zhang, J.; Zhao, Z.; Sun, K.; Liu, J.; Zhang, C. Deep gradient projection networks for pan-sharpening. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1366–1375.
9. Zhou, H.; Liu, Q.; Wang, Y. PanFormer: A transformer based model for pan-sharpening. In Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, 18–22 July 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6.
10. Zhou, M.; Huang, J.; Fang, Y.; Fu, X.; Liu, A. Pan-sharpening with customized transformer and invertible neural network. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022; Volume 36, pp. 3553–3561.
11. Zhou, M.; Huang, J.; Yan, K.; Yu, H.; Fu, X.; Liu, A.; Wei, X.; Zhao, F. Spatial-frequency domain information integration for pan-sharpening. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Cham, Switzerland, 2022; pp. 274–291.
12. Yuan, Q.; Wei, Y.; Meng, X.; Shen, H.; Zhang, L. A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 978–989. [\[CrossRef\]](#)
13. Laben, C.A.; Brower, B.V. Process for Enhancing the Spatial Resolution of Multispectral Imagery Using Pan-Sharpening. U.S. Patent 6,011,875, 4 January 2000.
14. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; Timofte, R. Swinir: Image restoration using swin transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 1833–1844.
15. He, X.; Cao, K.; Yan, K.; Li, R.; Xie, C.; Zhang, J.; Zhou, M. Pan-mamba: Effective pan-sharpening with state space model. *arXiv* **2024**, arXiv:2402.12192. [\[CrossRef\]](#)
16. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Liu, Y. VMamba: Visual State Space Model. *arXiv* **2024**, arXiv:2401.10166.
17. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv* **2024**, arXiv:2401.09417.
18. Zhang, K.; Wang, A.; Zhang, F.; Wan, W.; Sun, J.; Bruzzone, L. Spatial-spectral dual back-projection network for pansharpening. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5402216. [\[CrossRef\]](#)
19. Li, H.; Nie, R.; Cao, J.; Jin, B.; Han, Y. MPEFNet: Multilevel Progressive Enhancement Fusion Network for Pansharpening. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 9358–9368. [\[CrossRef\]](#)
20. Zhang, K.; Yang, G.; Zhang, F.; Wan, W.; Zhou, M.; Sun, J.; Zhang, H. Learning deep multiscale local dissimilarity prior for pansharpening. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5406015. [\[CrossRef\]](#)
21. Choi, J.; Yu, K.; Kim, Y. A new adaptive component-substitution-based satellite image fusion by using partial replacement. *IEEE Trans. Geosci. Remote Sens.* **2010**, *49*, 295–309. [\[CrossRef\]](#)

22. Schowengerdt, R.A. Reconstruction of multispatial, multispectral image data using spatial frequency content. *Photogramm. Eng. Remote Sens.* **1980**, *46*, 1325–1334.
23. Aiazzi, B.; Alparone, L.; Baronti, S.; Garzelli, A. Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis. *IEEE Trans. Geosci. Remote Sens.* **2002**, *40*, 2300–2312. [[CrossRef](#)]
24. Ballester, C.; Caselles, V.; Igual, L.; Verdera, J.; Rougé, B. A variational model for P+ XS image fusion. *Int. J. Comput. Vis.* **2006**, *69*, 43–58. [[CrossRef](#)]
25. Jia, S.; Zhu, S.; Wang, Z.; Xu, M.; Wang, W.; Guo, Y. Diffused convolutional neural network for hyperspectral image super-resolution. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5504615. [[CrossRef](#)]
26. Peng, S.; Guo, C.; Wu, X.; Deng, L.J. U2net: A general framework with spatial-spectral-integrated double u-net for image fusion. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 3219–3227.
27. Yang, J.; Fu, X.; Hu, Y.; Huang, Y.; Ding, X.; Paisley, J. PanNet: A deep network architecture for pan-sharpening. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5449–5457.
28. Lv, Z.; Zhang, P.; Sun, W.; Benediktsson, J.A.; Lei, T. Novel land-cover classification approach with nonparametric sample augmentation for hyperspectral remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4407613. [[CrossRef](#)]
29. Qiao, S.; Chen, L.C.; Yuille, A. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 10213–10224.
30. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
31. Liao, Z.; Zhang, W.; Chu, Q.; Ding, H.; Hu, Y. Multispectral remote sensing image deblurring using auxiliary band gradient information. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5403418. [[CrossRef](#)]
32. Gu, A.; Goel, K.; Ré, C. Efficiently modeling long sequences with structured state spaces. *arXiv* **2021**, arXiv:2111.00396.
33. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752.
34. Mehta, H.; Gupta, A.; Cutkosky, A.; Neyshabur, B. Long range language modeling via gated state spaces. *arXiv* **2022**, arXiv:2206.13947.
35. He, X.; Yan, K.; Zhang, J.; Li, R.; Xie, C.; Zhou, M.; Hong, D. Multiscale dual-domain guidance network for pan-sharpening. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5403213. [[CrossRef](#)]
36. Gu, A.; Johnson, I.; Goel, K.; Saab, K.; Dao, T.; Rudra, A.; Ré, C. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 572–585.
37. Xie, G.; Nie, R.; Cao, J.; Li, H.; Li, J. A Deep Multi-Resolution Representation Framework for Pansharpening. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5517216. [[CrossRef](#)]
38. Alparone, L.; Wald, L.; Chanussot, J.; Thomas, C.; Gamba, P.; Bruce, L.M. Comparison of pansharpening algorithms: Outcome of the 2006 GRS-S data-fusion contest. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 3012–3021. [[CrossRef](#)]
39. Pushparaj, J.; Hegde, A.V. Evaluation of pan-sharpening methods for spatial and spectral quality. *Appl. Geomat.* **2017**, *9*, 1–12. [[CrossRef](#)]
40. Yuhas, R.H.; Goetz, A.F.; Boardman, J.W. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm. In *Summaries of the Third Annual JPL Airborne Geoscience Workshop*; Volume 1: AVIRIS Workshop; JPL: La Cañada Flintridge, CA, USA, 1992.
41. Wang, Z.; Bovik, A.C. A universal image quality index. *IEEE Signal Process. Lett.* **2002**, *9*, 81–84. [[CrossRef](#)]
42. Garzelli, A.; Nencini, F. Hypercomplex quality assessment of multi/hyperspectral images. *IEEE Geosci. Remote Sens. Lett.* **2009**, *6*, 662–665. [[CrossRef](#)]
43. Alparone, L.; Aiazzi, B.; Baronti, S.; Garzelli, A.; Nencini, F.; Selva, M. Multispectral and panchromatic data fusion assessment without reference. *Photogramm. Eng. Remote Sens.* **2008**, *74*, 193–200. [[CrossRef](#)]
44. Tang, Y.; Li, H.; Xie, G.; Liu, P.; Li, T. Multi-Frequency Spectral-Spatial Interactive Enhancement Fusion Network for Pan-Sharpening. *Electronics* **2024**, *13*, 2802. [[CrossRef](#)]
45. Gillespie, A.R.; Kahle, A.B.; Walker, R.E. Color enhancement of highly correlated images. II. Channel ratio and “chromaticity” transformation techniques. *Remote Sens. Environ.* **1987**, *22*, 343–365. [[CrossRef](#)]
46. Vivone, G.; Alparone, L.; Chanussot, J.; Dalla Mura, M.; Garzelli, A.; Licciardi, G.A.; Restaino, R.; Wald, L. A critical comparison among pansharpening algorithms. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 2565–2586. [[CrossRef](#)]
47. Zhang, Y.; Liu, C.; Sun, M.; Ou, Y. Pan-sharpening using an efficient bidirectional pyramid network. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5549–5563. [[CrossRef](#)]
48. Wang, Y.; Deng, L.J.; Zhang, T.J.; Wu, X. SSconv: Explicit spectral-to-spatial convolution for pansharpening. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, China, 20–24 October 2021; pp. 4472–4480.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.