Cross-Scale Domain Adaptation with Comprehensive Information for Pansharpening

Meiqi Gong[1], **Hao Zhang**[1], **Hebaixu Wang**[1], **Jun Chen**[2], **Jun Huang**[1], **Xin Tian**[1] and **Jiayi Ma**[1*]
[1]Electronic Information School, Wuhan University, Wuhan 430072, China
[2]School of Automation, China University of Geosciences, Wuhan 430074, China
meiqigong@whu.edu.cn, jyma2010@gmail.com

Abstract

Deep learning-based pansharpening methods typically use simulated data at the reduced-resolution scale for training. It limits their performance when generalizing the trained model to the full-resolution scale due to incomprehensive information utilization of panchromatic (PAN) images at the full-resolution scale and low generalization ability. In this paper, we adopt two targeted strategies to address the above two problems. On the one hand, we introduce a cross-scale comprehensive information capture module, which improves the information utilization of the original PAN image through fully-supervised reconstruction. On the other hand, we pioneer a domain adaptation strategy to tackle the problem of low generalization across different scales. Considering the instinct domain gap between different scales, we leverage the maximum mean discrepancy loss and the inherent pixel-level correlations between features at different scales to reduce the scale variance, thus boosting the generalization ability of our model. Experiments on various satellites demonstrate the superiority of our method over the state-of-the-arts in terms of information retention. Our code is publicly available at https://github.com/Meiqi-Gong/SDIPS.

1 Introduction

Low-resolution multi-spectral (MS) images and panchromatic (PAN) images are two distinct modalities of remote sensing images captured by satellites [Zhang *et al.*, 2021; Gong *et al.*, 2023]. The former possess low spatial resolution but high spectral resolution, while the latter exhibit the opposite characteristics. Since single MS images or PAN images limit their applications in downstream tasks such as land cover mapping [Garzelli *et al.*, 2018] and environment monitoring [Zhang *et al.*, 2014], high-resolution multi-spectral (HRMS) images are in high demand. Pansharpening aims to merge HRMS images from MS images and PAN images and has become a hot topic due to low technical cost [Ghassemian, 2016].

Existing pansharpening methods can be classified into two main types: traditional methods and deep learning-based methods. Early research primarily focused on traditional methods. These methods can be further divided into detail injection (DI)-based methods and variation optimization (VO)-based methods. Generally, DI-based methods inject the spatial details from the PAN image to the MS image through image decomposition and component substitution [Tu *et al.*, 2001; Aiazzi *et al.*, 2002]. VO-based methods treat the pansharpening process as an optimization problem and leverage an objective function to find the suitable solution in the optimization scheme [Fu *et al.*, 2019; Li *et al.*, 2013]. In stark contrast to traditional methods, deep learning-based approaches exploit the powerful feature extraction of neural networks to extract features that are friendly for the pansharpening task. They have achieved significantly improved performance compared to traditional algorithms. According to the training ways, deep learning-based methods can be further classified into two categories: supervised methods and unsupervised methods. Most deep learning-based methods belong to the supervised category. For example, the first deep learning-based pansharpening method is PNN [Masi *et al.*, 2016], which is inspired by SRCNN [Gao *et al.*, 2017]. PNN constructs a simple three-layer convolutional neural network framework, achieving impressive performance and high efficiency. Recently, Gong *et al.* [Gong *et al.*, 2022] introduced a method that utilizes the advantages of ConvLSTM in transmitting long-term information to implement effective information transfer across multiple layers and scales, achieving promising results in both spatial and spectral information. All these supervised methods follow the Wald's protocol [Wald *et al.*, 1997] to generate simulated data for training, ignoring the domain gap between different scales. Compared to supervised methods, unsupervised methods do not rely on simulated data while constructing observation models to constrain the network. However, in many cases, observation models are unknown, and their inverse operation has countless solutions. Both of them result in inaccurate results obtained by unsupervised methods. One typical example of unsupervised pansharpening methods is PanGAN [Ma *et al.*, 2020], which employs dual discriminators to continually reduce the distribution distance between the HRMS results and the two source images. To fit the distribution of PAN images, PanGAN exploits a linear weighted

*Corresponding author

strategy. However, the real spectral observation model is not simply based on linear weighted [Jiang *et al.*, 2015].

Although existing pansharpening methods have achieved promising performance, there are still critical issues that require urgent attention. Due to the lack of the ground truth, researchers generally downsample the FMS images (*i.e.*, MS images at the full-resolution scale) and FPAN images (*i.e.*, PAN images at the full-resolution scale) to obtain images at the reduced-resolution scale. The downsampled images are used as training data and FMS images are treated as the ground truth. This strategy has two main limitations. On the one hand, FPAN images are disregarded during training, while they contain the richest spatial detail information. On the other hand, existing methods directly generalize models trained at the reduced-resolution scale to the full-resolution scale, while the domain gap between the two scales can significantly affect their performance at the full-resolution scale. Therefore, addressing these issues is crucial for further advancing the performance of pansharpening methods at the full-resolution scale.

To tackle the above challenges, in this work we propose the cross-**S**cale **D**omain adaptation with comprehensive **I**nformation for **P**an**S**harpening (SDIPS), which addresses the problem from two aspects: cross-scale information utilization and model generalization. Firstly, regarding the information utilization of both reduced-resolution scale and full-resolution scale, we introduce a cross-scale comprehensive information capture (CSCIC) module. Unlike traditional feature extractors limited to be effective at the reduced-resolution scale, the feature extractor of the CSCIC module combines the pansharpening process at the reduced-resolution scale and the reconstruction process at both two scales. Thus, we obtain a robust feature extractor with cross-scale comprehensive information capture, where the "robustness" attributes to the ability to simultaneously utilize information from both scales. Meanwhile, it address the issue of under-utilizing FPAN image information during training. Secondly, to address the model generalization issue at the full-resolution scale, we adopt the cross-scale domain adaptation (CSDA) strategy from the reduced-resolution scale to the full-resolution scale. Specifically, we employ two loss functions to bridge the domain gap between the two scales: the maximum mean discrepancy (MMD) [Gretton *et al.*, 2012] loss and the loss related to pixel-level correlations between the two scales, which captures the inherent prior that the reduced-resolution images are obtained through downsampling the full-resolution images. The MMD loss is treated as a regularization term based on the pixel-level prior. It enables the model to adapt to different scales, thus improving its performance at the full-resolution scale. It is worth noting that, it is the first time to address the generalization issue of pansharpening models across different scales from the perspective of domain adaptation.

The contributions of our work can be summarized as the following two aspects:

- We design a cross-scale comprehensive information capture module to incorporate the FPAN images into the network training process for pansharpening, which improves the utilization of accessible information.
- We adopt a novel domain adaptation perspective to address the model generalization issue across different scales. By leveraging the MMD loss and the inherent pixel-level relationships between the two scales, we reduce the domain gap and improve the performance at the full-resolution scale.

2 Method

As mentioned above, existing methods commonly generalize the model trained at the reduced-resolution scale to the full-resolution scale. The neglected information of the FPAN image and the inherent scale domain lead to undesirable performance at the full-resolution scale. To meet the above challenges, we propose the SDIPS method, which introduces a CSCIC module to incorporate information from both scales, and adopts a CSDA strategy to facilitate the model generalization. In what follows, details about the above modules will be described.

2.1 Cross-Scale Comprehensive Information Capture

As we know, Wald's protocol is a common approach in pansharpening to train the networks in a supervised manner. Specifically, FMS images (I_{FM}) and FPAN images (I_{FP}) are first downsampled to obtain MS images (I_{MS}) and PAN images (I_{PAN}). Consequently, pairs of I_{MS} and I_{PAN} are treated as training data, and the corresponding I_{FM} is considered as the ground truth to realize supervised learning. Apparently, the information of I_{FP} is partly discarded. Some works attempt to compensate for this deficiency by incorporating the pansharpening process at the full-resolution scale during training like $PS(PS(I_{MS}, I_{PAN}), I_{FP})$, where $PS(\cdot)$ represents the pansharpening model. However, the lack of ground truth at the original full-resolution scale limits the exploitation of the FPAN information.

To cope with this challenge, we introduce a CSCIC module to enhance the utilization of information from different scales. As shown in Figure 1, in the pansharpening process at the reduced-resolution scale, the MS encoder block (EMB) and PAN encoder block (EPB) are employed to extract features from I_{MS} and I_{PAN}, respectively, written as:

$$f_{MS} = EMB(I_{MS}), \;\; f_{PAN} = EPB(I_{PAN}). \quad (1)$$

Subsequently, the extracted features f_{MS} and f_{PAN} are concatenated to generate HRMS results I_H, and this process is formalized as:

$$I_H = DPSB(f_{MS}, f_{PAN}), \quad (2)$$

where $DPSB(\cdot)$ represents the decoder block of the pansharpening network.

To address the limitation of existing methods that do not utilize FPAN information, we design a reconstruction network at the full-resolution scale to better utilize comprehensive information at this scale. This process is defined as:

$$\hat{I}_{FM} = DMB(EMB(I_{FM})), \;\; \hat{I}_{FP} = DPB(EPB(I_{FP})), \quad (3)$$

where DMB and DPB are decoder blocks for MS and PAN of the reconstruction network, respectively. Further, to ensure

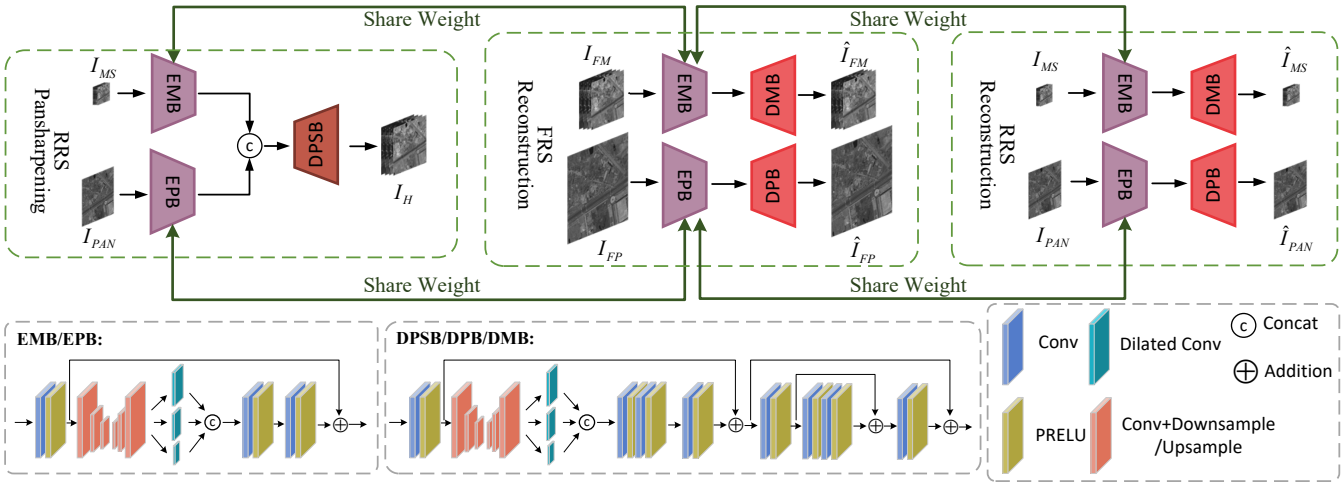

Figure 1: The network of the CSCIC module, including the pansharpening network at the reduced-resolution scale (RRS), the reconstruction network at the full-resolution scale (FRS), and some network details including the encoder blocks and the decoder blocks.

that the features extracted by the reconstruction network and that of the pansharpening network are mutually beneficial, we also add the reconstruction process of the reduced-resolution images, formalized as:

$$\hat{I}_{MS} = DMB(EMB(I_{MS})), \quad \hat{I}_{PAN} = DPB(EPB(I_{PAN})). \tag{4}$$

By sharing parameters between the encoder blocks of the reconstruction network and the pansharpening network, the encoder ensures the perceptual capability of effective features from images on both scales.

Regarding network architectures, the Unet-like structure is employed for better learning of details. Additionally, considering small objects in remote sensing images, we incorporate dilated convolution structure into the network to enlarge the receptive field without altering the resolution. The decoder block of the pansharpening network (*i.e.*, DPSB) and that of the reconstruction network (*i.e.*, DMB and DPB) enjoy the same network architecture (nuances only in input channels) but different parameters due to discrepant tasks.

Obviously, both the pansharpening task and the reconstruction task are fully-supervised. The loss function of the pansharpening network can be formalized as:

$$\mathcal{L}_{ps} = \left\| I_H - I_{FM} \right\|_1 + 1 - SSIM(I_H, I_{FM}), \tag{5}$$

where $SSIM(\cdot)$ means the structure similarity.

Similarly, the loss functions of the reconstruction network are defined as:

$$\begin{aligned} \mathcal{L}_{fs} = &\left\| \hat{I}_{FM} - I_{FM} \right\|_1 + 1 - SSIM(\hat{I}_{FM}, I_{FM}) \\ &+ \left\| \hat{I}_{FP} - I_{FP} \right\|_1 + 1 - SSIM(\hat{I}_{FP}, I_{FP}), \end{aligned} \tag{6}$$

$$\begin{aligned} \mathcal{L}_{rs} = &\left\| \hat{I}_{MS} - I_{MS} \right\|_1 + 1 - SSIM(\hat{I}_{MS}, I_{MS}) \\ &+ \left\| \hat{I}_{PAN} - I_{PAN} \right\|_1 + 1 - SSIM(\hat{I}_{PAN}, I_{PAN}), \end{aligned} \tag{7}$$

where $\mathcal{L}_{fs}$ and $\mathcal{L}_{rs}$ represent the reconstruction loss of the full-resolution scale and the reduced-resolution scale, respectively. Since the pansharpening network and the reconstruction network are jointly trained, the loss of this joint training process can be expressed concisely as:

$$\mathcal{L}_1 = \mathcal{L}_{ps} + \lambda_1 \mathcal{L}_{fs} + \lambda_2 \mathcal{L}_{rs}, \tag{8}$$

where λ_1 and λ_2 are hyper-parameters to control the balance between the pansharpening task on the reduced-resolution scale and the reconstruction task on two scales.

2.2 Cross-Scale Domain Adaptation

In the previous section, we introduce robust encoder blocks that have the perceptual capability to capture effective information at both two scales. However, the domain gap between different scales has not been addressed yet. As is well known, existing deep learning-based methods often perform worse at the full-resolution scale compared to the reduced-resolution scale. This performance degradation is attributed to the domain differences between the reduced-resolution scale and the full-resolution scale. As such, directly generalizing the model trained at the reduced-resolution scale to the full-resolution scale leads to unsatisfactory results.

To address this issue, we pioneer an approach by considering model degradation from the perspective of domain adaptation. If the features extracted from input images at different scales are domain-invariant, the DPSB trained on one scale can be effectively generalized to the other scale. We refer to this process as "cross-scale domain adaptation".

Specifically, as shown in Figure 2, we employ maximum mean discrepancy (MMD) and the pixel-level correlation prior to constrain distributions $P_{rs}(f)$ and $P_{fs}(F)$ of two features, where f and F denote the features from the reduced scale and the full scale, respectively, and $P_{rs}(\cdot)$ and $P_{fs}(\cdot)$ represent the corresponding distributions.

Firstly, we introduce the details of computing the MMD loss. The MMD loss is commonly used in the classification area to narrow the domain discrepancy between the real image domain and the synthetic image domain, thereby achieving model generalization across different domains [Van Opbroek *et al.*, 2018; Zhang *et al.*, 2023; Zhu *et al.*, 2021]. Essentially, the MMD loss is an approach based on kernel techniques. It measures the distribution differences by computing

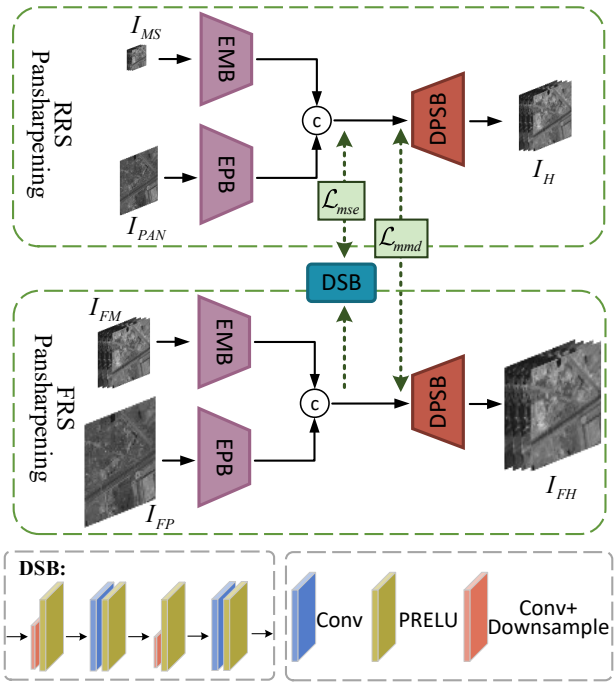

Figure 2: The diagram of cross-scale domain adaptation.

the average distribution discrepancy between samples from two domains in the Hilbert feature space. Specifically, it can be mathematically formalized as:

$$\mathcal{L}_{MMD} = \left\| \frac{1}{N_f} \sum_{i=1}^{N_f} \phi(f_i) - \frac{1}{N_F} \sum_{i=1}^{N_F} \phi(F_i) \right\|_H^2, \quad (9)$$

where N_f and N_F are the number of samples from the reduced-resolution scale domain and the full-resolution scale domain. $\|\cdot\|_H$ means the norm of the Reproducing Kernel Hilbert Space (RKHS), and ϕ represents the mapping from the original feature space to the RKHS.

In existing implementations of the MMD loss, the inputs typically consist of images of the same size [Bonev *et al.*, 2008; Zhu *et al.*, 2019]. However, in this study, we use the MMD loss to measure the similarity of features at different scales with a size ratio of 4. Specifically, to maintain the characteristics of the specific scale domain, we employ a 3×3 window to simultaneously slide over the features of both scales. The features obtained from the sliding window are then averaged to generate samples for computing the MMD loss. The process of generating samples is shown in Figure 3.

Secondly, we describe the prior pixel-level correlation between features from two scales. In the pansharpening task, features from the full-resolution scale can be downsampled to establish a certain pixel-level correspondence with that from the reduced-resolution scale. As such, we additionally introduce a downsampling block (DSB) to utilize this prior relationship, and the MMD loss serves as a regularization term for the domain adaptation purpose. The overall formulation of the CSDA part is defined as:

$$\mathcal{L}_2 = \left\| DSB(F) - f \right\|_2 + \eta \mathcal{L}_{MMD}, \quad (10)$$

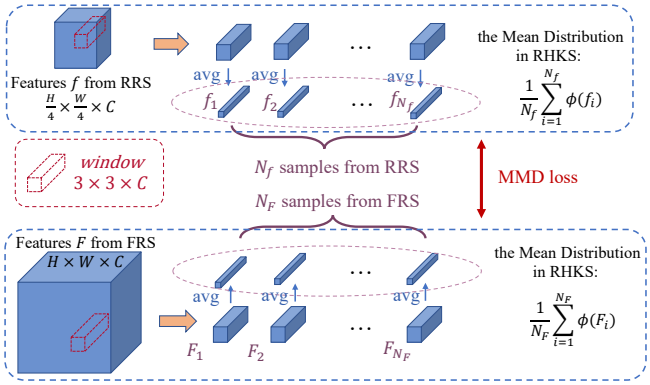

Figure 3: The process of generating samples at different scales for computing the MMD loss.

Algorithm 1: The training process of SDIPS

```
Initialization;
while epoch ≤ 50 do
    if epoch ≤ 40 then
        Input: I_{MS_i}, I_{PAN_i}, i = 1, ..., N
        Output: I_{H_i}
        Update parameters of the pansharpening
          network (EMB, EPB, DPSB) by minimizing
          L_ps with learning rate 5 × 10^{-5}.
    else
        Input: I_{MS_i}, I_{PAN_i}, I_{FM_i}, I_{FP_i}, i=1,...,N
        Output: I_{H_i}, Î_{MS_i}, Î_{PAN_i}, Î_{FM_i}, Î_{FP_i}
        Update parameters of the pansharpening
          network (EMB, EPB, DPSB), reconstruction
          network (EMB, EPB, DMB, DPB), and DSB
          by minimizing L_all with learning rate 10^{-5}.
    end
end
```

where $DSB(\cdot)$ is the downsampling block, the details of which are shown in Figure 2, and η is the regularization parameters to balance the two terms.

In general, the loss function of the whole part can be written as follows:

$$\mathcal{L}_{all} = \mathcal{L}_1 + \sigma \mathcal{L}_2, \quad (11)$$

where σ is the hyper-parameter to ensure the balance.

3 Experiments

3.1 Experimental Settings

Datasets. We conduct experiments on three satellites: QuickBird (QB), GaoFen-2 (GF2) and WorldView-II (WV2). QB and GF2 capture 4-band MS and 1-band PAN images, while WV2 captures 8-band MS and 1-band PAN images. To expand the training dataset, we crop the FMS images to sizes of $80 \times 80 \times 4/80 \times 80 \times 8$, and correspondingly crop the FPAN images to a size of 320×320. Subsequently, all images are downsampled by 4 times following the Wald's protocol to obtain images at the reduced-resolution scale, with sizes of $20 \times 20 \times 4/20 \times 20 \times 8$ for MS images and 80×80 for

Datasets	RRS				FRS					
	QB		GF2		QB			GF2		
Methods	ERGAS↓	SSIM↑	ERGAS↓	SSIM↑	D_λ↓	D_s↓	QNR↑	D_λ↓	D_s↓	QNR↑
CAIS	3.1942	0.9075	2.8914	0.8525	0.0730	0.0450	0.8881	0.2113	0.0556	0.7461
GLP-Reg FS	3.1000	0.8932	2.6133	0.8568	0.0827	0.0502	0.8729	0.1234	0.0343	0.8472
BDSD-PC	2.8668	0.9099	3.1063	0.8220	0.0632	0.0336	0.9080	0.1375	0.0177	0.8475
CDIF	4.5143	0.8071	3.0520	0.8553	0.0267	**0.0154**	**0.9583**	0.0487	**0.0086**	0.9432
PanGAN	3.1219	0.8725	3.1735	0.7735	0.0388	0.0289	0.9348	0.1785	0.0570	0.7749
D2TNet	1.3036	0.9721	0.8085	0.9826	0.0426	0.0254	0.9332	0.0378	0.0132	0.9498
LPPN	1.2654	0.9719	1.0152	0.9713	0.0462	0.0298	0.9260	0.0318	0.0120	0.9567
MMNet	1.2263	0.9731	0.7515	0.9823	0.0343	0.0206	0.9460	0.0354	0.0114	0.9538
STP-SOM	1.4989	0.9619	1.8196	0.9120	0.0267	0.0246	0.9498	0.0426	0.0112	0.9469
Ours	**1.0486**	**0.9825**	**0.6422**	**0.9891**	**0.0262**	0.0191	0.9553	**0.0238**	0.0104	**0.9662**

Table 1: Quantitative results on QB and GF2 datasets. ↑ and ↓ represent the higher the better and the lower the better, respectively. **Bold** represents the best value, and underline represents the sub-optimal value.

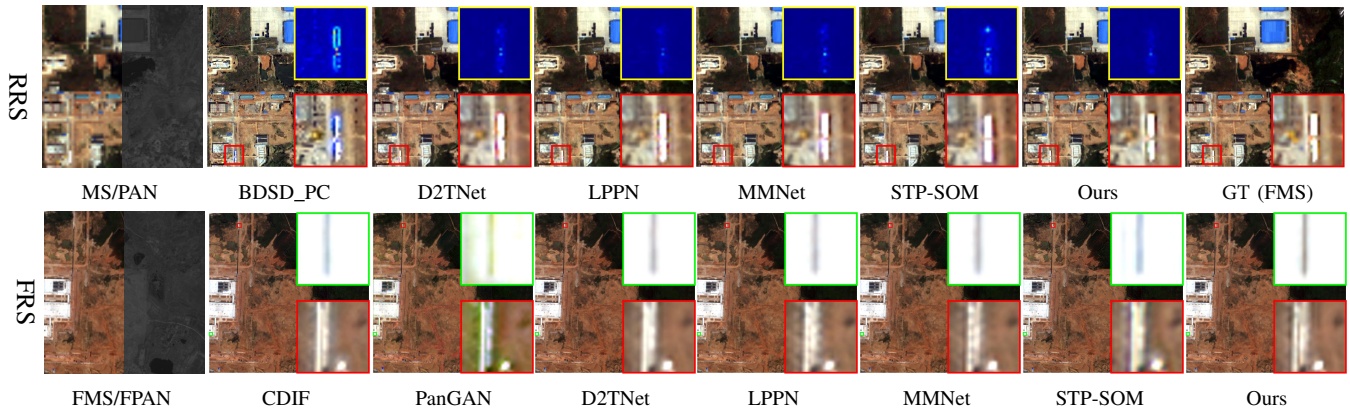

Figure 4: Qualitative comparison on the QB dataset. The yellow box shows the residual map with respect to the ground truth (FMS), and the red/green boxes show the zoomed-in view of the corresponding positions in the result.

PAN images. We obtain a total of 10000 pairs of images for the training set by applying rotations, adding noise, and other means. During testing, for better assessment, MS images at the reduced-resolution scale are of size $66\times66\times4/66\times66\times8$, and PAN images are of size 264×264. At the full-resolution scale, FMS images are of size $264 \times 264 \times 4/264 \times 264 \times 8$, and FPAN images are of size 1056×1056. The testing dataset consists of 200 pairs of images in total.

Training Details. The overall training process is reported in Algorithm 1. We first train the pansharpening network with the learning rate 5e-4 for 40 epochs. The network is optimized by minimizing $\mathcal{L}_{ps}$ using the Adam optimizer with a decay rate of 0.95 every five epochs. After that, the pansharpening network, the reconstruction network and DSB are jointly optimized by minimizing $\mathcal{L}_{all}$ using the Adam optimizer with a decay rate of 0.95 every epoch, while the learning rate is set as 1e-4 and this part continues for 10 epochs. As the pansharpening network and the reconstruction network are equally important to train the EN block, λ_1 and λ_2 are set as 1. η and σ are set as 0.5 and 0.1 to minimize the potential performance degradation of the pansharpening network. α in the PRELU activation function is set as 0.25. Experiments are implemented on the PyTorch platform using a 3.5-GHz Intel Core i9-9920X CPU and NVIDIA Titan RTX GPU.

Comparison Methods. We select nine comparison methods, including four traditional ones (CAIS [Restaino *et al.*, 2017], GLP-Reg FS [Vivone *et al.*, 2018], BDSD-PC [Vivone, 2019] and CDIF [Xiao *et al.*, 2022]) and five deep learning-based ones (PanGAN [Ma *et al.*, 2020], D2TNet [Gong *et al.*, 2022], LPPN [Jin *et al.*, 2022], STP-SOM [Zhang and Ma, 2023] and MMNet [Zhou *et al.*, 2023]). To ensure fairness, all deep learning-based methods are trained using the publicly available code with the same training dataset.

Evaluation Metrics. Since the ground truth exists at the reduced-resolution scale, we use two comprehensive metrics, relative dimensionless global error in synthesis (ERGAS) and structural similarity index measure (SSIM), for quantitative evaluation. At the full-resolution scale, where the ground truth is absent, we employ commonly used quality with no reference (QNR) [Vivone *et al.*, 2014], spectral distortion index (D_λ) and spatial distortion index (D_s) for assessment.

3.2 Experimental Results

QB Dataset. We present quantitative results on the QB dataset in Table 1. At the reduced-resolution scale, our method achieves significantly superior performance on the metrics ERGAS and SSIM, indicating that the proposed

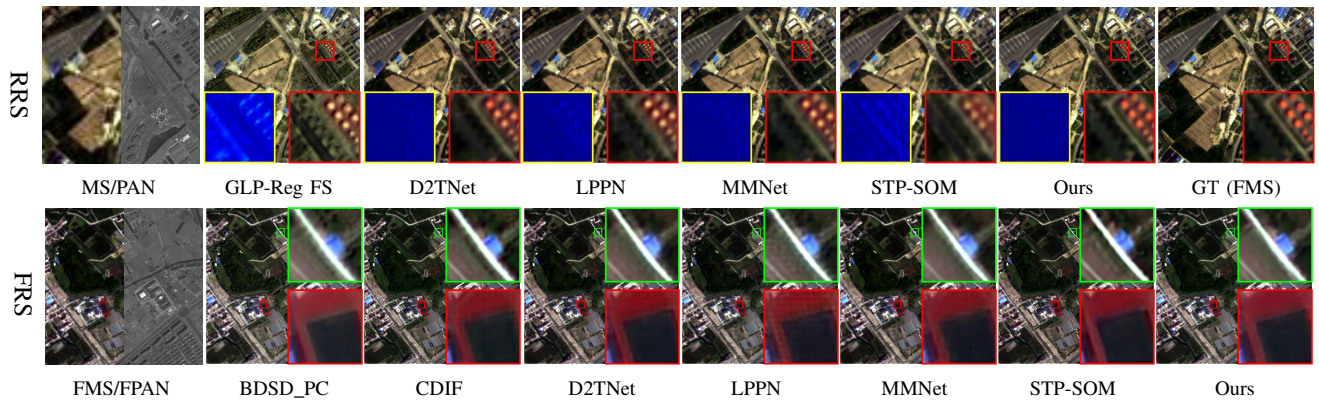

Figure 5: Qualitative comparison on the GF2 dataset.

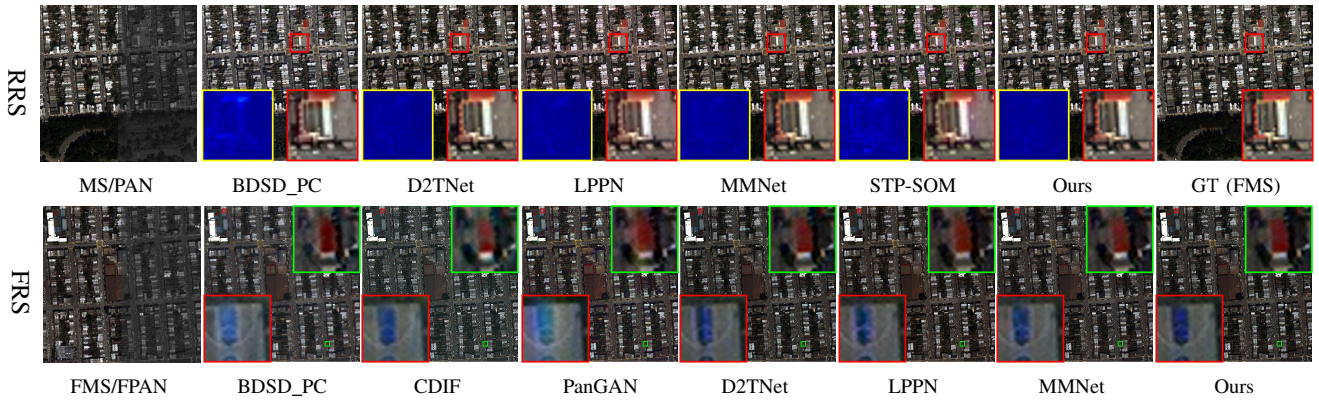

Figure 6: Qualitative comparison on the WV2 dataset.

SDIPS can produce results closer to the ground truth. At the full-resolution scale, our method presents outstanding performance among deep learning-based methods, although slightly inferior to CDIF, which employs some priors friendly to the metrics. It proves the generalization ability of our method across different scales. In contrast, LPPN performs better than D2TNet at the reduced-resolution scale but is inferior to D2TNet at the other scale.

Further, we select the top 5 methods based on metrics at the reduced-resolution scale and the top 6 methods based on metrics at the full-resolution scale, providing qualitative experimental results shown in Figure 4. At the reduced-resolution scale, the enlarged parts in the red box demonstrate that the proposed method exhibits higher consistency in both spectral and spatial aspects with the ground truth. The residual maps further validate this observation. At the full-resolution scale, the enlarged parts demonstrate that our method possesses clearer contours and more normal spectral information. For example, the gap between two lines within the red box is significantly more distinct in our method. The details of the buildings within the green box appear enhanced details and real spectral information, while some methods show blurred details like D2TNet, LPPN, MMNet, and other methods appear spectral distortion like CDIF, PanGAN and STP-SOM.

GF2 Dataset. Quantitative results on the GF2 dataset are also presented in Table 1. At the reduced-resolution scale, our method achieves consistent results with QB, indicating the superior results of the proposed SDIPS. At the full-resolution scale, our method presents significant outstanding performance among all comparable methods, showing performance gains in terms of both spectral distortion and spatial retaining.

Also, qualitative experimental results of the top few methods are provided in Figure 5. At the reduced-resolution scale, the enlarged parts and the residual maps demonstrate that the proposed method exhibits higher consistency in both spectral and spatial aspects with the ground truth. At the full-resolution scale, the enlarged parts demonstrate that our method possesses clearer details and higher contrast. More specifically, in the green box, the proposed SDIPS shows the clearest details, while other methods show blurred details and obvious artifacts. In the red box, the proposed SDIPS shows the most normal spectral information.

WV2 Dataset. Quantitative results on the WV2 dataset are presented in Table 2. Our method exhibits significant advantages both at the reduced-resolution scale and the full-resolution scale. It indicates that despite the simplicity of our network architecture, the strong generalization ability across different scales enables consistent results.

From a visual perspective, Figure 6 presents qualitative results on WV2 dataset. Results at the reduced resolution scale demonstrate the higher consistency of our method with the ground truth. Meanwhile, results at the full resolution

Methods	RRS		FRS		
	ERGAS↓	SSIM↑	D_λ↓	D_s↓	QNR↑
CAIS	5.3448	0.9188	0.0632	0.0612	0.8825
GLP-Reg FS	4.9777	0.9314	0.0612	0.0784	0.8675
BDSD-PC	4.7127	0.9388	0.0416	0.0421	0.9213
CDIF	5.7045	0.9184	0.0432	0.0360	0.9231
PanGAN	6.0751	0.8987	0.0383	0.0352	0.9289
D2TNet	3.2317	0.9821	0.0306	0.0297	0.9440
LPPN	3.4941	0.9753	0.0355	0.0348	0.9359
MMNet	2.9965	0.9848	**0.0254**	0.0287	0.9493
STP-SOM	4.2274	0.9591	0.0387	0.0480	0.9152
Ours	**2.8537**	**0.9870**	0.0263	**0.0248**	**0.9531**

Table 2: Quantitative results on the WV2 dataset.

DataSets		Metrics	PS	+RS	+RS+MSE	+RS+DA
QB	RRS	ERGAS↓	**0.9578**	1.0291	1.0300	1.0486
		SSIM↑	**0.9860**	0.9833	0.9833	0.9825
	FRS	QNR↑	0.9495	0.9519	0.9522	**0.9553**
		D_λ↓	0.0291	0.0276	0.0275	**0.0262**
		D_s↓	0.0221	0.0212	0.0210	**0.0191**
GF2	RRS	ERGAS↓	0.6401	**0.6400**	0.6401	0.6422
		SSIM↑	0.9891	**0.9892**	0.9891	0.9891
	FRS	QNR↑	0.9586	0.9620	0.9632	**0.9662**
		D_λ↓	0.0292	0.0281	0.0254	**0.0238**
		D_s↓	0.0266	0.0126	0.0105	**0.0104**
WV2	RRS	ERGAS↓	**2.5087**	2.6406	2.6732	2.8537
		SSIM↑	**0.9896**	0.9882	0.9864	0.9870
	FRS	QNR↑	0.9444	0.9475	0.9483	**0.9531**
		D_λ↓	0.0313	0.0291	0.0283	**0.0263**
		D_s↓	0.0289	0.0272	0.0266	**0.0248**

Table 3: Ablation study for the CSCIC module and CSDA strategy.

scale show that the proposed SDIPS possesses clearer details. In contrast, other methods exhibit apparent artifacts like CDIF, D2TNet, LPPN, and MMNet, or spectral distortion like BDSD-PC and PanGAN.

3.3 Ablation Studies

In this section, we conduct ablation experiments to study the impact of the specific designs including the CSCIC module and the CSDA strategy on the final results. All ablation experiments are conducted under the same training settings and environment. The same metrics as above are employed to evaluate the performance at two scales. Table 3 reports the corresponding quantitative results. In the table, the column PS, +RS, +RS+MSE and +RS+DA represent results with the CSCIC module, results with the CSCIC module and the downsampling block, results with the CSCIC module and the CSDA strategy, and our proposed results, respectively.

Effectiveness of The CSCIC Module. As shown in the table, it is evident that by incorporating the CSCIC module (the column +RS), the model significantly improves the performance at the full-resolution scale while only incurring a small loss at the reduced-resolution scale. Figure 7 shows qualitative results on the QB dataset and the WV2 dataset. Taking the results on the QB dataset as an example, at the reduced-resolution scale, the final result is similar to that of the ablation results. However, at the full-resolution scale, the result of the original PS displays abnormal artifacts, while the result of the +RS shows normal spectral information. The advantages can be attributed to the powerful ability of information capture at both two scales.

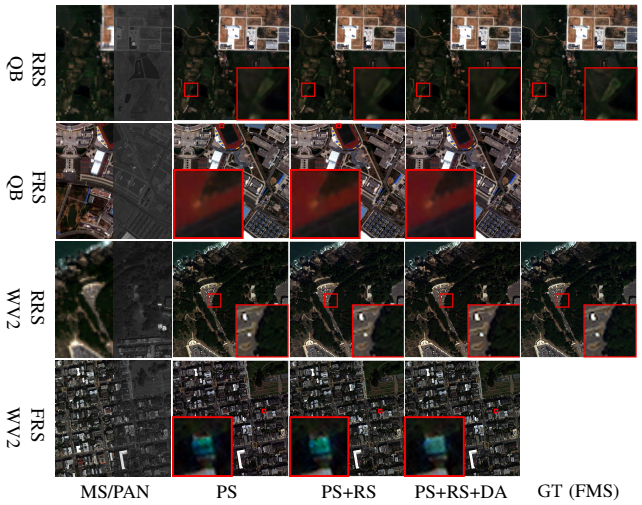

Figure 7: Qualitative comparison of ablation experiments on the QB dataset (the first two rows) and the WV2 dataset (the last two rows).

Effectiveness of The CSDA Strategy. Similarly to the above situation, the addition of the CSDA strategy also achieves certain performance gains at the full-resolution scale as expected, while generating minimal performance loss at the reduced-resolution scale. From Figure 7, further incorporating domain adaptation strategies can not only remove artifacts but also enhance the clarity of the resulting details. The advantages can be attributed to the ability to reduce the domain gap between the two scales.

4 Conclusion

In this paper, we propose an effective pansharpening method SDIPS based on comprehensive information utilization with strong model generalization. On the one hand, through employing a cross-scale comprehensive information capture module, we obtain robust extractors to extract effective features at both scales and achieve comprehensive exploitation of available information. On the other hand, we introduce the cross-scale domain adaptation strategy, employing the inherent pixel-wise relationships between features at two scales as a prior and the MMD loss as a regulation term to reduce the scale domain gap. This strategy improves the ability of model generalization. Overall, our method effectively addresses the shortcomings of under-utilization of the original PAN information and low ability of model generalization across different scales. The experiments conducted on various datasets demonstrate the superiority of our method and validate the effectiveness of the proposed designs.

Ethical Statement

There are no ethical issues.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62276192).

References

[Aiazzi *et al.*, 2002] Bruno Aiazzi, Luciano Alparone, Stefano Baronti, and Andrea Garzelli. Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 40(10):2300–2312, 2002.

[Bonev *et al.*, 2008] Boyan Bonev, Francisco Escolano, and Miguel Cazorla. Feature selection, mutual information, and the classification of high-dimensional patterns: Applications to image classification and microarray data analysis. *Pattern Analysis and Applications*, 11:309–319, 2008.

[Fu *et al.*, 2019] Xueyang Fu, Zihuang Lin, Yue Huang, and Xinghao Ding. A variational pan-sharpening with local gradient constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10265–10274, 2019.

[Gao *et al.*, 2017] Yunxing Gao, Hengjian Li, Jiwen Dong, and Guang Feng. A deep convolutional network for medical image super-resolution. In *Proceedings of the Chinese Automation Congress*, pages 5310–5315, 2017.

[Garzelli *et al.*, 2018] Andrea Garzelli, Bruno Aiazzi, Luciano Alparone, Simone Lolli, and Gemine Vivone. Multispectral pansharpening with radiative transfer-based detail-injection modeling for preserving changes in vegetation cover. *Remote Sensing*, 10(8):1308, 2018.

[Ghassemian, 2016] Hassan Ghassemian. A review of remote sensing image fusion methods. *Information Fusion*, 32:75–89, 2016.

[Gong *et al.*, 2022] Meiqi Gong, Jiayi Ma, Han Xu, Xin Tian, and Xiao-Ping Zhang. D2tnet: A convlstm network with dual-direction transfer for pan-sharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

[Gong *et al.*, 2023] Meiqi Gong, Hao Zhang, Han Xu, Xin Tian, and Jiayi Ma. Multipatch progressive pansharpening with knowledge distillation. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.

[Gretton *et al.*, 2012] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[Jiang *et al.*, 2015] Yiyong Jiang, Xinghao Ding, Delu Zeng, Yue Huang, and John Paisley. Pan-sharpening with a hyper-laplacian penalty. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 540–548, 2015.

[Jin *et al.*, 2022] Cheng Jin, Liang-Jian Deng, Ting-Zhu Huang, and Gemine Vivone. Laplacian pyramid networks: A new approach for multispectral pansharpening. *Information Fusion*, 78:158–170, 2022.

[Li *et al.*, 2013] Shutao Li, Haitao Yin, and Leyuan Fang. Remote sensing image fusion via sparse representations over learned dictionaries. *IEEE Transactions on Geoscience and Remote Sensing*, 51(9):4779–4789, 2013.

[Ma *et al.*, 2020] Jiayi Ma, Wei Yu, Chen Chen, Pengwei Liang, Xiaojie Guo, and Junjun Jiang. Pan-gan: An unsupervised pan-sharpening method for remote sensing image fusion. *Information Fusion*, 62:110–120, 2020.

[Masi *et al.*, 2016] Giuseppe Masi, Davide Cozzolino, Luisa Verdoliva, and Giuseppe Scarpa. Pansharpening by convolutional neural networks. *Remote Sensing*, 8(7):594, 2016.

[Restaino *et al.*, 2017] Rocco Restaino, Mauro Dalla Mura, Gemine Vivone, and Jocelyn Chanussot. Context-adaptive pansharpening based on image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2):753–766, 2017.

[Tu *et al.*, 2001] Te-Ming Tu, Shun-Chi Su, Hsuen-Chyun Shyu, and Ping S Huang. A new look at ihs-like image fusion methods. *Information Fusion*, 2(3):177–186, 2001.

[Van Opbroek *et al.*, 2018] Annegreet Van Opbroek, Hakim C Achterberg, Meike W Vernooij, and Marleen De Bruijne. Transfer learning for image segmentation by combining image weighting and kernel learning. *IEEE Transactions on Medical Imaging*, 38(1):213–224, 2018.

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

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

[Vivone, 2019] Gemine Vivone. Robust band-dependent spatial-detail approaches for panchromatic sharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 57(9):6421–6433, 2019.

[Wald *et al.*, 1997] Lucien Wald, Thierry Ranchin, and Marc Mangolini. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. *Photogrammetric Engineering and Remote Sensing*, 63(6):691–699, 1997.

[Xiao *et al.*, 2022] Jin-Liang Xiao, Ting-Zhu Huang, Liang-Jian Deng, Zhong-Cheng Wu, and Gemine Vivone. A new context-aware details injection fidelity with adaptive coefficients estimation for variational pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022.

[Zhang and Ma, 2023] Hao Zhang and Jiayi Ma. Stp-som: Scale-transfer learning for pansharpening via estimating spectral observation model. *International Journal of Computer Vision*, 131(12):3226–3251, 2023.

[Zhang *et al.*, 2014] Lefei Zhang, Liangpei Zhang, Dacheng Tao, Xin Huang, and Bo Du. Hyperspectral remote sensing

image subpixel target detection based on supervised metric learning. *IEEE Transactions on Geoscience and Remote Sensing*, 52(8):4955–4965, 2014.

[Zhang *et al.*, 2021] Hao Zhang, Han Xu, Xin Tian, Junjun Jiang, and Jiayi Ma. Image fusion meets deep learning: A survey and perspective. *Information Fusion*, 76:323–336, 2021.

[Zhang *et al.*, 2023] Yuxiang Zhang, Wei Li, Mengmeng Zhang, Ying Qu, Ran Tao, and Hairong Qi. Topological structure and semantic information transfer network for cross-scene hyperspectral image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 34(6):2817–2830, 2023.

[Zhou *et al.*, 2023] Man Zhou, Keyu Yan, Jinshan Pan, Wenqi Ren, Qi Xie, and Xiangyong Cao. Memory-augmented deep unfolding network for guided image super-resolution. *International Journal of Computer Vision*, 131(1):215–242, 2023.

[Zhu *et al.*, 2019] Yongchun Zhu, Fuzhen Zhuang, Jindong Wang, Jingwu Chen, Zhiping Shi, Wenjuan Wu, and Qing He. Multi-representation adaptation network for cross-domain image classification. *Neural Networks*, 119:214–221, 2019.

[Zhu *et al.*, 2021] Yongchun Zhu, Fuzhen Zhuang, Jindong Wang, Guolin Ke, Jingwu Chen, Jiang Bian, Hui Xiong, and Qing He. Deep subdomain adaptation network for image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 32(4):1713–1722, 2021.