Domain-irrelevant Feature Learning for Generalizable Pan-sharpening

Yunlong Lin†
School of Informatics, Xiamen
University, Xiamen, China
lyl047853860@gmail.com

Zhenqi Fu†
School of Informatics, Xiamen
University, Xiamen, China
fuzhenqi@stu.xmu.edu.cn

Ge Meng
School of Informatics, Xiamen
University, Xiamen, China
mengg@stu.xmu.edu.cn

Yingying Wang
Institute of Artificial Intelligence,
Xiamen University, Xiamen, China
wangyingying7@stu.xmu.edu.cn

Yuhang Dong
School of Informatics, Xiamen
University, Xiamen, China
dongyh@stu.xmu.edu.cn

Hedeng Yu
School of Informatics, Xiamen
University, Xiamen, China
yuhedeng@stu.xmu.edu.cn

Xinghao Ding†
School of Informatics, Xiamen
University, Xiamen, China
dxh@xmu.edu.cn

ABSTRACT
Pan-sharpening aims to spatially enhance the low-resolution multispectral image (LRMS) by transferring high-frequency details from a panchromatic image (PAN) while preserving the spectral characteristics of LRMS. Previous arts mainly focus on how to learn a high-resolution multispectral image (HRMS) on the i.i.d. assumption. However, the distribution of training and testing data often encounters significant shifts in different satellites. To this end, this paper proposes a generalizable pan-sharpening network via domain-irrelevant feature learning. On the one hand, a structural preservation module (STP) is designed to fuse high-frequency information of PAN and LRMS. Our STP is performed on the gradient domain because it consists of structure and texture details that can generalize well on different satellites. On the other hand, to avoid spectral distortion while promoting the generalization ability, a spectral preservation module (SPP) is developed. The key design of SPP is to learn a phase fusion network of PAN and LRMS. The amplitude of LRMS, which contains ‘satellite style’ information is directly injected in different fusion stages. Extensive experiments have demonstrated the effectiveness of our method against state-of-the-art methods in both single-satellite and cross-satellite scenarios. Code is available at: https://github.com/LYL1015/DIRFL.

CCS CONCEPTS
• Computing methodologies → Hyperspectral imaging

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM ’23, October 29–November 3, 2023, Ottawa, ON, Canada.
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0108-5/23/10...$15.00
https://doi.org/10.1145/3581783.3611894

1 INTRODUCTION
High-resolution multispectral images (HRMS) can be beneficial in a broad range of remote sensing tasks such as military systems, environmental monitoring, mapping services, and scene interpretation. Nevertheless, due to technological and physical limitations, satellites will often only capture a high-resolution panchromatic image (PAN) and a low-resolution multispectral image (LRMS). To obtain
HRMS, various pan-sharpening techniques have been developed to fuse LRMS and PAN [48, 57, 62, 63].

Traditional pan-sharpening methods adopt component substitution, multi-resolution analysis, and variational approaches to transform spatial details from PAN to LRMS. Due to the inaccessibility of sensor characteristics and improper modeling of prior knowledge, traditional methods commonly fail to restore precise spatial and spectral details of HRMS. Recently, deep convolutional neural networks have been introduced for pan-sharpening and have shown significant progress. The pioneering one refers to PNN [33], which adapts a three-layer convolution operation to directly learn specific features and thus significantly improve the generalization capability of the network. As a result, the proposed method can avoid learning domain-specific features from large quantities of annotated data, they cannot always generalize well when the input distribution changes. Unfortunately, most existing pan-sharpening networks train and evaluate the model on the same satellite dataset. These solutions are susceptible to the domain shift issue because the distribution of LRMS can be significantly different. For verification, we show an example of the domain shift issue in Figure 1, where all methods are trained on WV3 and evaluated on both WV3 and GF2. As can be seen, state-of-the-art (SOTA) approaches suffer from sharp performance degradation when their models are directly applied to new satellite data. Therefore, the generalization issue of deep learning-based pan-sharpening methods poses challenges to practical applications.

In this paper, we propose a novel pan-sharpening framework via learning domain-irrelevant representations to promote the generalization capability of the network. The proposed model consists of three distinct designs, i.e., a structural preservation module (STP), a spectral preservation module (SPP), and a complementary information fusion module (CIF). Specifically, STP aims at learning spatial structure information. STP is performed in the gradient domain to fuse the gradient features and enforce structural consistency between PAN and MS. Since gradient features are naturally consistent across different domains, STP is insensitive to the domain shift. To preserve spectral information, we propose SPP, which first decouples the phase and amplitude of PAN and MS in the frequency domain. Then, phase features are fused to further enhance the structural consistency. Note that SPP does not use the amplitude feature of PAN because it is uninformative for pan-sharpening. The amplitude of MS contains domain-specific features, which are directly injected into the network in different fusion stages. The final prediction of our method is obtained via a simple fusion module (i.e., CIF), which integrates the complementary information of structure and spectra. Note that CIF does not learn any amplitude features of MS. As a result, the proposed method can avoid learning domain-specific features and thus significantly improve the generalization ability.

In summary, the contributions of this work are as follows:

- We propose a novel pan-sharpening framework that improves the generalization capability of CNN-based fusion networks. To the best of our knowledge, this is the first attempt to improve the generalization capability for pan-sharpening.
- To learn generalized features, we use gradient and phase features to enforce structural consistency. Besides, the amplitude features of LRMS are directly injected into the network in different fusion stages to avoid learning domain-specific features.
- Extensive experiments on different satellite datasets demonstrate that the proposed method achieves SOTA performance in both single-satellite and cross-satellite scenarios.

2 RELATED WORK

2.1 Classic pan-sharpening methods

Classic pan-sharpening methods can be roughly classified into three categories, including component substitution (CS) [5, 10, 35], multi-resolution analysis (MRA) [12, 36, 38], and variational optimization (VO) [11, 14]. PCA [29], IHS [5] and Brovey [19] are three typical CS methods. These methods improved the spatial resolution by projecting the MS image into a new space and replacing the spatial information with the PAN image. To solve the spectral distortion problems, MRA used multi-resolution decomposition methods such as laplacian pyramid [37] and decimated wavelet transform [32] to extract the spatial information of the PAN image, and then inject it into the up-sampled MS images. VO methods [2, 25] designed specific optimization functions based on various prior assumptions for pan-sharpening. Overall, classic pan-sharpening approaches rely on hand-crafted features. These methods often generate spatial and spectral distortions due to the limited representation ability of the applied priors.

2.2 Deep learning based methods

With the highly nonlinear mapping capability of deep convolutional neural networks (CNN), numerous researchers have explored the use of this technology for image restoration [7–9, 16, 17, 30, 31, 51–53], hyperspectral images [4, 15, 23, 24] and remote sensing images [47, 56, 58, 59]. Recently, the paradigm of pan-sharpening has gradually shifted to data-driven approaches based on deep learning. For example, Masi et al. [33] were the first to apply CNN to address the problem of pan-sharpening and achieved a significant improvement by comparison with the classical methods. Yang et al. [49] adopted the resblock [20] and trained the network in the high-frequency domain. The LRMS image is directly added to the network output. This method can generalize to new satellites to a certain degree. However, the spectral knowledge of LRMS is not exploited. Yuan et al. [54] added the multi-scale module into the fundamental CNN architecture, and Cai et al.[3] also refers to the design idea of the single-image super-resolution network SRCNN [13]. Wu et al. [43] deployed many parallel branches to continuously integrate features with varied sizes into the network’s backbone. Besides, some model-driven methods with physical constraints have been presented. Xie et al. [44] and Xu et al. [46] first used prior knowledge to formulate optimization problems for pan-sharpening task. Then the authors replaced the steps in the algorithm with deep neural networks. Zhou et al. [60, 61] trained a pan-sharpening network in both spatial and frequency domains.
Figure 2: The overall framework of our proposed pan-sharpening framework. It consists of three distinct designs: 1) structural preservation module (STP), 2) spectral preservation module (SPP), and 3) complementary information fusion module (CIF). Specifically, STP is performed in the gradient domain to fuse the gradient features and enforce structural consistency between PAN and LRMS. To preserve spectral information, we propose SPP, which first decouples the phase and amplitude of PAN and LRMS in the Fourier domain. Then, phase features are fused to further enhance structural consistency. Meanwhile, the amplitude of LRMS contains domain-specific features, which is directly injected into the network at different stages. Finally, CIF integrates the complementary of structure and spectra to output the prediction.

3 METHODS
The goal of pan-sharpening is to combine the complementary information of PAN ($P \in \mathbb{R}^{H \times W \times 1}$) and LRMS ($LM \in \mathbb{R}^{H/4 \times W/4 \times C}$) image to generate HRMS ($HM \in \mathbb{R}^{H \times W \times C}$) image, where $H$ and $W$ are the height and width of the image, $C$ refers to spectral bands. The ratio of spatial resolution between PAN and the corresponding LRMS is equal to 4, i.e., $r = 4$. The overall framework of our proposed method is presented in Figure 2. It consists of three distinct designs: 1) structural preservation module, 2) spectral preservation module, and 3) complementary information fusion module.

3.1 Motivation
Deep learning-based approaches have achieved notable progress in pan-sharpening. However, those solutions commonly suffer from a lack of generalization ability. As shown in Figure 1, deep neural networks trained on WV3 perform poorly when evaluated on other satellites. We argue that the poor generalization performance of pan-sharpening networks may result from the training method, which leads to the model overfitting the ‘styles information’ in the LRMS image. Therefore, how to address the domain-specific features in LRMS is the key to generalizable pan-sharpening.

In Figure 3, to enforce structural consistency and promote generalization performance, PanNet is trained in the high-frequency domain, which is insensitive to the domain shift. Inspired by it, the gradient features of PAN and LRMS are utilized in our method to enhance structural consistency and learn domain-irrelevant features. For spectra and domain-specific features preservation, PanNet propagates the LRMS image to the network output. Nevertheless, directly mapping of the spectral information contained in the LRMS may lead to inaccurate predictions. This is because the spectral knowledge of LRMS is not well exploited and used. As the domain-specific features can be disentangled in the frequency domain[6, 41, 45, 50], which motivates us to design frequency-based feature integration networks for generalizable pan-sharpening. As illustrated in Figure 3, the domain-irrelevant knowledge (i.e., the phase) of LRMS is sufficiently leveraged in our method to enhance the structural consistency. Meanwhile, the amplitude of LRMS is directly injected in different fusion stages to ensure the generalization ability and avoid spectral distortion.

3.2 Structural Preservation
Since gradients contain rich image structure details and they are inherently consistent across domains, we perform the structural preservation on the gradient domain. Specifically, the gradients of
PAN and LRMS are first calculated by:

\[ P_x, P_y = \nabla (P) \]

\[ M_x, M_y = \nabla (LM \uparrow) \]

where \( P \) refers to the PAN image, and \( LM \uparrow \) denotes the upsampled version of LRMS. \( LM \uparrow \) and \( P \) have the same spatial resolution. \( \nabla \) indicates the gradient operation, \( P_x \) and \( P_y \) represent the horizontal and vertical gradient images of PAN, respectively. Similarly, \( M_x \) and \( M_y \) represent the horizontal and vertical gradient images of upsampled LRMS, respectively.

To facilitate the interaction between gradient features, a high-order information interaction module (HOFI) is employed (shown in Figure 4). Concretely, we first use a linear projection layer \( \varphi_{in} (\cdot) \) to obtain a set of projected feature components \( K_0 \) and \( \{ Q_k \}_{k=0}^{n-1} \).

\[
\left[ \begin{array}{c}
K_0^{H \times W \times C}, Q_0^{H \times W \times C}, \ldots, Q_{n-1}^{H \times W \times C_{n-1}}
\end{array} \right] = \varphi_{in}(F_H) \tag{2}
\]

We set the interaction orders (i.e., the \( n \) in \( g^n \text{Conv} \) [34]) as 3 by default. \( F_H \) refers to the joint shallow gradient features of PAN and upsampled LRMS, \( \varphi_{in}(F_H) \in \mathbb{R}^{H \times W \times (C_0 + \sum_{k=1}^{n-1} C_k)} \). The channel dimension in each order as \( C_k = \frac{C}{2^{n-k}}, 0 \leq k \leq n-1 \). We then perform the gated convolution recursively by:

\[
K_{k+1} = \begin{cases} 
    f_k(Q_k) \odot K_k, & k = 0 \\
    f_k(Q_k) \odot g_k(K_k), & 1 \leq k \leq n-1 
\end{cases} \tag{3}
\]

The coarse fused gradient component \( K_{k+1} \) is derived through the element-wise multiplication of \( f_k(Q_k) \) and \( g_k(K_k) \). Where \( f_k \) and \( g_k \) are a set of depth-wise convolutions (i.e., \( 7 \times 7 \) convolution) for deep feature extraction. \( \{ g_k \} \) are a set of \( 1 \times 1 \) convolutions used to match the dimension in different orders.

Finally, we feed the last recursion step fused gradient component \( K_n \) to the projection layer \( \varphi_{out} (\cdot) \) to obtain the result of high-order interactions. And we then model the relationship between feature channels by channel attention \( CA(\cdot) \) to obtain a multispectral image \( H_G \in \mathbb{R}^{H \times W \times C} \) that contains structural information:

\[
H_G = CA(\varphi_{out}(K_n)) \tag{4}
\]

Figure 4: The architecture of our proposed high-order information interaction module (HOFI). We also provide the detailed implementation of \( g^n \text{Conv} \) [34] with the interaction order \( n = 3 \) (below).
expressed as:

\[ H_f = DFM(H_G, H_S) \quad (9) \]

\(DFM\) is based on the residual channel attention mechanism and its detailed structure is presented in Figure 6. Notably, to avoid the network learning domain-specific features, we merely estimate the phase of the fused image \(H_f\) and discard the amplitude:

\[ \mathcal{A}(H_f), \mathcal{P}(H_f) = \mathcal{F}(H_f) \quad (10) \]

Finally, we inject the amplitude of LRMS again via inverse DFT. Moreover, a residual learning mechanism is adopted by adding \(H_G\) to the final prediction \(HM\):

\[ HM = \mathcal{F}^{-1}(\mathcal{A}(LM \uparrow), \mathcal{P}(H_f)) + H_G \quad (11) \]

3.4 Loss Function

To generate high-quality results, we propose a joint structural-spectral loss to train the network. For structural consistency, we adopt the \(L_1\) loss:

\[ L_1 = \|HM - H_{gt}\|_1 \quad (12) \]

where \(HM\) and \(H_{gt}\) denote the network output and the corresponding ground truth, respectively. To further supervise the structural consistency, we employ the DFT to convert \(HM\) and \(H_{gt}\) into Fourier space, where the \(L_1\)-norms of phase difference are calculated:

\[ L_P = \| \mathcal{P}(HM) - \mathcal{P}(H_{gt}) \|_1 \quad (13) \]

Simultaneously, we employ a spectral loss function to enhance the spectral information consistency among \(HM\) and \(H_{gt}\). The spectral loss function is defined as follows:

\[ L_s = \frac{1}{N} \sum_{i=1}^{N} \arccos \left( \frac{HM \cdot H_{gt}}{\|HM\|_2 \|H_{gt}\|_2} \right) \quad (14) \]

where \(N\) denotes the number of pixels in each multispectral image band. Finally, the full objective function for our method is a weighted sum of all sub-loss terms:

\[ \mathcal{L} = L_1 + \lambda_1 L_P + \lambda_2 L_s \quad (15) \]

where \(\lambda_1\) and \(\lambda_2\) denote the weights for balancing the three terms. \(\lambda_1, \lambda_2\) are empirically set to 0.1 and 0.2 in all experiments.

4 EXPERIMENTS

4.1 Datasets

In order to show the effectiveness of the proposed model, we conduct experiments over the widely-used datasets including WorldView-II (WV2), GaoFen2 (GF2) and WorldView-III (WV3). As the paired training samples are not available, we construct the training datasets using the Wald protocol [40] to generate paired images. To be specific, given an origin MS image \(HM \in R^{H \times W \times C}\) and its corresponding PAN image \(P \in R^{rH \times rW \times C}\), both of them are downsampled with ratio \(r\) to obtain image pairs \(LM \in R^{H/r \times W/r \times C}\) and \(P \in R^{H/r \times W/r \times C}\). \(LM\) and \(P\) are treated as inputs, while \(HM\) serves as the ground truth. Furthermore, the PAN images are cropped into patches with the size of \(128 \times 128 \times 1\), while the MS images are cropped into patches with the size of \(32 \times 32 \times 4\).

For the WV3 satellite dataset, approximately 90% of the data are allocated for training and 10% for validation, whereas the remaining two satellite datasets are solely utilized for validation. The detailed composition of each dataset is reported in Table 1.

<table>
<thead>
<tr>
<th>Satellite</th>
<th>WorldView-III</th>
<th>WorldView-II</th>
<th>GaoFen2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training number</td>
<td>2150</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Testing number</td>
<td>200</td>
<td>840</td>
<td>2912</td>
</tr>
</tbody>
</table>

4.2 Implementation Details

All experiments are conducted on a single NVIDIA GeForce GTX 2080Ti GPU, and the PyTorch framework is used to construct our networks. During the training phase, we employ an ADAM optimizer with \(\beta_1 = 0.9, \beta_2 = 0.999\) to update the network parameters for 1000 epochs with a batch size of 4. The learning rate is initialized with \(8 \times 10^{-4}\). In parallel, a StepLR learning rate adjustment strategy is employed to reduce the learning rate by half after every 150 iterations.

4.3 Evaluation Metrics

We evaluate the algorithm performance using the following four widely used image quality assessment metrics: peak signal-to-noise ratio (PSNR) [22], structural similarity index (SSIM) [42], relative dimensionless global error in synthesis (ERGAS) [39], spectral angle mapper (SAM) [55]. The first three metrics measure the spatial distortion and the fourth one measures the spectral distortion. An image is better if its PSNR and SSIM are higher, and SAM and
Table 2: Quantitative comparison. Highlighted in red and underlined respectively indicate the first and second best results.

<table>
<thead>
<tr>
<th>Method</th>
<th>Worldview III</th>
<th>Worldview II</th>
<th>GaoFen2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PSNR ↑ SSIM ↑ SAM ↓ ERGAS ↓</td>
<td>PSNR ↑ SSIM ↑ SAM ↓ ERGAS ↓</td>
<td>PSNR ↑ SSIM ↑ SAM ↓ ERGAS ↓</td>
</tr>
<tr>
<td>PNN [33]</td>
<td>29.9418 0.9121 0.0824 3.3206</td>
<td>26.0051 0.8212 0.1461 6.5339</td>
<td>24.1737 0.7709 0.1556 4.9988</td>
</tr>
<tr>
<td>PanNet [49]</td>
<td>30.6840 0.9072 0.0851 3.4263</td>
<td>30.3104 0.8369 0.0707 3.3140</td>
<td>30.9604 0.7971 0.0502 3.0355</td>
</tr>
<tr>
<td>MSDCNN [54]</td>
<td>30.3038 0.9184 0.0782 3.1884</td>
<td>28.0123 0.8076 0.1025 5.0103</td>
<td>27.9701 0.7621 0.0736 4.2616</td>
</tr>
<tr>
<td>DiCNN [21]</td>
<td>29.7445 0.9090 0.0830 3.3902</td>
<td>26.1971 0.8176 0.1237 6.3263</td>
<td>28.4187 0.7869 0.0635 3.4095</td>
</tr>
<tr>
<td>SRPPNN [3]</td>
<td>30.4346 0.9202 0.0770 3.1553</td>
<td>28.2124 0.8183 0.1005 4.9073</td>
<td>28.5696 0.7622 0.0817 3.3873</td>
</tr>
<tr>
<td>GPPNN [46]</td>
<td>30.1785 0.9175 0.0776 3.2593</td>
<td>26.1912 0.8146 0.1069 6.0099</td>
<td>25.0855 0.6960 0.1100 5.2588</td>
</tr>
<tr>
<td>BAM [26]</td>
<td>30.3845 0.9188 0.0773 3.1670</td>
<td>27.4929 0.8127 0.1086 5.3810</td>
<td>27.9112 0.7237 0.0656 3.9072</td>
</tr>
<tr>
<td>LAGConv [27]</td>
<td>30.2933 0.9147 0.0782 3.2050</td>
<td>30.1302 0.8365 0.0761 3.8513</td>
<td>24.0323 0.7262 0.1404 7.9625</td>
</tr>
<tr>
<td>SFIIN [61]</td>
<td>30.5439 0.9228 0.0745 3.1097</td>
<td>29.5121 0.8360 0.0693 4.0178</td>
<td>28.2132 0.7029 0.0646 3.6545</td>
</tr>
<tr>
<td>Ours</td>
<td>30.4397 0.9207 0.0768 3.1478</td>
<td>31.1664 0.8464 0.0643 3.0163</td>
<td>32.8621 0.7874 0.0500 3.0117</td>
</tr>
</tbody>
</table>

Table 3: Quantitative comparison. Highlighted in red and underlined respectively indicate the first and second best results.

<table>
<thead>
<tr>
<th>Method</th>
<th>WorldView-III</th>
<th>WorldView-II</th>
<th>GaoFen2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$D_A$ ↓ $D_S$ ↓ QNR ↑</td>
<td>$D_A$ ↓ $D_S$ ↓ QNR ↑</td>
<td>$D_A$ ↓ $D_S$ ↓ QNR ↑</td>
</tr>
<tr>
<td>PNN [33]</td>
<td>0.0460 0.0933 0.8654</td>
<td>0.1186 0.1228 0.7741</td>
<td>0.2778 0.1487 0.6153</td>
</tr>
<tr>
<td>PanNet [49]</td>
<td>0.0474 0.0942 0.8634</td>
<td>0.1092 0.1227 0.7826</td>
<td>0.1517 0.1253 0.7412</td>
</tr>
<tr>
<td>MSDCNN [54]</td>
<td>0.0432 0.0878 0.8732</td>
<td>0.1329 0.1228 0.7621</td>
<td>0.2916 0.1270 0.6191</td>
</tr>
<tr>
<td>DiCNN [21]</td>
<td>0.0469 0.0910 0.8666</td>
<td>0.1098 0.1156 0.7880</td>
<td>0.2564 0.1669 0.6204</td>
</tr>
<tr>
<td>SRPPNN [3]</td>
<td>0.0414 0.0909 0.8719</td>
<td>0.1371 0.1273 0.7543</td>
<td>0.2390 0.1092 0.6779</td>
</tr>
<tr>
<td>GPPNN [46]</td>
<td>0.0438 0.0936 0.8671</td>
<td>0.1193 0.1196 0.7764</td>
<td>0.2343 0.1391 0.6594</td>
</tr>
<tr>
<td>BAM [26]</td>
<td>0.0469 0.0910 0.8666</td>
<td>0.1314 0.1207 0.7649</td>
<td>0.3320 0.1292 0.5823</td>
</tr>
<tr>
<td>LAGConv [27]</td>
<td>0.0444 0.0886 0.8711</td>
<td>0.1248 0.1162 0.7743</td>
<td>0.2574 0.1864 0.6045</td>
</tr>
<tr>
<td>SFIIN [61]</td>
<td>0.0413 0.0876 0.8722</td>
<td>0.1209 0.1196 0.7751</td>
<td>0.3076 0.1201 0.6097</td>
</tr>
<tr>
<td>Ours</td>
<td>0.0420 0.0877 0.8716</td>
<td>0.0995 0.1195 0.7937</td>
<td>0.1798 0.0774 0.7562</td>
</tr>
</tbody>
</table>

ERGAS are lower. Furthermore, due to the lack of ground-truth MS images, we also quantify the model performance with three no-reference image quality assessment measures, i.e., the spectral distortion index $D_A$ [28], the spatial distortion index $D_S$ and the quality with (QNR) [1].

### 4.4 Comparison with state-of-the-art methods

To prove the superior generalization capability of our approach, we compare its performance with several representative deep learning-based pan-sharpening methods: including PNN [33], PanNet [49], MSDCNN [54], DiCNN [21], SRPPNN [3], GPPNN [46], BAM [26], LAGConv [27], and SFIIN [61].

We’ve motivated the proposed method as being more robust to differences across satellites because it focuses on gradient and phase features for structural consistency, and the amplitude features of upsampled LRMS are directly injected into the network for spectral preservation. This strategy can help the network learn domain-irrelevant knowledge, and so networks trained on one satellite can generalize better to new satellites. To empirically show this, we train all comparison methods on WorldView-III dataset, but test them on multiple types of satellite datasets (e.g., WorldView-III, WorldView-II and GaoFen2).

Quantitative Comparison. Table 2 and Table 3 summarize the average assessment metrics for three datasets, with the best results highlighted in red. The results show that our model outperforms all the other models in terms of generalization performance on the WorldView-II and GaoFen2 datasets. Compared to the second-best results, our method improves PSNR by 0.86dB and 1.90dB across the two datasets, respectively. Furthermore, our method also leads all algorithms by a large margin in other metrics in addition to PSNR. As the domain shift increases, our model exhibits a slower performance degradation and thus demonstrates better generalization. In contrast, other models suffer from significant performance drops when dealing with more severe domain discrepancies. At the same time, our approach achieves comparable performance with state-of-the-art methods on the WorldView-III dataset.

Visual Comparison. Additionally, we also provide the visual comparison with other advanced algorithms on WorldView-III, WorldView-II and GaoFen2 datasets, respectively, as shown in Figure 7, Figure 8 and Figure 9. The last row of the figure shows the mean squared error (MSE) between the ground truth and the fused pan-sharpened images. In Figures 8 and 9, we can clearly observe that our method achieves the best visually appealing results, and the fused high-quality pan-sharpened image is closer to ground truth even though it is trained only on the WorldView-III dataset. Specifically, upon amplifying local regions, we observe that our proposed method preserves spectral information (less color difference in smooth regions) and achieves better spatial resolution (clearer structures around edge regions) compared to other methods. It’s worth mentioning that our proposed strategy is more accurate than other comparison methods in terms of MSE residues, thus further demonstrating the ability of our proposed method to generalize effectively across multiple satellites. These results confirm and support our motivation to learn domain-irrelevant representations to
promote the generalization capability of the network. In contrast, existing models tend to overfit the training satellite data distribution and fail when facing unseen satellites. More quantitative and qualitative results can be found in the supplementary material.

4.5 Comparison Over Full-Resolution Scenes
To compare the generalization of models in full-resolution scenes, we apply a pre-trained model built on WorldView-III data to an additional real-world full-resolution GaoFen2 dataset for evaluation. Table 4 provides an overview of the experimental findings for all approaches. From Table 4, we can observe that our devised technique performs almost at the top of all the indices, which suggests that it has superior generalization capacity than other deep learning-based methods. Please refer to the supplementary material to see visual results.

4.6 Ablation experiments
In this subsection, several ablation experiments are performed to verify the effectiveness of the proposed key insights, including: a) the phase and amplitude of PAN and LRMS are disentangled in the frequency domain, which can help network learn domain-irrelevant features and improve the model generalization capability; b) the gradient features are insensitive to the domain shift, which can also promote the model generalization ability. We train our model on the WorldView-III dataset, and evaluate the performance on the GaoFen2 dataset. The commonly used IQA measures, such as PSNR, SSIM, SAM, ERGAS index, \( D_1 \), \( D_5 \), and QNR, are utilized to evaluate all of the experimental data.

(I) Effect of phase-amplitude decoupling in the frequency domain: In the spectral preservation phase, compared to previous
Table 4: Evaluation on the real-world full-resolution scenes from GaoFen2. Highlighted in red indicates the best results.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_\lambda$</td>
<td>0.2443</td>
<td>0.1959</td>
<td>0.1825</td>
<td>0.1643</td>
<td>0.1601</td>
<td>0.2454</td>
<td>0.262</td>
<td>0.2455</td>
<td>0.1401</td>
<td>0.0696</td>
</tr>
<tr>
<td>$D_\delta$</td>
<td>0.1419</td>
<td>0.2743</td>
<td>0.3226</td>
<td>0.3595</td>
<td>0.3246</td>
<td>0.3252</td>
<td>0.2983</td>
<td>0.2969</td>
<td>0.3258</td>
<td>0.2456</td>
</tr>
<tr>
<td>QNR</td>
<td>0.5090</td>
<td>0.6907</td>
<td>0.5563</td>
<td>0.5379</td>
<td>0.5700</td>
<td>0.5117</td>
<td>0.5200</td>
<td>0.5317</td>
<td>0.5825</td>
<td>0.7025</td>
</tr>
</tbody>
</table>

Table 5: Ablation studies comparison on GaoFen2. Highlighted in red indicates the best results.

<table>
<thead>
<tr>
<th>Ablation</th>
<th>PSNR ↑</th>
<th>SSIM ↑</th>
<th>SAM ↓</th>
<th>GaoFen2</th>
<th>ERGAS ↓</th>
<th>$D_\lambda$ ↓</th>
<th>$D_\delta$ ↓</th>
<th>QNR ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>(I)</td>
<td>26.6153</td>
<td>0.6887</td>
<td>0.0677</td>
<td>4.2745</td>
<td>0.3379</td>
<td>0.1189</td>
<td>0.5838</td>
<td></td>
</tr>
<tr>
<td>(II)</td>
<td>28.6974</td>
<td>0.7325</td>
<td>0.0809</td>
<td>3.7741</td>
<td>0.2457</td>
<td>0.1143</td>
<td>0.6684</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>32.8621</td>
<td>0.7874</td>
<td>0.0500</td>
<td>3.0117</td>
<td>0.1798</td>
<td>0.0774</td>
<td>0.7562</td>
<td></td>
</tr>
</tbody>
</table>

Figure 9: Visual comparison of all methods on Gaofen2.

works, we aim to enhance model generalization ability by decoupling the phase and amplitude of PAN and LRMS in the frequency domain. To demonstrate its effectiveness, we compare it to a variant model that directly fuses MS and PAN images in the spatial domain to preserve spectral information. As shown in Table 5, due to training-testing inconsistency, the quantitative performance of variant model degrades significantly, which nearly loses its generalization ability. This can be attributed to the direct fusion of LRMS and PAN images in the spatial domain, which inevitably guides the network parameters to learn domain-specific features contained in the amplitude of LRMS, thus degrading the generalization performance.

(II) The gradient features are insensitive to the domain shift: In the structural preservation phase, the gradient features of PAN and MS are utilized to enforce the structural consistency and learn the domain-irrelevant features. To assess its impact, we conduct an experiment that directly fuse the PAN and upsampled LRMS in the spatial domain instead of the gradient domain for structural consistency. Observing the results from Table 5, it can be clearly figured out that the variant model generalization performance has obtained considerable degradation in terms of all the IQAs when removing the gradient domain. It is because the gradient features are naturally consistent across different domains and are insensitive to the domain shift.

5 CONCLUSION

In this paper, we propose a generalizable pan-sharpening framework. Two core designs are devised to equip the network, i.e., a structural preservation module (STP) and a spectral preservation module (SPP). Concretely, STP is designed to fuse gradient information and enforce structural consistency between PAN and LRMS. SPP is presented to learn a phase fusion network of PAN and LRMS. The amplitude of LRMS that contains domain-specific features is directly injected into the network, which benefits the generalization capability of the network. Extensive experiments demonstrate that our proposed framework achieves better generalization performance than existing state-of-the-art pan-sharpening methods over multiple satellite datasets.

ACKNOWLEDGMENTS

This work was supported partly by the National Natural Science Foundation of China under Grants (No.82172033, No.U19B2031, No.61971369, No.52105126).


