DRFormer: Learning Disentangled Representation for Pan-Sharpening via Mutual Information-Based Transformer

Feng Zhang, Kai Zhang, *Member,* IEEE, Jiande Sun, Jian Wang, Lorenzo Bruzzone, *Fellow,* IEEE

***Abstract*—In this paper, we propose a new pan-sharpening method that disentangles low spatial resolution multispectral (LRMS) and panchromatic (PAN) images in terms of sensor-specific features and common features. These features are obtained by defining mutual information-based transformers designed to achieve disentangled learning. In the proposed method, LRMS and PAN images are cross-reconstructed by cross-coupled transformers to facilitate the disentanglement of the common features and sensor-specific features. To ensure compatibility among the disentangled features, self-reconstructions of LRMS and PAN images are imposed on them, and source images are reconstructed by self-coupled transformers. In addition to the reconstruction-guided disentangled learning, we maximize the mutual information (MI) between the common features of LRMS and PAN images to improve the correlation of the common features from different images. We also minimize the MI between the common features and sensor-specific features from the same image to reduce the redundancy among them. Through the reconstruction and disentangled representation of source images, sensor-specific features and common features can be decomposed efficiently. Finally, all disentangled features are integrated by a fusion transformer to generate the high spatial resolution multispectral image. Experiments on different datasets demonstrate that the proposed method produces competitive fusion results. The code is available at https://github.com/RSMagneto/DRFormer.**

***Index Terms*—Pan-sharpening, disentangled representation, mutual information, transformer, remote sensing image.**

I. Introduction

REMOTE sensing images with high spatial and spectral resolutions contain abundant information about ground objects. The rich texture, shape, and color characteristics in these images offer sufficient foundations for the efficient implementations of downstream tasks, such as environmental monitoring [1] and land survey [2]. However, owing to the technical limitations of imaging sensors, it is difficult to acquire remote sensing images with high spatial and spectral resolutions [3]. Most of the satellites, such as QuickBird and WorldView-3, can collect a high spatial resolution panchromatic (PAN) image and a low spatial resolution multispectral (LRMS) image at the same time. To generate the high spatial resolution multispectral (HRMS) image, pan-sharpening techniques are advanced to integrate the spatial and spectral information in PAN and LRMS images.

At present, the deep neural network (DNN) era has brought a new generation of pan-sharpening methods, and various advanced DNNs are applied to this task. For example, Sheng *et al*. [4] proposed a unified pansharpening model based on few-shot learning. Zheng *et al*. [5] introduced an uncertainty mechanism to capture the spatial-variant distributions between source images. In addition, variation optimization is also incorporated with DNNs for better modeling of spatial and spectral priors [6]-[9]. For instance, Wu *et al*. [6] derived an iterative algorithm for the variational model and introduced a DNN to learn the details to be injected. Yan *et al*. [7] built a variation model for pan-sharpening and unfolded its optimization algorithm as a cascaded memory-augmented network. Yang *et al*. [8] proposed a variational network for fusion, in which the deep prior regularized variational fusion model was solved by the half-quadratic splitting algorithm. Thanks to the powerful learning capability, DNN-based methods characterize the spatial and spectral features in LRMS and PAN images more efficiently and produce state-of-the-art fusion results. However, the information content existing in extracted features is not analyzed insightfully.

As LRMS and PAN images are collected from the same scene, both of them tend to contain some common features, such as contours and shapes. In Fig. 1, we can see that the common features from LRMS and PAN images are highly correlated. In addition to the common features, LRMS and PAN images contain sensor-specific features since they represent the scene information from different perspectives and are acquired by different imaging sensors. The abundant spatial details in the PAN image are regarded as sensor-specific features. For instance, the sensor-specific features of the PAN image in Fig. 1 contain a large number of subtle edges or textures. Compared with the PAN image, the sensor-specific features of the LRMS image are related to spectral information.

This work was supported in part by the Natural Science Foundation of Shandong Province of China (ZR2023MF066), the Natural Science Foundation of China (61901246), the China Postdoctoral Science Foundation Grant (2019TQ0190, 2019M662432), the China Scholarship Council (202008370035), the Scientific Research Leader Studio of Ji'nan (2021GXRC081), and Joint Project for Smart Computing of Shandong Natural Science Foundation (ZR2020LZH015). (*Corresponding author: Kai Zhang*.)

F. Zhang is with the School of Information Science and Engineering, University of Jinan, Ji'nan 250022, China (e-mail: fengzhangpl@163.com).

K. Zhang, J. Sun, and J. Wang are with the School of Information Science and Engineering, Shandong Normal University, Ji'nan 250358, China (e-mail: zhangkainuc@163.com, jiandesun@hotmail.com, jwang@sdnu.edu.cn).

L. Bruzzone is with the Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy (e-mail: lorenzo.bruzzone@unitn.it).

The sensor-specific features and common features are distinct in terms of both information content and redundancy among them. Therefore, to reduce the redundancy among features extracted by DNNs, it is necessary to disentangle the common features and sensor-specific features in LRMS and PAN images for a better reconstruction of the fused image.

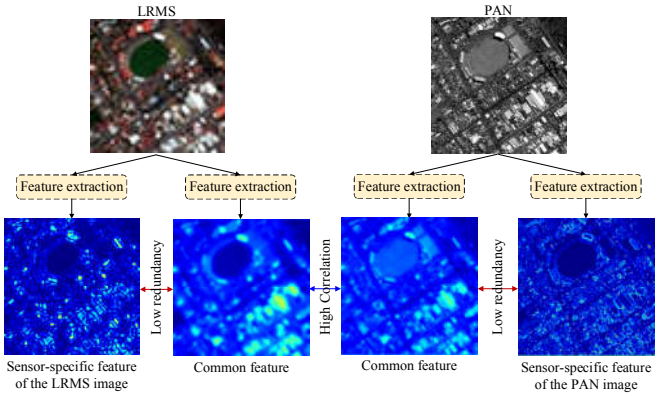


Fig. 1. Example of sensor-specific features and common features in LRMS and PAN images.

Disentangled representation aims to explore the low-dimensional representations of data and discover different underlying factors of data variations. Thanks to its interpretability, disentangled representation has been developed a lot [10]-[11]. In the image fusion field, Xu *et al*. [12] separated the visible and infrared images into the scene- and sensor-related representations through different encoders. Hong *et al*. [13] proposed a decoupled-and-coupled network to divide MS and hyperspectral images into common components and sensor-specific components. In [14], LRMS and PAN images were decomposed as the common features and unique features under the framework of convolutional sparse coding. In these methods, the disentanglement of the common features and sensor-specific features is performed in a self-reconstruction or cross-reconstruction manner. Besides, Zhou *et al*. [15] minimized the mutual information between features of source images to obtain the sensor-specific features. Then, these features are considered to generate the desired HRMS image. Although good fusion results are produced, only sensor-specific features of LRMS and PAN images are used for reconstruction. The common features between source images are discarded. However, some information of the observed information still exists in common features. The absence of common features results in the information loss of the observed scene, which limits the fusion performance of this model.

To effectively disentangle the features in source images and exploit their global properties, we propose a novel mutual information-based transformer for disentangled representation (DRFormer) to sharpen the LRMS image by using the PAN image. In the proposed method, the reconstructions of LRMS and PAN images are introduced to facilitate the disentanglement of the common and sensor-specific features. In particular, to ensure the disentanglement of these features, the features extracted by encoders are recombined to achieve the cross-reconstruction of source images. The self-reconstructions of source images are imposed on the disentangled features to improve the compatibility among them. In addition to the reconstruction-guided disentangled representation, we maximize the mutual information (MI) between the common features of LRMS and PAN images to force consistency in terms of distribution. For each source image, the MI between the common features and sensor-specific features is minimized to eliminate the high coupling among them. Considering the semantic gap between the common features and sensor-specific features, the attention mechanism in the transformer is utilized to obtain coupled features when reconstructing source images. Based on the coupled features, source images can be reconstructed better by decoders. Finally, all disentangled features are integrated by the cross-attention in the transformer to produce the fused HRMS image. The experimental results demonstrate that the proposed DRFormer has a competitive performance compared to some state-of-the-art methods on GeoEye-1 and QuickBird datasets.

The main contributions of the proposed DRFormer are summarized as follows.

1) It introduces a disentangled learning representation into the pan-sharpening task to decompose source images into common features and sensor-specific features.

2) It adopts the image cross- and self-reconstruction techniques to balance the disentanglement and compatibility in the extracted features. The transformer is considered to enhance the coupling between the common features and sensor-specific features when source images are reconstructed.

3) It exploits an MI disentanglement loss to separate common features and sensor-specific features in source images efficiently. In the loss, the MI between the same kinds of features is maximized. Meanwhile, the MI between different kinds of features is minimized.

The remainder of the paper is organized as follows. Section II summarizes different kinds of pan-sharpening methods. Section III introduces the proposed DRFormer in detail, describing the reconstruction-guided disentanglement, the disentanglement via mutual information, the fused image reconstruction, the network details, and the optimization strategy. Section IV demonstrates the experimental results on reduced-scale and full-scale datasets. Conclusions are given in Section V.

II. Related Work

Over the past three decades, researchers have developed various pan-sharpening methods and produced satisfactory fused images [16]-[18]. According to the widely adopted taxonomy, these methods are clustered into traditional methods and DNN-based methods.

A. Traditional Methods

As typical pan-sharpening paradigms, component substitution (CS) and multiresolution analysis (MRA) have been extensively developed. CS methods generally substitute the spatial component of the LRMS image with the PAN image to generate the desired HRMS image. For example, the intensity-hue-saturation (IHS) method [19], the Gram-Schmidt (GS) method [20], and the principal component analysis (PCA) [21] fall into this category. Moreover, some adaptive methods,

such as adaptive IHS [22] and band-dependent spatial detail (BDSD) [23], are developed to generate the spatial component more accurately. MRA methods inject the spatial details extracted from the PAN image into the up-sampled LRMS image under the assumption that the spatial details are absent in the LRMS image. Considering the complex spatial structures in source images, various MRA techniques, such as wavelets [24], curvelets [25], and support value transform (SVT) [26], are used for the extraction of spatial details.

In traditional methods, the model-based paradigm also attracts a lot of attention. This kind of methods [27]-[28] assumes that LRMS and PAN images are viewed as the counterparts of the fused image after spatial and spectral degradations, respectively. Then, the degradation relationships are described quantitatively according to a specific sensor model. Due to the ill-conditioned problem, the fused HRMS image cannot be accurately reconstructed by exclusively solving the degradation models. Therefore, the solution space of the degradation models is further regularized by many efficient priors existing in the images, such as sparsity [29], low-rank property [30]-[31], and non-negativity [32]. Finally, variational algorithms [33]-[34] are employed to optimize these prior-regularized models for the generation of fused images. In addition, some model-based methods [35]-[36] are presented to improve the robustness against misregistration.

B. DNN-Based Methods

For DNN-based methods, they can learn the mapping relationships between source images and the fused image efficiently. Nowadays, the common convolution neural network (CNN) has been used for the learning of nonlinear mapping. For example, Masi *et al*. [37] proposed a pan-sharpening neural network (PNN) that adopted a three-layer convolutional architecture. Subsequently, Scarpa *et al*. [38] presented an advanced version of PNN (A-PNN) by exploring the influences of architectural variations on the fused result. To enhance the spatial details in the LRMS image, Cai *et al*. [39] employed progressive residual networks to extract the high-frequency information in the PAN image. A progressive injection framework was also adopted in [40] to improve the spatial resolution of the LRMS image. Yang *et al*. [41] introduced residual modules into a dual-stream CNN to integrate the residuals at different levels. Fu *et al*. [42] combined residual learning with multiscale dilated blocks to preserve the spatial structures in the fused HRMS image. In addition, generative adversarial networks (GANs) [43]-[44] are considered to better reconstruct the fused image in terms of the data distribution. For instance, Zhou *et al*. [43] constructed two discriminators to improve the generalization of the trained model on the full-scale image pairs.

With network types growing exponentially, researchers are shifting their focus to the attention mechanism. Lei *et al*. [45] proposed a nonlocal attention residual network to characterize the similarity among all pixels in source images. Qu *et al*. [46] exploited the self-attention mechanism to achieve the unsupervised sharpening of the LRMS image. To capture the global properties in images, the recently proposed neural network based on transformer utilized the self-attention mechanism [47]. For example, Zhang *et al*. [48] used the multiscale transformer to learn the features in source images, which are then merged by a spatial-spectral interaction attention module. Similarly, a pyramid transformer [49] was also designed to model the global features in images. Sun *et al*. [50] considered the transformer as a backbone to extract spatial and spectral information in LRMS and PAN images.

To simultaneously use the decent interpretability of model-based methods and the good learning ability of DNNs, researchers develop model-driven DNNs for better fusion of LRMS and PAN images. For instance, Yan *et al*. [51] integrated model-driven and data-driven techniques and advanced a network, named MD^{3}Net, for pan-sharpening. Li *et al*. [52] unfolded the optimization of nonlocal similarity-regularized fusion model different network modules to learn the priors in images. Zhang *et al*. [53] proposed a dual back-projection network consisting of spatial and spectral projections to reconstruct the fused image. Zhou *et al*. [54] proposed a memory-augmented deep unfolding network derived from the alternating optimization of the fusion model and applied it to pan-sharpening. Li *et al*. [55] introduced cross-attention into the unfolding iteration network to improve its representation capability.

C. Disentangled Representation

Nowadays, disentangled representation has been widely explored in various tasks. For example, Ji *et al*. [56] used disentangled representation to separate the content and haze information for image dehazing. Disentangled representation is also applied to the image fusion task and different loss functions are advanced for efficient disentanglement. For instance, Luo *et al*. [57] designed contrastive constraints to decompose the common and private components in images. Gao *et al*. [58] disentangled images as content and modal features via a cycle adversarial loss. In [59] and [60], the disentangled features are extracted by cross-reconstruction losses, in which the disentangled features are recombined to approximate source images. Besides, Zhao *et al*. [61] obtained the common and sensor-specific features by minimizing the correlation between them. In the proposed DRFormer, we employ mutual information as a metric for disentangled learning and minimize it to disentangle source images as common and sensor-specific features. Meanwhile, we also maximize the mutual information between common features of LRMS and PAN images to improve their consistency. The self- and cross-reconstructions of source images are also introduced to ensure the compatibility between these features. Compared to the above methods, common and sensor-specific are disentangled better by the proposed DRFormer, and their redundancy is further reduced.

III. Proposed DRFormer

An overview of the proposed DRFormer can be seen in Fig. 2. Given the up-sampled LRMS image $\mathbf{L} \in \mathbb{R}^{H \times W \times B}$ and the PAN image $\mathbf{P} \in \mathbb{R}^{H \times W}$ to be fused, we first employ two encoders to

embed them into the feature domain rather than disentangling them in the original domain. H and W are the height and width of the image. B denotes the number of bands in the MS image. Then, the extracted features are fed into two convolutional blocks for disentangled representation. For $\mathbf{P}$, the common feature $C_P \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 128}$ and the sensor-specific feature $S_P \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 128}$ are produced after the corresponding convolutions. Similarly, we obtain the common feature $C_L \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 128}$ and the sensor-specific feature $S_L \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 128}$ of $\mathbf{L}$. By recombing the two categories of features, we decouple the information content in source images by synthesizing the cross-reconstructed PAN image $\mathbf{P}_C \in \mathbb{R}^{H\times W}$ and LRMS image $\mathbf{L}_C \in \mathbb{R}^{H\times W\times B}$. The self-reconstructions of the PAN image $\mathbf{P}_S \in \mathbb{R}^{H\times W}$ and the LRMS image $\mathbf{L}_S \in \mathbb{R}^{H\times W\times B}$ are considered to ensure the compatibility of these features. The reconstruction-guided disentanglement of source images will be explained in detail in Section II.A. The disentanglement via mutual information, described in Section II.B, is further implemented on the disentangled features to eliminate the coupling among them. After the disentangled learning, C_L and C_P are added to generate $C \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 128}$. For the reconstruction of the fused image $\mathbf{H} \in \mathbb{R}^{H\times W\times B}$, S_L, S_P, and C are integrated into the fusion transformer described in Section II.C. Finally, the reconstruction losses, MI loss, and fusion loss are minimized for training.

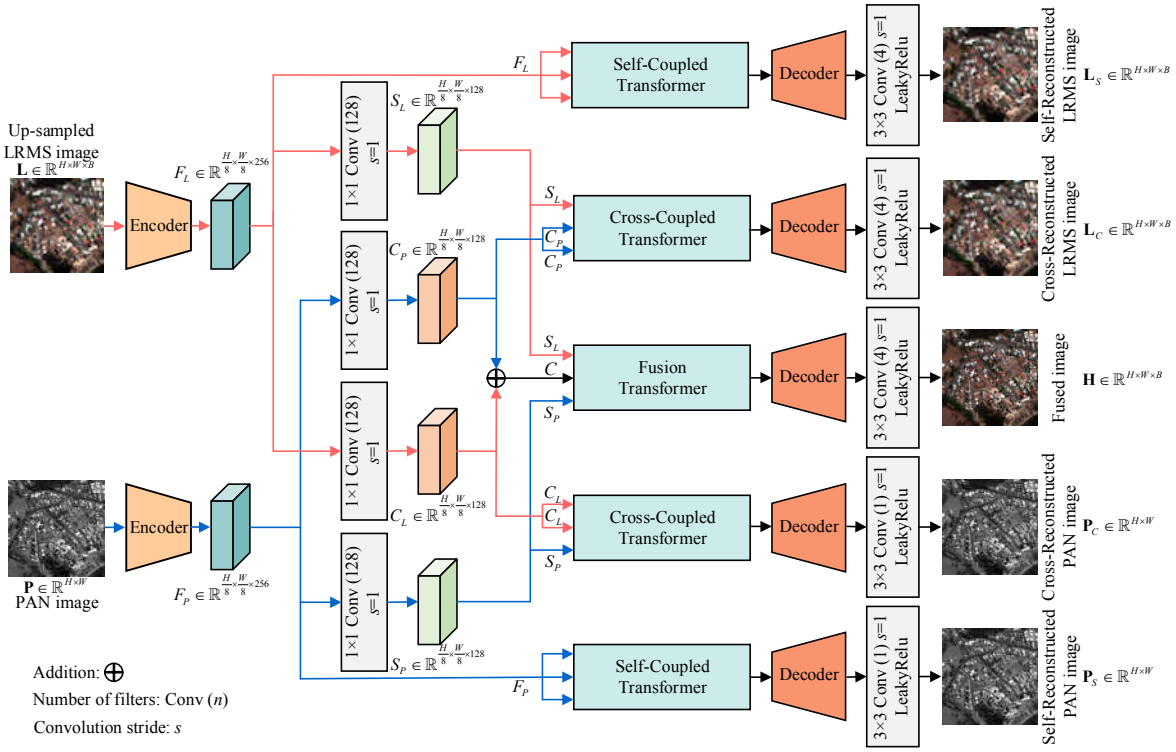


Fig. 2. Architecture of the proposed DRFormer.

A. Reconstruction-Guided Disentanglement

In the proposed DRFormer, the cross- and self-reconstructions of source images are simultaneously used to facilitate their disentangled learning. Through the reconstructions of source images, the common features from different images will be the same in terms of information content. Meanwhile, the sensor-specific features are modeled to reflect the attribute gap between LRMS and PAN images.

1) *Cross-Reconstruction Disentanglement*: For the cross-reconstruction of the LRMS image, the sensor-specific features S_L and the common features C_P are combined. Because S_L and C_P are from different images, the cross-coupled transformer shown in Fig. 3 is designed to improve the coupling between them. In the cross-coupled transformer, the inputs are first projected by the corresponding convolution blocks. Then, the projected features are coupled by the multiscale multi-head attention (MMHA). A residual block (ResBlock) is introduced to refine the coupled features. The operation performed by the cross-coupled transformer is formulated as:

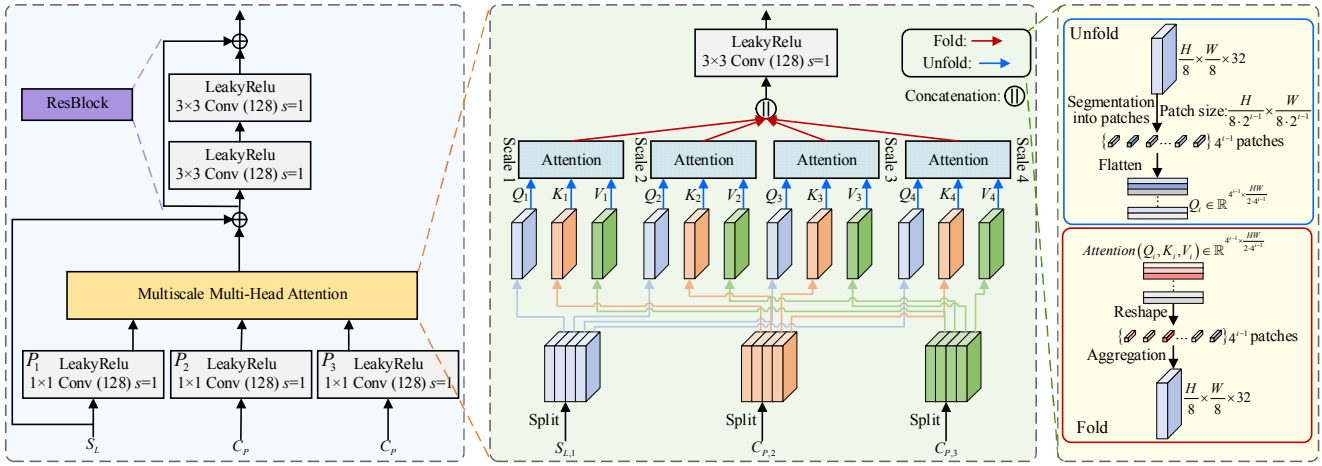
$$\begin{aligned} S_{L,1} &= P_1(S_L)\ C_{P,2} = P_2(C_P)\ C_{P,3} = P_3(C_P) \\ M_L &= S_L + \mathrm{MMHA}(S_{L,1}, C_{P,2}, C_{P,3}) \\ H_L &= M_L + \mathrm{ResBlock}(M_L) \end{aligned} \tag{1}$$

where $P_1(\cdot)$, $P_2(\cdot)$ and $P_3(\cdot)$ stand for the convolution blocks in Fig. 3. H_L is the output of the cross-coupled transformer which is fed into the corresponding decoder for the cross-reconstruction of the LRMS image $\mathbf{L}_C$.

Fig. 3(a) shows the MMHA module designed to exploit the global dependencies of source images at different scales. Fig. 3(b) shows the structure of the MMHA module. After the projections of P_1, P_2, and P_3 we can obtain the embedded $S_{L,1}$, $C_{P,2}$, and $C_{P,3}$. Each of them is split into four parts along the channel dimension in the MMHA module. For each part, the size is $\frac{H}{8}\times\frac{W}{8}\times 32$. Then, the split features are unfolded as the matrix form. The unfolding and folding operations are displayed in Fig. 3(c). Taking one split part from $S_{L,1}$ as an example, it is first segmented as non-overlapping patches with the size of $\frac{H}{8\cdot 2^{i-1}}\times\frac{W}{8\cdot 2^{i-1}}$. i=1, 2, 3, 4 is the index of scale in Fig. 3(b) and the patch size varies with the index of scale to model the multiscale spatial information in images. After segmentation, 4^{i-1} patches are obtained and each patch is flattened as one vector with the length of $\frac{HW}{2\cdot 4^{i-1}}$ considering all 32 channels in this patch. Then, all vectors corresponding to 4^{i-1} patches are used to form the matrix $Q_i\in\mathbb{R}^{4^{i-1}\times\frac{HW}{2\cdot 4^{i-1}}}$. For the split features of $C_{P,2}$ and $C_{P,3}$, they are also unfolded in the same way to produce K_i and V_i for following attention estimation among them:

$$Attention\left(Q_i,K_i,V_i\right)=softmax\left(\frac{Q_iK_i^T}{\sqrt{d_i}}\right)V_i \tag{2}$$

where d_i is the number of columns in Q_i. Then, we fold the result of (2) as the feature with the size of by the corresponding inverse operation as shown in Fig. 3(c). Through the MMHA module, we get the weighted features by folding and concatenation. The semantic gap among these features is reduced by MMHA. Then, these features are refined by the ResBlock. Finally, the output of the cross-coupled transformer is sent into the corresponding decoder for the generation of the cross-reconstructed LRMS image $\mathbf{L}_C$. Through the MMHA module, we can reconstruct the LRMS better, due to the coupling among the weighted features.



(a) Cross-coupled transformer (b) MMHA (c) Unfold and fold.

Fig. 3. Detailed illustration of the proposed cross-coupled transformer. (a) Cross-coupled transformer, (b) MMHA, (c) Unfold and fold.

Similarly, we integrate the sensor-specific feature S_P and the common feature C_L for the cross-reconstruction of the PAN image. S_P and C_L are fed into the cross-coupled transformer, which has the same structure as that in Fig. 3(a). Specifically, the inputs S_L and C_P in Fig. 3(a) are replaced by S_P and C_L, respectively. Then, S_L and C_P are coupled by:

$$\begin{aligned} S_{P,1}=P_1\left(S_P\right)\ C_{L,2}=P_2\left(C_L\right)\ C_{L,3}=P_3\left(C_L\right)\\ M_P=S_P+\mathrm{MMHA}\left(S_{P,1},C_{L,2},C_{L,3}\right)\\ H_P=M_P+\mathrm{ResBlock}\left(M_P\right)\end{aligned} \tag{3}$$

where H_P is the output and also the input of the corresponding decoder for the cross-reconstruction of the PAN image $\mathbf{P}_C$.

According to the cross-reconstructions of $\mathbf{L}$ and $\mathbf{P}$, their losses are written as:

$$\mathcal{L}_D=\left\|\mathbf{L}-\mathbf{L}_C\right\|_F^2+\left\|\mathbf{P}-\mathbf{P}_C\right\|_F^2 \tag{4}$$

When the source images are cross-reconstructed, the common and sensor-specific features are disentangled effectively.

2) *Self-Reconstruction Constraint*: In the above part, the cross-reconstructions of source images improve the complementarity between the common and sensor-specific features from different images. However, it is difficult to preserve the compatibility between the common and sensor-specific features from the same image. So, we introduce the self-reconstruction tasks to constrain the extracted features. To improve the compatibility between S_L and C_L, a self-coupled transformer is built when the LRMS image is self-reconstructed. Here, the structure of the self-coupled transformer is the same as that of the cross-coupled transformer except for inputs. In the self-coupled transformer, the inputs are all F_L. Specifically, when all inputs in Fig. 3(a) are replaced by F_L, the cross-coupled transformer becomes the self-coupled counterpart. Then, the output of the self-coupled transformer is used to synthesize $\mathbf{L}_S$ by the corresponding decoder.

In the same way, the self-reconstruction of the PAN image is also completed to make S_P and C_P more compatible. Then, the self-reconstruction losses are defined as:

$$\mathcal{L}_S = \left\|\mathbf{L}-\mathbf{L}_S\right\|_F^2 + \left\|\mathbf{P}-\mathbf{P}_S\right\|_F^2 \tag{5}$$

Through the self-reconstructions of $\mathbf{L}$ and $\mathbf{P}$, the compatibility between disentangled features is enhanced, which avoids the information loss caused by the feature decomposition.

B. Disentanglement via Mutual Information

The disentanglement in Section II.A is implemented by the image reconstruction in terms of the information content. In this part, we further use the MI regularization to reduce the redundancy among the common features and the sensor-specific features from the same image. Meanwhile, the MI between the common features C_L and C_P is also maximized to achieve a high correlation in terms of distributions. For the MI of S_L and C_L, it is calculated as:

$$MI\left(S_L,C_L\right) = E\left(S_L\right)+E\left(C_L\right)-E\left(S_L,C_L\right) \tag{6}$$

where $E(\cdot)$ denotes the entropy. $E\left(S_L\right)$ and $E\left(C_L\right)$ are marginal entropies, respectively. $E\left(S_L,C_L\right)$ stands for the joint entropy of S_L and C_L. The conditional entropies between S_L and C_L can be computed by the Kullback-Leibler divergence (KL):

$$\begin{aligned} KL\left(S_L\left\|C_L\right.\right) &= E_{C_L}\left(S_L\right)-E\left(S_L\right) \\ KL\left(C_L\left\|S_L\right.\right) &= E_{S_L}\left(C_L\right)-E\left(C_L\right) \end{aligned} \tag{7}$$

where $E_{C_L}\left(S_L\right)$ and $E_{S_L}\left(C_L\right)$ are the cross-entropies. By the sum of (6) and (7), we obtain:

$$\begin{aligned} MI\left(S_L,C_L\right) = {} & E_{C_L}\left(S_L\right)+E_{S_L}\left(C_L\right)-E\left(S_L,C_L\right) \\ & -\left(KL\left(S_L\left\|C_L\right.\right)+KL\left(C_L\left\|S_L\right.\right)\right) \end{aligned} \tag{8}$$

Considering the nonnegativity of $E\left(S_L,C_L\right)$, we optimize the following loss to minimize the MI between S_L and C_L:

$$\begin{aligned} \mathcal{L}_{MI}\left(S_L,C_L\right) = {} & E_{C_L}\left(S_L\right)+E_{S_L}\left(C_L\right) \\ & -\left(KL\left(S_L\left\|C_L\right.\right)+KL\left(C_L\left\|S_L\right.\right)\right) \end{aligned} \tag{9}$$

By minimizing the loss, the redundancy between S_L and C_L is reduced. Similarly, the MI loss between S_P and C_P is also defined according to the form in (9). In addition, we maximize the MI of C_L and C_P to enforce the distribution similarity among them. Then, the final MI loss is defined as:

$$\mathcal{L}_{MI} = \mathcal{L}_{MI}\left(S_L,C_L\right)+\mathcal{L}_{MI}\left(S_P,C_P\right)-\mathcal{L}_{MI}\left(C_L,C_P\right) \tag{10}$$

Given the source images, the minimization of (10) achieves the disentanglement of the features with different attributes and the consistency between the common features simultaneously.

C. Fused Image Reconstruction

After the disentangled learning described in Sections III.A and III.B, we use the decomposed features to produce the fused image. In Fig. 2, S_L, S_P, and C flow into a fusion transformer to improve their coupling. The proposed fusion transformer is shown in Fig. 4. It includes two MMHA modules to mix the common features and the disentangled features. First, C is projected by two convolution blocks O_2 and O_3. S_L and S_P are also embedded by the projections O_1 and O_4, respectively. Then, the coupling in the MMHA modules is written as:

$$\begin{aligned} G_1 &= \mathrm{MMHA}\left(O_1\left(S_L\right),O_2\left(C\right),O_3\left(C\right)\right) \\ G_2 &= \mathrm{MMHA}\left(O_4\left(S_P\right),O_3\left(C\right),O_2\left(C\right)\right) \end{aligned} \tag{11}$$

where G_1 and G_2 are the outputs of the two MMHA modules. Then, G_1 and G_2 are concatenated together and processed by a ResBlock. Finally, we send the output of the fusion transformer into the corresponding decoder for the generation of the fused image. The reconstruction error of the fused image is modeled as:

$$\mathcal{L}_F = \left\|\mathbf{H}-\mathbf{R}\right\|_F^2 \tag{12}$$

where $\mathbf{R}\in\mathbb{R}^{H\times W\times B}$ is the reference image. Through the optimization of (12), the information in S_L, S_P, and C is integrated by the fusion transformer and the corresponding decoder.

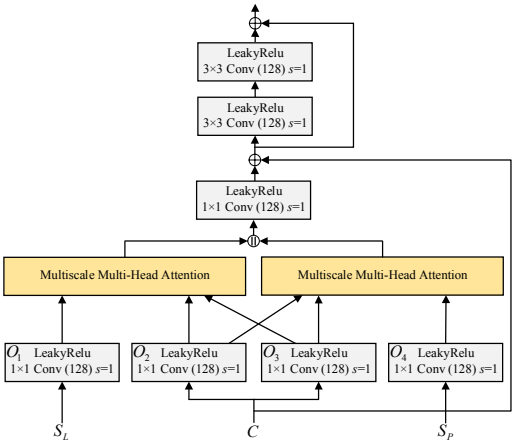
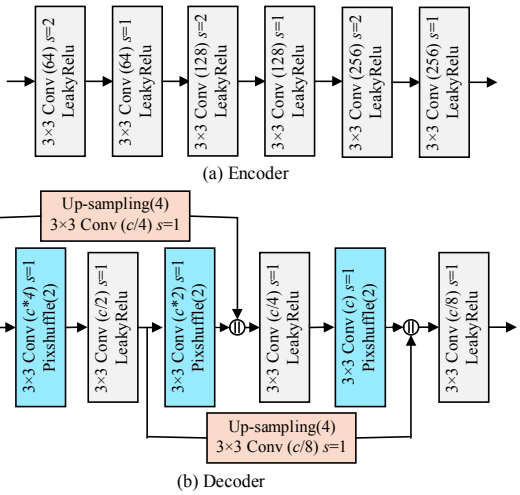


Fig. 4. Architecture of the proposed fusion transformer.



(a) Encoder

(b) Decoder

Fig. 5. Architectures of (a) encoder and (b) decoder.

D. Encoders and Decoders

In the proposed DRFormer, the disentangled learning of source images is implemented in feature domain. Therefore, we adopt two encoders with the same structure to extract the features from the LRMS and the PAN images. Fig. 5(a) shows the architecture of the encoders used in Fig. 2. The encoders are made up of 6 cascaded convolution blocks. The features are gradually down-sampled by strided convolutions to reduce the computational complexity of the disentangled representation. For the reconstructions of the source images and the fused image, the decoder in Fig. 5(b) is introduced into the proposed DRFormer. The five decoders in Fig. 2 have similar structures. In these decoders, feature maps are progressively up-sampled to the size of the fused image. In the skip connections of the decoder, we introduce an up-sampling operator with a ratio of 4 to concatenate the feature maps from shallow layers. For the self-reconstructions of source images, c in the two decoders is set to 256. In the other three decoders, we set c as 128.

E. Training and Test

In the final step, we optimize the overall loss to train the proposed DRFormer. The overall loss is calculated as:

$$\mathcal{L} = \mathcal{L}_D + \mathcal{L}_S + \lambda \mathcal{L}_{MI} + \mathcal{L}_F \tag{13}$$

where λ is a tradeoff parameter. We empirically set λ as 0.0001. We train the proposed DRFormer on the PyTorch framework by the Adam optimizer. The batch size and the learning rate are set to 4 and 0.0001, respectively. The training is completed after 1000 epochs. In the test stage, the reconstructions of source images are removed. The disentangled features are just fed into the fusion transformer to obtain the fused image.

TABLE I. DATASETS FOR TRAINING, VALIDATION, AND TEST.

Satellite	Image	Size	# Image pairs for training	#Image pairs for validation	#Image pairs for test	
					Reduced scale	Full scale
GeoEye-1	LRMS	64×64×4	1800	20	30	30
	PAN	256×256				
QuickBird	LRMS	64×64×4	960	20	30	30
	PAN	256×256				

IV. EXPERIMENTS

In this section, comparison experiments are conducted on different datasets to demonstrate the effectiveness of the proposed method. In the ablation study, the contribution of each part in the DRFormer is also explored. Besides, we also investigate the influences of different architectures on the fusion results.

A. Experimental Settings

The experiments are implemented on two datasets from GeoEye-1 and QuickBird satellites. In these datasets, the original PAN and MS images are smoothed and down-sampled with a ratio of 4 to obtain the image pairs to be fused. Then, the original MS image is considered the reference image. When the training is completed, the proposed DRFormer is tested on the reduced- and full-scale image pairs. More details about the constructed datasets are reported in Table I.

We compare the proposed DRFormer with 10 methods, including BDSD [23], AWLP [24], SVT [26], LRP [31], VPLGC [62], TFNet [63], PanNet [64], GPPNN [65], M-GAN [66], and MIP [15]. All DNN-based methods are trained and tested on a server with Intel Core i7-9700 processors, 3.0GHz, an NVIDIA 2080Ti GPU, and 128G memory. Then, Q4 [16], spectral angle mapper (SAM) [16], and *erreur relative globale adimensionnelle de synthèse* (ERGAS) [16] are utilized to evaluate the fusion results on reduced-scale datasets. The fusion results at full scale are assessed by D_λ, D_S, and quality without reference (QNR) [67].

B. Experiments on Reduced-Scale Datasets

In this section, we conduct pan-sharpening experiments on the reduced-scale datasets from GeoEye-1 and QuickBird satellites. Fig. 6 displays the pan-sharpened images of all methods on the GeoEye-1 dataset. A local region from the fused image is enlarged for a more detailed analysis. In addition, we also plot the difference map between the reference image and each fused image in Fig. 6. From the analysis of the images, one can see that some spatial details are injected excessively into the result of BDSD. The results of LRP and VPLGC are affected by blurring effects. For example, the edges of the building in the enlarged areas are smoothed when we compare the fused images in Figs. 6(e) and 6(f) with the reference one. Traditional pan-sharpening methods have larger differences than DNN-based methods. Compared with the fusion results of other DNN-based methods, the M-GAN result suffers from obvious spectral distortions and large errors can also be found in its difference map. The MIP result shows better spatial details thanks to the separation of redundant features. The result of the proposed DRFormer is close to the reference image in terms of visual performance. Moreover, the reconstruction errors of the proposed DRFormer are very small, which demonstrates the effectiveness of the disentangled representation.

Table II gives the average values of all indexes on 30 LRMS and PAN image pairs from the GeoEye-1 dataset. We use bold font to highlight the best value for each index. The proposed DRFormer has the best performance in terms of all indexes. In addition, we can find that the index values of DRFormer are slightly better than those of MIP. The performance gain results from the common information between LRMS and PAN images are retained more reasonably. In MIP, the common information among source images is omitted by the minimization of mutual information.

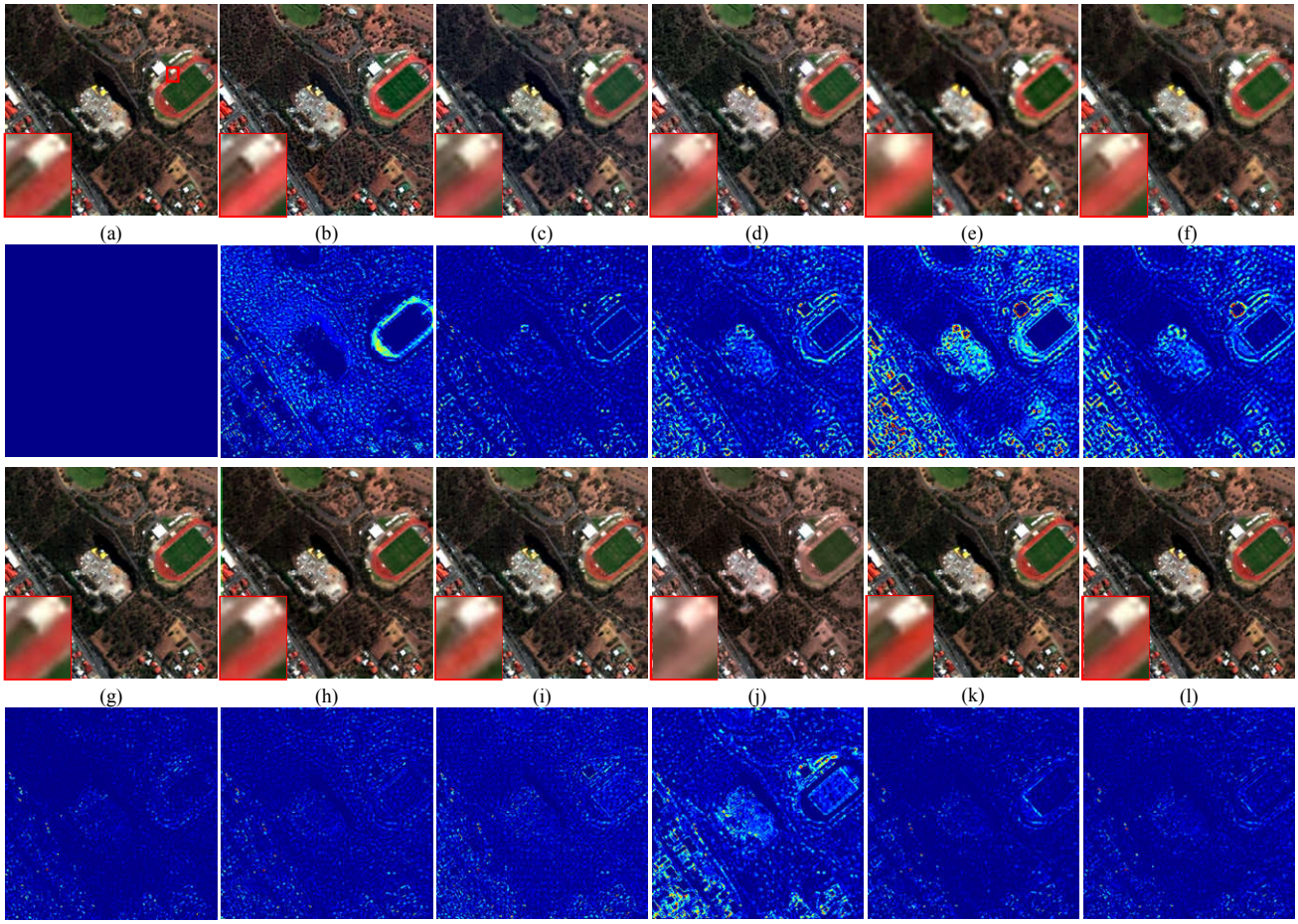


Fig. 6. Qualitative evaluations on the reduced-scale GeoEye-1 dataset. (a) Reference image; (b) BDSD; (c) AWLP; (d) SVT; (e) LRP; (f) VPLGC; (g) TFNet; (h) PanNet; (i) GPPNN; (j) M-GAN; (k) MIP; (l) Proposed DRFormer.

TABLE II. QUANTITATIVE EVALUATIONS OF THE REDUCED-SCALE GEOEYE-1 DATASET

Index	BDSD	AWLP	SVT	LRP	VPLGC	TFNet	PanNet	GPPNN	M-GAN	MIP	Proposed DRFormer
Q4	0.8271	0.8403	0.7950	0.6930	0.7871	0.8638	0.8549	0.8567	0.8221	0.8672	**0.8681**
SAM	5.0332	4.1979	4.9870	4.4823	3.9308	2.8672	3.5864	3.2569	5.0713	2.8252	**2.8163**
ERGAS	1.6718	1.3804	1.5322	2.0664	1.4757	0.9417	1.2129	1.0462	1.7801	0.8962	**0.8954**

Fig. 7 shows the fused images of all methods on the reduced-scale QuickBird dataset. The magnified regions and difference maps are also presented for a more detailed analysis. The fused images in Fig. 7 contain many buildings and abundant texture information. However, traditional methods cannot restore the texture information well, as one can see in the difference maps. For example, there are some contours of buildings in the difference maps of LRP and VPLGC. Moreover, some spectral distortions exist in the results of BDSD and SVT. DNN-based pan-sharpening methods reconstruct the fused image better than traditional methods. However, the errors in difference maps cannot be ignored. For instance, we can see some blocking effects in the results of PanNet and GPPNN. Compared with other methods, the proposed DRFormer enhances the texture information better, because the redundancy among features from sub-networks is reduced by the disentangled representation.

Besides, 30 image pairs from the QuickBird dataset are tested and evaluated. The average results of all fused images are reported in Table III. The proposed method provides the best Q4, SAM, and ERGAS, which implies that the proposed DRFormer preserves spatial and spectral information better.

C. *Experiments on Full-Scale Datasets*

In this section, the full-scale pan-sharpening experiments are implemented on the datasets from GeoEye-1 and QuikBird satellites. Fig. 8 shows the fusion results of all methods on the full-scale GeoEye-1 dataset. An interesting area containing a swimming pool is selected and represented at the bottom left corner of the fused image. The results of AWLP and SVT have similar appearances in terms of buildings, and the color of roofs in the SVT result is different from that in the results of other methods. Blurring effects are observed in the result of LRP, in which the subtle structures may be destroyed by the invalid assumption in LRP. The spatial details in the result of PanNet are over-enhanced. From the magnified region of PanNet, one

can see that some spatial artifacts arise in the flat area of the swimming pool. The hue of the M-GAN result is also different from those of other methods, which may be caused by the excessive constraint on the SAM loss in M-GAN. The proposed DRFormer enhances spatial information better and avoids the introduction of spectral distortions.

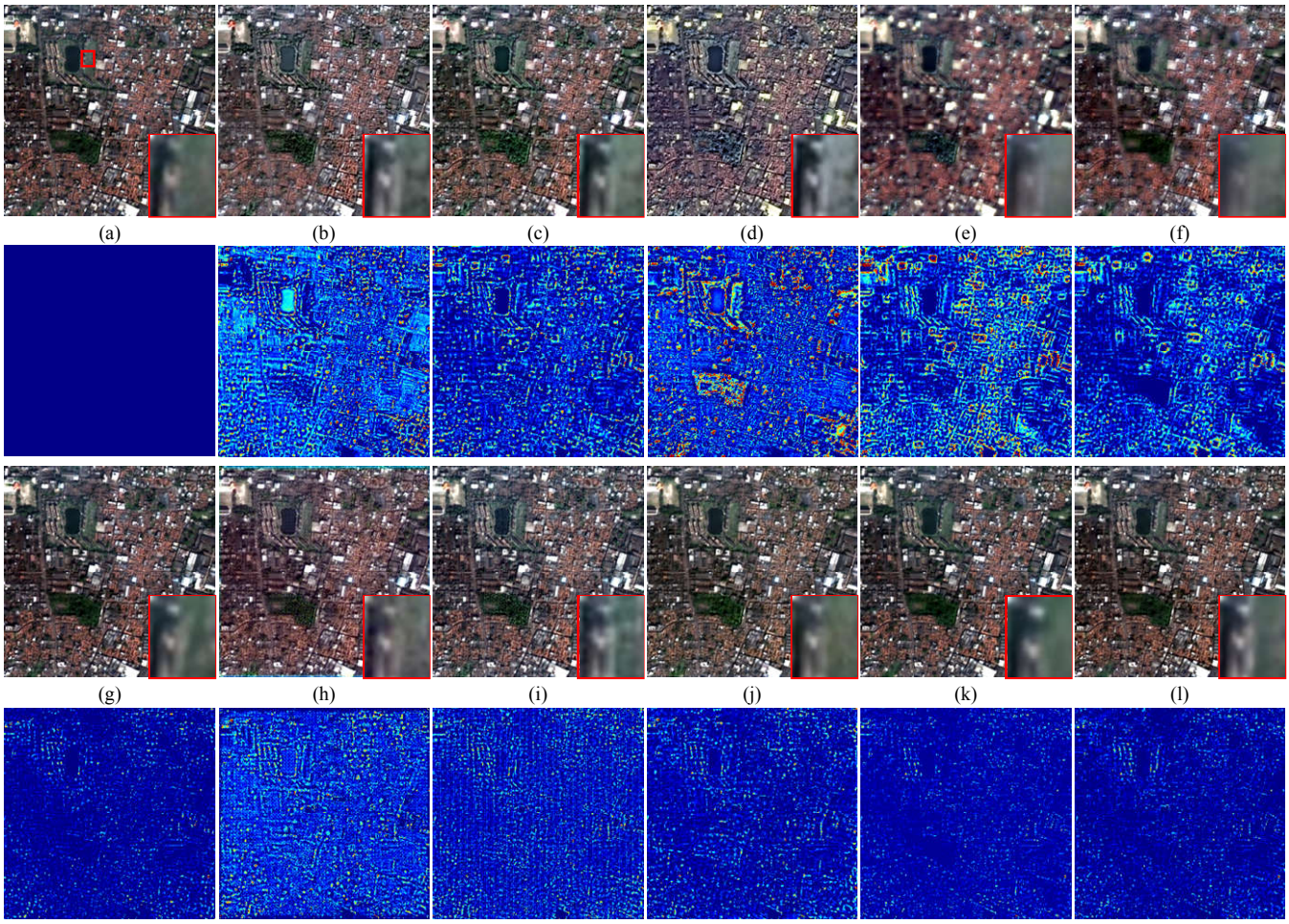


Fig. 7. Qualitative evaluations on the reduced-scale QuickBird dataset. (a) Reference image; (b) BDSD; (c) AWLP; (d) SVT; (e) LRP; (f) VPLGC; (g) TFNet; (h) PanNet; (i) GPPNN; (j) M-GAN; (k) MIP; (l) Proposed DRFormer.

TABLE III. QUANTITATIVE EVALUATIONS OF THE REDUCED-SCALE QUICKBIRD DATASET

Index	BDSD	AWLP	SVT	LRP	VPLGC	TFNet	PanNet	GPPNN	M-GAN	MIP	Proposed DRFormer
Q4	0.8910	0.9250	0.8788	0.7528	0.8510	0.9450	0.9069	0.9335	0.9407	0.9487	**0.9528**
SAM	3.5645	2.3342	4.4927	3.6586	2.6213	1.7102	3.7227	2.6668	2.0677	1.5521	**1.4864**
ERGAS	1.2415	0.8954	1.5196	1.3356	1.0448	0.6235	1.3138	0.8420	0.7288	0.6061	**0.5929**

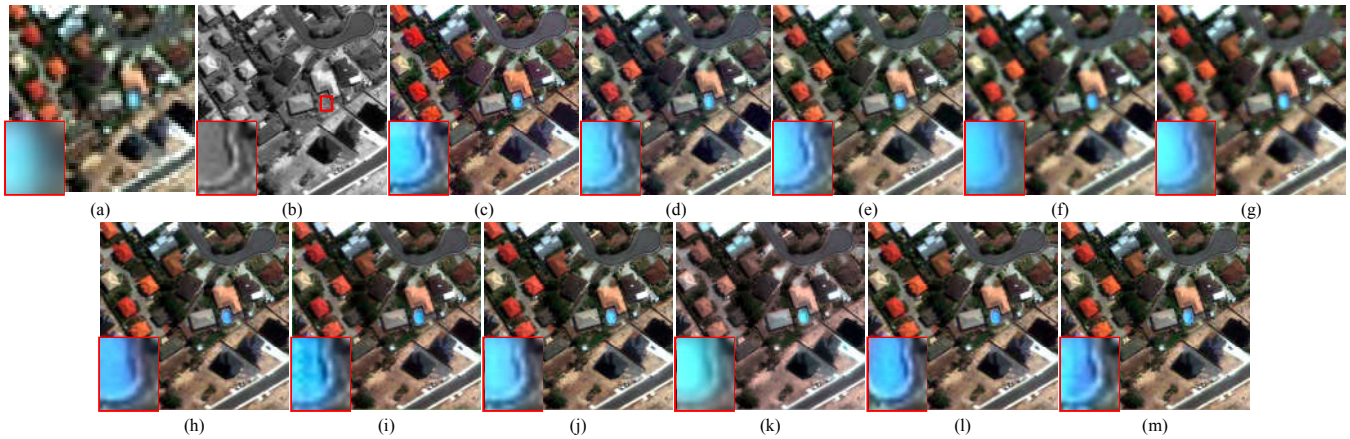


Fig. 8. Qualitative evaluations on the full-scale GeoEye-1 dataset. (a) LRMS image; (b) PAN image (c) BDSD; (d) AWLP; (e) SVT; (f) LRP; (g) VPLGC; (h) TFNet; (i) PanNet; (j) GPPNN; (k) M-GAN; (l) MIP; (m) Proposed DRFormer.

TABLE IV. QUANTITATIVE EVALUATIONS OF THE FULL-SCALE GEOEYE-1 DATASET

Index	BDSD	AWLP	SVT	LRP	VPLGC	TFNet	PanNet	GPPNN	M-GAN	MIP	Proposed DRFormer
D_λ	0.0819	0.0962	0.0615	**0.0380**	0.0421	0.0553	0.0495	0.0545	0.0736	0.0555	0.0392
D_S	0.0444	0.0452	0.0386	0.0709	0.0620	0.0315	0.0310	0.0319	0.0355	0.0320	**0.0298**
QNR	0.8777	0.8632	0.9025	0.8938	0.8994	0.9151	0.0736	0.9154	0.8936	0.9143	**0.9322**

We list the average values of all indexes on the full-scale GeoEye-1 dataset in Table IV. The dataset is composed of 30 LRMS and PAN image pairs. The best spectral index, D_λ, is given by LRP, but the D_λ of the proposed DRFormer is the second-best and very close to that of LRP. Moreover, the best values of other indexes are from DRFormer, which confirms the effectiveness of disentanglement learning.

Fig. 9 illustrates the fusion results on the full-scale QuickBird dataset. Some vegetation regions are included in the source images to be fused. For SVT, it is difficult to preserve the color information of the vegetation regions and serious spectral distortions appear. We can also see some spectral distortions in the magnified area of LRP. The results of TFNet and PanNet suffer from some blocking effects, which may result from the skip connections between shallow layers and deep layers. The results of GPPNN and MIP are slightly blurry. For instance, the textures in its magnified region are lost. The blurring effects in the magnified region of MIP may result from the information loss of the observed scene because the common features between LRMS and PAN images are removed by the constraint of mutual information and not introduced into the reconstruction of the fused image. Compared with other DNN-based methods, the proposed DRFormer has a better balance between spatial and spectral information. Table V reports the index values on 30 LRMS and PAN image pairs. The proposed DRFormer produces the best numerical results on this dataset.

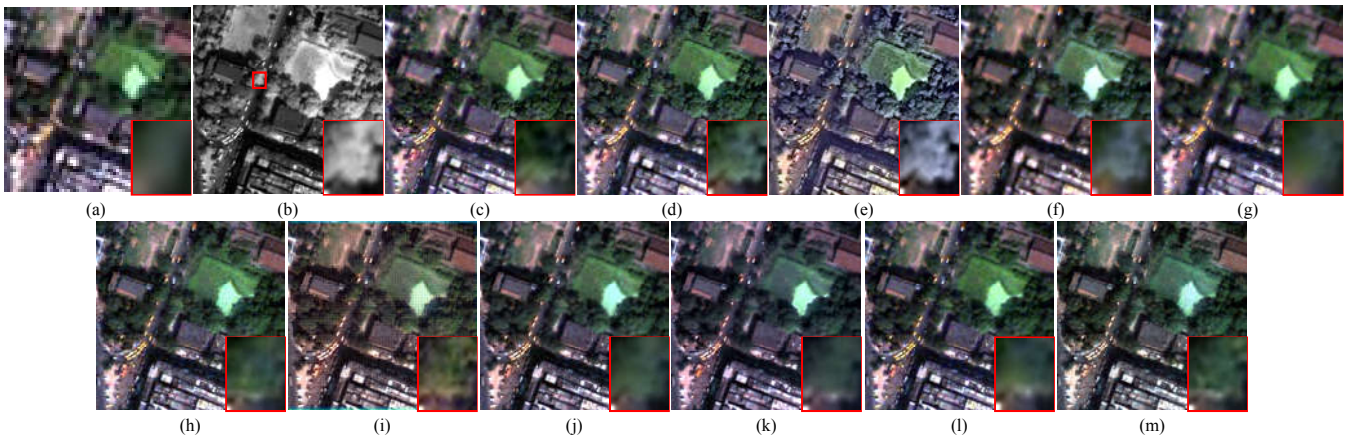


Fig. 9. Qualitative evaluations on the full-scale QuickBird dataset. (a) LRMS image; (b) PAN image (c) BDSD; (d) AWLP; (e) SVT; (f) LRP; (g) VPLGC; (h) TFNet; (i) PanNet; (j) GPPNN; (k) M-GAN; (l) MIP; (m) Proposed DRFormer.

TABLE V. QUANTITATIVE EVALUATIONS OF THE FULL-SCALE QUICKBIRD DATASET

Index	BDSD	AWLP	SVT	LRP	VPLGC	TFNet	PanNet	GPPNN	M-GAN	MIP	Proposed DRFormer
D_λ	0.0491	0.0649	0.1844	0.0477	0.0539	0.0438	0.0589	0.0470	0.0494	0.0510	**0.0397**
D_S	0.0418	0.0500	0.1756	0.0537	0.0486	0.0282	0.0347	0.0327	0.0534	0.0326	**0.0280**
QNR	0.9113	0.8888	0.6738	0.9008	0.8997	0.9293	0.9088	0.9221	0.8994	0.9184	**0.9335**

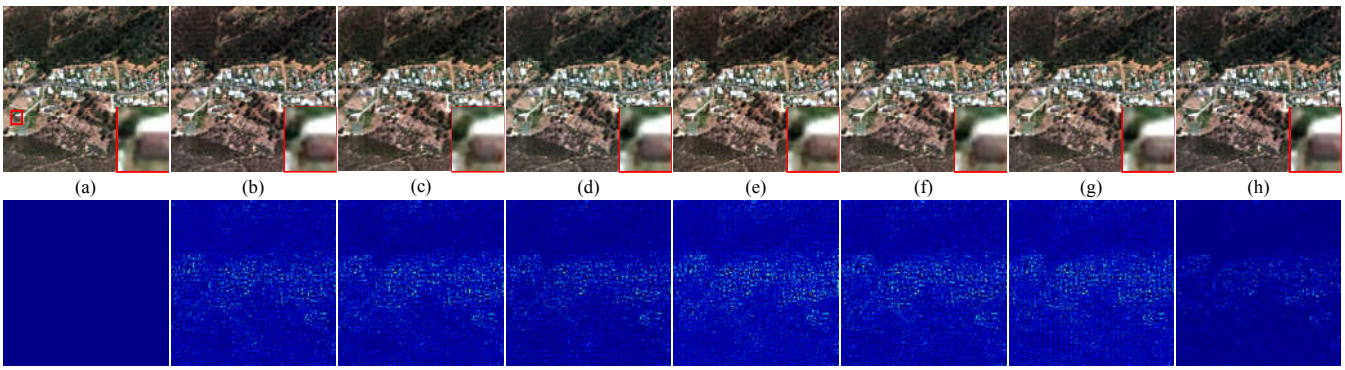


Fig. 10. Ablation study on the reduced-scale GeoEye-1 dataset. (a) Reference image; (b) Case 1; (c) Case 2; (d) Case 3; (e) Case 4; (f) Case 5; (g) Case 6; (h) DRFormer.

TABLE VI. ABLATION STUDY OF LOSS FUNCTIONS ON THE REDUCED-SCALE GEOEYE-1 DATASET.

Loss	$\|\mathbf{L}-\mathbf{L}_C\|_F^2$	$\|\mathbf{P}-\mathbf{P}_C\|_F^2$	$\|\mathbf{L}-\mathbf{L}_S\|_F^2$	$\|\mathbf{P}-\mathbf{P}_S\|_F^2$	$\mathcal{L}_{MI}(S_L,C_L)$	$\mathcal{L}_{MI}(S_P,C_P)$	$\mathcal{L}_{MI}(C_L,C_P)$	Q4	SAM	ERGAS
Case 1	×	×	✓	✓	✓	✓	✓	0.8554	3.3387	1.0669
Case 2	✓	✓	×	×	✓	✓	✓	0.8568	3.2846	1.0537
Case 3	✓	✓	✓	✓	×	×	✓	0.8576	3.2666	1.0499
Case 4	✓	✓	✓	✓	×	×	×	0.8569	3.2511	1.0426
Case 5	×	✓	×	✓	✓	✓	✓	0.8554	3.2968	1.1418
Case 6	✓	×	✓	×	✓	✓	✓	0.8545	3.3509	1.0929
DRFormer	✓	✓	✓	✓	✓	✓	✓	**0.8681**	**2.8163**	**0.8954**

D. Ablation Study

In this section, we investigate the influences of loss functions on the fused images. In (13), $\mathcal{L}_D$ and $\mathcal{L}_S$ control the cross-reconstructions and self-reconstructions of source images, respectively. The correlations among spatial and spectral features are optimized by $\mathcal{L}_{MI}$. Thus, the three terms are removed to show their influence on the fusion results. The related fusion results are shown in Figs. 10(b), 10(c), and 10(e), which correspond to case 1, case 2, and case 4. In case 3, we retain the MI loss between C_L and C_P in the total loss to analyze the influence of common features. In addition, we only remove the losses of the LRMS image in case 5 to demonstrate the influences in terms of spectral information. Similarly, the losses of the PAN image are also omitted in case 6. The fusion results of the two cases are displayed in Figs. 10(f) and 10(g). We can see that the fusion results of different loss functions have similar visual performance. However, DRFormer with the total loss function produces the best-reconstructed result when we compare the difference maps in the second row of Fig. 10.

Table VI presents the average results of DRFormer with different losses on 30 LRMS and PAN image pairs from the reduced-scale GeoEye-1 dataset. When the cross- or self-reconstruction loss is removed, the index values become inferior. Compared with case 4, the Q4 of case 3 is better, which means better overall performance, because the loss of common features is kept. In cases 5 and 6, the losses of the LRMS or PAN image are removed, and the spatial and spectral quality of the fused images becomes poor. When all losses are considered, the proposed DRFormer obtains the best values.

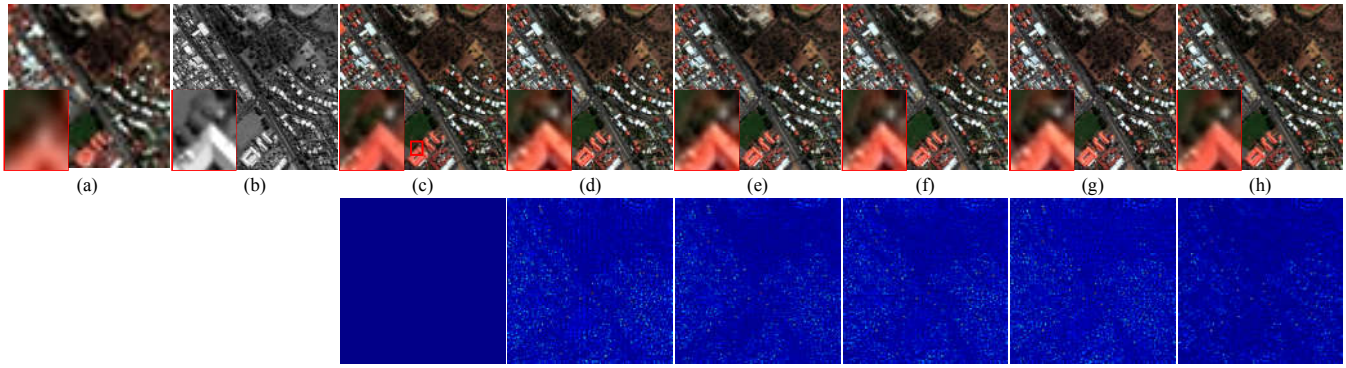


Fig. 11. Qualitative comparison of the fused images produced by different network architectures. (a) LRMS image; (b) PAN image; (c) Reference image; (d) w/o multiscale; (e) w/o self-coupled transformer; (f) w/o cross-coupled transformer; (g) w/o fusion transformer; (h) Complete DRFormer.

E. Analysis of Network Architectures

In the proposed DRFormer, some modules have significant influences on the fusion result. For example, the multiscale formulation in the MMHA module is introduced to exploit the global properties of source images at different scales. Self-coupled and cross-coupled transformers are employed to produce highly coupled features for better reconstructions of source images. In addition, features from different sub-networks are coupled by the fusion transformer. To verify their effectiveness, these modules are removed from DRFormer in this section.

Fig. 11 shows the fusion results of DRFormer with different architectures. A region containing one building is also chosen for specific comparisons. In the magnified region of Fig. 11(d), one can find that the edges of the building are distorted when the multiscale setting is removed from DRFormer. Results in Figs. 11(e) and 11(f) show some blurring artifacts in the magnified areas due to the lack of constraint of self-coupled or cross-coupled transformers. Fig. 11(g) shows that the spectral information is slightly distorted. We can see that the spatial and spectral information is preserved better in the result of the complete DRFormer. The error maps in the second row of Fig. 11 also demonstrate that the complete DRFormer can reconstruct the fused image better.

TABLE VII. QUANTITATIVE EVALUATIONS OF DIFFERENT NETWORKS ON THE REDUCED-SCALE GEOEYE-1 DATASET.

	Q4	SAM	ERGAS
w/o multiscale	0.8545	3.2953	1.0623
w/o self-coupled transformer	0.8574	3.2392	1.0361
w/o cross-coupled transformer	0.8581	3.2234	1.0382
w/o fusion transformer	0.8552	3.3500	1.0682
Complete DRFormer	**0.8681**	**2.8163**	**0.8954**

Table VII provides the average results on the reduced-scale GeoEye-1 dataset. Multiscale formulation and fusion

transformer have greater influences on the fusion results than self-coupled and cross-coupled transformers. The reason for this may be that the multiscale formulation and the fusion transformer are contained in the sub-network for the reconstruction of the fused image. They directly affect the quality of the fused image. Self-coupled and cross-coupled transformers are used to reconstruct the source images, which have a lesser impact on the fusion result. However, the ablation of self-coupled and cross-coupled transformers still degrades the numerical performance. So, the values in Table VII show the effectiveness of each module in the proposed DRFormer.

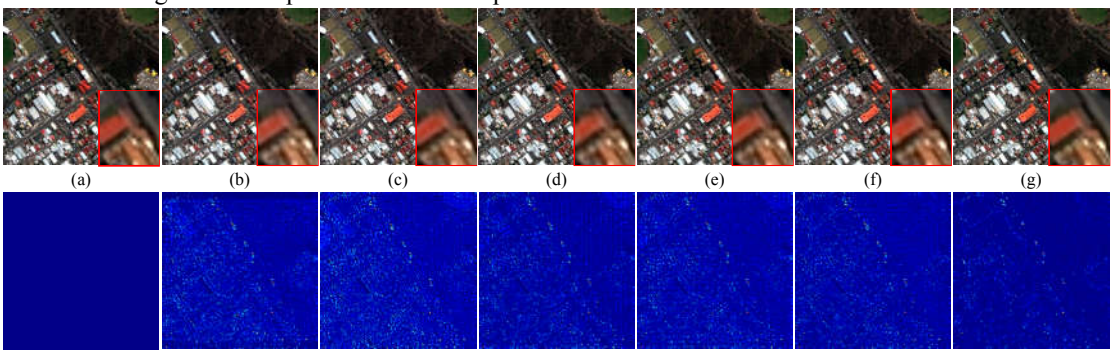


Fig. 12. Qualitative comparison of the fused images produced by different tradeoff parameters. (a) Reference image; (b) $0.1\mathcal{L}_D + 0.1\mathcal{L}_S + 0.0001\mathcal{L}_{MI} + \mathcal{L}_F$; (c) $0.01\mathcal{L}_D + 0.01\mathcal{L}_S + 0.0001\mathcal{L}_{MI} + \mathcal{L}_F$; (d) $\mathcal{L}_D + \mathcal{L}_S + 0.1\mathcal{L}_{MI} + \mathcal{L}_F$; (e) $\mathcal{L}_D + \mathcal{L}_S + 0.01\mathcal{L}_{MI} + \mathcal{L}_F$; (f) $\mathcal{L}_D + \mathcal{L}_S + 0.001\mathcal{L}_{MI} + \mathcal{L}_F$; (g) $\mathcal{L}_D + \mathcal{L}_S + 0.0001\mathcal{L}_{MI} + \mathcal{L}_F$.

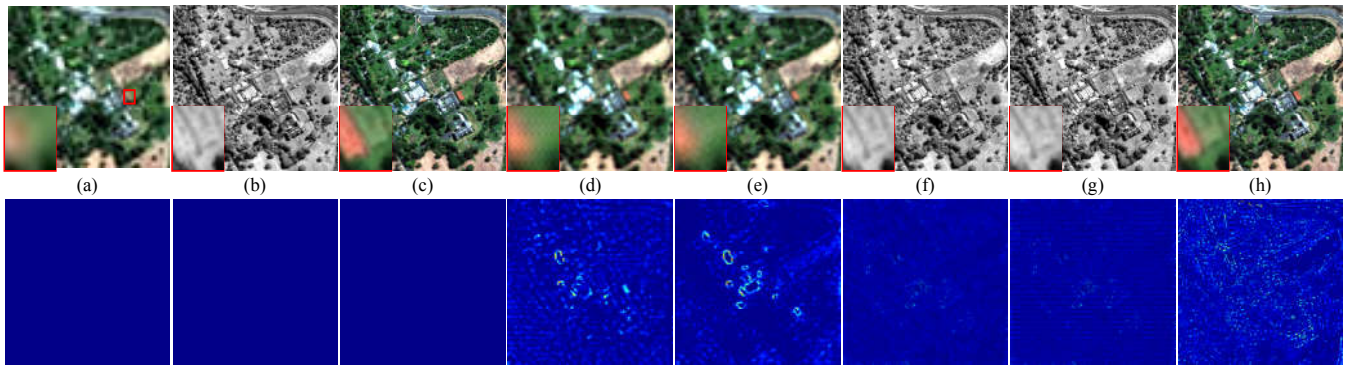


Fig. 13. Qualitative comparison of the reconstructed source images and the fused image on the reduced-scale GeoEye-1 dataset. (a) LRMS image; (b) PAN image; (c) Reference image; (d) Cross-reconstructed LRMS image; (e) Self-reconstructed LRMS image; (f) Cross-reconstructed PAN image; (g) Self-reconstructed PAN image; (h) Fused image.

F. Analysis of Loss Functions

In this part, we analyze the influences of loss functions with different tradeoff parameters on the GeoEye-1 dataset. In (13), we keep the tradeoff parameter of $\mathcal{L}_F$ as 1 and adjust the settings of other terms, including $\mathcal{L}_D$, $\mathcal{L}_S$, and $\mathcal{L}_{MI}$. Because $\mathcal{L}_D$ and $\mathcal{L}_S$ are both introduced for the constraint of disentangled features, the two terms adopt the same tradeoff parameters. Specifically, the tradeoff parameters on $\mathcal{L}_D$ and $\mathcal{L}_S$ vary from 0.01 to 1. For $\mathcal{L}_{MI}$, its tradeoff parameter decreases from 0.1 to 0.0001. Fig. 12 and Table VIII show the fusion results and evaluation results of the proposed DRFormer with different tradeoff parameters, respectively. In Fig. 12, the error maps of fusion results are also shown for a more direct comparison. Although the visual performance of the fused images in Fig. 12 is close, reconstruction error maps demonstrate the influences of different parameter settings. One can find that the fused image is reconstructed better when the tradeoff parameters of $\mathcal{L}_D$, $\mathcal{L}_S$, and $\mathcal{L}_{MI}$ are set as 1, 1, and 0.0001. In Table VIII, the best values are labeled in bold and we can find that the variations of the tradeoff parameters on $\mathcal{L}_D$ and $\mathcal{L}_S$ have larger influences on Q4 and SAM. It proves that sufficient constraints on $\mathcal{L}_D$ and $\mathcal{L}_S$ can guarantee the compatibility between disentangled features and these features are integrated better in the fused image. For $\mathcal{L}_{MI}$, the influence of the tradeoff parameter is smaller but over-constrained $\mathcal{L}_{MI}$ may have a negative effect on the compatibility of disentangled features. Finally, we set the tradeoff parameters of $\mathcal{L}_D$, $\mathcal{L}_S$, and $\mathcal{L}_{MI}$ as 1, 1, and 0.0001, respectively.

TABLE VIII. QUANTITATIVE EVALUATIONS OF LOSS FUNCTIONS WITH DIFFERENT TRADEOFF PARAMETERS ON THE REDUCED-SCALE GEOEYE-1 DATASET.

$\mathcal{L}_D$	$\mathcal{L}_S$	$\mathcal{L}_{MI}$	Q4	SAM	ERGAS
0.1	0.1	0.0001	0.8551	3.3290	1.0790
0.01	0.01	0.0001	0.8552	3.2827	1.0750
1	1	0.1	0.8620	2.9283	1.0662
1	1	0.01	0.8606	2.9221	1.0633
1	1	0.001	0.8634	2.8967	1.0350
1	1	0.0001	**0.8681**	**2.8163**	**0.8954**

G. Reconstructions of Source Images

In the proposed method, source images are reconstructed for

the disentanglement of sensor-specific and common features. In this section, we compare the reconstructed source images with the original source images. Fig. 13 shows the reconstructed source images and the fused image on the reduced-scale GeoEye-1 dataset. For the reconstructions of the LRMS image, some larger errors can be found in difference maps. These errors are focused on the building areas with high exposure. Slight block effects appear in the magnified areas of the reconstructed LRMS images. The reason for this may be the down-sampling and up-sampling operations in the proposed network. However, the overall performance of the reconstruction of the LRMS image is satisfactory. In addition, we can find that the PAN image is reconstructed better than the LRMS image. Indeed, the errors in difference maps are very small. Moreover, the fused image is also very close to the reference image. Thus, the introduction of auxiliary reconstruction tasks contributes to the quality improvement of the fused image.

H. Running Time and Model Size

Table IX presents the training time and model sizes of DNN-based pan-sharpening methods. The model size of the proposed DRFormer is close to that of the M-GAN and larger than those of other methods. For M-GAN, most parameters are from the generator because a very deep network is adopted to ensure the pan-sharpening performance. For the proposed DRFormer, its model size is much larger than those of TFNet, PanNet, and GPPNN. The reason for this is that four sub-networks are introduced as auxiliary means to reconstruct LRMS and PAN images for disentangled representation. The extra sub-networks increase the model size of DRFormer significantly. Meanwhile, with the introduction of four sub-networks, its computational complexity is also further boosted. However, the reconstruction tasks are removed when the proposed method is in the test phase. The extracted features need only to be fed into the corresponding decoder to generate the final fused image. Moreover, the training time of the proposed method is smaller than that of M-GAN. Besides, the proposed method also behaves better than GPPNN in terms of training time, although its model size is larger than that of GPPNN. In GPPNN, many blocks are cascaded to synthesize the fusion result.

TABLE IX. NUMBER OF PARAMETERS AND TRAINING TIME OF DIFFERENT METHODS.

	TFNet	**PanNet**	**GPPNN**	**M-GAN**	**DRFormer**
#Para. (MB)	2.36	0.15	0.12	15.4	15.1
Training time (h)	50.8	6.2	84.9	103.3	59.4

V. CONCLUSION

In this paper, we have presented a novel pan-sharpening method, DRFormer, to disentangle the sensor-specific features and common features in LRMS and PAN images. In the proposed method, LRMS and PAN images are decomposed as sensor-specific and common features by cross- and self-reconstructions in the feature space. Cross-coupled transformers and self-coupled transformers are designed to integrate the disentangled features, by which source images can be better reconstructed. In addition, the maximization of MI is imposed on the common features of LRMS and PAN images to ensure consistency among them. The MI between sensor-specific features and common features is also minimized to separate them effectively. Finally, the fused image is produced by combining all disentangled features. We have conducted extensive experiments on GeoEye-1 and QuickBird datasets. Experimental results show that the proposed DRFormer obtains state-of-the-art pan-sharpening results. Owing to the introduction of four auxiliary sub-networks, the model size and complexity of the proposed DRFormer are boosted significantly. For future work, we will design more efficient disentangled representation techniques to separate the sensor-specific and common features.

REFERENCES

[1] S. Jia, Z. Lin, B. Deng, J. Zhu, and Q. Li, "Cascade superpixel regularized Gabor feature fusion for hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst*., vol. 31, no. 5, pp. 1638-1652, May 2020.

[2] C. Paris, L. Bruzzone, D. Fernández-Prieto, "A novel approach to the unsupervised update of land-cover maps by classification of time series of multispectral images," *IEEE Trans. Geosci. Remote Sens*., vol. 57, no. 7, pp. 4259-4277, Jul. 2019.

[3] W. Diao, F. Zhang, J. Sun, Y. Xing, K. Zhang, L. Bruzzone, "ZeRGAN: Zero-reference GAN for fusion of multispectral and panchromatic images," *IEEE Trans. Neural Netw. Learn. Syst*., early access, 2022, doi: 10.1109/TNNLS.2021.3137373.

[4] Z. Sheng, F. Zhang, J. Sun, Y. Tan, K. Zhang, L. Bruzzone, "A unified two-stage spatial and spectral network with few-shot learning for pan-sharpening," *IEEE Trans. Geosci. Remote Sens*., vol. 61, pp. 5403517, 2023.

[5] K. Zheng, J. Huang, M. Zhou, D. Hong, F. Zhao, "Deep adaptive pansharpening via uncertainty-aware image fusion," *IEEE Trans. Geosci. Remote Sens*., vol. 61, pp. 5403715, 2023.

[6] Z. Wu, T. Huang, L. Deng, J. Hu, G. Vivone, "VO+Net: An adaptive approach using variational optimization and deep learning for panchromatic sharpening," *IEEE Trans. Geosci. Remote Sens*., vol. 60, pp. 5401016, 2022.

[7] K. Yan, M. Zhou, L Zhang, C. Xie, "Memory-augmented model-driven network for pansharpening," in Proc. *ECCV*, Nov. 2022, pp. 306-322.

[8] J. Yang, L. Xiao, Y. Zhao, J. Chan, "Variational regularization network with attentive deep prior for hyperspectral–multispectral image fusion," *IEEE Trans. Geosci. Remote Sens*., vol. 60, pp. 5508817, 2022.

[9] J. Xiao, T. Huang, L. Deng, Z. Wu, X. Wu, G. Vivone, "Variational pansharpening based on coefficient estimation with nonlocal regression," *IEEE Trans. Geosci. Remote Sens*., vol. 61, pp. 5406115, 2023.

[10] X. Wang, H. Chen, Y. Zhou, J. Ma, W. Zhu, "Disentangled representation learning for recommendation," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 45, no. 1, pp. 408 - 424, 2023.

[11] X. Liu, P. Sanchez, S. Thermos, A. O'Neil, S. Tsaftaris, "Learning disentangled representations in the imaging domain," *Medical Image Anal*., vol. 80, pp. 102516, 2022.

[12] H. Xu, X. Wang, J. Ma, "DRF: Disentangled representation for visible and infrared image fusion," *IEEE Trans. Instrum. Meas*., vol. 70, pp. 5006713, 2021.

[13] D. Hong, J. Yao, D. Meng, N. Yokoya, and J. Chanussot, "Decoupled-and-coupled networks: Self-supervised hyperspectral image super-resolution with subpixel fusion," 2022, arXiv: 2205.03742. [Online]. Available: https://arxiv.org/abs/2205.03742.

[14] X. Cao, X. Fu, D. Hong, Z. Xu, and D. Meng, "PanCSC-Net: A model-driven deep unfolding method for pansharpening," *IEEE Trans. Geosci. Remote Sens*., vol. 60, pp. 5404713, 2022.

[15] M. Zhou, K. Yan, J. Huang, Z. Yang, X. Fu, F. Zhao, "Mutual information-driven pan-sharpening," in Proc. *IEEE CVPR*, Jun. 2022, pp. 1788-1798.

[16] G. Vivone, M. D. Mura, and A. Garzelli *et al*., "A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods," *IEEE Geosci. Remote Sens. Mag*., vol. 9, no. 1, pp. 53-81, Mar. 2021.

[17] K. Zhang *et al*., "Panchromatic and multispectral image fusion for remote sensing and Earth observation: Concepts, taxonomy, literature review, evaluation methodologies and challenges ahead," *Inf. Fusion*, vol. 93, pp. 227-242, May 2023.

[18] G. Vivone *et al*., "A critical comparison among pansharpening algorithms," *IEEE Trans. Geosci. Remote Sens*., vol. 53, no. 5, pp. 2565-2586, May 2015.

[19] T. Tu, P. Huang, C. Hung, and C. Chang, "A fast intensity-hue-saturation fusion technique with spectral adjustment for IKONOS imagery," *IEEE Geosci. Remote Sens. Lett*., vol. 1, no. 4, pp. 309-312, Oct. 2004.

[20] C. A. Laben and B. V. Brower, "Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening," U.S. Patent 6011875, 2000.

[21] P. S. Chavez Jr., S. C. Sides, and J. A. Anderson, "Comparison of three different methods to merge multiresolution and multispectral data: Landsat TM and SPOT panchromatic," *Photogramm. Eng. Remote Sens.*, vol. 57, no. 3, pp. 295-303, Mar. 1991.

[22] S. Rahmani, M. Strait, D. Merkurjev, M. Moeller, T. Wittman, "An adaptive IHS pan-sharpening method," *IEEE Geosci. Remote Sens. Lett*., vol. 7, no. 4, pp. 746-750, Apr. 2010.

[23] A. Garzelli, F. Nencini, and L. Capobianco, "Optimal MMSE pan sharpening of very high resolution multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 1, pp. 228-236, Jan. 2008.

[24] X. Otazu, M. González-Audícana, O. Fors, and J. Nunez, "Introduction of sensor spectral response into image fusion methods. Application to wavelet-based methods," *IEEE Trans. Geosci. Remote Sens*., vol. 43, no. 10, pp. 2376-2385, Oct. 2005.

[25] F. Nencini, A. Garzelli, S. Baronti, and L. Alparone, "Remote sensing image fusion using the curvelet transform," *Inf. Fus*., vol. 8, no. 2, pp. 143-156, Sept. 2007.

[26] S. Zheng, W. Shi, J. Liu, and J. Tian, "Remote sensing image fusion using multiscale mapped LS-SVM," *IEEE Trans. Geosci. Remote Sens*., vol. 46, no. 5, pp. 1313-1322, May 2008.

[27] K. Zhang, M. Wang, S. Yang, and L. Jiao, "Convolution structure sparse coding for fusion of panchromatic and multispectral images," *IEEE Trans. Geosci. Remote Sens*., vol. 57, no. 2, pp. 1117-1130, Feb. 2019.

[28] H. A. Aly and G. Sharma, "A regularized model-based optimization framework for pan-sharpening," *IEEE Trans. Image Process*., vol. 23, no. 6, pp. 2596-2608, Apr. 2014.

[29] S. Li and B. Yang, "A new pan-sharpening method using a compressed sensing technique," *IEEE Trans. Geosci. Remote Sens*., vol. 49, no. 2, pp. 738-746, Feb. 2011.

[30] F. Zhang, H. Zhang, K. Zhang, Y. Xing, J. Sun, Q. Wu, "Exploiting low-rank and sparse properties in strided convolution matrix for pansharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens*., vol. 14, pp. 2649-2661, 2021.

[31] S. Yang, K. Zhang, and M. Wang, "Learning low-rank decomposition for pan-sharpening with spatial-spectral offsets," *IEEE Trans. Neural Netw. Learn. Syst*., vol. 29, no. 8, pp. 3647-3657, Aug. 2018.

[32] K. Zhang, M. Wang, S. Yang, Y. Xing, and R. Qu, "Fusion of panchromatic and multispectral images via coupled sparse nonnegative matrix factorization," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens*., vol. 9, no. 12, pp. 5740-5747, Dec. 2016.

[33] F. Fang, F. Li, C. Shen, G. Zhang, "A variational approach for pan-sharpening," *IEEE Trans. Image Process*., vol. 22, no. 7, pp. 2822-2834, Jul. 2013.

[34] X. Tian, Y. Chen, C. Yang, J. Ma, "Variational pansharpening by exploiting cartoon-texture similarities," *IEEE Trans. Geosci. Remote Sens*., vol.60, pp. 5400416, 2022.

[35] J. Duran, A. Buades, B. Coll, C. Sbert, and G. Blanchet, "A survey of pansharpening methods with a new band-decoupled variational model," *ISPRS J. Photogramm. Remote Sens*., vol. 125, pp. 78-105, Mar. 2017.

[36] C. Chen, Y. Li, W. Liu, and J. Huang, "SIRF: Simultaneous satellite image registration and fusion in a unified framework," *IEEE Trans. Image Process*., vol. 24, no. 11, pp. 4213-4224, Nov. 2015.

[37] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sens*., vol. 8, no. 7, pp. 594:1-594:22, July 2016.

[38] G. Scarpa, S. Vitale, and D. Cozzolino, "Target-adaptive CNN-based pansharpening," *IEEE Trans. Geosci. Remote Sens*., vol.56, no. 9, pp. 5443-5457, Sept. 2018.

[39] J. Cai and B. Huang, "Super-resolution-guided progressive pansharpening based on a deep convolutional neural network," *IEEE Trans. Geosci. Remote Sens*., vol. 59, no. 6, pp. 5206-5220, Jun. 2021.

[40] T. Zhang, L. Deng, T. Huang, J. Chanussot, G. Vivone, "A triple-double convolutional neural network for panchromatic sharpening," *IEEE Trans. Neural Netw. Learn. Syst*., early access, 2022, doi: 10.1109/TNNLS.2022.3155655.

[41] Y. Yang, W. Tu, S. Huang, H. Lu and B. Huang *et al*., "Dual-stream convolutional neural network with residual information enhancement for pansharpening," *IEEE Trans. Geosci. Remote Sens*., vol. 60, pp. 5402416, 2021.

[42] X. Fu, W. Wang, Y. Huang, X. Ding, and J. Paisley, "Deep multiscale detail networks for multiband spectral image sharpening," *IEEE Trans. Neural Netw. Learn. Syst*., vol. 32, no. 5, pp. 2090-2104, May 2021.

[43] H. Zhou, Q. Liu, Y. Wang, "PGMAN: An unsupervised generative multiadversarial network for pansharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens*., vol. 14, pp. 6316-6327, 2021.

[44] Q. Liu, H. Zhou, Q. Xu, X. Liu, and Y. Wang, "PSGAN: A generative adversarial network for remote sensing image pan-sharpening," *IEEE Trans. Geosci. Remote Sens*., vol. 59, no. 12, pp. 10227-10242, Dec. 2021.

[45] D. Lei, H. Chen, L. Zhang, and W. Li, "NLRNet: An efficient nonlocal attention ResNet for pansharpening," *IEEE Trans. Geosci. Remote Sens*., vol. 60, pp. 5401113, 2022.

[46] Y. Qu, R. Baghbaderani, H. Qi, C. Kwan, "Unsupervised pansharpening based on self-attention mechanism," *IEEE Trans. Geosci. Remote Sens*., vol. 59, no. 4, pp. 3192-3208, Apr. 2021.

[47] X. Meng, N. Wang, F. Shao, S. Li, "Vision transformer for pansharpening," *IEEE Trans. Geosci. Remote Sens*., vol. 60, pp. 5409011, 2022.

[48] F. Zhang, K. Zhang, J. Sun, "Multiscale spatial-spectral interaction transformer for pan-sharpening," *Remote Sens*., vol. 14, no. 7, pp. 1736, 2022.

[49] S. Li, Q. Guo, A. Li, "Pan-sharpening based on CNN+pyramid transformer by using no-reference loss," *Remote Sens*., vol. 14, no. 3, pp. 624, 2022.

[50] X. Sun, J. Li, Z. Hua, "Transformer-based regression network for pansharpening remote sensing images," *IEEE Trans. Geosci. Remote Sens*., vol. 60, pp. 5407423, 2022.

[51] Y. Yan, J. Liu, S. Xu, Y. Wang, X. Cao, "MD³Net: Integrating model-driven and data-driven approaches for pansharpening," *IEEE Trans. Geosci. Remote Sens*., vol. 60, pp. 5411116, 2022.

[52] X. Li, Y. Li, G. Shi, L. Zhang, W. Li, D. Lei, "Pansharpening method based on deep nonlocal unfolding," *IEEE Trans. Geosci. Remote Sens*., vol. 61, pp. 5404111, 2023.

[53] K. Zhang, A. Wang, F. Zhang, W. Wan, J. Sun. L. Bruzzone, "Spatial-spectral dual back-projection network for pansharpening," *IEEE Trans. Geosci. Remote Sens*., vol. 61, pp. 5402216, 2023.

[54] M. Zhou, K. Yan, J. Pan, W. Ren, Q. Xie, X. Cao, "Memory-augmented deep unfolding network for guided image super-resolution," *Int. J. Comput. Vis.*, vol. 131, pp. 215-242, 2023.

[55] Z. Li, J. Li, F. Zhang, L. Fan, "CADUI: Cross-attention-based depth unfolding iteration network for pansharpening remote sensing images," *IEEE Trans. Geosci. Remote Sens*., vol. 61, pp. 5402420, 2023.

[56] J. Li, Y. Li, L. Zhou, L. Kuang, T. Wu, "USID-Net: Unsupervised single image dehazing network via disentangled representations," *IEEE Trans. Multimedia*, vol. 25, pp. 3587-3601, 2023.

[57] X. Luo, Y. Gao, A. Wang, Z. Zhang, X. Wu, "IFSepR: A general framework for image fusion based on separate representation learning," *IEEE Trans. Multimedia*, vol. 25, pp. 608-623, 2023.

[58] Y. Gao, S. Ma, J. Liu, "DCDR-GAN: A densely connected disentangled representation generative adversarial network for infrared and visible image fusion," *IEEE Trans. Circ. Syst. Vid. Tech.*, vol. 33, no. 2, pp. 549-561, Feb. 2023.

[59] L. Tang, X. Xiang, H. Zhang, M. Gong, J. Ma, "DIVFusion: Darkness-free infrared and visible image fusion," *Inf. Fusion*, vol. 91, pp. 477-493, 2023.

[60] G. Yang, K. Zhang, F. Zhang, J. Wang, J. Sun, "Cross-resolution semi-supervised adversarial learning for pansharpening," *IEEE Trans. Geosci. Remote Sens*., vol. 61, pp. 5402617, 2023.

[61] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, L. Gool, "CDDFuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion," in Proc. *IEEE CVPR*, Jun. 2023, pp. 5906-5916.

[62] X. Fu, Z. Lin, Y. Huang, and X. Ding, "A variational pan-sharpening with local gradient constraints," in Proc. *IEEE CVPR*, Jun. 2019, pp. 10265-10274.

[63] X. Liu, Q. Liu, Y. Wang, "Remote sensing image fusion based on two-stream fusion network," *Inf. Fusion*, vol. 55, pp. 1-15, 2020.

[64] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, "PanNet: A deep network architecture for pan-sharpening," in *Proc. IEEE ICCV*, Oct. 2017, pp. 5449-5457.

[65] S. Xu, J. Zhang, Z. Zhao, K. Sun, J. Liu, and C. Zhan, "Deep gradient projection networks for pan-sharpening," in Proc. *IEEE CVPR*, Jun. 2021, pp. 1366-1375.

[66] A. Gastineau, J. Aujol, Y. Berthoumieu, and C. Germain, "Generative adversarial network for pansharpening with spectral and spatial discriminators," *IEEE Trans. Geosci. Remote Sens*., vol. 60, pp. 1–11, 2022.

[67] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva, "Multispectral and panchromatic data fusion assessment without reference," *Photogramm. Eng. Remote Sens*., vol. 74, no. 2, pp. 193-200, Feb. 2008.