Pan-sharpening with Customized Transformer and Invertible Neural Network

Man Zhou[1†], Jie Huang[1†], Yanchi Fang[2], Xueyang Fu[1], Aiping Liu[1*]
[1] **University of Science and Technology of China, China**
[2] **University of Toronto, Canada**
manman,hj0117@mail.ustc.edu.cn, aipingl@ustc.edu.cn

Abstract

In remote sensing imaging systems, pan-sharpening is an important technique to obtain high-resolution multispectral images from a high-resolution panchromatic image and its corresponding low-resolution multispectral image. Owing to the powerful learning capability of convolution neural network (CNN), CNN-based methods have dominated this field. However, due to the limitation of the convolution operator, long-range spatial features are often not accurately obtained, thus limiting the overall performance. To this end, we propose a novel and effective method by exploiting a customized transformer architecture and information-lossless invertible neural module for long-range dependencies modeling and effective feature fusion in this paper. Specifically, the customized transformer formulates the PAN and MS features as queries and keys to encourage joint feature learning across two modalities while the designed invertible neural module enables effective feature fusion to generate the expected pan-sharpened results. To the best of our knowledge, this is the first attempt to introduce transformer and invertible neural network into pan-sharpening field. Extensive experiments over different kinds of satellite datasets demonstrate that our method outperforms state-of-the-art algorithms both visually and quantitatively with fewer parameters and flops. Further, the ablation experiments also prove the effectiveness of the proposed customized long-range transformer and effective invertible neural feature fusion module for pan-sharpening.

Introduction and Related Work

With the rapid development of satellite sensors, satellite images have been used in a wide range of applications like military systems, environmental monitoring, and mapping services. However, due to the technological and physical limitation of imaging devices, satellites are usually equipped with both multispectral (MS) and panchromatic (PAN) sensors to simultaneously measure the complementary images, MS images with low spatial resolution and high spectral resolution, and PAN images with low spectral resolution and high spatial resolution. To obtain the images with both high spectral and high spatial resolutions, the pan-sharpening technique that fuses the low-resolution MS images and high spatial PAN images to break the technological limits, has drawn much attention from either image processing and remote sensing communities.

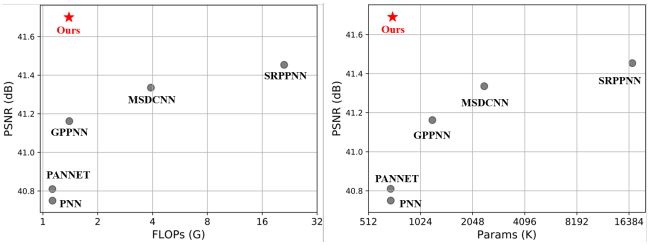

Figure 1: Trade-off between PSNR, number of parameters and FLOPs over worldview-II dataset.

In the past few decades, a deal of pan-sharpening algorithms has been proposed and obtained promising results. The traditional algorithms include component substitutes (Aiazzi and Selva 2007; Choi and Kim 2011; Kang and Benediktsson 2014), multiresolution analysis (Aiazzi et al. 2003; Kaplan and Erer 2012; Yokoya et al. 2012; Shah, Younan, and King 2008) and model-based methods (Ghahremani et al. 2016; Garzelli, Nencini, and Capobianco 2008). However, all of them are generally based on handcrafted features, with limited capacity to reconstruct the missing information in the MS images. Very recently, to overcome the aforementioned shortcomings, researchers focus on exploiting the powerful feature representation capability of convolution neural networks (CNNs) to construct numerous CNNs-based pan-sharpening methods (Wang et al. 2021a,b; Xu et al. 2021a; Peng et al. 2021; Benzenati, Kallel, and Kessentini 2021; Hu et al. 2021; Liu et al. 2020; Xu et al. 2021b; Cai and Huang 2021), which outperforms previous state-of-the-art methods by a large margin. However, existing CNN-based methods remain some limitations: 1) lacking the modeling of long-range dependency owing to the local neighbor reception characterize of convolution operator, 2) ineffective feature extraction and fusion. Both result in the loss of some essential feature that might be useful for an exemplary pan-sharpened image.

Long-range dependency modeling. Transformer architecture is firstly proposed and has achieved a remarkable performance in the natural language processing field (Vaswani et al. 2017). Different from the local reception characterize of convolution operator, transformer architecture is naturally

*Corresponding author. †Co-first authors contributed equally.

good at capturing the long-range dependencies by employing the multi-head global attention mechanism among different ordered input feature parts. Afterward, motivated by their success, many researchers have begun to introduce the transformer structure into computer vision (Yuan et al. 2021; Li et al. 2021). The pioneering work is the visual transformer (ViT) (Vaswani et al. 2017) for the image recognition task, which obtains excellent results compared with state-of-the-art CNN-based methods. Since then, transformer-based methods have emerged to successfully work in other computer vision problems like object detection (Dosovitskiy et al. 2020; Carion et al. 2020; Zhu et al. 2020), image segmentation (H. Wang and Chen 2021) and image restoration (Wang et al. 2021c; Chen et al. 2021a) as well. However, it has not been explored in the pan-sharpening task. In addition, existing transformer architectures are designed to find the self-similarity in a single image. The goal of pan-sharpening is to seek the interactive information between two kinds of modality images, MS image and PAN image. To achieve this, inspired by (Yang et al. 2020), we redevelop a customized pan-sharpening transformer architecture. Specifically, the proposed transformer formulates the PAN and MS features as queries and keys to encourage joint feature learning across two modalities for searching the long-range features, shown in Figure 4.

Effective feature extraction and fusion. The goal of the pan-sharpening task is to fuse the complementary information from MS image and PAN image to generate high spatial resolution MS image. As recognized, how to effectively extract and fuse complementary information is crucial for pan-sharpening performance. Specifically, most of the existing pan-sharpening methods directly concatenate the MS and PAN image in the image space and then feed them into a single-stream shared convolution encoder for feature extraction and fusion. The remaining methods adapt two-stream independent convolution encoders to provide the modality-specific feature maps from MS and PAN images, and then concatenate the obtained feature maps for fusion in the feature space. However, the above methods have not fully investigated the feature extraction and fusion potentials. To this end, we design two schemes: 1) local and long-range feature extraction module; 2) densely-connected invertible neural network fusion module. Specifically, the former consists of two branches, local convolution branch and long-range transformer branch. Both of them receive the MS image and PAN image as input for local and long-range feature extraction. Due to the natural information lossless capability of invertible neural architecture (Dinh, Krueger, and Bengio 2015; Laurent Dinh and Bengio. 2017), different from existing methods adapting pure convolution layers to achieve fusion, we design a new densely-connected invertible neural network for effective feature fusion. The implementation details can refer to Figure 3.

In a word, we propose a novel effective pan-sharpening method by combining the advantages of long-range dependencies modeling of transformer architecture and information-lossless invertible neural network in this paper. To the best of our knowledge, this is the first attempt to introduce transformer and invertible neural network into pan-sharpening field. As shown in Figure 2, our method consists of three procedures: 1) local and long-range feature extraction by convolution and transformer, 2) effective local and long-range feature fusion by densely-connected invertible neural module, and 3) high-resolution MS image reconstruction. Extensive experiments over different kinds of satellite datasets demonstrate that our method outperforms state-of-the-art algorithms both visually and quantitatively. Further, the ablation experiments also prove the effectiveness of the proposed customized long-range modeling of transformer and effective invertible neural feature fusion module.

Our contributions can be summarized as follows:

- We propose a novel pan-sharpening method by combining advantages of long-range dependencies modeling of transformer architecture and effective feature fusion capability of invertible neural network in this paper. To the best of our knowledge, this is the first attempt to introduce transformer and invertible neural network into the pan-sharpening field.
- We design a customized Transformer architecture for pan-sharpening and a new densely-connected invertible neural module. The ablation experiments also prove the effectiveness of the proposed transformer and invertible neural feature fusion module.
- Extensive experiments over different kinds of satellite datasets demonstrate that our method outperforms state-of-the-art algorithms both visually and quantitatively with fewer parameters and running flops.

Methodology

In this section, we first illustrate the overall architecture of our pan-sharpening network. It has two core designs to make it suitable for pan-sharpening, local and long-range feature extraction module, and effective densely-connected invertible feature fusion neural module. The details will be illustrated below.

Overall Network Architecture.

The overall structure is shown in Figure 2. It takes the MS image and PAN image as input and integrates the texture details of the high-resolution (HR) PAN images with the spectral information from low-resolution (LR) MS images to generate HR-MS images. To be specific, given the PAN image $P \in R^{1\times H\times W}$ and MS image $M \in R^{C\times \frac{H}{4}\times \frac{W}{4}}$, our method firstly applies two independent 3×3 convolution layers to project the up-sampling MS with four times and PAN image into feature space with modality-specific features, P_0 and M_0. Next, the feature maps P_0, and M_0 are passed through two-stream local and long-range feature extraction module. The local branch consists of several convolution layers and provides the local-range feature maps, while the transformer branch takes advantage of multi-head attention to generates the long-range features between the flatten feature patches from P_0 and M_0. The obtained local and long-range features are remarked as L_0 and G_0. Followed by, these two features are further propagated to densely-connected invertible feature fusion neural module

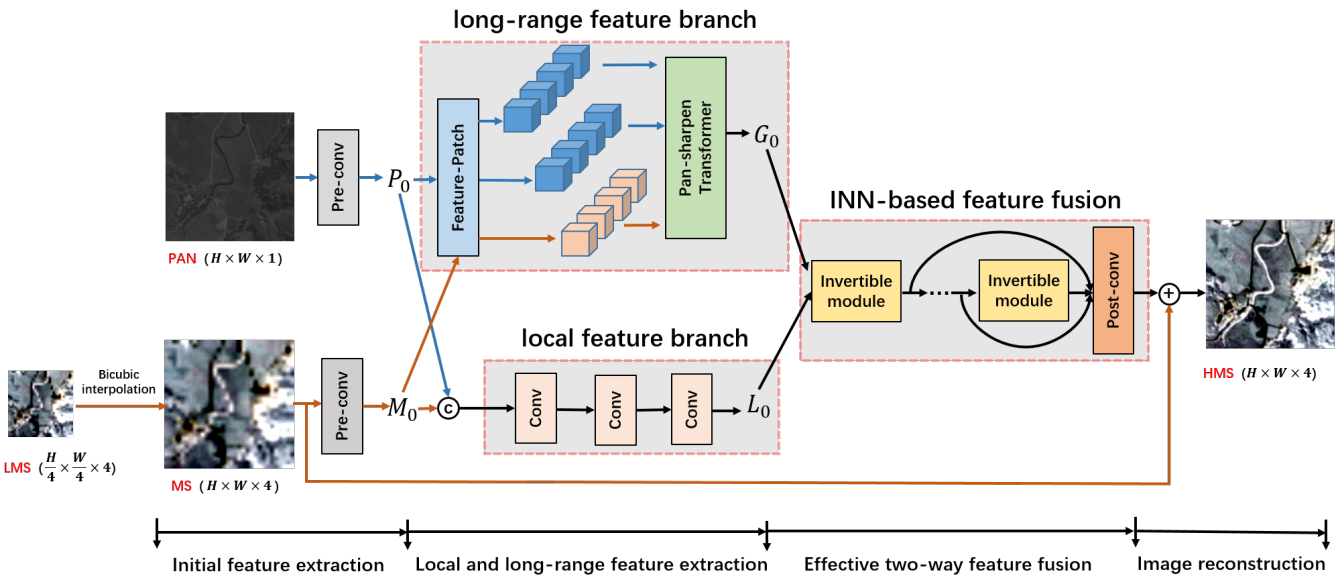

Figure 2: The overall structure of our proposed method. The MS and PAN image pair is firstly projected into modality-aware features by pre-convolution layer. Then, above features are fed into a customized Transformer-based long-range feature modeling and CNN-based local feature extraction module. Next, the obtained long and local-term features are passed through a newly-designed INN-based module for effective feature fusion. Finally, the fused feature combined with skip-connection MS image is exploited for reconstruction.

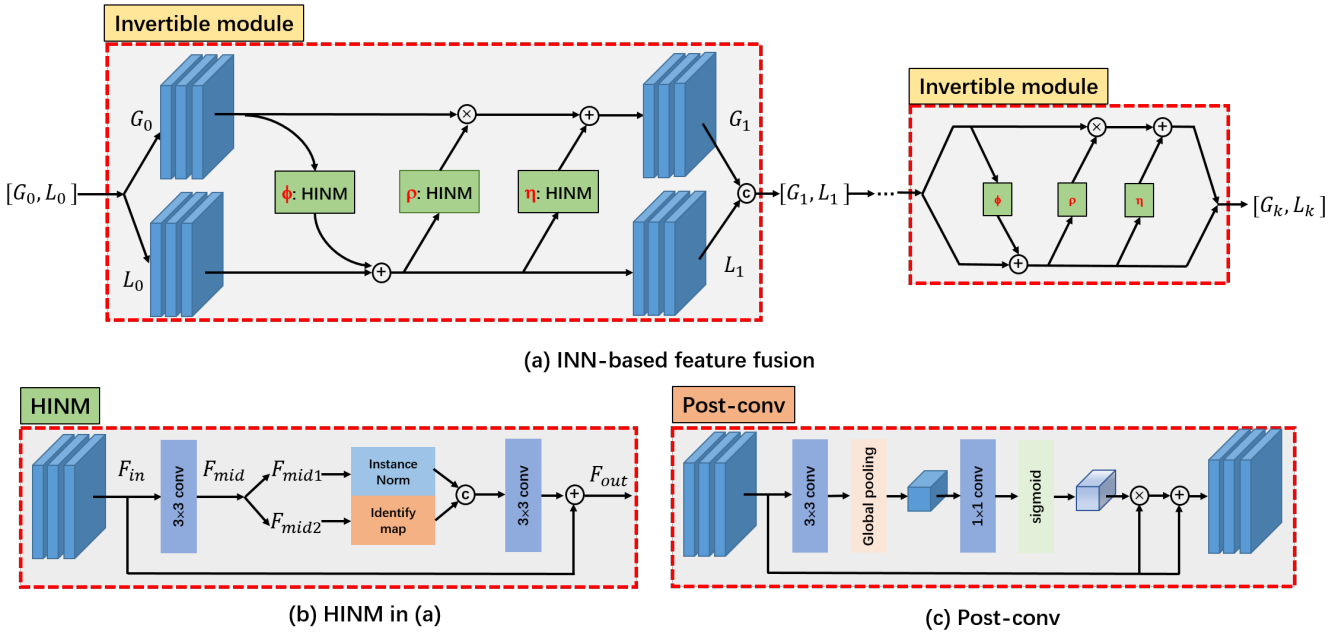

Figure 3: Architecture of the proposed densely-connected invertible feature fusion module. The sub-figure (a) and (c) detail the invertible unit and Post-conv of Figure 2 respectively, while the sub-figure (b) deepens into the HINM of sub-figure (a).

to achieve effective fusion. Specifically, these two kinds of features interact with each other to enhance their representation. Then the enhanced representation is transformed to the same size and channel of the upsampling MS images. Finally, we construct the HR-MS images by adding the upsampling MS images to the transformed representation with skip-connection. The pan-sharpening process can be described as:

$$H = (M) \uparrow_s + f([P, (M) \uparrow_s]). \quad (1)$$

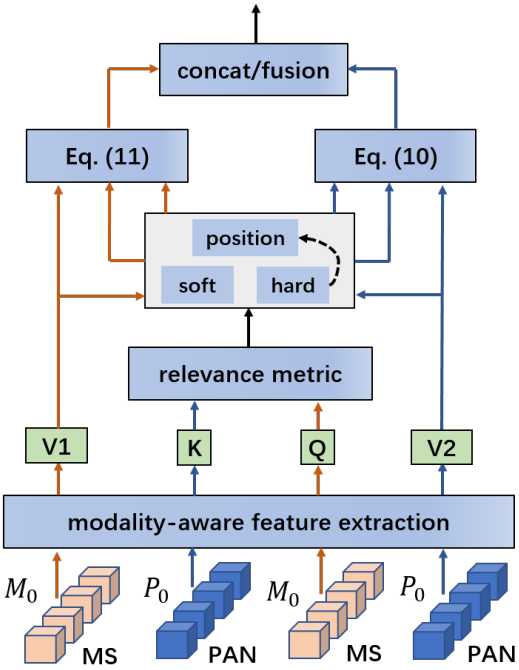

Figure 4: The structure of pan-sharpening transformer.

Note that the direct output of our network $f(.)$ is the residual high-frequency details, which is a common technique used in existing methods to ease learning.

Local and long-range feature extraction.

As shown in Figure 2, our designed feature extraction module consists of two branches, the local-range feature branch by convolution layer and the long-range feature branch by transformer architecture. To preserve the initial features of the MS image and PAN image, the up-sampling MS images $\hat{M} \in R^{C\times H\times W}$ and PAN image $P \in R^{1\times H\times W}$ are firstly fed into two independent 3×3 convolution layers to obtain shallow features, $M_0 \in R^{8\times H\times W}$ and $P_0 \in R^{8\times H\times W}$ respectively. Then, concatenating M_0 and P_0 by channel dimension is passed into the above two branches.

Specifically, the local feature branch is implemented by a 3×3 convolution layer, and receives the full-resolution feature maps M_0 and P_0 to extract the local-range feature L_0. In the long-range feature branch, a newly-designed transformer is used for generating the long-range dependency. As recognized, the standard transformer is designed to capture the long-range self-similarity dependency among all the tokens of single image. Owing to the nature of pan-sharpening that needs to integrate the complementary information between two kinds of images, the MS image and PAN image, directly applying standard Transformer architecture for pan-sharpening task is not suitable. The structure of our transformer is shown in Figure 4. The transformer takes the divided MS and PAN feature patches $M^1, \ldots, M^n$ and $P^1, \ldots, P^n$ with 16×16 pixel size as input from shallow full-resolution features M_0 and P_0.

Firstly, we use several convolution layers to project the MS and PAN feature patches $M^1, \ldots, M^n$ and $P^1, \ldots, P^n$ with 16×16 pixel size to the texture features, Q (query), K (key), and V (value) of three basic elements inside a transformer. Different from standard transformer, we expand the V (value) with two component, V1 and V2 as

$$Q = Conv([M^1, \ldots, M^n]), \tag{2}$$
$$K = Conv([P^1, \ldots, P^n]), \tag{3}$$
$$V1 = Conv([M^1, \ldots, M^n]), \tag{4}$$
$$V2 = Conv([P^1, \ldots, P^n]) \tag{5}$$

where $Conv$ and [.] represent the convolution operation and concatenation by channel dimension, respectively. Then, K and Q will be used in our **relevance metric module** to estimating the similarity. We unfold both K and Q into patches, denoted as q_i ($i \in [1, H \times W]$) and k_j ($j \in [1, H \times W]$). Then for each patch q_i in Q and k_j in K, we calculate the relevance $r_{i,j}$ between these two patches by normalized inner product as $r_{i,j} = (\frac{q_i}{||q_i||}, \frac{k_j}{||k_j||})$. The whole relevance matrix is remarked,

$$R = Q^T K. \tag{6}$$

Then, we further use the relevance matrix R to generate the hard-attention and soft-attention map. Different from traditional attention mechanism takeing a weighted sum of V for each query q_i, we propose hard-attention and soft-attention module to transfer the image texture features V to the HR-MS image. More specifically, we first calculate a hard-attention map H in which the $i-th$ element h_i ($i \in [1, H \times W]$) is calculated from the relevance $r_{i,j}$: $h_j = \underset{j}{argmax}(r_{i,j})$. Then we apply an index selection operation to the unfolded patches of $V1$ and $V2$ using the **hard-attention** map as the index:

$$t_i^1 = v_{h_i}^1, \tag{7}$$
$$t_i^2 = v_{h_i}^2 \tag{8}$$

where t_i^1 and t_i^2 denote the value selected from the h_i-th position of $V1$ and $V2$. As a result, we obtain a HR feature representation $T1$ and $T2$ for the PAN feature and MS feature by **position index attention**. Furthermore, we calculate the **soft-attention** map as,

$$S = softmax(R), \tag{9}$$

where $softmax$ is the softmax function in mathematical. Finally, we obtain the enhanced long-range features G_0 by integrating the soft-attention and hard-attention map with the PAN features

$$G_0^1 = P_0 + Conv([P_0, T1]) \odot S, \tag{10}$$
$$G_0^2 = M_0 + Conv([M_0, T2]) \odot S, \tag{11}$$
$$G_0 = Conv([G_0^1, G_0^2]) \tag{12}$$

where P_0, M_0, [.] and $Conv$ represent the PAN features and MS feature from pre-convolution, concatenation operation by channel dimension and convolution layer respectively.

Invertible neural module for feature fusion.

Different from pure convolution layer, invertible networks are information-lossless during the transformation (Liu et al. 2021; Zhang et al. 2021; Xing, Qian, and Chen 2021; Lu et al. 2021; Paschalidou et al. 2021). For the invertible model, the input needs to be divided into two parts. In our work, the input of our invertible module naturally consists of two parts, local and long-range features L_0 and G_0, which exactly match the splitting of input. To take advantage of invertible networks for preserving the extracted features, we design a densely-connected invertible feature fusion neural module with the composition of a stack of invertible basic units. As shown in Figure 3 (a), each basic unit we follow in this work is the affine coupling layer.

To increase the representational capacity of the network, two kinds of schemes are proposed, 1) immediate sequential features of each invertible unit are propagated to the final unit by skip-connection and then concatenated to enhance its representation, 2) effective transformation operation between two parts is designed. To be specific, we use an additive transformation for the long-range branch, and employ an enhanced affine transformation for the local-range branch. Take the first affine coupling layer for example, given local and long-range features L_0 and G_0, the output will be calculated as

$$L_1 = L_0 + \phi(G_0), \tag{13}$$

$$G_1 = G_0 \odot \exp(\rho(L_1)) + \eta(L_1) \tag{14}$$

where $\exp(.)$ is Exponential function in mathematical, and $\rho(.)$ and $\eta(.)$ represent the scale and translation functions from the channels of local feature L_0 to the channels of long-range feature G_0, respectively. $\phi(.)$ performs the inverse function as $\rho(.)$ and $\eta(.)$. $\odot$ is the Hadamard product. Note that the scale and translation functions are not necessarily invertible, and thus we realize them by neural networks. By doing so, the other $k-1$ invertible blocks receive the output of the previous and generate the results. All the outputs $L_0/G_0, \ldots, L_k/G_k$ of each invertible unit are concatenated to generate the high-frequency details by using the residual channel attention block and then added with the input low-spatial MS image to obtain HR-MS image by skip-connection

$$H = (M) \uparrow_s + RCAB([L_0, G_0, \ldots, L_k, G_k]). \tag{15}$$

where $RCAB$ (Zhang et al. 2018) and $[.]$ represent the residual channel attention and concatenation by the channel dimension. The k is the number of our stacked invertible neural units and set as 3 to reduce the computational cost.

In addition, to enhance the interaction with two-part features, we implement the transformation operation $\rho(.)$, $\eta(.)$ and $\phi(.)$ with two cascaded Half Instance Normalization blocks (HIN) (Chen et al. 2021b). As shown in Figure 3 (b), HIN block firstly employs 3×3 convolution to project input features $F_{in} \in R^{C_{in} \times H \times W}$ to intermediate features $F_{mid} \in R^{16 \times H \times W}$. Then, the features F_{mid} are divided into two parts ($F_{mid_1}/F_{mid_2} \in R^{8 \times H \times W}$). The first part F_{mid_1} is normalized by Instance Normalization (IN) and then concatenates with F_{mid_2} in channel dimension. HIN blocks use Instance Normalization (IN) on the half of the channels and keep context information by the other half of the channels. After the concatenation operation, the obtained features F_{res} are passed through one 3×3 convolution layer and two leaky ReLU layers. Finally, HIN blocks output the enhanced feature F_{out} by adding F_{res} with shortcut features (obtained after 1×1 convolution) as

$$F_{mid} = Conv_{3*3}(F_{in}), \tag{16}$$

$$F_{mid_1}, F_{mid_2} = split(F_{mid}), \tag{17}$$

$$F_{res} = concat(IN(F_{mid_1}), F_{mid_2}), \tag{18}$$

$$F_{out} = F_{res} + F_{in}. \tag{19}$$

where $Conv_{3*3}$ represents the 3×3-kernel convolution operator. The $split(.)$ and $concate$ is the splitting and concatenation function in channel dimension. IN is the Instance Normalization.

Network loss function

We adopt the the mean absolute error ($L1$ loss) to optimize our proposed method

$$\mathcal{L} = \sum_{i=1}^{K} \left\| \mathbf{H}_i - \mathbf{H}_{gt,i} \right\|_1, \tag{20}$$

where K is the number of training data, H_i and $H_{gt,i}$ denote the output high-resolution MS image and ground truth, respectively.

Experiments

Baseline methods

To verify the effectiveness of the proposed method, a series of experiments are carried out between our proposed method and ten state-of-the-art pan-sharpening algorithms. To be specific, five representative deep learning based methods are selected for comparison, namely, PNN (Masi et al. 2016), PANNET (Yang et al. 2017), MSDCNN (Yuan et al. 2018), SRPPNN (Cai and Huang 2021), and GPPNN (Xu et al. 2021b). Our method is also compared with five classic methods, including, SFIM (Liu. 2000), Brovey (Gillespie, Kahle, and Walker 1987), GS (Laben and Brower 2000), IHS (Haydn et al. 1982) and GFPCA (Liao et al. 2017).

Implementation details

We implement all our networks in PyTorch framework on the PC with a single NVIDIA GeForce GTX 2080Ti GPU. In the training phase, they are optimized by Adam optimizer over 1000 epochs with a learning rate of 8×10^{-4} and a batch size of 4. When reaching 200 epochs, the learning rate is decayed by multiplying 0.5. The paired training samples are unavailable in practice. When we construct the training set, the Wald protocol is employed to generate the paired samples. For example, given the MS image $H \in R^{M \times N \times C}$ and the PAN image $P \in R^{rM \times rN \times b}$, both of them are downsampled with ratio r, and the downsampled versions are denoted by $L \in R^{M/r \times N/r \times C}$ and $p \in R^{M \times N \times b}$. In the training set, L and p are regarded as the inputs, while H is the ground truth.

Method	Num of Params	worldview II				GaoFen2				worldview III			
		PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	EGAS↓	PSNR↑	SSIM↑	SAM↓	EGAS↓
SFIM	-	34.1297	0.8975	0.0439	2.3449	36.906	0.8882	0.0318	1.7398	21.8212	0.5457	0.1208	8.973
Brovey	-	35.8646	0.9216	0.0403	1.8238	37.7974	0.9026	0.0218	1.372	22.506	0.5466	0.1159	8.2331
GS	-	35.6376	0.9176	0.0423	1.8774	37.226	0.9034	0.0309	1.6736	22.5608	0.547	0.1217	8.2433
IHS	-	35.2962	0.9027	0.0461	2.0278	38.1754	0.91	0.0243	1.5336	22.5579	0.5354	0.1266	8.3616
GFPCA	-	34.5581	0.9038	0.0488	2.1411	37.9443	0.9204	0.0314	1.5604	22.3344	0.4826	0.1294	8.3964
PNN	0.689	40.7550	0.9624	0.0259	1.0646	43.1208	0.9704	0.0172	0.8528	29.9418	0.9121	0.0824	3.3206
PANNET	0.688	40.8176	0.9626	0.0257	1.0557	43.0659	0.9685	0.0178	0.8577	29.684	0.9072	0.0851	3.4263
MSDCNN	2.390	41.3355	0.9664	0.0242	0.994	45.6874	0.9827	0.0135	0.6389	30.3038	0.9184	0.0782	3.1884
SRPPNN	17.114	41.4538	0.9679	0.0233	0.9899	47.1998	0.9877	0.0106	0.5586	30.4346	0.9202	0.077	3.1553
GPPNN	1.198	41.1622	0.9684	0.0244	1.0315	44.2145	0.9815	0.0137	0.7361	30.1785	0.9175	0.0776	3.2593
Ours	0.706	**41.6903**	**0.9704**	**0.0227**	**0.9514**	**47.3528**	**0.9893**	**0.0102**	**0.5479**	**30.5365**	**0.9225**	**0.0747**	**3.0997**

Table 1: The four metrics on test datasets. The best and the second best values are highlighted by the red bold and underline, respectively. The up or down arrow indicates higher or lower metric corresponds to better images.

Dataset and evaluation metrics

Remote sensing images acquired by three satellites are used in our experiments, including worldview II, worldview III , and GaoFen2, the basic information of which are listed in supplementary materials. For each satellite, we have hundreds of image pairs, and they are divided into two parts for training and test. In the training set, the MS images are cropped into patches with the size of 128×128 , and the corresponding PAN patches are with the size of 32×32. For numerical stability, each patch is normalized by dividing the maximum value to make the pixels range from 0 to 1.

Several widely used image quality assessment (IQA) metrics are employed to evaluate the performance, including the relative dimensionless global error in synthesis (ERGAS) (Alparone et al. 2007), the peak signal-to-noise ratio (PSNR), the spectral angle mapper (SAM) (J. R. H. Yuhas and Boardman 1992).

Comparison with SOTA methods

The evaluation metrics on three datasets are reported in Table 1, where the values highlighted by red color represent the best results. It is clearly found that our method surpasses other comparative algorithms in all evaluation metrics on three satellites. In addition, We also show the comparison of the visual results to testify the effectiveness of our method in Figure 5. Images in the last row are the MSE residues between the pan-sharpened results and the ground truth. To be specific, other comparison methods suffer from severe spatial and spectral distortion. However, our method has the most minor spatial and spectral distortions. Specifically, from the amplified local regions, we observe that our proposed method has finer-grained textures and coarser-grained structures compared with other methods. As for the MSE residues, we can figure out that our proposed method is the closest to the ground truth than other comparison methods.

Flops and Parameter Numbers

In this section, we investigate the complexity of the proposed method, including the flops and the number of parameters (in 10 M). Comparisons on parameter numbers and model performance (representation by PSNR) are shown in Table 3 and in Figure 1. It can be seen that our network can achieve a good trade-off between calculation and performance compared to other deep learning-based methods. We use the tensor with $1 \times 4 \times 32 \times 32$ and $1 \times 1 \times 128 \times 128$ to represent the MS and PAN roles for evaluation.

Ablation experiments

Since the transformer module and densely-connected invertible neural network fusion module are the core of our method, to investigate their necessity and effectiveness, a series of ablation experiments are carried out. There are 3 different configurations for the corresponding network variants of our proposed method and the results of ablation experiments are shown in Table 2.

Transformer module. The Transformer module is responsible for capturing long-range dependency, which is critical for pan-sharpening performance. In the first experiment, we delete the Transformer module to verify its necessity while expanding the feature channels of local-range branch for fair comparison. Table 2 shows that deleting Transformer module will degrade all metrics dramatically. Therefore, Transformer module plays a significant role in our network.

Invertible neural network fusion module. In the second experiment, to verify the effectiveness of densely-connected invertible neural network fusion module, we replace it with its transformation units . In other words, the extracted long and local-range features are concatenated and then fed into pure densely-connected architecture. For fair comparison, we keep the above two comparisons with the same number of parameters. The results in Table 2 demonstrate that removing the invertible fusion module will weaken our network's performance. Therefore, invertible neural network

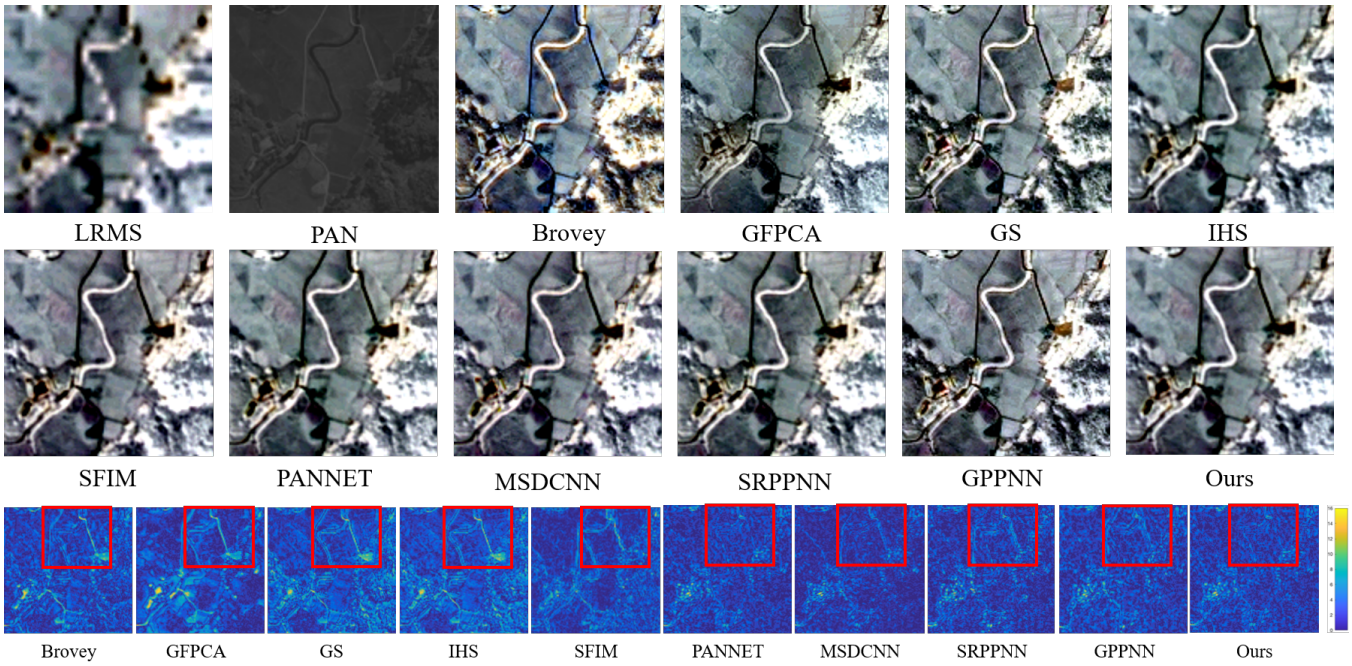

Figure 5: Qualitative comparison of our method with nine counterparts on a typical satellite image pair from the GaoFen-2 dataset. Images in the last row visualize the MSE residues between the pan-sharpened results and the ground truth.

Configurations	Transformer	Invertible Fusion	worldview II				GaoFen2				worldview III			
			PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	EGAS↓	PSNR↑	SSIM↑	SAM↓	EGAS↓
(I)	×	✓	41.1932	0.9684	0.0238	1.0059	46.7804	0.9879	0.011	0.5835	30.2943	0.9193	0.0784	3.1882
(II)	✓	×	41.2232	0.9683	0.0238	1.0049	46.9368	0.988	0.0106	0.5728	30.2588	0.9181	0.0785	3.2053
Ours	✓	✓	**41.6903**	**0.9704**	**0.0227**	**0.9514**	**47.3528**	**0.9893**	**0.0102**	**0.5479**	**30.5365**	**0.9225**	**0.0747**	**3.0997**

Table 2: The results of ablation experiments over three datasets. The best values are highlighted by the red bold. The up or down arrow indicates higher or lower metric corresponds to better images.

	PNN	PANNET	MSDCNN	SRPPNN	GPPNN	Ours
params	0.689	0.688	2.390	17.114	1.198	0.706
flops	1.1289	1.1275	3.9158	21.1059	1.3967	1.3907

Table 3: Comparisons on flops and parameter numbers.

fusion module is critical in our method.

Our complete network. In the last row of Table 2, we can clearly find that compared with above two variants, taking worldview-II dataset for example, adding the Transformer module achieves an improvement of 0.5 dB and 0.01 on average PSNR and SSIM, respectively. Similarly, the invertible fusion module improves the baseline by 0.47 dB and 0.01. Other datasets also keep consistent as above in model performance. This is because the two modules are beneficial to capture the long-range dependency spatially and effectively fuse the features for the pan-sharpening task. The best results can be obtained by combining the two modules.

Conclusion and future work

In this paper, we propose a novel and effective pan-sharpening method by integrating long-range dependencies modeling of Transformer architecture and Information-lossless invertible neural network in this paper. To the best of our knowledge, this is the first attempt to introduce transformer and invertible neural network into pan-sharpening field. Extensive experiments over different kinds of satellite datasets demonstrate that our method outperforms state-of-the-art algorithms both visually and quantitatively.

In the future, we will explore the potential of our proposed customized transformer and invertible module into existing pan-sharpening methods.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (61701158) and the USTC Research Funds of the Double First-Class Initiative under Grants YD2100002004.

References

Aiazzi, B.; Alparone, L.; Baronti, S.; Garzelli, A.; and Selva, M. 2003. An MTF-based spectral distortion minimizing model for pan-sharpening of very high resolution multispectral images of urban areas. In *Joint Workshop on Remote Sensing and Data Fusion over Urban Areas*.

Aiazzi, B. S., B.; and Selva, M. 2007. Improving Component Substitution Pansharpening Through Multivariate Regression of MS +Pan Data. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10): 3230–3239.

Alparone, L.; Wald, L.; Chanussot, J.; Thomas, C.; Gamba, P.; and Bruce, L. M. 2007. Comparison of Pansharpening Algorithms: Outcome of the 2006 GRS-S Data Fusion Contest. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10): 3012–3021.

Benzenati, T.; Kallel, A.; and Kessentini, Y. 2021. Two Stages Pan-Sharpening Details Injection Approach Based on Very Deep Residual Networks. *IEEE Transactions on Geoscience and Remote Sensing*, 59(6): 4984–4992.

Cai, J.; and Huang, B. 2021. Super-Resolution-Guided Progressive Pansharpening Based on a Deep Convolutional Neural Network. *IEEE Transactions on Geoscience and Remote Sensing*, 59(6): 5206–5220.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In *European Conference on Computer Vision*.

Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; and Gao, W. 2021a. Pre-Trained Image Processing Transformer. In *IEEE Conference on Computer Vision and Pattern Recognition*, 12299–12310.

Chen, L.; Lu, X.; Zhang, J.; Chu, X.; and Chen, C. 2021b. HINet: Half Instance Normalization Network for Image Restoration. arXiv:2105.06086.

Choi, Y. K., J.; and Kim, Y. 2011. A New Adaptive Component-Substitution-Based Satellite Image Fusion by Using Partial Replacement. *IEEE Transactions on Geoscience and Remote Sensing*, 49(1): p.295–309.

Dinh, L.; Krueger, D.; and Bengio, Y. 2015. NICE: Nonlinear Independent Components Estimation. *International Conference on Learning Representations*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; and Houlsby, N. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.

Garzelli, A.; Nencini, F.; and Capobianco, L. 2008. Optimal MMSE Pan Sharpening of Very High Resolution Multispectral Images. *IEEE Transactions on Geoscience and Remote Sensing*, 46(1): 228–236.

Ghahremani; Morteza; Ghassemian; and Hassan. 2016. A Compressed-Sensing-Based Pan-Sharpening Method for Spectral Distortion Reduction. *IEEE Transactions on Geoscience and Remote Sensing*.

Gillespie, A. R.; Kahle, A. B.; and Walker, R. E. 1987. Color enhancement of highly correlated images. II. Channel ratio and "chromaticity" transformation techniques - ScienceDirect. *Remote Sensing of Environment*, 22(3): 343–365.

H. Wang, H. A. A. Y., Y. Zhu; and Chen, L.-C. 2021. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Haydn, R.; Dalke, G. W.; Henkel, J.; and Bare, J. E. 1982. Application of the IHS color transform to the processing of multisensor data and image enhancement. *National Academy of Sciences of the United States of America*, 79(13): 571–577.

Hu, J.; Hu, P.; Kang, X.; Zhang, H.; and Fan, S. 2021. Pan-Sharpening via Multiscale Dynamic Convolutional Neural Network. *IEEE Transactions on Geoscience and Remote Sensing*, 59(3): 2231–2244.

J. R. H. Yuhas, A. F. G.; and Boardman, J. M. 1992. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm. *Proc. Summaries Annu. JPL Airborne Geosci. Workshop*, 147–149.

Kang, L. S., X.; and Benediktsson, J. A. 2014. Pansharpening With Matting Model. *IEEE Transactions on Geoscience and Remote Sensing*, 52(8): 5088–5099.

Kaplan, N. H.; and Erer, I. 2012. Bilateral pyramid based pansharpening of multispectral satellite images. In *Geoscience and Remote Sensing Symposium*.

Laben, C.; and Brower, B. 2000. Process for Enhancing the Spatial Resolution of Multispectral Imagery Using Pan-Sharpening. *US Patent 6011875A*.

Laurent Dinh, J. S.-D.; and Bengio., S. 2017. Density estimation using real NVP. *ICLR*.

Li, Y.; Zhang, K.; Cao, J.; Timofte, R.; and Gool, L. V. 2021. LocalViT: Bringing Locality to Vision Transformers.

Liao, W.; Xin, H.; Coillie, F. V.; Thoonen, G.; and Philips, W. 2017. Two-stage fusion of thermal hyperspectral and visible RGB image by PCA and guided filter. In *Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*.

Liu., J. G. 2000. Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details. *International Journal of Remote Sensing*, 21(18): 3461–3472.

Liu, Q.; Zhou, H.; Xu, Q.; Liu, X.; and Wang, Y. 2020. PSGAN: A Generative Adversarial Network for Remote Sensing Image Pan-Sharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 1–16.

Liu, Y.; Qin, Z.; Anwar, S.; Ji, P.; Kim, D.; Caldwell, S.; and Gedeon, T. 2021. Invertible Denoising Network: A Light Solution for Real Noise Removal. In *IEEE Conference on Computer Vision and Pattern Recognition*, 13365–13374.

Lu, S.-P.; Wang, R.; Zhong, T.; and Rosin, P. L. 2021. Large-Capacity Image Steganography Based on Invertible Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 10816–10825.

Masi, G.; Cozzolino, D.; Verdoliva, L.; and Scarpa, G. 2016. Pansharpening by convolutional neural networks. *Remote Sensing*, 8(7): 594.

Paschalidou, D.; Katharopoulos, A.; Geiger, A.; and Fidler, S. 2021. Neural Parts: Learning Expressive 3D Shape Abstractions With Invertible Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3204–3215.

Peng, J.; Liu, L.; Wang, J.; Zhang, E.; Zhu, X.; Zhang, Y.; Feng, J.; and Jiao, L. 2021. PSMD-Net: A Novel Pan-Sharpening Method Based on a Multiscale Dense Network. *IEEE Transactions on Geoscience and Remote Sensing*, 59(6): 4957–4971.

Shah, V. P.; Younan, N. H.; and King, R. L. 2008. An Efficient Pan-Sharpening Method via a Combined Adaptive PCA Approach and Contourlets. *IEEE Transactions on Geoscience and Remote Sensing*, 46(5): 1323–1335.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. *arXiv*.

Wang, D.; Bai, Y.; Wu, C.; Li, Y.; Shang, C.; and Shen, Q. 2021a. Convolutional LSTM-Based Hierarchical Feature Fusion for Multispectral Pan-Sharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 1–16.

Wang, J.; Shao, Z.; Huang, X.; Lu, T.; and Zhang, R. 2021b. A Dual-Path Fusion Network for Pan-Sharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 1–14.

Wang, Z.; Cun, X.; Bao, J.; and Liu, J. 2021c. Uformer: A General U-Shaped Transformer for Image Restoration.

Xing, Y.; Qian, Z.; and Chen, Q. 2021. Invertible Image Signal Processing. In *IEEE Conference on Computer Vision and Pattern Recognition*, 6287–6296.

Xu, H.; Ma, J.; Shao, Z.; Zhang, H.; Jiang, J.; and Guo, X. 2021a. SDPNet: A Deep Network for Pan-Sharpening With Enhanced Information Representation. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5): 4120–4134.

Xu, S.; Zhang, J.; Zhao, Z.; Sun, K.; Liu, J.; and Zhang, C. 2021b. Deep Gradient Projection Networks for Pan-sharpening. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1366–1375.

Yang, F.; Yang, H.; Fu, J.; Lu, H.; and Guo, B. 2020. Learning Texture Transformer Network for Image Super-Resolution. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang, J.; Fu, X.; Hu, Y.; Huang, Y.; Ding, X.; and Paisley, J. 2017. PanNet: A deep network architecture for pan-sharpening. In *IEEE International Conference on Computer Vision*, 5449–5457.

Yokoya, N.; Member, S.; IEEE; Yairi, T.; and Iwasaki, A. 2012. Coupled Nonnegative Matrix Factorization Unmixing for Hyperspectral and Multispectral Data Fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 50(2): 528–537.

Yuan, K.; Guo, S.; Liu, Z.; Zhou, A.; Yu, F.; and Wu, W. 2021. Incorporating Convolution Designs into Visual Transformers.

Yuan, Q.; Wei, Y.; Meng, X.; Shen, H.; and Zhang, L. 2018. A Multiscale and Multidepth Convolutional Neural Network for Remote Sensing Imagery Pan-Sharpening. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(3): 978–989.

Zhang, S.; Zhang, C.; Kang, N.; and Li, Z. 2021. iVPF: Numerical Invertible Volume Preserving Flow for Efficient Lossless Compression. In *IEEE Conference on Computer Vision and Pattern Recognition*, 620–629.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018. Image super-resolution using very deep residual channel attention networks. In *European Conference on Computer Vision*, 286–301.

Zhu, X.; Su, W.; Lu, L.; Li, B.; and Dai, J. 2020. Deformable DETR: Deformable Transformers for End-to-End Object Detection.