Pansformers: Transformer-Based Self-Attention Network for Pansharpening

Nithin G R, Nitish Kumar M, Rajanikanth Kakani, Venkateswaran N, Ankur Garg, and Ujjwal Kumar Gupta

***Abstract*—Pansharpening is the task of creating a High-Resolution Multi-Spectral Image (HRMS) by extracting and infusing pixel details from the High-Resolution Panchromatic Image into the Low-Resolution Multi-Spectral (LRMS). With the boom in the amount of satellite image data, researchers have replaced traditional approaches with deep learning models. However, existing deep learning models are not built to capture intricate pixel-level relationships. Motivated by the recent success of self-attention mechanisms in computer vision tasks, we propose *Pansformers*, a transformer-based self-attention architecture, that computes band-wise attention. A further improvement is proposed in the attention network by introducing a *Multi-Patch Attention* mechanism, which operates on non-overlapping, local patches of the image. Our model is successful in infusing relevant local details from the Panchromatic image while preserving the spectral integrity of the MS image. We show that our Pansformer model significantly improves the performance metrics and the output image quality on imagery from two satellite distributions IKONOS and LANDSAT-8.**

***Index Terms*—Pansharpening, Multispectral, Panchromatic, Pransformers, Multi-Patch Attention**

I. Introduction

Spaceborn satellites such as IKONOS and LANDSAT-8 provide two different complementary types of images: a high-spatial and low-spectral resolution Panchromatic Image, and a high-spectral and low-spatial resolution Multispectral (MS). Due to the constraints on signal transmission broadband and imaging sensor storage, it is very difficult to acquire a high spatial resolution MS image directly. Pansharpening aims at injecting the details from the Panchromatic image into the Multispectral image, to generate a high spatial/spectral resolution Multispectral (MS) image. The fusing process has become a key preprocessing step in many applications such as feature detection, land-cover classification, and also in making high-resolution maps. This makes Pansharpening an central task in the field of Satellite Remote Sensing.

In the past, traditional have been proposed for Pansharpening which include the Intensity–Hue–Saturation (IHS)[12], Principal Component Analysis (PCA)[1] and the Brovey Transform method. However, the problem with the traditional methods were the spectral distortions caused in the bands of the pansharpened image. To address the problem of spectral distortion, many deep learning methods has been recently adopted by researchers. The authors in [16] proposed a deep network called PanNet, which preserves spectral and spatial structure using the concept of 'spectra mapping'. Another deep neural network was also proposed in [5].

A fair share of CNN-based models has also been proposed for the task of Pansharpening, which includes residual networks. In [9], the authors introduced PCNN, a simple three-layer convolutional network to perform Pansharpening. The authors of [2] proposed a deep convolutional network that learns an end-to-end mapping between the low and high-resolution images. In [3], SCRNN was proposed, which had a lightweight structure with only little extra pre/post-processing required. Yuan et al. [18] included the multiscale feature extraction and residual learning into the basic convolutional neural network (CNN) architecture. Similarly, the authors of [15] proposed DRPNN, which is robust architecture and performs high-quality fusion. Other CNN-based methods were also proposed in [7], [13], [6], [17].

Recently, Attention mechanisms have been adopted to improve the performance on the Pansharpening task. The authors in [8] proposed a multi-scale channel attention residual network (MSCARN) to comprehensively extract the coarse structures and high-frequency details through a squeeze-and-excitation block. In [11], Qu et al. proposed a self-attention based method to perform Pansharpening in an unsupervised setting. Transformer networks [14], which were originally proposed for sequence-to-sequence tasks have been successfully adapted to Computer Vision applications like Image Superresolution [10] and Recognition [4]. However, straightforward adaptations of these networks do not give the required performance in Pansharpening because of their highly task-specific nature.

In this letter, inspired by the Transformer networks, we create an architecture called *Pansformers*, which uses PCNN combined with channel-wise Self-Attention to improve the performance of the task. Since Pansharpening is a region-based fusion task, we create a *Multi-Patch Attention* mechanism, which divides the input image into smaller, non-overlapping patches and computes attention on those individual patches to capture the local level details important for fusion. To the best of our knowledge, ours is the first work to draw motivation from Transformer-based architectures for Pansharpening in the field of Satellite Remote Sensing.

II. Problem Definition and Preprocessing

In this section, we formally define the problem and introduce the notation for the input and the output images. Consider three images - a single band Panchromatic image

Nithin G R, Nitish Kumar M, Rajanikanth Kakani and Venkateswaran N are with the Electronics and Communication Department (ECE) at Sri Sivasubramaniya Nadar College of Engineering (Autonomous),Chennai, India.

Ankur Garg and Ujjwal Kumar Gupta are with the Space Applications Centre at Indian Space Research Organization (ISRO).

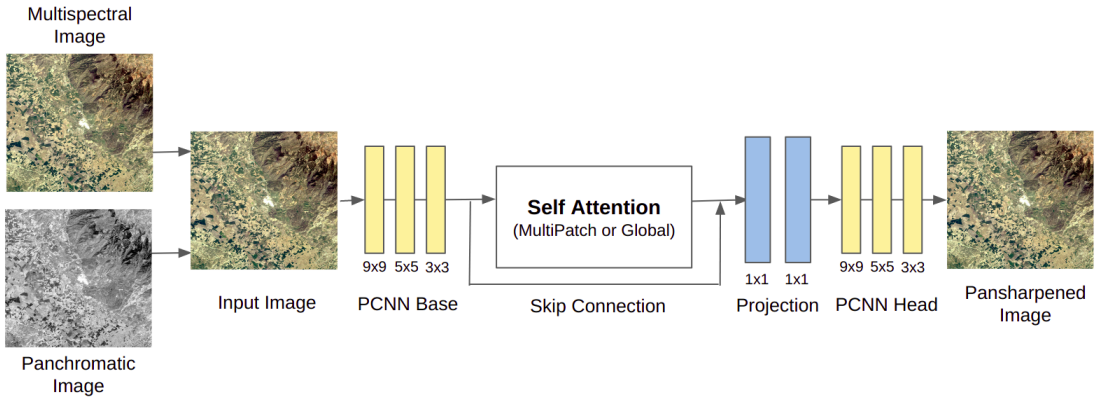

Fig. 1. The proposed Pansformers Architecture. Here, PCNN Base, Self-Attention Layer, Projection layers, and PCNN Head are the 4 components. The PCNN Base produces a feature map of the input image, which is to the self-attention layer to infuse details into different bands. The Projection layer and the PCNN Head layer process the Self-Attention layer outputs to produce the Pansharpened Image. The Self-Attention Layer shown here can either be Global Attention or Multi-Patch Attention. We also include a skip-connection from the PCNN Base to the end of the Self-Attention layer.

$H_{Pan} \in \mathbb{R}^{X \times Y}$, a low resolution n-band Multispectral image $H_{LRMS} \in \mathbb{R}^{N \times X \times Y}$, as same size as H_{Pan}, and a ground truth Multispectral image H_{HRMS}. Here, H_{Pan} is concatenated to H_{LRMS} to form the input to the learning model $H_{Input} \in \mathbb{R}^{N+1 \times X \times Y}$. Given H_{Input} and the groundtruth image H_{HRMS}, the task of the learning model is to produce a multispectral image $H_{Pred} \in \mathbb{R}^{N \times X \times Y}$ which closely matches the ground truth H_{HRMS}. The H_{Pred} contains one less band than H_{Input} image, but matches the order of the bands as H_{LRMS}. In other words, the objective of the deep learning model is to minimize the loss function L(H_{Pred}, H_{HRMS}) by learning to produce a higher resolution image with both spectral and spatial characteristics of the images preserved.

However, satellite data distribution is only provided with one MS image and a corresponding Panchromatic image and no groundtruth image is readily available. So, we use Wald's protocol to generate the input-groundtruth pairs from the original low-resolution Multispectral (MS) image MS_{Org} and the high-resolution Panchromatic image P. According to Wald's protocol, MS_{Org} is set to be the groundtruth image H_{HRMS}. First, we use a 3×3 Gaussian Blur on MS_{Org} to create MS_{Blur} Then, P and MS_{Blur} are downsampled by a factor K, determined by the inherent resolution of the satellite, to create H_{Pan} and MS_{Down}. Finally, MS_{Down} alone is interpolated by the same factor K to produce the input MS image H_{LRMS}. Due to the downsampling and interpolating, the H_{LRMS} will be of a lower resolution compared to H_{HRMS}. Hence, H_{HRMS} will serve as the ground-truth image while H_{LRMS} and H_{Pan} comprise the input image H_{Input} to the model. The model learns a mapping between H_{Input} and H_{HRMS} in a manner that is independent of the resolution of the images ensuring that the model works on higher-resolution satellite images. We finally use a standard tiling procedure used in [9] to divide the large satellite image into numerous 64×64 tiles for training.

III. Pansformers

We develop a Transformer-based architecture called *Pansformers* to tackle the problem of Pansharpening. The main highlight of our architecture lies in our channel-wise self-attention network, which operates across different bands to extract information on the relative importance of each band in fusing details from the Panchromatic image. We experiment with two different types of self-attention, namely *Global Attention* and *Multi-Patch Attention*. Our architecture is shown in Fig 1. The basic structure of our Pansformers architecture involves the self-attention network sandwiched between two *PCNN* blocks which act as a 'base' and 'head' to extract enriched information. The structure of our architecture is explained in the sections below.

A. PCNN

The Pansharpening Convolutional Neural Network (Masi et al. [9]) is a baseline CNN network proposed for Pansharpening. The PCNN is a shallow, fully convolutional network that consists of three layers having filter sizes 9×9, 5×5 and 3×3 respectively with 64, 32, and 4 filters. We hypothesized that combining multiple PCNN block can aid the performance of the self-attention layers, therefore, boosting the performance significantly. Hence, in our architecture, we use two PCNN network blocks, named $PCNN_{Base}$ and $PCNN_{Head}$, as a basic processing block at the input and output sides to facilitate the learning of relevant features in the self-attention layer. The $PCNN_{Base}$ layer operates on the input image H_{Input} and produces H_{Base}. In $PCNN_{Base}$ alone, we alter the last convolutional layer to contain 5 filters instead of 4, to preserve the number of input channels.

$$H_{Base} = PCNN_{Base}(H_{Input}), \tag{1}$$

B. Image Self-Attention

Self-attention, also known as intra-attention, is an attention mechanism that computes a representation of the individual elements of the inputs by learning to 'attend' to the most relevant details present in the elements. In our self-attention module, the multiple bands of the input image are considered as the elements, and the self-attention process computes the attention scores across the bands through the equation:

$$SelfAtt = Softmax(Q \cdot K^T) \cdot V, \tag{2}$$

where Q, K and V correspond to *Query*, *Key* and *Value* are independent projections of the image learnt through three single 1×1 convolutional layers denoted by C^Q, C^K, and C^V.

The self-attention layer operates on H_{Base} to form a condensed, information-rich attention map H_{Att}. The attention map not only helps in extracting intricate, local-level information from H_{Pan}, but also determines the relative importance of each band in the optimal infusion of the details into H_{LRMS}. In our model, we propose two different forms for self-attention, namely Global Attention and Multi-Patch Attention, which are described and compared below.

Global Attention: In the Global Attention layer (Ga), the attention is computed on the entire image as a whole. In other words, global-level features are extracted across the bands of the entire image. In this process, only one Query Q_{Ga}, Key K_{Ga} and Value V_{Ga} per input image will be produced using just three 1×1 convolution layers C^Q_{Ga}, C^K_{Ga} and C^K_{Ga}. The projections are of the same spatial size as H_{Base}.

$$GlobalAtt = Softmax(Q_{Ga} \cdot K^T_{Ga}) \cdot V_{Ga} \tag{3}$$

$$H_{Att} = GlobalAtt(H_{Base}) \tag{4}$$

However, we hypothesize that in global attention, the finer-grained features are not captured since the attention scores are calculated for the entire image and not region-wise. Since the region of operation is large, the single 1×1 convolution layers extract only the top-level, coarse details present in the image. Also, computation of self-attention across the entire image slows down the training speed significantly, since the time complexity of computation of self-attention scores scales quadratically with an increase in input size ([14]).

Since Pansharpening is a region-based fusion task, the intricate local-level pixel details are extremely important for increasing the resolution. Motivated by this, we developed a Multi-Patch Attention (Mpa) module, which calculates attention scores separately on multiple, smaller patches and is explained in the section below:

Multi-Patch Attention: In Multi-Patch Attention, the attention is calculated separately on several smaller, non-overlapping patches of the input image. As the first step, the input image of the previous layer is divided into the m number of non-overlapping patches of spatial size $X/m \times Y/m$, where $X \times Y$ is the spatial size of the input image. Then, self-attention is computed on each of the patches separately. Formally defining, given H_{Base} of size 64×64, we divide it into 16 patches, each of size 16×16, denoted by the set of $H_{BasePatch} = \{H^{11}_{Base}, H^{12}_{Base}, ...H^{21}_{Base}, H^{22}_{Base}....H^{mm}_{Base}\}$ where we number the patches according to conventional matrix notation. The self-attention function is applied to all the patches separately in which each patch will have a set of three 1×1 convolution layers denoted by C^Q_{ij}, C^K_{ij}, and C^V_{ij} which give rise to Q_{ij}, K_{ij} and V_{ij}, where i and j denote the position of the patch in the original image. The query, key and value projections of each patch will be the same size as the patch.

$$MultiPatchAtt = Concat(SelfAtt(H^{ij}_{Base})) \tag{5}$$

$$SelfAtt(H^{ij}_{Base}) = Softmax(Q_{ij} \cdot K^T_{ij}) \cdot V_{ij} \tag{6}$$

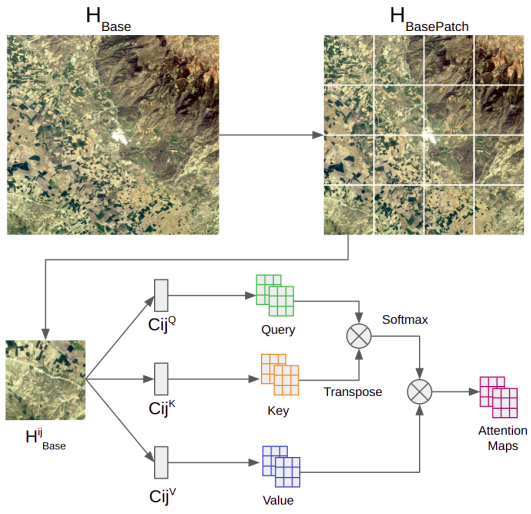

Fig. 2. The Multi-Patch Attention Mechanism. H_{Base} is divided into 16 patches ($H_{BasePatch}$). Then, the self-attention is computed on each of the individual patches H^{ij}_{Base} (Only single patch shown for clarity). The self-attention maps are then concatenated to form H_{Att} (Not Shown).

where i and j both range from 1 to m.

$$H_{Att} = MultiPatchAtt(H_{BasePatch}) \tag{7}$$

After computing self-attention, the attention maps of each patch are concatenated (or re-stitched) back together to form the output $64x \times 64$ image.

In multi-patch attention, the region of operation is significantly reduced. The local, more intricate pixel-level details that are important for obtaining the highest resolution possible, are captured. Our multi-patch attention module does not increase the number of trainable parameters in the architecture. The increase in the number of 1x1 convolution layers for patch-wise computation is canceled out by the reduction in the size of the operation. Another interesting feature of the multi-patch attention mechanism is the attention computation process becomes significantly faster due to the reduction in the input size, leading to faster training times. Even though the attention computation is sequential and not parallel, the speed-up achieved due to the smaller size outweighs the increase in the number of self-attention computations.

C. Skip-Connection

Due to the local-region computations in the attention layers, we predicted that there may occur a distortion in the long-range spatial coherence of the image. In our architecture, we also include a skip-connection, from the end of the $PCNN_{Base}$ to H_{Att}.

$$H_{Skip} = H_{Att} + H_{Base} \tag{8}$$

This is to ensure that both the spatial and the spectral characteristics of the image are preserved intact. The skip-connection provides the model with an alternative route to learn the required high-resolution output image by smoothing out any distortions caused in the attention maps. This skip connection also ensures unhindered gradient flow to the initial

TABLE I
IMAGE CHARACTERSITICS OF IKONOS AND LANDSAT-8

Characteristics	IKONOS	LANDSAT-8
Panchromatic Resolution	0.82 m	15 m
Multispectral Resolution	3.28 m	30 m
Bit Size	11 bit	16 bit

$PCNN_{Base}$ layer thereby contributing to the stability of the model.

D. Projection Layers and $PCNN_{Head}$

The output of the attention layers when combined with the skip connection, is raw and unprocessed, and require further processing. We use same-size convolution *Projection* layers to process the attention maps.

$$H_{Proj} = P2(P1(H_{Skip})) \tag{9}$$

where P1 and P2 are the two projection layers containing 20 and 5 filters respectively. The projection layers increase the number of channels from 5 to 20, then decrease it back to 5 channels again. We conjectured that this bottleneck structure helps in learning the final high-resolution details in a higher-dimensional space with more channels. This process helps in processing the raw attention and skip-connection outputs to produce a coherent, higher resolution output image. Finally, we use the $PCNN_{Head}$ block and discard the final band in H_{Proj} that corresponds to H_{Pan}.

$$H_{Pred} = PCNN_{Head}(H_{Proj}) \in \mathbb{R}^{N \times X \times Y} \tag{10}$$

It is to be noted that our architecture is directly inspired by Transformer networks([14]). The intuition behind the Multi-Patch Attention layers is to process parts of the input separately, similar to the Multi-Head Attention in Transformer Networks. Also, our projection layers P1 and P2, which learn relevant details in higher dimensional space, were derived from the working of feed-forward layers in Transformers.

IV. EXPERIMENTAL SETTINGS AND METRICS

We used Pytorch to construct our models and utilized an Nvidia Tesla P100 GPU on Google Colab Pro for training. A total of 6 and 4 images from IKONOS and LANDSAT-8 are used to create the training and testing sets. For both the datasets, we used 80% of the images for training and the remaining 20% equally for validation and testing sets. The Mean Squared Error(MSE) Loss with an Adam Optimizer was used to train the model with a learning rate of $10e^{-3}$. Apart from Global Attention and Multi-Patch Attention, we also trained one version containing a combination of both the attention modules in a two-streamed fashion. For visualizing the images, we upsample the original MS image MS_{Org} to the size of the original Panchromatic image P, and perform tiling, then re-stitch the pansharpened tiles to form the final image.

We evaluated our pansharpened image against the ground-truth image, by computing a series of performance metrics. While Peak Signal to Noise Ratio (PSNR) provides the ratio of the maximum possible power to the corrupting noise present, Universal Quality Index (UQI) measures image distortion. While Structural Similarity Index (SSIM) and Spatial Correlation Coefficient (SCC) calculates the structural and spatial similarity between the images, Spectral Angle Mapper (SAM) measures the angle of spectral distortion caused across the bands.

TABLE II
COMPARISON OF PERFORMANCE METRICS ON IKONOS DATA SET

Method/Model	PSNR	UQI	SAM	SCC	SSIM
Brovey	26.582	0.849	0.021	0.894	0.881
IHS	30.04	0.981	0.0268	0.809	0.979
PCNN	36.308	0.984	0.074	0.969	0.958
Global	43.210	0.997	0.043	0.984	0.986
Multi-Patch	**45.205**	**0.998**	**0.032**	**0.988**	**0.992**
Global+Multi-Patch	42.912	0.997	0.039	0.985	0.986

TABLE III
COMPARISON OF PERFORMANCE METRICS ON LANDSAT-8 DATA SET

Method/Model	PSNR	UQI	SAM	SCC	SSIM
Brovey	10.210	0.813	0.007	0.905	0.78
IHS	11.32	0.997	0.0115	0.877	0.847
PCNN	40.308	0.990	0.059	0.936	0.947
Global	47.593	0.999	0.024	0.978	0.995
Multi-Patch	**48.045**	**0.999**	**0.022**	**0.979**	**0.996**
Global+Multi-Patch	47.87	0.999	0.029	0.979	0.994

V. RESULTS AND DISCUSSION

Performance metrics have been calculated for the three variants of our architecture on IKONOS and LANDSAT-8 are given in Tables II and III. From the tables, it is evident that our model has recorded better performance metrics values when compared with the traditional methods Brovey Transforms and Intensity Hue Saturation (IHS), and the basic deep learning model PCNN. All the three variants of our proposed model have attained excellent values of SAM and SCC, highlighting the low spatial and spectral distortion in the pansharpened images, which suggests that the models preserve the characteristics of the image adequately. A low SAM value suggests that the pansharpened image differs very little in the angle of spectral distortion compared to the groundtruth image. The high values of SSIM also suggest the similarity achieved between the pansharpened image and the high-resolution groundtruth. We conjecture that the achieved performance is due to the relative simpleness and effectiveness of our architecture, reinforcing our belief that deeper networks introduce more distortions. The multipatch attention model generated pansharpened images have been visualized in Figs 3 and 4. From the pansharpened images, it is evident that the proposed model produces sharper images and higher-resolution Multispectral images.

Out of the three variants, the Multi-Patch Attention network has constantly achieved better performance in terms of the metrics and image quality. The intricate, local details captured in the region-wise attention computation in the Multi-Patch Attention layer is responsible for superior performance. The fact that the proposed Multi-Patch attention network trains

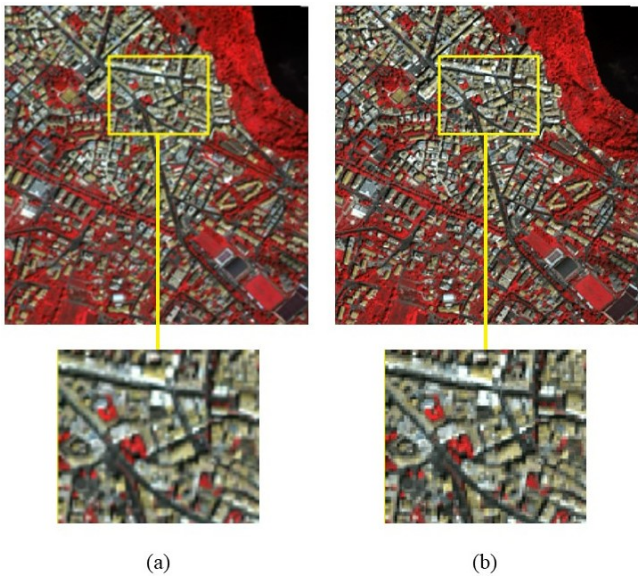

Fig. 3. Visual comparision of IKONOS Image (a) Interpolated MS image (b) Multi-Patch Attention Pansharpened MS image

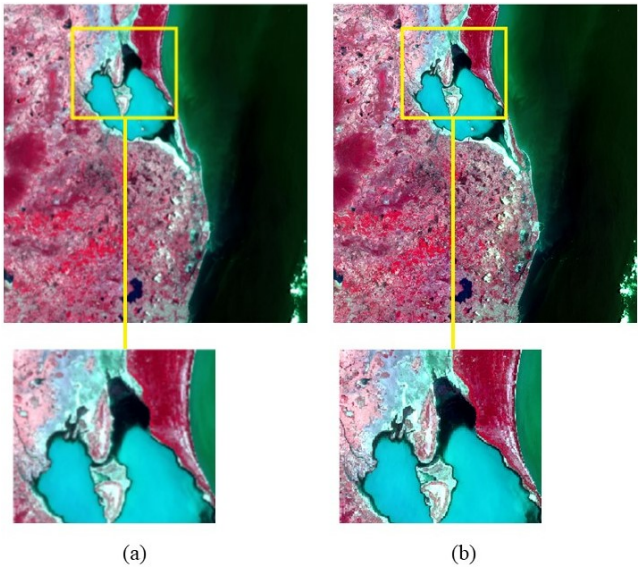

Fig. 4. Visual comparision of LANDSAT-8 Image (a) Interpolated MS image (b) Multi-Patch Attention Pansharpened MS image

faster than other networks makes it well suited to perform Pansharpening on large satellite images. To achieve further quality, the models can be trained on larger tiles, like 256×256, to take into account longer-range spatial dependencies.

VI. Conclusion and Future Directions

In this letter, we propose Pansformers, a channel-attention-based architecture for Pansharpening. The proposed architecture contains a novel Multi-Patch Attention module which computes attention on multiple, smaller patches to capture the intricate local-level details required for fusion. The proposed architecture is effective and able to produce high-quality pansharpened MS images as evaluated through the number of quality performance metrics and also outperforms the previous state-of-the-art methods. Further, as future work, it is suggested that an attempt for selective attention mechanism which computes attention separately over the MS and Panchromatic images may be carried out instead of combining them in the input image.

VII. Acknowledgments

The authors are grateful to Space Applications Centre, Indian Space Research Organization, Department of Space, India for supporting this work under RESPOND Scheme.

References

[1] P. Chavez, S. C. Sides, J. A. Anderson, et al. Comparison of three different methods to merge multiresolution and multispectral data-landsat tm and spot panchromatic. *Photogrammetric Engineering and remote sensing*, 57(3):295–303, 1991.

[2] C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014.

[3] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[5] W. Huang, L. Xiao, Z. Wei, H. Liu, and S. Tang. A new pan-sharpening method with deep neural networks. *IEEE Geoscience and Remote Sensing Letters*, 12(5):1037–1041, 2015.

[6] M. Jiang, H. Shen, J. Li, Q. Yuan, and L. Zhang. A differential information residual convolutional neural network for pansharpening. *ISPRS Journal of Photogrammetry and Remote Sensing*, 163:257–271, 2020.

[7] M. E. A. Larabi, M. S. Karoui, S. Chaib, K. Bakhti, and M. I. Tchenar. Multibranch cnn-based pansharpening with skip connection. In *2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS)*, pages 137–140. IEEE, 2020.

[8] X. Li, F. Xu, X. Lyu, Y. Tong, Z. Chen, S. Li, and D. Liu. A remote-sensing image pan-sharpening method based on multi-scale channel attention residual network. *IEEE Access*, 8:27163–27177, 2020.

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

[10] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran. Image transformer. In *International Conference on Machine Learning*, pages 4055–4064. PMLR, 2018.

[11] Y. Qu, R. K. Baghbaderani, H. Qi, and C. Kwan. Unsupervised pansharpening based on self-attention mechanism. *IEEE Transactions on Geoscience and Remote Sensing*, 59(4):3192–3208, 2020.

[12] S. Rahmani, M. Strait, D. Merkurjev, M. Moeller, and T. Wittman. An adaptive ihs pan-sharpening method. *IEEE Geoscience and Remote Sensing Letters*, 7(4):746–750, 2010.

[13] Z. Shao and J. Cai. Remote sensing image fusion with deep convolutional neural network. *IEEE journal of selected topics in applied earth observations and remote sensing*, 11(5):1656–1669, 2018.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[15] Y. Wei and Q. Yuan. Deep residual learning for remote sensed imagery pansharpening. In *2017 International Workshop on Remote Sensing with Intelligent Processing (RSIP)*, pages 1–4. IEEE, 2017.

[16] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley. Pannet: A deep network architecture for pan-sharpening. In *Proceedings of the IEEE international conference on computer vision*, pages 5449–5457, 2017.

[17] W. Yao, Z. Zeng, C. Lian, and H. Tang. Pixel-wise regression using u-net and its application on pansharpening. *Neurocomputing*, 312:364–371, 2018.

[18] Q. Yuan, Y. Wei, X. Meng, H. Shen, and L. Zhang. 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, 2018.