Fusformer: A Transformer-based Fusion Network for Hyperspectral Image Super-resolution

Jin-Fan Hu, Ting-Zhu Huang, *Member, IEEE*, Liang-Jian Deng, *Member, IEEE*, Hong-Xia Dou, Danfeng Hong, *Senior Member, IEEE*, Gemine Vivone, *Senior Member, IEEE*

***Abstract*—Hyperspectral image super-resolution (HISR) is to fuse a low-resolution hyperspectral image (LR-HSI) and a high-resolution multispectral image (HR-MSI), aiming to obtain a high-resolution hyperspectral image (HR-HSI). Recently, various convolution neural network (CNN) based techniques have been successfully applied to address the HISR problem. However, these methods generally only consider the relation of a local neighborhood by convolution kernels with a limited receptive field, thus ignoring the global relationship in a feature map. In this paper, we design a transformer-based architecture (called Fusformer) for the HISR problem, which is the first attempt to apply the transformer architecture to this task to the best of our knowledge. Thanks to the excellent ability of feature representations, especially by the self-attention in the transformer, our approach can globally explore the intrinsic relationship within features. Considering the specific HISR problem, since the LR-HSI holds the primary spectral information, our method estimates the spatial residual between the upsampled LR-MSI and the desired HR-HSI, reducing the burden of training the whole data in a smaller mapping space. Various experiments show that our approach outperforms current state-of-the-art HISR methods. The code is available at https://github.com/J-FHu/Fusformer.**

***Index Terms*—Hyperspectral image super-resolution, Transformer, Image fusion, Remote sensing.**

I. Introduction

HYPERSPECTRAL images can provide more abundant spectral characteristics than standard red-green-blue (RGB) images or multispectral images. As a result, the hyperspectral images have many practical applications, such as classification [1]–[3], and remote sensing [4], [5], thus showing a quite important role. However, limited by the current physical imaging system, there is an unavoidable issue, *i.e.*, it is impossible to generate an image with a high spatial resolution and a high spectral resolution at the same time [6]. Therefore, the HISR becomes a promising technique to produce the desired HR-HSI.

Many methods have been proposed from various perspectives in the last few years to address the HISR problem. They can be roughly divided into two classes, *i.e.*, traditional methods and deep learning (DL) based approaches.

For traditional methods, many researchers propose different prior knowledge in their models to exploit intrinsic properties under the maximum a posteriori (MAP) framework. This prior knowledge, such as sparsity, low rankness and self-similarity, has been applied to many computer vision and image processing tasks, see, *e.g.*, remote sensing pansharpening [7]–[9], and HISR [10]–[13]. However, those handcrafted priors are usually limited to depicting all the latent characteristics. Furthermore, their optimal algorithm parameters sometimes need to be tuned carefully for different datasets.

Recently, DL-based methods, especially CNN techniques, have been exploited to deal with the HISR problem and showed promising performance, see *e.g.*, [14]–[21]. CNNs are more flexible and comprehensive than traditional approaches using manual priors. Nevertheless, each neuron in CNN architecture has a limited receptive field; thus, the CNN only observes the input feature in a local neighborhood. Besides, excellent results are usually accompanied by an excessive amount of parameters. Due to the above-mentioned issues, the further development of CNN methods has been limited.

It is worth to be remarked that the transformer architecture proposed by Vaswani *et al*. in [22] and its various modifications [23]–[26] have obtained outstanding achievements in many tasks. This novel technique brings a powerful capability of extracting global information from the whole feature map. Moreover, Gu *et al*. [27] showed that a wider range of features can achieve better performance in image super-resolution. Based on the excellent property of the transformer module, a network architecture (called Fusformer) is designed for the HISR problem. Our method integrates a self-attention mechanism that can exploit more global relationships among pixels than standard convolution operations with a limited receptive field.

To sum up, this paper designs an efficient network architecture to solve the HISR problem. The contributions of this paper are summarized as follows:

1) To the best of our knowledge, this work represents the first attempt to utilize the transformer for solving the HISR problem. The self-attention mechanism in the transformer enables our network to represent more global information than previous CNN architectures. A

The work is supported by National Natural Science Foundation of China (Grant No. 12171072, 61702083 and 61876203), Key Projects of Applied Basic Research in Sichuan Province (Grant No. 2020YJ0216), and National Key Research and Development Program of China (Grant No. 2020YFA0714001)

J.-F. Hu, T.-Z. Huang, and L.-J. Deng are with the School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731, China (e-mails: hujf0206@163.com; tingzhuhuang@126.com; liangjian.deng@uestc.edu.cn).

H.-X. Dou is with the School of Science, Xihua University, Chengdu, 610039, China (e-mail: hongxia.dou@mail.xhu.edu.cn).

D. Hong is with the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, 100094, China (e-mail: hongdf@aircas.ac.cn).

G. Vivone is with the Institute of Methodologies for Environmental Analysis, National Research Council, 85050 Tito Scalo, Italy (e-mail: gemine.vivone@imaa.cnr.it).

preliminary version of the paper can be found in the preprint website[1].

2) The proposed approach focuses on learning in the residual domain instead of the image domain, which leads to a smaller mapping space for more accessible training.
3) Only a few parameters are involved in the given lightweight network, making our approach more practical. Furthermore, the network is plain and easy to follow. Thus, future researchers can easily improve our simple yet effective architecture.

II. Proposed Method

A. Background

Previous CNN-based methods have obtained state-of-the-art (SOTA) performance in recent years. The core elements in the CNN framework are various convolution kernels with limited sizes, resulting in the receptive field being restricted within a small (local) area. As a result, the global structure containing valuable information is neglected in the CNN framework. Considering the limitation of the standard convolution, better ways to extract and understand global information are challenges of paramount importance.

The transformer model was proposed first by Vaswani *et al*. in 2017 [22]. The transformer outperforms other methods and has proven crucial and effective in natural language processing tasks. Motivated by the success of the transformer architecture, Dosovitskiy *et al*. [23] proposed the vision transformer (ViT) for image classification. After that, Chen *et al*. [24] designed the image processing transformer (IPT) to address low-level vision tasks. The achievements of the transformer in various tasks inspired us to design a network utilizing its superior ability to capture long-term information and relationships in the HISR problem.

B. Network Architecture

According to the analysis of Gu *et al*. [27], a wider range of involved pixels usually brings better performance. However, the global relationship among all the pixels is hard to be obtained due to the limitation of the standard convolution operation in the CNN architecture. Hence, we integrate the transformer into our network for effectively and globally exploiting that information. The overall flowchart of our network structure is presented in Fig. 1. In this figure, the LR-HSI, $\mathcal{Y} \in \mathbb{R}^{h\times w\times S}$, holds a similar spectral structure as the ground-truth HR-HSI, $\mathcal{X} \in \mathbb{R}^{H\times W\times S}$, where $h = f \cdot H$, $w = f \cdot W$, and f is the scaling factor, besides, S represents the band number of the HSI.

1) Inputs: From Fig. 1, it is clear that the inputs of the given architecture are the upsampled LR-HSI, $\mathcal{Y}^U \in \mathbb{R}^{H\times W\times S}$ and the HR-MSI, $\mathcal{Z} \in \mathbb{R}^{H\times W\times s}$. s represents the band number of the MSI. The HR-MSI, $\mathcal{Z}$, is concatenated first with $\mathcal{Y}^U$ along the spectral dimension to get the data cube, $\mathcal{D} \in \mathbb{R}^{H\times W\times (S+s)}$, containing the spectral and spatial information. Then we unfold the tensor $\mathcal{D}$ to a matrix $\mathbf{D} \in \mathbb{R}^{HW\times (S+s)}$ due to the input's dimension requirement of the transformer model. It is worth noting that each row vector in the matrix $\mathbf{D}$ has its own significance. Namely, the vector $d = [d_1, d_2] \in \mathbb{R}^{1\times (S+s)}$ can be viewed as two vectors $d_1 \in \mathbb{R}^{1\times S}$ and $d_2 \in \mathbb{R}^{1\times s}$, which denote the tube pixels of the hyperspectral and multispectral images, respectively. Note that other transformer-based methods for computer vision tasks, such as [23], [24], reshape a small image patch (*e.g.*, 16×16 in ViT) into a vector instead of a pixel. On one hand, the hyperspectral image contains more spectral bands than natural RGB images, thus the vector reshaped from a hyperspectral image patch leads to a heavy computation load. On other hand, pixel-wise information (instead of patch-wise) embedded to a vector is also suitable for our pixel-wise super-resolution problem. Hence, the transformer model is quite consistent with the characteristics of the HISR problem. Every pixel can naturally be represented as a vector, and the transformer architecture enables the network to discover and consider the global relationships among all the pixels. With a simple fully connected layer, the matrix $\mathbf{D} \in \mathbb{R}^{HW\times (S+s)}$ is then embedded to the matrix $\mathbf{E} \in \mathbb{R}^{HW\times F}$, where F denotes the number of feature channels. After getting the inputs, we send the embedded features to the transformer model.

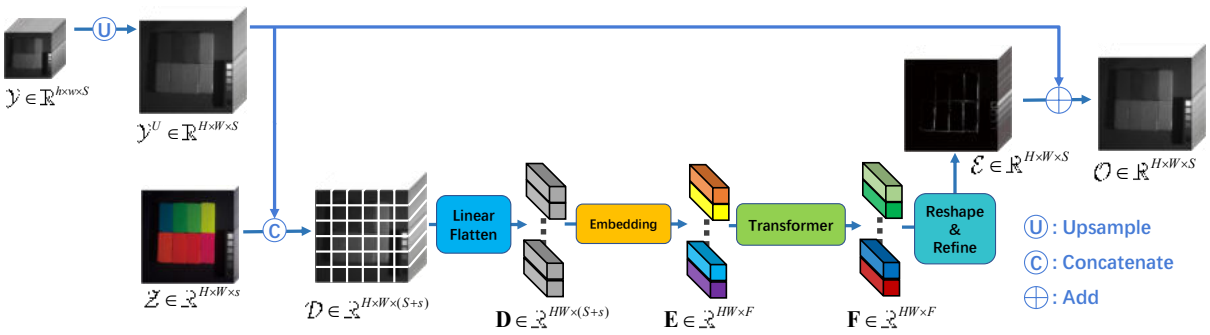
2) Main Network Architecture: The transformer model is the main part of our architecture, which is shown in Fig. 1-(b). We use both the encoder and decoder parts of the original ViT. For the encoder shown in the top of Fig. 1-(b), a layer normalization (LN), widely used in transformer-based methods [22]–[24], is employed for the training's stability. Then, the use of the multi-head attention, mainly constructed by the self-attention (SA) mechanism with multi-heads, enables the network to capture the long-term information and the global relationship in the HISR problem. The calculation of the self-attention of the input $\mathbf{X} \in \mathbb{R}^{HW\times F}$, *i.e.*, $A = \mathrm{SA}(\mathbf{X})$, is given as follows:

$$\begin{aligned}
[\mathbf{Q}, \mathbf{K}, \mathbf{V}] &= \mathbf{X}\left[\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v\right], \\
A &= \mathrm{softmax}\left(\frac{\mathbf{X}\mathbf{W}_q\mathbf{W}_k^T\mathbf{X}^T}{\sqrt{d_k}}\right)\mathbf{X}\mathbf{W}_v \\
&= \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \\
&= \mathbf{S}\mathbf{V},
\end{aligned} \tag{1}$$

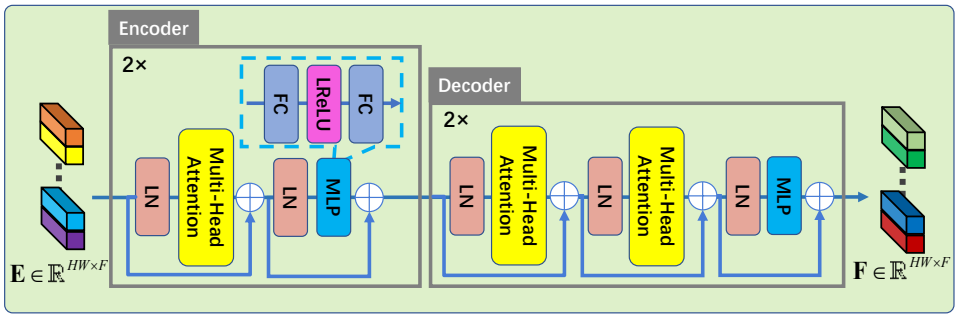
where $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v \in \mathbb{R}^{F\times b}$ denote the corresponding learnable weights of $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ (*i.e.*, the query, the key, and the value matrices), respectively. Besides, b indicates the number of feature channels, d_k represents the dimension of $\mathbf{K}$ for scaling, and $\cdot^T$ is the transpose operator. Note that the score of $\mathbf{S} \in \mathbb{R}^{HW\times HW}$ defines the degree of similarity among all the pixels in the data to some extent. A larger value of s_{mn} $(m, n = 1, \cdots, HW)$ in the similarity matrix, $\mathbf{S}$, represents a stronger relationship between the m-th and n-th pixel in $\mathbf{X}$. Next, the introduction of $\mathbf{V}$ gives learning flexibility to the network to extract the intrinsic features of $\mathbf{X}$. The detailed encoder is described as follows:

$$\begin{aligned}
&\mathbf{X}_0 = \mathbf{E}, \\
&\mathbf{X}'_i = \mathrm{MHA}(\mathrm{LN}(\mathbf{X}_{i-1})), \\
&\mathbf{X}_i = \mathbf{X}_{i-1} + \mathbf{X}'_i, \\
&\mathbf{X}'_i = \mathrm{MLP}(\mathrm{LN}(\mathbf{X}_i)), \\
&\mathbf{X}_i = \mathbf{X}_i + \mathbf{X}'_i, \qquad i = (1, 2),
\end{aligned} \tag{2}$$

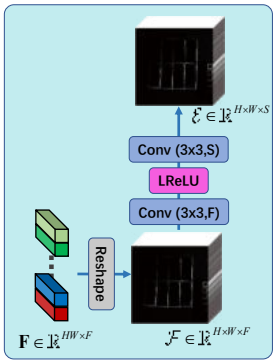
[1]https://arxiv.org/abs/2109.02079



(a) The sketch of our Fusformer architecture



(b) Transformer module



(c) Reshape & Refine

Fig. 1. (a) Overview of the proposed Fusformer architecture. (b) The transformer module used in the network. MLP denotes the multi-layer perception, LN represents the layer normalization, FC represents the fully connected layer and LReLU indicates the leaky ReLU activation function. (c) An illustration of the Reshape & Refine module.

where $\mathrm{MHA}(\cdot) = (\mathrm{SA}_1(\cdot), \cdots, \mathrm{SA}_3(\cdot))$ denotes the multi-head attention module, $\mathrm{MLP}(\cdot)$ defines the multi-layer perception, and $\mathrm{LN}(\cdot)$ indicates the layer normalization. The decoder can be described in a similar way. Finally, the feature matrix $\mathbf{F} \in \mathbb{R}^{HW \times F}$ is obtained by the transformer module, and then it is reshaped into a tensor $\mathcal{F} \in \mathbb{R}^{H \times W \times F}$ in the Reshape & Refine module to generate the residual $\mathcal{E} \in \mathbb{R}^{H \times W \times S}$.

3) Loss function: With a skip connection, we finally add the learned residual $\mathcal{E} \in \mathbb{R}^{H \times W \times S}$ to the upsampled LR-HSI, $\mathcal{Y}^U$, to obtain the HR-HSI, $\mathcal{O} \in \mathbb{R}^{H \times W \times S}$. Then, the ℓ_1 loss function is employed to train the network:

$$\begin{aligned} \mathcal{L}_i &= \frac{1}{N} \sum_{n=1}^{N} \left\| \mathcal{Y}^U_{(n)} + \mathcal{E}_{(n)} - \mathcal{X}_{(n)} \right\|_1 \\ &= \frac{1}{N} \sum_{n=1}^{N} \left\| \mathcal{O}_{(n)} - \mathcal{X}_{(n)} \right\|_1, \end{aligned} \tag{3}$$

where N indicates the number of training pairs, and $\|\cdot\|_1$ represents the ℓ_1 norm that has shown its superiority in preserving edges and textures [28].

III. Experiments

To verify the effectiveness of our Fusformer, we compare it with representative SOTA HISR methods, including 1) traditional approaches: the fast fusion based on solving Sylvester equation (FUSE) approach [29], the generalized Laplacian pyramid for hypersharpening (GLP-HS) technique [7], the coupled sparse tensor factorization (CSTF) method [10] and the fusion with CNN denoiser (CNN-FUS) approach [30]; 2) DL-based approaches: the spatial-spectral reconstruction network (SSRNet) [20], the residual two-stream fusion network, (ResTFNet) [21], the MS/HS image fusion network (MHF-Net) [15], and the hyperspectral image super-resolution network (HSRnet) [31], on three benchmark hyperspectral image datasets, *i.e.*, CAVE dataset [32], Harvard dataset [33] and Chikusei dataset [34]. Note that all the networks are only trained on the CAVE dataset and tested on both CAVE and Harvard datasets, thus, the experiments on the Harvard images can be viewed as a test for the network generalization, which is of crucial importance for DL-based methods. Furthermore, the Chikusei dataset is selected for remote sensing hyperspectral images experiments. It consists of 2517 × 2335 pixels and has 128 bands. We regard the original data as the ground-truth HR-HSI and simulate the LR-HSI in the same way as the previous experiments. The corresponding RGB image for the HR-MSI is obtained by Canon EOS 5D Mark II coupled with the HR-HSI. Following that, we choose the top-left area with a spatial size of 1000 × 2200 for training and cut 64 × 64 overlapping regions from the training part as the ground-truth HR-HSI patches. Furthermore, the input HR-MSI and LR-HSI patches are 64 × 64 × 3 and 16 × 16 × 128, respectively. For the testing data, we extract the same 6 non-overlapping 680 × 680 × 128 areas from the leftover Chikusei dataset as the experiment setting in [31]. For fair comparison, the used datasets are the same as in [31]. Moreover, four widely used quality indexes (QIs), *i.e.*, the peak signal-to-noise ratio (PSNR), the spectral angle mapper (SAM) [35], the erreur relative globale adimensionnelle de synthèse (ERGAS) [36], and the structure similarity (SSIM) [37] are

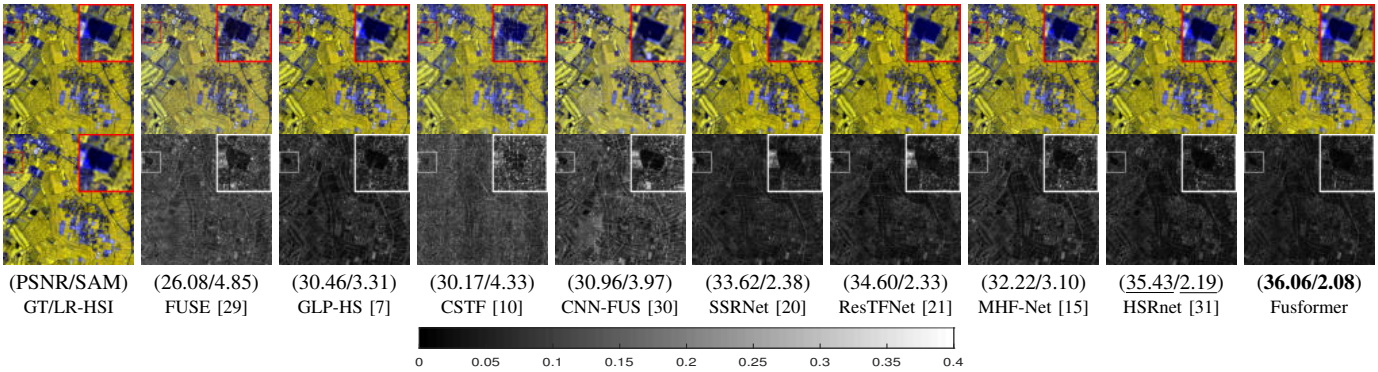


Fig. 2. The top/bottom of the first column shows the GT image/corresponding LR-HSI of testing image ($680 \times 680 \times 128$) from Chikusei dataset with pseudo-color display (R-81, G-76, B-2). The 2nd-10th columns: the pseudo-color results and the corresponding residual maps of *area* obtained by the methods in the benchmark, pointing out some close-ups to facilitate the visual analysis.

used for quantitative evaluation. Furthermore, all the DL-based networks are trained using Python 3.8.5, Pytorch 1.7.1, and an NVIDIA GPU GeForce GTX 2080Ti on Windows operating system. To minimize the loss function (3), we utilized the Adam optimizer with a dynamic learning rate starting from 1×10^{-3} and multiplying by 0.1 for every 200 epochs (totally 1000 epochs). In addition, the batch size is set as 3 for training. The feature channels F and b of $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v \in \mathbb{R}^{F \times b}$ in the self-attention are set to 48 and 16, respectively.

TABLE I
AVERAGE QUANTITATIVE RESULTS ON CAVE (11 TESTING IMAGES), HARVARD (10 TESTING IMAGES) AND CHIKUSEI DATASETS (6 TESTING IMAGES) AND THE CORRESPONDING AVERAGE RUNNING TIMES AND NUMBER OF PARAMETERS. THE BEST VALUES ARE HIGHLIGHTED IN BOLDFACE, THE SECOND-BEST VALUES ARE UNDERLINED. C INDICATES CPU, G INDICATES GPU, AND M INDICATES A MILLION.

Dataset	Method	PSNR	SAM	ERGAS	SSIM	Running Time (s)	# parameters
CAVE	FUSE [29]	39.72±3.5	5.83±2.0	4.18±3.1	0.975±0.02	1.357 (C)	/
	GLP-HS [7]	37.81±3.1	5.36±1.8	4.66±2.7	0.972±0.01	5.621 (C)	/
	CSTF [10]	42.14±3.0	9.92±4.1	3.08±1.6	0.964±0.03	22.887 (C)	/
	CNN-FUS [30]	42.66±3.5	6.44±2.3	2.95±2.2	0.982±0.01	6.772 (C+G)	/
	SSRNet [20]	45.28±3.1	4.72±1.8	2.06±1.3	0.990±0.00	**0.002** (G)	**0.03M**
	ResTFNet [21]	45.35±3.7	3.76±1.3	1.98±1.6	0.993±0.002	0.003 (G)	2.26M
	MHF-Net [15]	46.32±2.7	4.33±1.8	1.74±1.2	0.992±0.00	0.224 (G)	3.63M
	HSRnet [31]	47.82±2.7	2.66±0.9	1.34±0.8	**0.995**±0.00	0.068 (G)	1.90M
	Fusformer	**48.56**±3.0	**2.52**±0.8	**1.30**±0.9	**0.995**±0.00	0.173 (G)	0.10M
Harvard	FUSE [29]	42.06±2.9	3.23±0.9	3.14±1.5	0.977±0.01	4.935 (C)	/
	GLP-HS [7]	40.14±3.2	3.52±1.0	3.74±1.4	0.966±0.01	20.578 (C)	/
	CSTF [10]	42.97±3.5	3.30±1.2	**2.43**±1.1	0.972±0.02	26.346 (C)	/
	CNN-FUS [30]	43.61±4.7	3.32±1.2	2.78±1.6	0.978±0.02	25.355 (C+G)	/
	SSRNet [20]	39.87±4.2	5.40±2.3	5.44±2.2	0.963±0.02	**0.003** (G)	**0.03M**
	ResTFNet [21]	38.39±4.3	5.85±2.5	6.98±2.4	0.957±0.02	**0.003** (G)	2.26M
	MHF-Net [15]	40.37±3.7	4.64±1.8	24.17±46.7	0.966±0.01	0.864 (G)	3.63M
	HSRnet [31]	44.28±3.0	**2.66**±0.7	2.45±0.8	**0.984**±0.01	0.267 (G)	1.90M
	Fusformer	**44.42**±3.2	**2.66**±0.7	2.48±1.0	**0.984**±0.01	0.247 (G)	0.10M
Chikusei	FUSE [29]	27.76±1.5	4.80±1.2	7.22±0.5	0.882±0.02	4.992 (C)	/
	GLP-HS [7]	31.60±1.3	3.29±0.3	5.69±0.3	0.919±0.01	39.784 (C)	/
	CSTF [10]	30.36±0.9	4.58±0.5	5.91±0.6	0.824±0.02	25.675 (C)	/
	CNN-FUS [30]	31.83±1.7	4.76±0.9	5.25±0.9	0.918±0.01	10.576 (C+G)	/
	SSRNet [20]	35.54±1.2	2.33±0.2	3.79±0.3	0.953±0.01	**0.003** (G)	**0.03M**
	ResTFNet [21]	36.70±1.5	2.20±0.2	3.66±0.3	0.949±0.01	**0.003** (G)	2.26M
	MHF-Net [15]	33.19±1.0	3.18±0.4	6.24±0.4	0.927±0.01	0.327 (G)	3.63M
	HSRnet [31]	**36.95**±1.1	2.08±0.2	**3.60**±0.3	0.952±0.01	0.352 (G)	1.90M
	Fusformer	36.34±0.9	**2.00**±0.2	3.68±0.3	**0.957**±0.01	0.276 (G)	0.10M

Tab. I reports the quantitative comparisons on the CAVE, Harvard and Chikusei datasets. The proposed Fusformer gets the best outcomes on almost all the QIs and only involves about 0.1 million parameters (differently from MHF-Net and HSRnet involving 3.63 million and 1.9 million, respectively), making our network more practical. In Fig. 2, we also show visual performance and the corresponding absolute residual maps of one selected sample from the Chikusei dataset. One can observe that the Fusformer obtains the smallest error, both in the overall outcome and in the zoomed-in area. Moreover, FUSE [29], GLP-HS [7], CSTF [10] and CNN-FUS [30] have more obvious residual maps than those deep learning-based approaches, and our Fusformer still obtains the highest PSNR and lowest SAM, also showing the darkest residuals among all these methods. These outcomes could be due to the powerful feature representation and extraction ability of Transformer (especially self-attention for depicting global dependencies).

Generalization: The generalization ability of DL-based methods is of crucial importance. As presented in Tab. I, our Fusformer is still satisfying, showing the best performance for almost all the QIs. We believe that the excellent generalization of the proposed network comes mainly from two aspects. i) The global feature extraction brought by the Transformer allows Fusformer not to be restricted by the specific content of a local region. ii) Fusformer is trained in the high-pass domain, rather than the image domain directly associated with a specific image content, yielding abstract residual information.

TABLE II
AVERAGE QIS AND RELATED STANDARD DEVIATIONS OF THE RESULTS ON THE CAVE DATASET USING THE PROPOSED METHOD WITH AND WITHOUT THE RESIDUAL LEARNING STRATEGY (RLS). THE BEST VALUES ARE HIGHLIGHTED IN BOLDFACE.

Method	PSNR	SAM	ERGAS	SSIM
W/o RLS	42.71±7.82	3.29±1.24	4.48±7.43	0.984±0.02
Fusformer	**48.56**±3.03	**2.52**±0.83	**1.30**±0.86	**0.995**±0.00

Residual Learning: Fusformer learns the residuals between the upsampled LR-HSI and the HR-HSI instead of directly reconstructing the HR-HSI. We conduct a simple experiment to verify the effectiveness of residual learning (RL). Tab. II shows the results of the architecture with or without the RL. It is clear that the adding of the upsampled LR-HSI, $\mathcal{Y}^U$, is important for the network to boost the performance and strengthen the stability.

IV. CONCLUSIONS

This paper proposed a transformer-based network architecture, called Fusformer, for the HISR problem. Compared with previous CNN-based methods, our method can consider the global information instead of focusing on the local neighborhood with a limited kernel size. To the best of our knowledge, it is the first attempt to adopt the transformer to the HISR problem. Experimental results demonstrated our method's SOTA performance using fewer network parameters and with a better network generalization.

REFERENCES

[1] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 6, pp. 1351–1362, 2005.

[2] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 5966–5978, 2020.

[3] M. E. Paoletti, J. M. Haut, X. Tao, J. Plaza, and A. Plaza, "Flop-reduction through memory allocations within cnn for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 5938–5952, 2021.

[4] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, "Advances in spectral-spatial classification of hyperspectral images," *Proceedings of the IEEE*, vol. 101, no. 3, pp. 652–675, 2012.

[5] J. M. Haut, M. E. Paoletti, S. Moreno-Álvarez, J. Plaza, J.-A. Rico-Gallego, and A. Plaza, "Distributed deep learning for remote sensing data interpretation," *Proceedings of the IEEE*, vol. 109, no. 8, pp. 1320–1349, 2021.

[6] R. Dian, S. Li, B. Sun, and A. Guo, "Recent advances and new guidelines on hyperspectral and multispectral image fusion," *Information Fusion*, vol. 69, pp. 40–51, 2020.

[7] M. Selva, B. Aiazzi, F. Butera, L. Chiarantini, and S. Baronti, "Hyper-sharpening: A first approach on SIM-GA data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 6, pp. 3008–3024, 2015.

[8] Z.-C. Wu, T.-Z. Huang, L.-J. Deng, G. Vivone, J.-Q. Miao, J.-F. Hu, and X.-L. Zhao, "A new variational approach based on proximal deep injection and gradient intensity similarity for spatio-spectral image fusion," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 6277–6290, 2020.

[9] L.-J. Deng, M. Feng, and X.-C. Tai, "The fusion of panchromatic and multispectral remote sensing images via tensor-based sparse modeling and hyper-laplacian prior," *Information Fusion*, vol. 52, pp. 76–89, 2019.

[10] S. Li, R. Dian, L. Fang, and J. M. Bioucas-dias, "Fusing hyperspectral and multispectral images via coupled sparse tensor factorization," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4118–4130, 2018.

[11] T. Xu, T.-Z. Huang, L.-J. Deng, X.-L. Zhao, and J. Huang, "Hyperspectral image superresolution using unidirectional total variation with tucker decomposition," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 4381–4398, 2020.

[12] R. Dian, S. Li, and L. Fang, "Learning a low tensor-train rank representation for hyperspectral image super-resolution," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 9, pp. 2672 –2683, 2019.

[13] R. Dian and S. Li, "Hyperspectral image super-resolution via subspace-based low tensor multi-rank regularization," *IEEE Transactions on Image Processing*, vol. 28, no. 10, pp. 5135–5146, 2019.

[14] R. Dian, S. Li, A. Guo, and L. Fang, "Deep hyperspectral image sharpening," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 11, pp. 5345–5355, 2018.

[15] Q. Xie, M. Zhou, Q. Zhao, Z. Xu, and D. Meng, "MHF-Net: An interpretable deep network for multispectral and hyperspectral image fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1457–1473, 2022.

[16] Y. Qu, H. Qi, and C. Kwan, "Unsupervised sparse dirichlet-net for hyperspectral image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2511–2520.

[17] J. Zhou, C. Kwan, and B. Budavari, "Hyperspectral image super-resolution: A hybrid color mapping approach," *Journal of Applied Remote Sensing*, vol. 10, no. 3, p. 035024, 2016.

[18] Z. Zhang, K. Gao, J. Wang, L. Min, J. Shijing, C. Ni, and D. Chen, "Gradient enhanced dual regression network: Perception preserving super-resolution for multi-sensor remote sensing imagery," *IEEE Geoscience and Remote Sensing Letters*, 2021, DOI:10.1109/LGRS.2021.3134798.

[19] J. Hu, Y. Tang, Y. Liu, and S. Fan, "Hyperspectral image super-resolution based on multi-scale mixed attention network fusion," *IEEE Geoscience and Remote Sensing Letters*, 2021, DOI:10.1109/LGRS.2021.3124974.

[20] X. Zhang, W. Huang, Q. Wang, and X. Li, "SSR-NET: Spatial-spectral reconstruction network for hyperspectral and multispectral image fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 5953–5965, 2021.

[21] X. Liu, Q. Liu, and Y. Wang, "Remote sensing image fusion based on two-stream fusion network," *Information Fusion*, 2020, DOI:10.1016/j.inffus.2019.07.010.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.

[23] 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.

[24] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 12 299–12 310.

[25] D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, and J. Chanussot, "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022, DOI: 10.1109/TGRS.2021.3130716.

[26] Z. Zhong, Y. Li, L. Ma, J. Li, and W.-S. Zheng, "Spectral-spatial transformer network for hyperspectral image classification: A factorized architecture search framework," *IEEE Transactions on Geoscience and Remote Sensing*, 2021, DOI:10.1109/TGRS.2021.3115699.

[27] J. Gu and C. Dong, "Interpreting super-resolution networks with local attribution maps," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 9199–9208.

[28] T. Huang, W. Dong, X. Yuan, J. Wu, and G. Shi, "Deep gaussian scale mixture prior for spectral compressive imaging," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 16 216–16 225.

[29] Q. Wei, N. Dobigeon, and J.-Y. Tourneret, "Fast fusion of multi-band images based on solving a Sylvester equation," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4109–4121, 2015.

[30] R. Dian, S. Li, and X. Kang, "Regularizing hyperspectral and multispectral image fusion by CNN denoiser," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 3, pp. 1124–1135, 2021.

[31] J.-F. Hu, T.-Z. Huang, L.-J. Deng, T.-X. Jiang, G. Vivone, and J. Chanussot, "Hyperspectral image super-resolution via deep spatiospectral attention convolutional neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, 2021, DOI:10.1109/TNNLS.2021.3084682.

[32] F. Yasuma, T. Mitsunaga, D. Iso, and S. Nayar, "Generalized assorted pixel camera: postcapture control of resolution, dynamic range, and spectrum," *IEEE Transactions on Image Processing*, vol. 19, no. 9, pp. 2241–2253, 2010.

[33] A. Chakrabarti and T. Zickler, "Statistics of real-world hyperspectral images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 193–200.

[34] N. Yokoya and A. Iwasaki, "Airborne hyperspectral data over chikusei," Space Application Laboratory, University of Tokyo, Japan, Tech. Rep. SAL-2016-05-27, 2016.

[35] R. Yuhas, J. Boardman, and A. Goetz, "Determination of semi-arid landscape endmembers and seasonal trends using convex geometry spectral unmixing techniques," in *Proceedings of the Annual JPL Airborne Geoscience Workshop*, 1993, pp. 620–636.

[36] L. Wald, *Data fusion: definitions and architectures: fusion of images of different spatial resolutions*. Presses des MINES, 2002.

[37] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.