High-Quality Multispectral Image Reconstruction for the Spectral Camera Based on Ghost Imaging via Sparsity Constraints Using CoT-Unet

Tao Hu, Jianxia Chen, Shu Wang, Jianrong Wu, Ziyan Chen, Zhifu Tian, Ruipeng Ma, and Di Wu

***Abstract*—To solve the problem of poor quality in ghost imaging via sparsity constraints (GISC) multispectral image reconstruction with correlation operations and compressed sensing algorithms under low sampling rate detection conditions, we propose an end-to-end deep-learning-based method. Based on the U-Net, Res2Net-SE-Conv is employed instead of convolutional blocks to extract local and global image features at a more fine-grained level while adaptively adjusting the channel feature response. The two-dimensional contextual transformer is constructed to fully use contextual correlation information to enhance the effectiveness of feature representations. We employ the two-dimensional contextual transformer in the decoder part, dubbed CoT-Unet, to reconstruct the desired 3D cube. The results show that compared with U-Net, TSA-Net based on spatial-spectral self-attention, the PSNR of reconstructed images by the CoT-Unet is improved by 5 dB and 3 dB, respectively, SSIM is improved by 0.23 and 0.07, and SAM is decreased by 0.06 and 0.58. Compared with conventional algorithms such as DGI and CS, our method significantly improves the quality of reconstructed images. Furthermore, the comparison results at 10%, 20%, and 30% sampling rates show that our approach has the best quality in reconstructing GISC multispectral images at low sampling rates.**

***Index Terms*—Multispectral image reconstruction, convolutional neural network, transformer, self-attention mechanism, ghost imaging.**

I. Introduction

GHOST imaging via sparsity constraints (GISC) spectral camera [1] is a new imaging system different from conventional spectral imaging, which has the advantages of transmissible media imaging, computational imaging, etc. The underlying principle is to use a spatial phase modulator [2] to modulate the image by phase throughout the spectral band, capturing a three-dimensional (3D) spectral data cube through a single two-dimensional(2D) measurement. At the same time, the system incorporates compressive sensing (CS) [3] to perform signal acquisition at frequencies lower than the Nyquist, thus improving the efficiency of optical channel capacity utilization and achieving information compression during the imaging acquisition process. Although the research on the principle and system of GISC spectral imaging has been more mature in recent years [4], the imaging quality of the system still needs to be improved at low sampling rates. For single-photon array detection, red outside array detection, and other small-scale surface array detectors, studying image reconstruction algorithms at low sampling rates provides technical support for their practical application.

Image reconstruction is an essential technical aspect of an imaging system, and the performance of the reconstruction algorithm directly affects the final imaging quality. With the development of deep learning (DL) in image processing, convolutional neural network (CNN) can use models with solid learning ability to establish end-to-end mapping functions from 2D measurements to 3D multispectral image data cubes. At present, many CNN-based reconstruction algorithms have been developed. Among them, image reconstruction algorithms based on U-Net [5] have achieved better results. $\lambda-$Net [6] adopts shallow U-Net to restore the spatial image details in the spectral channel. TSA-Net [7] embeds spatial and spectral self-attention models into the encoder-decoder structure to achieve reconstruction using spatial and spectral correlations of multispectral images. However, the limited convolutional field in CNN has some limitations [8], which cannot fully obtain the practical global information of the picture. Recently, the natural language processing (NLP) field has witnessed the rise of transformer with self-attention in robust language modeling architectures [9], [10] that triggers long-range interaction in a scalable manner. Inspired by this, there has been a steady momentum of breakthroughs [11] that push the limits of image reconstruction tasks by integrating CNN architecture with transformer modules. Nevertheless, these existing deep learning algorithms are not directly oriented to image reconstruction algorithms for GISC spectral cameras.

In GISC multispectral image reconstruction, correlation operations and compressed sensing algorithms are usually used. Among them, correlation algorithms such as differential

Manuscript received 10 February 2023; revised 10 May 2023; accepted 20 May 2023. Date of publication 24 May 2023; date of current version 7 June 2023. This work was supported by the National Science Foundation for Young Scientists of China under Grant 62001162. *(Corresponding author: Jianxia Chen.)*

Tao Hu, Jianxia Chen, Shu Wang, Zhifu Tian, and Di Wu are with the School of Institute of Data and Target Engineering, PLA Strategic Support Force Information Engineering University, Zhengzhou 450001, China (e-mail: hutaoengineering@163.com; 441142805@qq.com; 944689682@qq.com; tzhifu@qq.com; wudipaper@sina.com).

Jianrong Wu and Ziyan Chen are with the Key Laboratory for Quantum Optics of CAS, Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Shanghai 201800, China (e-mail: jrwu@siom.ac.cn; rzzyymzk@163.com).

Ruipeng Ma is with the School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou 450002, China (e-mail: 13164351610@163.com).

Digital Object Identifier 10.1109/JPHOT.2023.3279386

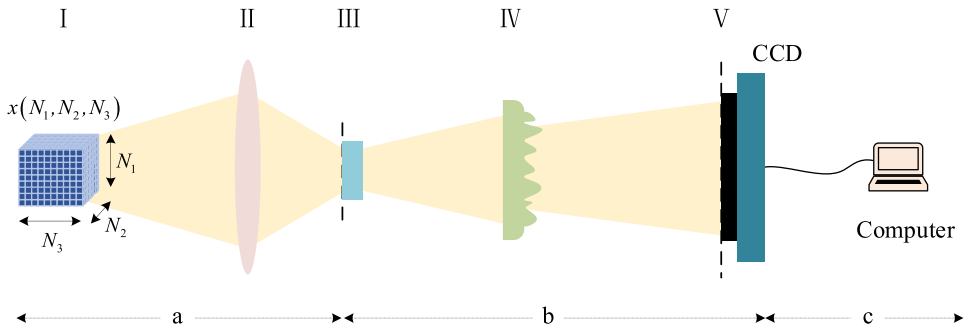


Fig. 1. Schematic of GISC spectral camera[1]: (I) the object, (II) the conventional imaging system, (III) the first image plane, (IV) the spatial phase modulator, (V) the speckles plane, (a) the front imaging module, (b) the modulation detection module, (c) the demodulation reconstruction module. N1: the height, N2: the width, N3: the spectral bands.

ghost imaging (DGI) [12] compute imaging by second-order correlation function, which has a simple computational process. Still, the quality of the reconstructed image is not high at low sampling rates. CS algorithms use the prior characteristics of multispectral images, such as low-rank and sparse modeling, to reconstruct images by solving underdetermined equations, improving image quality at low sampling rates. Still, its iterative operations increase computational complexity. Meanwhile, the lower the system's sampling rate, the more complex reconstructing a clear target image is. SSTU-Net3+ [13] obtained relatively good reconstruction results, but no further discussion was made on the sampling rate. Although existing deep learning methods have accepted good results in multispectral image reconstruction tasks, there are still the following problems in introducing them to GISC spectral imaging:

1) In GISC spectral imaging, the system measurement matrix size is enormous from reference [14], and the existing multispectral image reconstruction algorithms cannot be directly applied to GISC image reconstruction.
2) For the limited convolutional field in a standard convolutional kernel [8], extracting multispectral image features through a standard convolutional stacking network will lead to insufficient image feature extraction.
3) Difficult reconstruction of image detail information. In multispectral image reconstruction, the self-attention mechanism in the existing transformer relies on isolated key-value pairs during the computation and loses the correlation information between adjacent keys [15].
4) Insufficient discussion on the quality of reconstructed images under low sampling rate detection conditions.

To solve the above problems, we propose an effective end-to-end DL-based multispectral image reconstruction algorithm in this article. The main contributions of this article include the following four points:

- Adopting the correlation calculation results of DGI as the input of the CoT-Unet to avoid the measurement matrix increasing the complexity of network training.
- Based on the U-Net framework, we employ the Res2Net-SE-Conv module to replace convolution blocks to extract image features, which entirely use multispectral images' spatial and spectral correlation.
- Based on different spatial orientations, a two-dimensional contextual transformer module is constructed and embedded into the U-Net decoder side to represent multispectral images' spatial detail features better.
- The effectiveness of our algorithm in the GISC multispectral image reconstruction task at low sampling rates is verified by conducting comparison experiments at different sampling rates.

II. Related Works

A. Image Reconstruction for GISC Spectral Camera

The GISC spectral camera [1] contains three modules: (a) front imaging, (b) modulation detection, and (c) demodulation reconstruction. As shown in Fig. 1. First, the object x is projected onto the first image plane through the conventional imaging system. Then, the spatial random phase modulator acts as a random grating to disperse and modulate the image according to different wavelengths, generating a speckle pattern on the speckles plane. Next, the charge-coupled device (CCD) detector records the speckles by performing the speckle field data. Finally, the computer reconstructs the target image corresponding to the modulated band by the optimization algorithm.

For the GISC spectral camera, based on ghost imaging and combined with CS, the whole imaging process can be represented in the form of a matrix. Let $\boldsymbol{x} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_{N_3}] \in \mathbb{R}^{N_1 N_2 \times N_3}$ denote the object, where each column $\boldsymbol{x}_i$ is the vector form of the image of the i-th spectral band. To achieve a single detection of all spectral bands, each spectral image corresponds to a different measurement matrix $\boldsymbol{A}^i \in \mathbb{R}^{M \times N_1 N_2}$, the total measurement matrix $\boldsymbol{A} = (\boldsymbol{A}^1, \ldots, \boldsymbol{A}^i, \ldots, \boldsymbol{A}^{N_3}) \in \mathbb{R}^{M \times N}$, $N = N_1 \times N_2 \times N_3$. The whole detection process can be expressed as (1) shown at the bottom of next page: where $\boldsymbol{y} \in \mathbb{R}^{M \times 1}$ is the measurements vector, $\boldsymbol{x} \in \mathbb{R}^{N \times 1}$ is the vector form of the target multispectral image. M denotes the number of measurements, and N represents the total number of pixels of the multispectral image. The sampling rate (SR) can be defined as the ratio of the number of measurements to the total pixels of the multispectral image, i.e., SR $= M/N$. Reconstructing the 3D multispectral data cube of the target object from the 2D measurements is the process of solving $\boldsymbol{x}$ according to the

measurements and the pre-designed calibration measurement matrix.

B. CNN for Multispectral Image Reconstruction

CNN can well extract local features of images by stacking convolutional blocks. However, the regional connectivity of convolutional networks makes them lack the global receptive field. To improve the receptive field of convolutional networks, deeper network architectures are often used. DenseNet [16] enhances the efficiency of multi-scale feature representations by using convolutional layers with shortcut connections. Res2Net [17] represents multi-scale features at a granular level and increases the receptive fields for each layer by constructing hierarchical residual connections within one single residual block. SE-Net [18] uses the channel attention mechanism to aggregate global information and reallocate weights to each channel. However, these methods are all proposed for image segmentation and target detection tasks, where the network models are large and unfriendly to multispectral image reconstruction.

As a classic structure of CNN, U-Net is first proposed for medical image segmentation. Its variants [19], [20], [21] have verified the effectiveness of encoder-decoder architecture with skip-connection in image reconstruction tasks. U-Net mainly relies on convolution and pooling operations to extract image features. Due to a convolutional network's inherent receptive field defect, it is not sufficient to extract image details. Still, the encoder-decoder structure and skip connection can well fuse multi-scale features of images.

Bearing the above concerns and considering the model size, we do not use a very deep network for multispectral image reconstruction. Instead, based on U-Net, we use the Res2Net module embedded with a SE block to replace the traditional convolution operation of the U-Net, aiming to capture the local details of the desired spatial-spectral data cube.

C. Vision Transformer

At present, in computer vision tasks, such as image classification and object detection [22], [23], [24], the transformer used relies on the self-attention mechanism to achieve long-distance interaction between different elements in the sequence, which can capture long-distance correlation and non-local self-similarity of image information, becoming a current research hotspot. However, the key-value pair information in the conventional self-attention is computed in isolation. As a result, the correlation information between adjacent keys is lost in image feature learning, which is not conducive to low-level image tasks such as image reconstruction. CoT [15] further exploits the contextual information among input keys to facilitate self-attention learning, improving network representation properties. Sequentially, CoT is introduced to multispectral image reconstruction, and CCoT [11] combines convolution and the contextual transformer to extract more effective spectral features.

Inspired by CoT, in this article, considering the ability of the transformer to model long-distance information and combining the priori characteristics of multispectral images, we propose a contextual transformer module based on different spatial orientations to enhance the representation of spatial detail features of images. Embedding it into the U-Net decoder part, it can take full advantage of convolution and transformer to extract more effective spatial and spectral image features and better reconstruct the detailed information of multispectral images.

III. The Proposed Network

In this section, we introduce the proposed DL-based multispectral image reconstruction algorithm framework, CoT-Unet, which combines the advantages of convolution and transformer to reconstruct high-quality multispectral images with enjoyable model sizes. And in this part, we describe in detail the ingredients of Res2Net-SE-Conv and the two-dimensional CoT modules.

A. Overall Architecture

Inspired by the U-Net, CoT, Res2Net, and SE-Net, we propose a framework of CoT-Unet, as illustrated in Fig. 2. Firstly, we take the measurements recorded by CCD and the measurement matrix as the inputs, and then through DGI, which directly avoids the impact of the large-scale measurement matrix on the network training process. Then, the DGI results are used to train the network and further optimized to obtain the multispectral image.

To trade off the network size and reconstruction performance, CoT-Unet uses a five-layer U-Net encoder-decoder structure with the input channels of each layer C32, C64, C128, C256, and C512 (C32–C512 are half of the channels in Vanilla U-Net) to reduce the overall network parameters. We replace convolutional blocks in the first three layers of the shallow encoder-decoder network with the Res2Net-SE-Conv module (framed by the red dashed line in Fig. 2) to enhance the multi-scale feature representation of images from local and global, spatial, and spectral perspectives, respectively. Double-Conv extracts local features of the picture. Pooling and DeConv operations are used to compress and reduce feature maps. The two-dimensional CoT module (framed by the red dashed line in Fig. 2) is inserted at the end of the three decoder blocks to model the contextual information correlations of multispectral images in different spatial directions to enhance the network's contextual feature representation of the images. The model achieves the skip connection between two blocks through the internal connection between two sub-layers, which avoids the gradient disappearance during the network training.

$$\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \cdots \\ \boldsymbol{y}_M \end{pmatrix}_M = \boldsymbol{A}\boldsymbol{x} = \begin{pmatrix} \boldsymbol{A}_{11}^1 & \cdots & \boldsymbol{A}_{1(N_1,N_2)}^1 & \cdots & \cdots & \boldsymbol{A}_{11}^{N_3} & \cdots & \boldsymbol{A}_{1(N_1,N_2)}^{N_3} \\ \boldsymbol{A}_{21}^1 & \cdots & \boldsymbol{A}_{2(N_1,N_2)}^1 & \cdots & \cdots & \boldsymbol{A}_{21}^{N_3} & \cdots & \boldsymbol{A}_{2(N_1,N_2)}^{N_3} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \boldsymbol{A}_{M1}^1 & \cdots & \boldsymbol{A}_{M(N_1,N_2)}^1 & \cdots & \cdots & \boldsymbol{A}_{M1}^{N_3} & \cdots & \boldsymbol{A}_{M(N_1,N_2)}^{N_3} \end{pmatrix}_{M\times N} \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \cdots \\ \boldsymbol{x}_N \end{pmatrix}_N \tag{1}$$

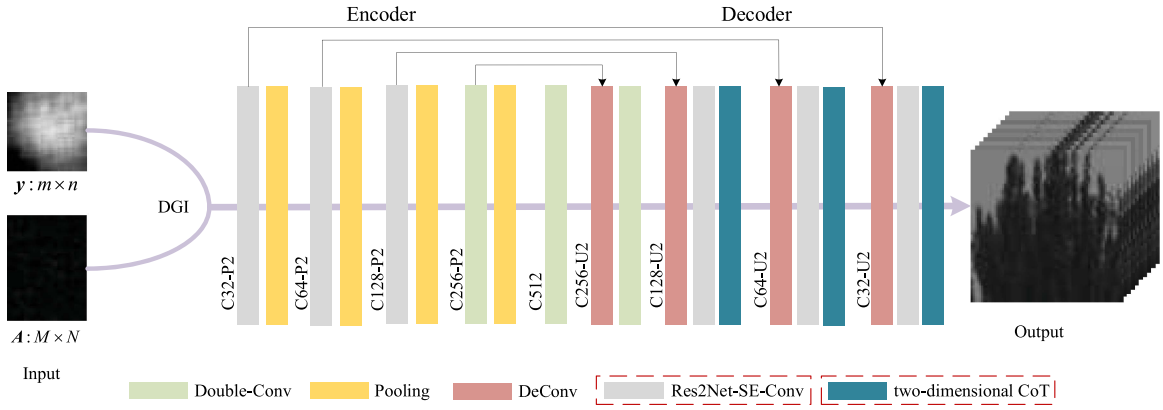


Fig. 2. CoT-Unet architecture. Each convolution layer adopts an 3×3 operator with stride one and outputs a C-channel cube. The size of pooling and upsampling is P and U. y represents the measurement with the size of $m \times n$, $\boldsymbol{A}$ means the measurement matrix with the size of $M \times N$, and $M = m \times n$.

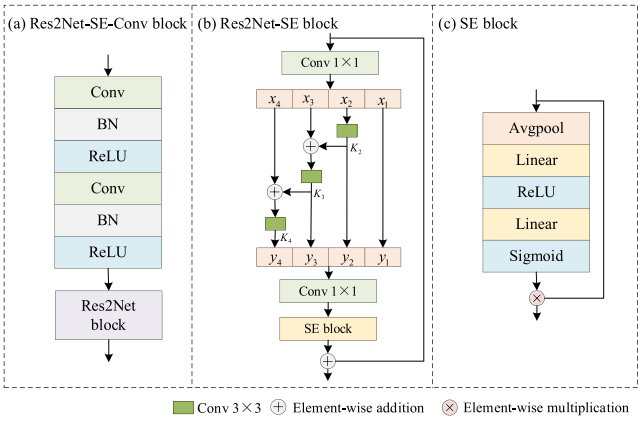


Fig. 3. (a) Res2net-SE-Conv block, (b) Res2net-SE block, (c) SE block.

Lastly, we define the loss function of the proposed model as follows:

$$Loss_{\text{MSE}} = \frac{1}{N_3}\sum_{i=1}^{N_3} \|x_i^* - x_i\|^2, \tag{2}$$

where $Loss_{\text{MSE}}$ represents the Mean Square Error (MSE) loss, x_i^* denotes the reconstructed multispectral image at the i-th spectral channel, x_i is the ground truth, and N_3 is the number of spectral bands of the multispectral image.

B. Res2Net-SE-Conv

Res2Net-SE-Conv contains two convolution blocks and a Res2Net-SE module, which first goes through two 3×3 convolutions to initially capture the shallow features of the input. Then, the Res2Net-SE module further captures the detailed characteristics of the image. As shown in Fig. 3, non-specifically labeled, all convolutional layers in the network use a convolutional kernel size of 3×3.

In the Res2Net module, after the input through 1×1 convolution, the input feature maps are divided into uniform subsets by channel. Compared with the input, each feature subset has the same space size, but the number of channels is $1/s$. Each feature subset is subject to 3×3 convolution processing except for the first feature subset. Each 3×3 convolution layer receives the information of all the previous feature map subsets due to the concatenation operation between subsets so that a larger receptive field can be obtained. The Res2Net module enables the network to represent multi-scale features at a finer granularity level through input feature splitting, multi-scale convolution, feature fusion, and other operations. As a result, it can more effectively process the features.

Res2Net modeling focuses on spatial image information and does not effectively use the image's channel features. However, exploiting the inter-spectral correlations of multispectral images facilitates enhancing the network's sensitivity to captured features. Since the squeeze and excitation (SE) module in SE-Net can adaptively recalibrate the feature response in the channel direction by explicitly modeling the interdependencies between channels. Therefore, the proposed network uses the Res2Net-SE module for feature mapping. The structure of the Res2Net-SE module is shown in Fig. 3(b), which adds the SE module to the Res2Net module.

The SE module first uses the global average pooling to squeeze global spatial information into a channel descriptor to achieve the aggregation of spatial information:

$$z(c) = \frac{1}{h \times w}\sum_{k=1}^{h}\sum_{j=1}^{w} u(c,k,j), \tag{3}$$

where $z(c)$ is the global average pooling result of channel c, $u(c,k,j)$ is the value of the feature map of channel c at space (k,j), and C is the spatial dimension of the feature map. $h \times w$ is the spatial size of the feature map, then we use ReLU and Sigmoid to capture channel-wise dependencies fully, that is:

$$s = \sigma(\boldsymbol{W}_2\delta(\boldsymbol{W}_1 z)), \tag{4}$$

where δ is the ReLU function, σ is the Sigmoid activation function, $\boldsymbol{W}_1 \in \mathbb{R}^{c/r \times c}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{c \times c/r}$ are the linear mapping function, and r is the compression ratio.

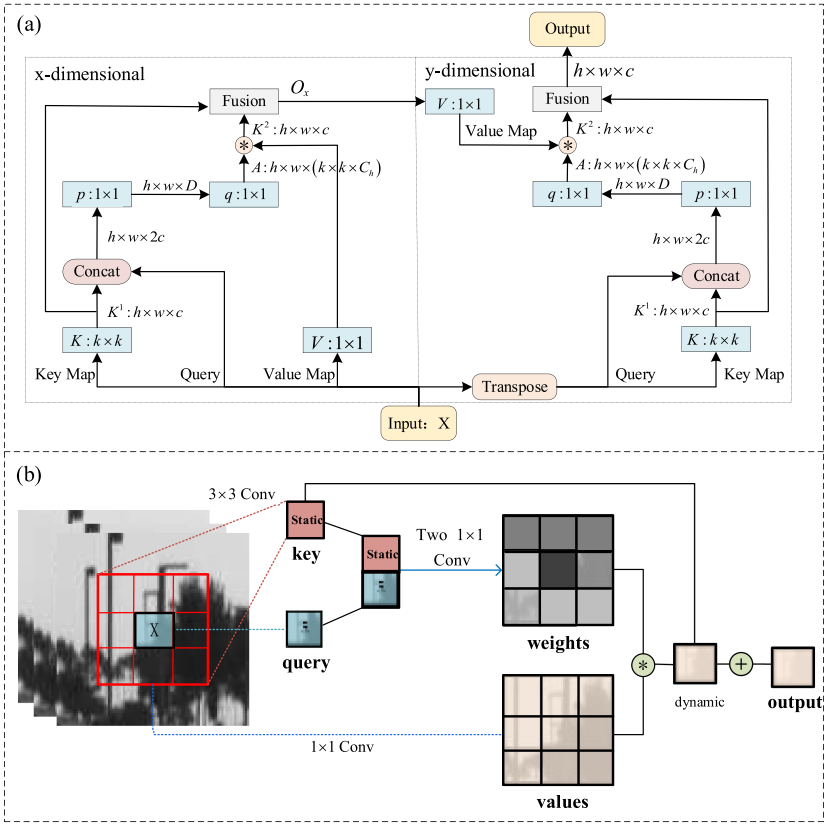


Fig. 4. (a) The detailed structures of the two-dimensional contextual transformer block, which involves the modeling for *x*-dimensional and *y*-dimensional separately and aggregation in an order-independent manner: the input is mapped to Q and K for each dimension: the size of the kernel is consistent, (b) the architecture of the CoT block [15].

Therefore, the Res2Net-SE-Conv module can capture local and global features at a more fine-grained level. The residual connectivity helps enhance contextual information and has multi-channel adaptive adjustment capability. In multispectral image reconstruction, different regions of each spectral image of the object and between each spectral segment are highly correlated. Embedding the Res2Net-SE-Conv module in the U-Net can effectively represent the multi-scale features of multispectral images and further improve the spatial resolution of the reconstructed images under the low sampling rate detection conditions.

C. Two-Dimensional CoT

CoT goes beyond the conventional self-attention mechanism by exploiting the contextual information among input keys to facilitate self-attention learning and strengthening the representative capacity of the output aggregated feature map. However, a single spectral image has two spatial dimensions, *x* and *y*, and the input keys of CoT vary with the different spatial-dimensional image features. Thus, we construct the two-dimensional CoT to fully utilize the image information with a negligible increase in model complexity.

The structure of the two-dimensional CoT is shown in Fig. 4(a). Suppose we have the input $X \in \mathbb{R}^{h\times w\times c}$. $K = X$, $Q = X$, and $V = XW_v$ denote the keys, queries, and values, respectively. The module first contextualizes each key representation by employing $k \times k$ group convolution over all neighboring keys within the $k \times k$ grid spatially to obtain the static context representation $K^1 \in h \times w \times c$ of the input X. Next, we concatenate the contextual information K^1 and queries Q and then pass two successive 1×1 convolutions to get the attention matrix $\boldsymbol{A} \in \mathbb{R}^{h\times w\times k\times k\times c_h}$:

$$\boldsymbol{A} = \left[K^1, Q\right] W_p W_q, \tag{5}$$

where c_h is the head number. In each head of the multi-head attention, each feature $\boldsymbol{A}^{(i)}$ at the i-th spatial location of $\boldsymbol{A}$ is a $k \times k \times c_h$-dimensional vector comprising c_h local query-contextualized key relation maps (size: 3×3) for all heads. Thus, the connection among various parts is strengthened through the guidance of context modeling, and the ability of self-attention learning is enhanced. Subsequently, the dynamic context $K^2 \in h \times w \times c$ is calculated by aggregating the attention matrix $\boldsymbol{A}$ with all the values V:

$$K^2 = V \otimes \boldsymbol{A}. \tag{6}$$

Then, the static and dynamic contexts are fused through the attention mechanism [24] as the output of the contextual transformer. Fig. 4(b) shows the contextual transformer module's architecture.

Finally, we calculate the values of y-dimensional image features according to the x-dimensional contextual features O_x to realize feature reuse. The output of the two-dimensional CoT is obtained:

$$O_{output} = V_y \otimes \boldsymbol{A}_y \oplus K_y^1. \tag{7}$$

In summary, based on learning contextual features in spatial x-dimensional, repeatedly mining contextual image information in y-dimensional and fusing it with x-dimensional image features can enhance the representation of detailed features of multispectral images. The two-dimensional CoT perceives long-distance information simultaneously, using the contextual information between neighboring keys in different spatial directions to enhance self-attention learning, which helps to obtain multi-dimensional information of multispectral images and improve the effectiveness of network training.

IV. Experiments and Analysis

In this section, under the same experimental configurations, we compare the performance of the proposed CoT-Unet with the traditional methods like DGI and CS and the DL-based methods like TSA-Net based on spatial-spectral self-attention and U-Net on the same datasets. The image reconstruction quality of CoT-Unet with the competitive methods is compared when SRs are 10%, 20%, and 30%. To quantitatively evaluate the algorithm's effectiveness in GISC spectral image reconstruction, the experiments and data models are based on GISC spectral camera [1].

A. Datasets

The experiments were conducted on the ICVL hyperspectral image datasets [25], including 201 natural scene images of $1390\times1300\times31$. The spectral bands of the ICVL datasets are ranged from 400 nm to 700 nm with 10 nm intervals. First, according to the imaging system, we chose ten channels with a spectral range from 610 nm to 700 nm in those datasets; then, the image data were extracted to obtain 201 copies of 10 spectral channels. To ensure the validity of the experimental data, we randomly select 199 scenes for training, one scene for validation, and one scene for testing. By enhancing the data for each scene, we obtained 5000 copies of image data of size $145\times145\times10$ as the ground truth.

B. Implementation Details and Comparison Metrics

First, according to the GISC spectral camera, the image with the size of $145\times145\times10$ is measured. We set the SRs as 10%, 20%, and 30%, respectively, and the corresponding 2D measurements are obtained according to these different SRs. Then, the 2D measurements and the system calibration measurement matrix are used as the input of the proposed model, the DGI result ($145\times145\times10$) is used as the training data, and the corresponding original image is used as the label to train the CoT-Unet model. The hyperparameters in the training process were consistent. The learning rate is 0.0004, the batch size is 10, and the network is trained for 200 epochs. The hardware device used for model training and experiments is NVIDIA Quadro RTX 6000 GPU with 24 GB video memory. In the training software environment, the Python version is 3.8.0, and the Pytorch version is 1.11.0.

The peak signal-to-noise ratio (PSNR), structured similarity index metrics (SSIM) [26], and spectral angle mapping (SAM) [27] were used to evaluate the image reconstruction effect. Among them, PSNR and SSIM are commonly used metrics to assess the image quality, and the higher the value, the better the quality. SAM measures the spectral similarity by considering the spectrum of each image element as a high-dimensional vector and calculates the angle between the reconstructed image and the high-dimensional vector formed by the image element at the corresponding position of the target image. The smaller the angle, the more similar the two spectra are, i.e., the closer the reconstructed image is to the target image.

C. Comparison Experiments

When SR = 20%, we compared our method with several competitive methods (DGI, CS, U-Net, and TSA-Net) on the same testing dataset. Table I shows the average PSNR, SSIM, and SAM of the reconstructed images of the seven representative test scenes in the test dataset by different algorithms. The bolded part is the optimal result in the comparison algorithm. We can see that the average PSNR of the reconstructed image by CoT-Unet is improved by 5 dB (22%), SSIM is improved by 0.232 (34%), and SAM is decreased by 0.066 (1%) compared with TSA-Net. Compared with U-Net, the average PSNR is improved by 3 dB (13%), SSIM is improved by 0.075 (8%), and SAM is decreased by 0.589 (11%). In addition, compared with the conventional algorithms DGI and CS, the average PSNR of the reconstructed images by the CoT-Unet is improved by more than 9 dB (40%), SSIM is enhanced by more than 0.3 (38%), and SAM has a significant decrease, indicating that the CoT-Unet has powerful feature representation capability in GISC multispectral image reconstruction. The standard deviation shows that due to different scenes, there are differences between various reconstructed image indicators, but overall, the reconstructed image quality of our method is the best among the comparison algorithms.

Meanwhile, we compare the parameters and FLOPs of the DL-based models in the compared algorithms, as shown in Table II, where the FLOPs are calculated with a uniform input (with the size of $145\times145\times10$, batch size = 1). We can see that the parameters in the CoT-Unet are reduced by 74% and 81% compared to U-Net and TSA-Net, respectively, and the FLOPs are reduced by 70% and 84%, respectively, which indicates that our CoT-Unet has lower model complexity while ensuring the optimal quality of the reconstructed images.

Fig. 5 shows part of the visualization results of Scene 1 using several multispectral image reconstruction algorithms and displays the zoom-in patches of the selected small white boxes in the subfigure. Enlarging the local area, we can see that our CoT-Unet can recover more edge details than other algorithms. Fig. 6 shows spectral density distribution curves at specified fields and the spectral correlation values between the ground

TABLE I
SR = 20%, THE PSNR/DB, SSIM, AND SAM/° BY DIFFERENT ALGORITHMS ON SEVEN SCENES

Algorithms	Metrics	Scene1	Scene2	Scene3	Scene4	Scene5	Scene6	Scene7	Average	Standard Deviation
DGI	PSNR	10.59	13.78	10.51	13.6	11.27	10.68	10.41	11.54	1.37
	SSIM	0.246	0.370	0.321	0.288	0.283	0.220	0.282	0.287	0.04
	SAM	21.493	10.631	7.331	19.147	20.011	21.614	20.688	17.273	5.37
CS	PSNR	15.49	15.75	6.45	15.48	15.98	15.04	15.78	14.28	3.20
	SSIM	0.161	0.179	0.13	0.216	0.192	0.16	0.197	0.176	0.02
	SAM	25.548	18.214	17.821	23.541	22.894	24.922	24.882	22.546	2.98
U-Net	PSNR	23.38	15.78	17.82	23.08	18.67	24.10	20.80	20.52	2.94
	SSIM	0.759	0.570	0.675	0.560	0.535	0.650	0.574	0.618	0.07
	SAM	4.887	3.459	5.041	7.077	4.256	4.057	6.631	5.058	1.24
TSA-Net	PSNR	18.52	15.90	11.74	19.30	16.41	27.27	20.04	18.45	4.41
	SSIM	0.471	0.442	0.373	0.478	0.338	0.570	0.431	0.443	0.06
	SAM	4.111	3.408	4.819	5.719	4.159	3.921	5.609	4.535	0.81
CoT-Unet	PSNR	25.31	23.10	20.98	18.89	20.54	31.88	25.92	**23.80**	4.05
	SSIM	0.828	0.492	0.725	0.544	0.662	0.780	0.697	**0.675**	0.11
	SAM	4.334	3.957	4.159	7.035	3.297	3.364	5.134	**4.469**	1.19

TABLE II
THE PARAMETERS AND FLOPS OF THE MODULE IN ALGORITHMS BASED ON DEEP LEARNING

Algorithms	Parameter/M	FLOPs/G
U-Net	31.04	17.45
TSA-Net	42.23	33.07
CoT-Unet	7.94	5.12

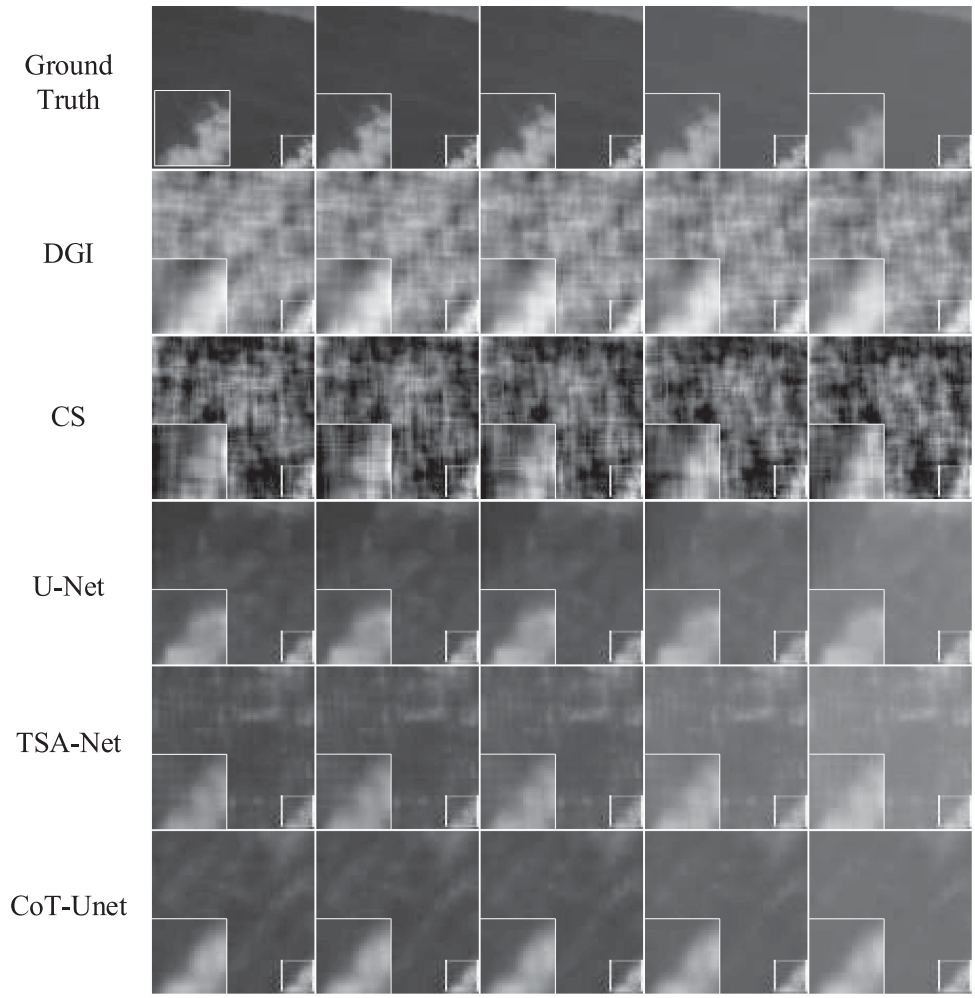


Fig. 5. The reconstructed images by five algorithms for scene 1. 5 (610, 630, 650, 670, and 690 nm) out of 10 spectral channels are shown to be compared with the ground truth.

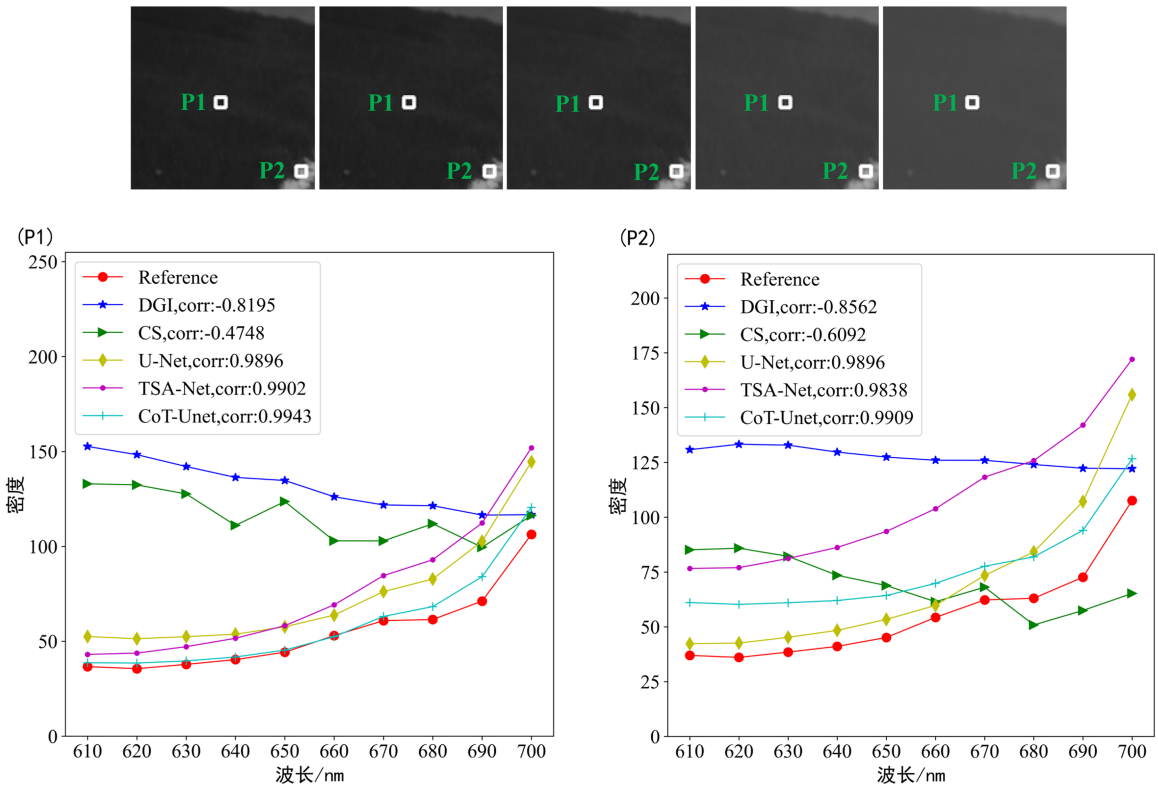


Fig. 6. The spectral curves and correlation values of Scene 1 are at the bottom of Fig. 6. Pictures P1 and P2 show the spectral curves, that is, the average density of the area plotted in scene 1 at different wavelengths. The small white boxes P1 and P2 of scene 1 are shown at the top of Fig. 6, from left to right, the visualized images corresponding to 5 (610, 630, 650, 670, and 690 nm) out of 10 spectral channels. The spectral correlation values (corr) in the symbol box calculate the spectral correlation between the ground truth and the reconstructed images.

truth and the reconstructed images calculated by the Pearson Correlation Coefficient.

As shown in Fig. 5, the images reconstructed by the traditional DGI and CS algorithms have much noise and blurred image contours. The spatial resolution of the reconstructed images by the deep neural network is higher than that of the traditional algorithm, which produces spatial blur due to the extensive offset range between each scattered spot. In contrast, CoT-Unet can effectively remove some of the noise and image artifacts, and the basic contours of the image are prominent. In addition, from the spectral curves at selected regions P1 and P2 and the spectral correlation values between the reconstructed image and the reference image shown in Fig. 6, it can be seen that the spatial fidelity and spectral recovery of the reconstructed images by the conventional algorithms such as DGI and CS are much less than those by CoT-Unet, compared with the reconstructed images by the TSA-Net and U-Net algorithms, the reconstructed images by CoT-Unet have much more transparent and more complete spatial details. The results further indicate that hierarchical residual convolution, channel feature aggregation, and contextual correlation modeling have significant advantages in multispectral image feature extraction and representation.

In summary, in GISC multispectral image reconstruction, both the visualization of the reconstructed images and the comparison of the reconstructed spectral curves at typical locations show the advantages of CoT-Unet.

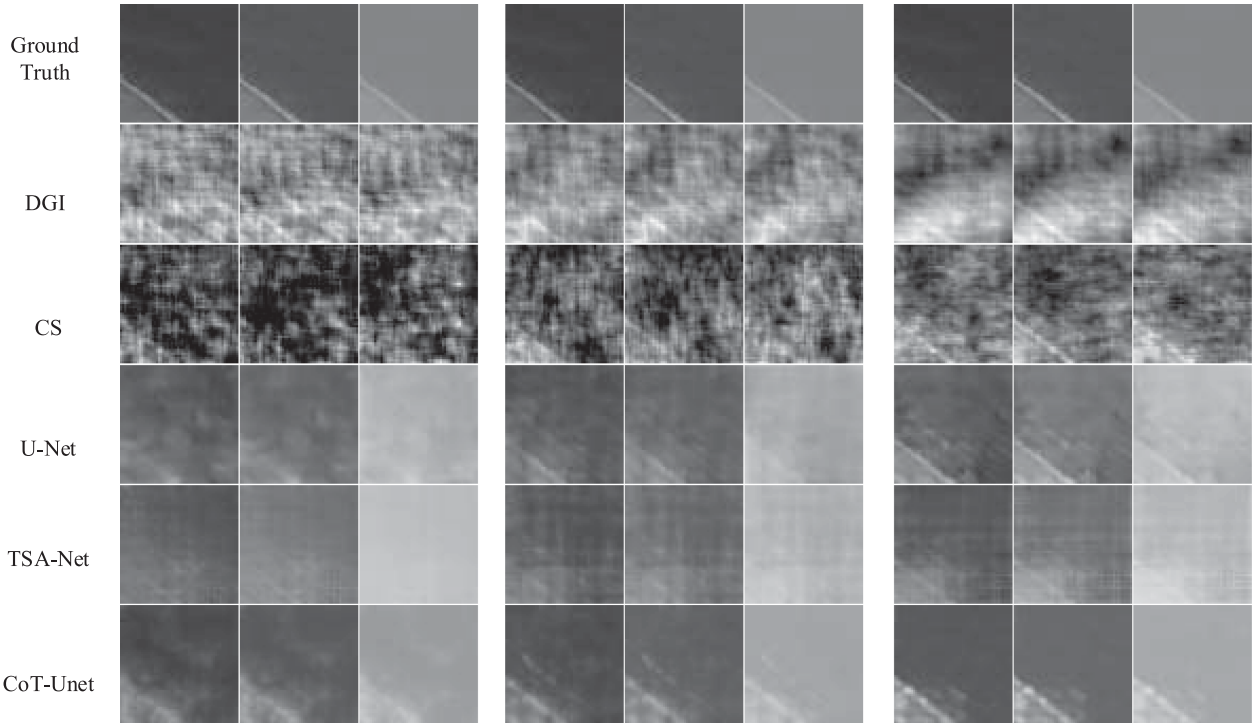
In addition, to verify the performance of CoT-Unet at low sampling rates, we compared the results of the competitive algorithms for GISC multispectral image reconstruction at different sampling rates of 10% and 30%, respectively. The average PSNR, SSIM, and SAM of the reconstructed images by each algorithm are shown in Table III, respectively, where the black bolded parts are the optimal results.

From the results in Table III, we can see that at a sampling rate of 30%, the average PSNR of the reconstruction image by CoT-Unet is improved by 9% and 6%, SSIM is enhanced by 33% and 2%, and SAM is reduced by 11% and 4% compared with TSA-Net and U-Net, at a sampling rate of 10%, the average PSNR of the reconstruction image by CoT-Unet is improved by 2.2% and 2.7%, SSIM is enhanced by 28% and 5%, and SAM is reduced by 21% and 8% compared with TSA-Net and U-Net. Compared with the reconstruction results of traditional algorithms CS and DGI. Our CoT-Unet has good reconstruction performance at a low sampling rate compared with the comparison algorithms.

The visualization comparison results of the reconstructed images of scene 5 by each algorithm at different SRs are shown in Fig. 7. Fig. 7 shows that the visualization results of the reconstructed images are consistent with the conclusions of the data obtained from Table III. CoT-Unet can see the general outline of the image details in its reconstructed images despite the sampling rate of 10%, and the image reconstruction quality is still the best among the compared algorithms.

TABLE III
THE AVERAGE PSNR/DB, SSIM, AND SAM/° OF THE RECONSTRUCTED IMAGES FOR THE SEVEN TEST SCENES, WHEN SR = 10% AND 30%, RESPECTIVELY

SR	CoT-Unet			TSA-Net			U-Net			CS			DGI		
	PSNR	SSIM	SAM	PSNR	SSIM	SAM	PSNR	SSIM	SAM	PSNR	SSIM	SAM	PSNR	SSIM	SAM
30%	**22.99**	**0.700**	**4.582**	21.04	0.526	5.20	21.70	0.683	4.784	15.35	0.268	20.919	12.28	0.386	17.402
10%	**19.57**	**0.612**	**4.690**	19.14	0.478	5.958	19.04	0.581	5.134	13.98	0.136	32.58	12.33	0.246	17.254



(a) CR=10% (b) CR=20% (c) CR=30%

Fig. 7. Comparison results of reconstructed images of scene 5 by each algorithm under different SRs. The corresponding spectral bands are 610, 660, and 700 nm from left to right in the three columns for each sampling rate: (a) the reconstruction results at the sampling rate of 10%, (b) the reconstruction results at the sampling rate of 20%, (c) the reconstruction results at the sampling rate of 30%. Please zoom in for a better view.

D. Ablation Study

Ablation experiments are performed on the proposed network. First, the two-dimensional CoT is removed on top of the CoT-Unet to examine the effect of contextual self-attention modeling on performance. Second, The Res2Net-SE-Conv is released on top of the CoT-Unet to explore the effect of multi-scale residual convolution on performance. Third, to investigate the impact of the combined effects of the Res2Net-SE-Conv and the two-dimensional CoT on network performance, we only conducted experiments based on Baseline, the model of CoT-Unet after removing the above two modules simultaneously. The results are shown in Table IV.

From the reconstruction results in Table IV, we can see that the PSNR and SSIM of the reconstructed image decrease by 3.72 dB (15%) and 0.022 (3%), respectively, when the two-dimensional CoT module is removed from CoT-Unet, which indicates that contextual self-attention can reconstruct the image well by modeling the image information correlation in the image reconstruction process. The PSNR of the reconstructed image decreases by 3.24 dB (13%), and the SSIM decreases by 0.032 (4%) when the Res2Net-SE-Conv module is removed from CoT-Unet, which indicates that the multi-scale residual convolution and SE module can reconstruct the image well by combining the prior properties of multispectral images. The PSNR decreases by 3.58 dB (15%), the SSIM is reduced by 0.06(8%), and the SAM is increased by 0.394(8%) when both the Res2Net-SE-Conv and the two-dimensional CoT module are removed from CoT-Unet, indicating that the two modules we proposed have

TABLE IV
THE AVERAGE PSNR/DB, SSIM, AND SAM/° BY DIFFERENT MODELS ON THE SAME SEVEN SCENES

Baseline	Res2Net-SE-Conv	two-dimensional CoT	PSNR	SSIM	SAM	Parameters/M	FLOPs
√			20.22	0.615	4.863	7.76	4.40
√	√		20.08	0.653	4.456	7.80	4.50
√		√	20.56	0.643	4.656	7.91	5.02
√	√	√	23.80	0.675	4.469	7.94	5.12

significant advantages for improving the quality of reconstructed images.

Regarding the issue of an increase in PSNR after removing the Res2Net-SE-Conv module from the network model based on Res2Net-SE-Conv and Baseline, on the one hand, due to the differences in scenes, the algorithm has differences in the reconstruction results of the target scene, and the PSNR here is the average value of the peak signal-to-noise ratio of the reconstructed scene images tested; on the other hand, 20.08 decreases by 0.69% compared to 20.22, while SSIM increases by 6.1% and SAM decreases by 8.3%, indicating that the Res2Net-SE-Conv module has significantly improved the overall effect of reconstructed images.

V. Conclusion

This article aims to propose an effective end-to-end model, called CoT-Unet, for the high-quality reconstruction of GISC multispectral images. In the network, the Res2Net-SE-Conv module is constructed to improve the model's ability to represent image features by learning and fusing the multi-scale elements of images through hierarchical residual connectivity and channel attention. The two-dimensional CoT module is constructed to enable the network to better understand image detail features and improve overall performance by modeling the image contextual information correlation. Experimental results show the network can reconstruct high-quality images even under low sampling rate conditions. Our end-to-end solution for GISC spectral image reconstruction may provide a reference and support for practical applications of ghost imaging. Meanwhile, due to the end-to-end nature of the network, its generalization performance is limited. The next step in the research is to improve the generalization performance of the network while maintaining low network parameters and reconstructing high-quality images under a low sampling rate to adapt to different target scenarios and enhance the practical application ability of the algorithm.

Acknowledgment

The authors wish to thank the anonymous reviewers for their valuable suggestions.

References

[1] Z. T. Liu, S. Y. Tan, J. R. Wu, E. R. Li, X. Shen, and S. S. Han, "Spectral camera based on ghost imaging via sparsity constraints," *Sci. Rep.*, vol. 6, no. 1, pp. 1–10, May 2016.

[2] R. Cerbino, L. Peverini, M. Potenza, A. Robert, P. Bosecke, and M. Giglio, "X-ray-scattering information obtained from near-field speckle," *Nature Phys.*, vol. 4, no. 3, pp. 238–243, Jan. 2008.

[3] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[4] C. Y. Chu, S. Y. Liu, Z. T. Liu, C. Y. Hu, Y. J. Zhao, and S. S. Han, "Spectral polarization camera based on ghost imaging via sparsity constraints," *Appl. Opt.*, vol. 60, no. 16, pp. 4632–4638, Jun. 2021.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Interv.*, 2015, vol. 9351, pp. 234–241.

[6] X. Miao, X. Yuan, Y. Pu, and V. Athitsos, "Lambda-net: Reconstruct hyperspectral images from a snapshot measurement," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4059–4069.

[7] Z. Meng, J. Ma, and X. Yuan, "End-to-end low cost compressive spectral imaging with spatial–spectral self-attention," in *Proc. Eur. Conf. Comput. Vis.*, 2020, vol. 12368, pp. 187–204.

[8] Y. H. Cai et al., "Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17481–17490.

[9] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 6000–6010.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[11] L. Wang, Z. Wu, Y. Zhong, and X. Yuan, "Spectral compressive imaging reconstruction using convolution and contextual transformer," *Photon. Res.*, vol. 10, no. 8, pp. 1848–1858, 2022.

[12] F. Ferri, D. Magatti, L. Lugiato, and A. Gatti, "Differential ghost imaging," *Phys. Rev. Lett.*, vol. 104, no. 25, 2010, Art. no. 253603.

[13] J. Chen et al., "Improved U-Net3+ with spatial–spectral transformer for multispectral image reconstruction," *IEEE Photon. J.*, vol. 15, no. 2, Apr. 2023, Art. no. 7800511.

[14] Z. Y. Chen et al., "Hyperspectral image reconstruction for spectral camera based on ghost imaging via sparsity constraints using v-dunet," 2022, *arXiv:2206.14199*.

[15] Y. Li, T. Yao, Y. Pan, and T. Mei, "Contextual transformer networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1489–1500, Feb. 2023.

[16] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261–2269.

[17] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.

[18] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[19] H. Wu et al., "Sub-Nyquist computational ghost imaging with deep learning," *Opt. Exp.*, vol. 28, no. 3, pp. 3846–3853, Feb. 2020.

[20] H. Wu et al., "Hybrid neural network-based adaptive computational ghost imaging," *Opt. Lasers Eng.*, vol. 140, no. 9, May 2021, Art. no. 106529.

[21] H. Chen et al., "Pre-trained image processing transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12294–12305.

[22] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.

[23] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.

[24] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 510–519.

[25] B. Arad and O. Ben-Shahar, "Sparse recovery of hyperspectral signal from natural RGB images," in *Proc. Eur. Conf. Comput. Vis.*, 2016, vol. 9911, pp. 19–34.

[26] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[27] F. A. Kruse et al., "The spectral image processing system (SIPS)—Interactive visualization and analysis of imaging spectrometer data," *Remote Sens. Environ.*, vol. 44, no. 2/3, pp. 145–163, May 1993.