IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 31 July 2023, accepted 20 August 2023, date of publication 29 August 2023, date of current version 7 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3309878

RESEARCH ARTICLE

Transformer-Based Reconstruction for Fourier Ptychographic Microscopy

LIN ZHAO[1], (Member, IEEE), XUHUI ZHOU[1], XIN LU[2], HAIPING TONG[2], AND HUI FANG[2]

[1]Hunan Institute of Science and Technology, Yueyang 414000, China
[2]Shenzhen Key Laboratory of Micro-Scale Optical Information Technology, Nanophotonics Research Center, Institute of Microscale Optoelectronics, Shenzhen University, Shenzhen 518060, China

Corresponding author: Hui Fang (fhui79@szu.edu.cn)

This work was supported in part by the Foundation of Shenzhen City through the Stable Support Plan Program under Grant 20200814164819001, in part by the National Natural Science Foundation of China under Grant 12074268, in part by the Natural Science Foundation of Hunan Province of China under Grant 2023JJ30284, and in part by the Graduate Research and Innovation Project of Hunan Province under Grant CX20221212.

ABSTRACT Fourier ptychographic microscopy (FPM) is a recently developed computational imaging technique which can perform complex amplitude imaging with both large field of view and high resolution by using a simple microscope setup. Here, we propose a transformer based neural network named as FP-transformer, which takes the low-resolution amplitude (LRA) images as the sequential input and uses self-attention mechanism to compute the relationship among them. The high-resolution FPM complex amplitude reconstruction is the end-to-end output of the FP-transformer. We apply the image library of div2k to generate the FPM LRA images with the physical model, and then perform the training and validation with this dataset containing ground truth. We also perform the validation with the experiment images and it is found that the high-quality FPM complex amplitude image pairs can be obtained. Therefore, the FP-transformer creates a new platform for the FPM deep learning reconstruction, which has the better dependability and adaptability.The code of this work will be available at https://github.com/zhaolin6/FPTransfomer for the sake of reproducibility.

INDEX TERMS FPM, transformer, attention mechanism, convolutional neural networks.

I. INTRODUCTION

Fourier ptychographic microscopy [1] (FPM) is a recently developed computational microscopic imaging technique combining synthetic aperture [2], [3] and phase retrieval [4]. It can image simultaneously by a large field of view and high-resolution that breaks the diffraction limit of the optical system, and has shown wide application prospects in biomedical diagnosis, industrial inspection and other fields [5].

The experimental system of FPM is relatively simple where only a LED array illumination is needed as an add-on in an general optical microscope. However, its high quality imaging depends on the reconstruction algorithm, which involves many important parameters (such as the position of the each LED illumination elements, the point spread function of the optical system including various possible aberrations) [6], [7]. Optimizing these parameters with high accuracy currently remains a challenge. The situation is further complicated by the fact that the LED oblique illumination in the optical system is not strictly plane wave illumination [8], and the vignetting influence is always unavoidable [9].

To address these problems, several iteration-based phase retrieval algorithms have been developed [10], [11], [12], [13], such as the Gauss-Newton method to obtain the reconstruction output [14], the simulated annealing algorithm, the LED intensity correction method [15], [16], and the introduction of the Fourier diffraction theory into FPM to realize 3D reconstruction [17].

Meanwhile, the FPM image reconstruction based on deep learning has also been developed rapidly in recent years, which can be roughly divided into two categories: one is the neural network based on the physical model, and the other is the end-to-end deep learning model. In the first category, the neural network is constructed by following the FPM forward

The associate editor coordinating the review of this manuscript and approving it for publication was Jinhua Sheng.

imaging model [18]. Its advantage is that the function of each network layer is clear and the ground truth is not needed. However, it basically needs to retrain the neural network for each image reconstruction task, being difficult to meet the requirements of real-time reconstruction. Moreover, it cannot take the advantage of the deep learning to improve the image reconstruction quality through feeding with a large amount of training data.

In the second category, the dataset with the ground truth is provided and the mapping relationship between input and output is established through supervised training. Once the training process is completed, the end-to-end reconstruction of high-resolution output can be done in real time. Initially, the convolution(CNN) taking the low resolution images as input is applied, such as Ptychnet [19]. Later on, a multi-scale fusion mechanism is added to enable the CNN to obtain receptive fields in different scales [20], [21] so that the local and the global information can be combined to improve the reconstruction accuracy. Other improvements include that: introducing the residual connections in the deep neural networks to alleviate the problem of vanishing gradients [22]; applying the generative adversarial networks [23], [24], [25], [26], [27] which demonstrated the clearer reconstruction results, although it is somewhat difficult to train; introducing the encoder-decoder structures with U-net to do denoising through its pixel-wise classification [28]; applying the conventional FPM algorithm to preprocess the LRA images for one iteration as the neural network input, which can reduce the neural network complexity although it also caused some loss of the original information.

Here, by working on the second category, we propose the transformer based deep learning model named FP-transformer which directly uses the LRA images as a sequential input. We noticed that the LRA images are highly correlated with each other and actually structured as a specific sequence based on their frequency information [29], and the transformer framework is specifically developed for sequential data processing. In our FP-transformer, the self-attention mechanism works by paying more attention to the spatial neighborhood correlation of pixels and the correlation between LRA features of different frequencies. Similar features are enhanced, and noise is suppressed through the fusion of different features based on their correlations [30]. This should surpass the recently proposed channel attention deep learning model [31], [32], [33] where more weight are attributed to the more important channels but the correlations between different features are ignored.

For the training and the validation, we used the dataset div2k [34] created for the image super-resolution task as the high-resolution ground truth, and got the LRA images by applying the FPM forward simulation model. Finally, in the form of transfer learning, we directly applied the trained FP-transformer to reconstruct the high resolution images of red blood cells from the experimentally collected LRA images, and the results are found to be comparable to those obtained from a few previous methods.

II. PRINCIPLE OF FPM

In the theoretical study of FPM imaging, the LRA image capturing process can be model by the so called forward imaging model while the high-resolution image reconstruction is fully computational depended and can be defined as the backward reconstruction model. As illustrated in Fig.1, the forward imaging model needs to consider the procedure that the sample is sequentially illuminated by each element of LED array (as shown in Fig. 1(a)), while the conventional backward reconstruction model is based on the alternative-projection algorithm running back and forth between the LRA image sequence and the whole frequency domain which contains the high-resolution image information (as shown in Fig. 1(b)). For comparison, as shown in Fig. 1(c), we proposed the deep learning backward reconstruction model based on transformer which takes the LRA image sequence as the input and produces the high-resolution image pair as the output.

To facilitate the FPM-transformer training, we applied the forward imaging model to create the LRA images from the original high resolution image pairs (one image works as the amplitude and the other works as the phase), such that the dataset with ground truth was built. Considering the LED array is far away enough from the sample such that the plane wave illumination approximation can be applied, the wave number vector can be expressed as follows:

$$k_i = (k_{xi}, k_{yi}) = \left(\frac{2\pi \cdot \sin\theta_{xi}}{\lambda}, \frac{2\pi \cdot \sin\theta_{yi}}{\lambda} \right) \tag{1}$$

where $(\theta_{xi}, \theta_{yi})$ represents the illumination angle of the i-th LED and $\boldsymbol{\lambda}$ represents the wavelength. Under each illumination, it is amount to shift the Fourier spectrum as:

$$O(k_x - k_{xi}, k_y - k_{yi}) = F\left\{ o(x, y) e^{j(k_{xi}x + k_{yi}y)} \right\} \tag{2}$$

where $F\{\cdot\}$ denotes the Fourier transform. Furthermore, because the objective lens works as a low-pass filter with its pupil function denoted as $P(\cdot)$, the captured intensity image $I_i(x, y)$ needs to be written as the following:

$$\begin{aligned} I_i(x, y) &= |g_i(x, y)|^2 \\ &= \left| F^{-1}\left\{ P\left(k_x, k_y\right) O\left(k_x - k_{xi}, k_y - k_{yi}\right) \right\} \right|^2 \end{aligned} \tag{3}$$

where $g_i(x, y)$ represents the complex amplitude image, $F^{-1}\{\cdot\}$ represents the inverse Fourier transform. In our model, we assumed the pupil function has the approximate form as:

$$\begin{aligned} P\left(k_x, k_y\right) &= CTF\left(k_x, k_y\right) \\ &= \begin{cases} 1, {k_x}^2 + {k_y}^2 \le (NA \cdot \dfrac{2\pi}{\lambda})^2 \\ 0, {k_x}^2 + {k_y}^2 > (NA \cdot \dfrac{2\pi}{\lambda})^2 \end{cases} \end{aligned} \tag{4}$$

where NA is numerical aperature.

In the conventional iteration FPM reconstruction model, as illustrated in Fig. 1(b), in each iteration the amplitude part is always replaced by the recorded low resolution FPM image while the phase part is continuously updated from the Fourier transform. Eventually, through alternative projection between

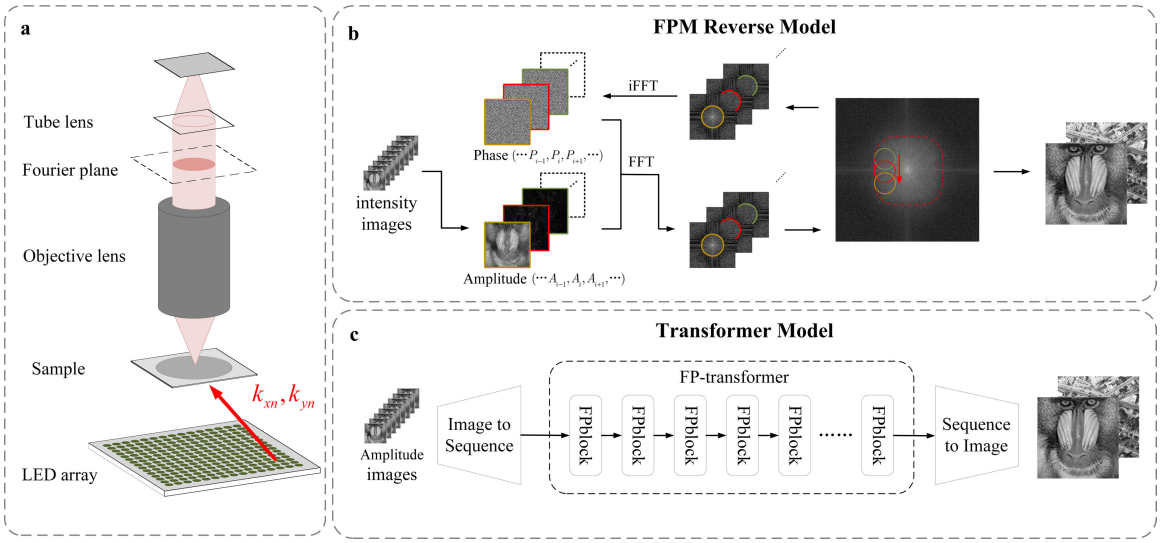


FIGURE 1. Illustration of the basic idea of using transformer model in the FPM. (a) Schematic of the FPM image capturing system. (b)Diagram to show the conventional iteration FPM reconstruction algorithm (c) Diagram to show the transformer model for the FPM image reconstruction.

the spatial and frequency domain, the iteration will reach the convergence condition and the Fourier domain information is fully recovered, from which the high resolution amplitude image as well as the phase image can be reconstructed. The above process can be express as:

$$g_i(x, y) = F^{-1}\left\{P\left(k_x, k_y\right) G_i(k_x, k_y)\right\} \tag{5}$$

$$T_i\left(f(k_x, k_y)\right) = f(k_x + k_{xi} - k_{xi-1}, y + k_{yi} - k_{yi-1}) \tag{6}$$

$$G_i(k_x, k_y) = T_i(P\left(k_x, k_y\right) F\left\{\sqrt{I_i(k_x, k_y)}\frac{g_{i-1}(x, y)}{|g_{i-1}(x, y)|}\right\} + \left(1 - P\left(k_x, k_y\right)\right) G_{i-1}(k_x, k_y)) \tag{7}$$

where, the operation T_i means to move the contained Fourier spectrum region to correct Fourier spectrum position corresponding to the i-th LED illumination.

III. FP-TRANSFORMER

Originated from Natural Language Processing(NLP) task, transformer has also been very successfully developed in Computer Vision(CV) research field due to its self-attention mechanism with long-range memory ability and excellent parallelism [35]. In CV research, the Swin-transformer [36] reduces computational costs by limiting the range of self-attention calculations to non-overlapping local windows, and then further uses feature fusion and also shifting windows for information interaction to address the limited field of view due to the local window.

By taking account that the LRA images actually form a sequence with their strong connection in frequency domain, we proposed the FP-transformer which is based on Swin-transformer framework to reconstruct the high-resolution amplitude image and phase image. As shown in Fig.2, the overall architecture of FP-Transformer includes three parts: Patch Embedding [37], UpSample Encode and Channel Fusion [38]. Given an image sequence comprising with N LRA images, denoted by $S = [I_1, I_2, \cdots, I_N]$, Patch Embedding will translate the image sequence S to the feature sequence X_1, UpSample Encode will then do encoding (by the Swin transformer blocks) and up sampling X_1, and Channel fusion module will fuse the coding results to obtain high-resolution amplitude image X_A and phase image X_P. Each module is detailed as shown in Fig.2.

A. PATCH EMBEDDING

Patch embedding is used to preprocess image sequences. It applies CNN layer to fix the dimension of the sequence S to that the transformer can handle. As well known, the transformer structure is better at capturing long-distance dependencies, but the convolution structure has stronger ability to extract shallow semantic features such as image texture [39]. Therefore essentially, we combine the CNN with Transformer to form a hybrid transformer. As the first step, the inputting image sequence S is stacked as a multi-channel feature maps $X_{Input} \in \mathbb{R}^{H\times W\times N}$ where H and W represent the spatial dimensions and N represents the channel dimension. Then, two 3×3 convolutions are applied to extract the initial feature and adjust the channel dimension of X_{Input} from N to a specific dimension N', thus creating a new feature map $X_{Input}{}' \in \mathbb{R}^{H\times W\times N'}$. Finally, the multi-channel feature map $X_{Input}{}'$ will be split into patches $X_1 \in \mathbb{R}^{\frac{H}{P}\times\frac{W}{P}\times N'}$ with each patch size of (P, P, N'). Here, the size of $X_{Input}{}'$ is $128\times128\times512$, and patch size takes $1 \times 1 \times 512$. So, X_1 contains total 128×128 patches.

B. UPSAMPLE ENCODE

UpSample Encode performs deep feature extraction and up sampling of the preprocessed feature sequence X_1, which is composed of 16 FP-blocks and 2 UpSamples. The FP-block is the feature extraction module, consisting of spitted parallel spatial and frequency self-attention (SFSA) as shown

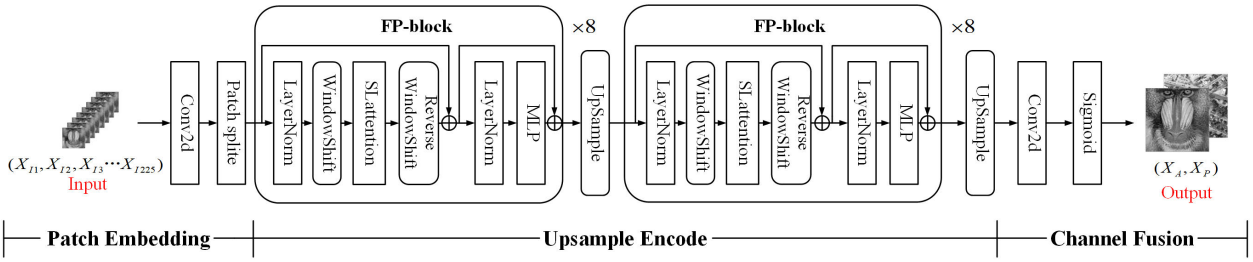


FIGURE 2. FP-transformer architecture diagram.

in Fig. 3 and a Multilayer Perceptron(MLP) feed forward network.

1) SPATIAL AND FREQUENCY SELF-ATTENTION

SFSA has structure as shown in Fig.3, consisting two parallel self-attention architects. One is the self-attention applied on the image dimensions thus defined as spatial self-attention(SSA). The other is the self-attention applied on the sequence dimension; because the sequence is tightly related in the Fourier domain, it is defined as frequency self-attention(FSA). In either SSA or FSA, the basic self-attention blocks are the same with the structure as shown in Fig. 3 (self-attention). Briefly, the self-attention calculates:

$$SSA = SoftMax\left(Q_s\left(X_n\right) * \frac{K_s{}^T\left(X_n\right)}{\sqrt{d_s}}\right) V_s\left(X_n\right) \tag{8}$$

$$FSA = SoftMax\left(Q_f\left(X_n{}^T\right) * \frac{K_f{}^T\left(X_n{}^T\right)}{\sqrt{d_f}}\right) V_f\left(X_n{}^T\right) \tag{9}$$

where in (8) Q_s, K_s, V_s represents the Query, Key and Value, respectively, and d_s represents the value of the channel dimension N', which is equal to 512; in (9) Q_f, K_f, V_f also represents the Query, Key and Value, respectively, but d_f represents the number of patches in the window.

The architect of SSA is exactly similar to the shifted window based self attention (W-MSA) in Swin transformer. By restricting the SSA calculation in local windows, the computational complexity is greatly reduced. For the SSA computation, the feature map X_n is evenly partitioned into several window blocks of size 16×16 in the spatial dimension, and each window block is flattened into a Transformer sequence of shape $256 \times N'$, which 256 represents the sequence length and N' represents the sequence feature dimension.

For the FSA computation, we first split the channel dimensions of feature map X_n into several local channel blocks of size $16 \times 16 \times 16$, each block is further split by channel into 16 channel patches of size $16 \times 16 \times 1$.

Finally, the SFSA output is obtained by not only combining the SSA and the FSA outputs and but also adding the residual connection directly from the input X_n, as:

$$SFSA(X_n) = SSA(X_n) + FSA(X_n) \tag{10}$$

2) WINDOWSHIFT

WindowShift is a technique invented in Swin-transformer which can efficiently reallocate patches of adjacent windows by introducing cross window information interaction. As illustrated in Fig. 4, in our design, the feature map is cyclically moved towards the upper left corner by half of the window size, and the moved feature map is reassigned to the divided windows to get the shifted output. The consecutive FP blocks are computed as:

$$\begin{aligned} X_{n-1}{}' &= SFSA(LN(X_{n-1})) + X_{n-1} \\ X_n &= MLP(LN(X_{n-1}{}')) + X_{n-1}{}' \\ X_n{}' &= WS^{-1}(SFSA(WS(LN(X_n)))) + X_n \\ X_{n+1} &= MLP(LN(X_n{}')) + X_n{}' \end{aligned} \tag{11}$$

where $WS(\cdot)$ indicates the window shift operation, $LN(\cdot)$ indicates the layer norm operation, and $WS^{-1}(\cdot)$ indicates the reverse window shift operation.

3) MLP

MLP is functioned to extract the global features from the sequence dimension. We first applied the Layer norm along the sequence dimensions of the input $X_n{}'$, and then calculated the output X_{n+1} by a Multilayer Perceptron as:

$$X_{n+1} = MLP(LN(X_n{}')) + X_n{}' \tag{12}$$

$$MLP(x) = Linear(GELU(Linear(x))) \tag{13}$$

where the MLP operation actually includes two Linear layers and one GELU hidden layer, Which $Linear(\cdot)$ represents the linear transformation of the input, $LN(\cdot)$ represents the layer normal operation, and $X_n{}'$ represents the n-th layer SFSA output.

4) UPSAMPLE

UpSample module performs the up-sampling of the input feature map. It is different from the Swin Transformer's Patch Merging: our UpSample model needs to expand the spatial dimension while compressing the channel dimension. It consists of a 3×3 convolutional 2d(Conv2d) layer and a pixel shuffle layer [40]. After the up-sampling, the features size is upgraded from $H \times W \times C$ in the input to $2H \times 2W \times 1/4C$ in the output.

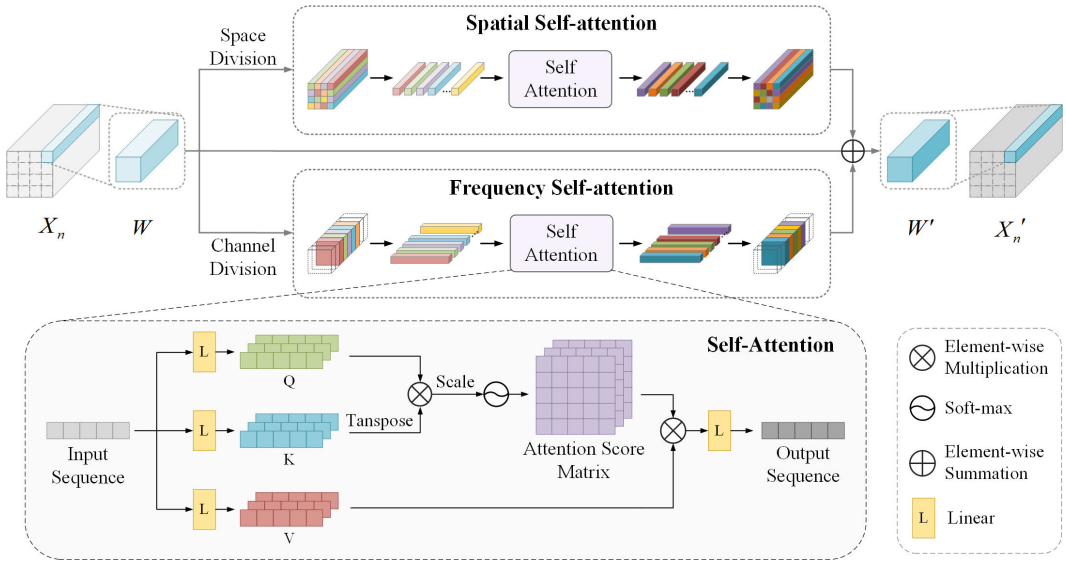


FIGURE 3. Schematic diagram of the SFSA.

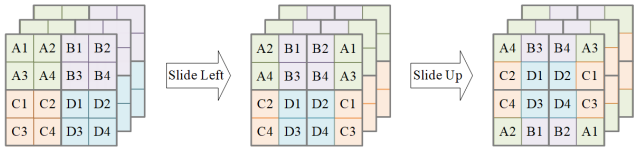


FIGURE 4. Window shift diagram.

C. CHANNEL FUSION

Channel Fusion performs the 3×3 Conv2d to the output of the UpSample Encode module, and then obtains high-resolution phase image X_P and amplitude image X_A through a sigmoid layer.

D. LOSS FUNCTION

Loss function of our FP-transformer consists of three parts: one L1-loss [41], one SSIM-loss [42] which are both performed on X_P and X_A output by comparing with the ground truth, another L1-loss which is performed on the Fourier spectrum calculated from X_P and X_A by also comparing with that calculated from the ground truth. The SSIM-loss evaluates the similarity between two images on the brightness, contrast and structure. The specific calculation formulas are:

$$\begin{aligned} loss = {} & w_1(L_1(X_A, Y_A) + L_1(X_P, Y_P)) \\ & + w_2(L_{SSIM}(X_A, Y_A) + L_{SSIM}(X_P, Y_P)) \\ & + w_3 L_1(F(X_A, X_P), F(Y_A, Y_P)) \end{aligned} \tag{14}$$

$$L_1(x, y) = \frac{1}{n}\sum_{i=1}^{n} |y_i - f(x_i)| \tag{15}$$

$$L_{SSIM}(x, y) = 1 - \frac{(2\mu_x\mu_y + c_1)(\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{16}$$

where $F(\cdot, \cdot)$ represents the Fourier transformation, μ_x, μ_y respectively represents the average value of x and y, σ_x, σ_y respectively represents the standard deviation of x and y, σ_{xy} represents the covariance of x and y, c_1, c_2 is a constant. Here, w_1, w_2, w_3 are hyperparameter, which is set as 1,1,0.05, respectively.

IV. EXPERIMENT RESULTS

We first trained and validated the FP-transformer with the simulation dataset which contains ground truth, and then applied the trained network directly on the experimental data where the ground truth results are not available (In the sense that the transfer learning approach based on data domain transferring is applied here). The reconstruction by the FP-Transformer is compared with the conventional alternative-projection reconstruction as well as the well-known two previous networks PtychNet and the PFM with Unet (The code used here is reproduced based on their research paper).

A. IMPLEMENTATION DETAILS

The network training was run on a NVIDIA GTX 2080Ti. Total 200 epoch training was taken. During the training process, we used an initial learning rate of 2×10^{-4}, and then reduced it by 25% for every 25 epochs. The optimizer we used is adaptive moment estimation (Adam). We used a batch size of 1 due to memory constraints. The total training time is about 7 hours.

B. DATASET CONSTRUCTION

The training data set consisted of 200 images randomly extracted from the super-resolution dataset div2k as the amplitude and phase high-resolution ground truth images, from which the LRA images are generated by the FPM forward simulation algorithm. For the simulation, the following parameters are set according to the experiment setup: the objective NA is 0.16, the optical magnification is 4, the optical wavelength is 0.625μm, and the number of LEDs is 15×15. The distance between adjacent LEDs is 2mm, the

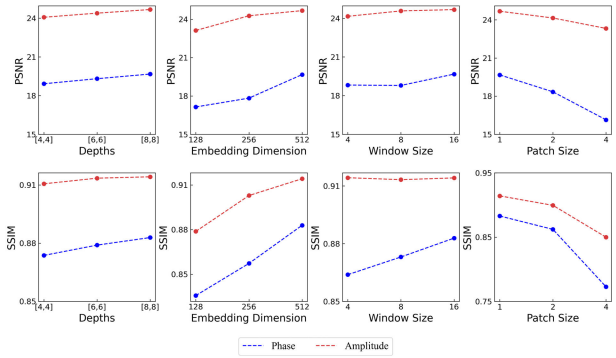


FIGURE 5. Visual comparison of the impact of different hyperparameters on reconstruction accuracy.

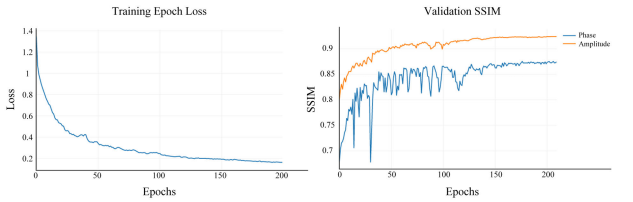


FIGURE 6. Training loss and validation accuracy variation chart.

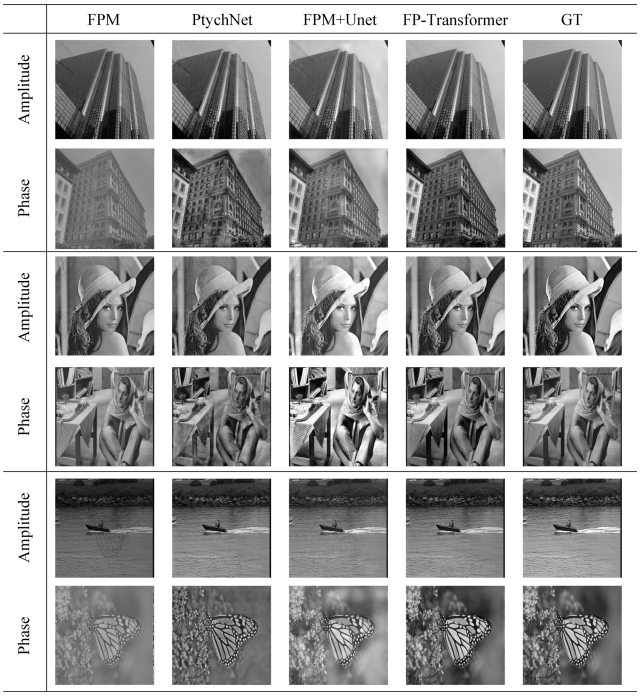


FIGURE 7. Visual comparison of partial results of test dataset.

imaging pixel size of the CCD camera is 4.4μm, and the distance between the sample and the LED array is 60 mm. The 225 LRA images have the same size of 128×128 pixels. The ground truth amplitude and phase image pairs have a size of 512×512 pixels.

For the test, the amplitude and phase ground truth images are taken from Set5, Set14, Urban100, Mange109, B100 super-resolution test sets, all of which are test datasets from div2K (The 225 LRA images are still obtained from the FPM forward simulation algorithm).

As evaluation metrics, we calculated both peak signal-to-noise ratio (PSNR) and structure similarity index measure (SSIM), which are both commonly used in image recovering and super-resolution research. PSNR is the logarithmic ratio of the mean squared error between the original image and processed image, divided by $(2^n - 1)^2$, where n represents the number of bits per sample value.

We also evaluated the reconstruction on the experimental data which includes the 225 low resolution images directly captured, while the ground truth amplitude and phase images are not available.

C. TRAINING DETAILS

In order to select the most appropriate hyperparameters, we compared the impact of different hyperparameters on the model reconstruction accuracy in Fig.5, where the baseline settings are window size of 16, embedding dimension of 512, patch size of 1, and depth of [8,8], where the depth represents the number of FP-blocks for two different stages in the Upsample Encode.

To prevent overfitting of the model, we employed an early stop strategy. Specifically, we took out 10% of the training set as a validation set, and selected the best of epoch according to the performance of the validation set. Fig.6 shows training loss and validation accuracy plotted against epochs.

D. RECONSTRUCTION ON SIMULATED DATA

We compared the reconstruction results of FP-transformer with the traditional FPM algorithm, PtychNet and FPM with Unet, where PtychNet is a typical CNN reconstruction network, and FPM with Unet is a reconstruction method that uses the results of one iteration of FPM as preprocessing and then uses Unet to optimize the output. We calculated the PSNR and SSIM in the amplitude and phase of the reconstruction of different test sets. Table.1 shows that the PSNR and SSIM of FP-transformer are better than those of traditional FPM, PtychNet and FPM with Unet in reconstruction of amplitude and phase.

Fig.7 displays the visualization effect of the four reconstruction of the test dataset. It can be seen that the reconstruction results of FP-transformer method show more details and are closer to the ground truth than the other methods.

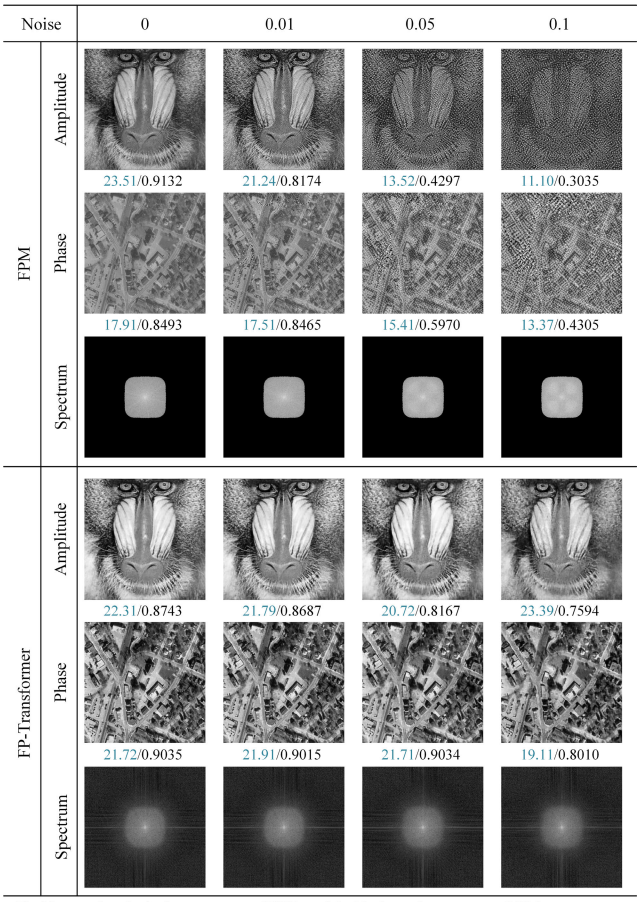
E. RECONSTRUCTION ON NOISY DATA

Considering the susceptibility of low-resolution images to noise interference during their acquisition in practical applications, we introduced random noise as data augmentation during model training to enhance the model's ability to handle noise and prevent overfitting. We compared the reconstruction performance of FP-Transformer and FPM on low-resolution images with Gaussian noise added, with mean 0 and variances of 0.01, 0.05, and 0.1 respectively. The results are shown in Fig.8. The experimental results show that our

TABLE 1. Different methods reconstruct results for different datasets.

Method		Set5		Set14		Urban100		B100	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
FPM	A	**27.37**	**0.9563**	20.68	0.9082	19.96	0.8810	**25.43**	**0.9460**
	P	14.61	0.8132	17.38	0.8661	14.70	0.7692	16.55	0.8399
PtychNet	A	24.24	0.9133	23.61	0.8991	22.91	0.8517	21.77	0.8858
	P	14.73	0.7646	17.33	0.8379	16.02	0.7873	16.53	0.8130
FPM + Unet	A	22.11	0.9059	24.38	0.9195	23.61	0.9046	24.13	0.9360
	P	18.12	0.8561	16.88	0.8913	17.29	0.8536	17.13	0.8440
FP-Transformer	A	25.42	0.9523	**25.08**	**0.9356**	**24.82**	**0.9156**	23.32	0.9217
	P	**19.98**	**0.8873**	**21.15**	**0.9049**	**19.70**	**0.8821**	**21.56**	**0.9056**

A stands for amplitude images and P for phase images, and the numbers marked in bold are the best measurements in each group



The blue numbers in the image represent PSNR, and the black numbers represent SSIM.

FIGURE 8. Comparison of reconstruction results under different noise intensities.

model are robust against noise, and can effectively reconstruct the spectrum.

F. RECONSTRUCTION ON REAL DATA

In order to verify the effectiveness of FP-transformer on real data, the red blood cell data collected by the experimental equipment is feed into the network. The reconstruction results of FP-transformer are displayed in Fig.9. FP-transformer shows a reasonable high-resolution reconstruction. FP-transformer reaches the similar results in amplitude recovery. In phase recovery, it has clearer cell boundaries and faster speed compared to traditional FPM algorithm and

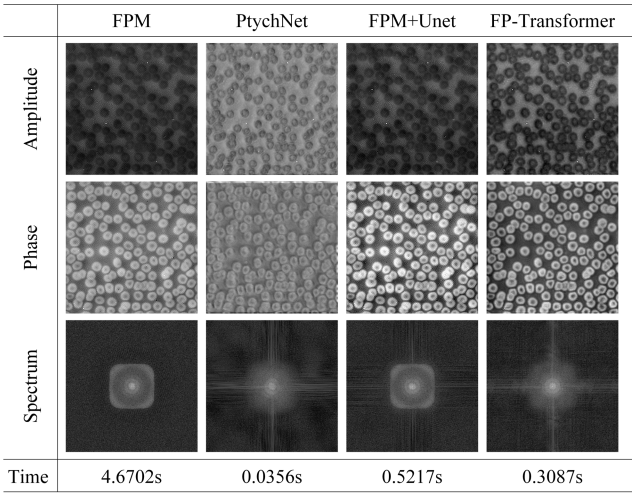


FIGURE 9. Comparison of reconstruction results of red blood cell experimental dataset.

the FPM with Unet, and shows more high value area than PtychNet. In general, FP-transformer has better visual effects.

G. ABLATION STUDIES

To further verify the necessity of several important modules in our FP-transformer network design, we performed the ablation studies by detecting the changes of PSNR and SSIM before and after eliminating each of these modules. Specifically, we studied the influence of convolution layer in patch embedding and the influence of SSA and FSA in SFSA. The ablation study procedure and the corresponding results are listed in Table 2.

The results demonstrated that both SSA and FSA improve the reconstruction with some definite degrees, especially for the phase image reconstruction. SFSA can fuse the pixels in the spatial dimension according to the attention scores in the frequency domain, and fuses the local frequency features in the channel dimension according to the attention scores in the spatial domain, so that the network extracts information in both spatial and frequency domain which is more in line with the physical principle of FPM image processing. As shown in Fig.10, the amplitude information and phase information in the sequence features exhibit different distributions by SFSA. Especially, the higher values corresponding to the red region in the feature map are associated with the amplitude informa-

TABLE 2. The ablation result of SFSA and convolution layer.

Conv2D in Patch embedding	Spatial Self-attention	Frequency Self-attention	A PSNR	A SSIM	P PSNR	P SSIM
✓	✓	✓	19.70	0.8821	24.82	0.9156
✓	✓	✗	18.45	0.8722	24.29	0.9116
✓	✗	✓	18.64	0.8689	24.17	0.9077
✓	✗	✗	17.17	0.8361	24.01	0.8932
✗	✓	✓	19.05	0.8612	24.01	0.8931

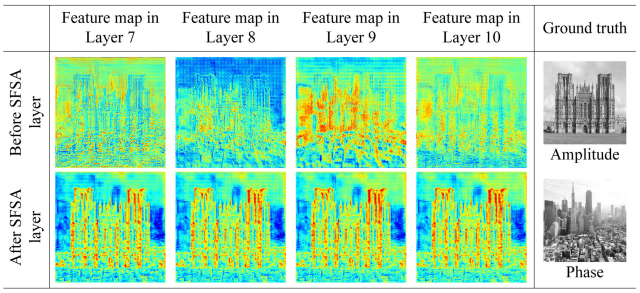


FIGURE 10. Comparison of input and output feature maps of SASF layer.

tion; while the lower values corresponding to the blue region are associated with the phase information. Consequently, SFSA effectively separates and denoises the amplitude and phase information, thereby improving the precision of super-resolution.

V. CONCLUSION

In this paper, we have detailed the architect of our proposed FP-transformer neural network designed for FPM image reconstruction. The main advantage of FP-transformer is that the low-resolution FPM amplitude images can be fed naturally as the input sequence for the Transformer. FP-transformer effectively captures long-range dependencies among low-resolution images. The SFSA module separates and reconstructs the amplitude and phase information by fully exploiting the spatial association as well as the correlation of different frequencies. This improves the accuracy and robustness of the model, enabling it to perform well in image reconstruction even in the presence of noise. We applied strategy of training and validating the FP-transformer with the simulation dataset created from the FPM forward modeling, while realizing the reconstruction for the experiment images by transfer learning. Both of the simulation validation and the experiment realization showed the results better than a few other reconstruction methods. And the ablation experiment demonstrated that the importance of the SFSA structure which can take the spatial relation and frequency relation in the equal footing in the self-attention process of FP-transformer. In the future works of FP-transformer, we will further explore the possibility to correct the misalignment of LED elements and to recover the real CTF with the optical aberration which are both important in the FPM study.

ACKNOWLEDGMENT

(Lin Zhao and Xuhui Zhou contributed equally to this work.)

REFERENCES

[1] G. Zheng, R. Horstmeyer, and C. Yang, "Wide-field, high-resolution Fourier ptychographic microscopy," *Nature Photon.*, vol. 7, no. 9, pp. 739–745, Sep. 2013.

[2] W. Luo, A. Greenbaum, Y. Zhang, and A. Ozcan, "Synthetic aperture-based on-chip microscopy," *Light, Sci. Appl.*, vol. 4, no. 3, p. e261, Mar. 2015.

[3] T. R. Hillman, T. Gutzler, S. A. Alexandrov, and D. D. Sampson, "High-resolution, wide-field object reconstruction with synthetic aperture Fourier holographic optical microscopy," *Opt. Exp.*, vol. 17, no. 10, pp. 7873–7892, 2009.

[4] J. M. Rodenburg and H. M. L. Faulkner, "A phase retrieval algorithm for shifting illumination," *Appl. Phys. Lett.*, vol. 85, no. 20, pp. 4795–4797, Nov. 2004.

[5] G. Zheng, C. Shen, S. Jiang, P. Song, and C. Yang, "Concept, implementations and applications of Fourier ptychography," *Nature Rev. Phys.*, vol. 3, no. 3, pp. 207–223, Feb. 2021.

[6] R. Eckert, Z. F. Phillips, and L. Waller, "Efficient illumination angle self-calibration in Fourier ptychography," *Appl. Opt.*, vol. 57, no. 19, pp. 5434–5442, 2018.

[7] P. Song, S. Jiang, H. Zhang, X. Huang, Y. Zhang, and G. Zheng, "Full-field Fourier ptychography (FFP): Spatially varying pupil modeling and its application for rapid field-dependent aberration metrology," *APL Photon.*, vol. 4, no. 5, May 2019, Art. no. 050802.

[8] T. Aidukas, L. Loetgering, and A. R. Harvey, "Addressing phase-curvature in Fourier ptychography," *Opt. Exp.*, vol. 30, no. 13, pp. 22421–22434, 2022.

[9] A. Pan, C. Zuo, Y. Xie, M. Lei, and B. Yao, "Vignetting effect in Fourier ptychographic microscopy," *Opt. Lasers Eng.*, vol. 120, pp. 40–48, Sep. 2019.

[10] R. Horstmeyer, G. Zheng, X. Ou, and C. Yang, "Modeling extensions of Fourier ptychographic microscopy," *Microsc. Microanal.*, vol. 20, no. S3, pp. 370–371, Aug. 2014.

[11] L. Bian, J. Suo, G. Zheng, K. Guo, F. Chen, and Q. Dai, "Fourier ptychographic reconstruction using Wirtinger flow optimization," *Opt. Exp.*, vol. 23, no. 4, pp. 4856–4866, 2015.

[12] S. Chen, T. Xu, J. Zhang, X. Wang, and Y. Zhang, "Optimized denoising method for Fourier ptychographic microscopy based on Wirtinger flow," *IEEE Photon. J.*, vol. 11, no. 1, pp. 1–14, Feb. 2019.

[13] X. Ou, G. Zheng, and C. Yang, "Embedded pupil function recovery for Fourier ptychographic microscopy," *Opt. Exp.*, vol. 22, no. 5, pp. 4960–4972, 2014.

[14] L. Tian, X. Li, K. Ramchandran, and L. Waller, "Multiplexed coded illumination for Fourier ptychography with an LED array microscope," *Biomed. Opt. Exp.*, vol. 5, no. 7, pp. 2376–2389, 2014.

[15] A. Pan, C. Zuo, and B. Yao, "High-resolution and large field-of-view Fourier ptychographic microscopy and its applications in biomedicine," *Rep. Prog. Phys.*, vol. 83, no. 9, Sep. 2020, Art. no. 096101.

[16] A. Pan, Y. Zhang, T. Zhao, Z. Wang, D. Dan, M. Lei, and B. Yao, "System calibration method for Fourier ptychographic microscopy," *J. Biomed. Opt.*, vol. 22, no. 9, Sep. 2017, Art. no. 096005.

[17] C. Zuo, J. Sun, and Q. Chen, "Adaptive step-size strategy for noise-robust Fourier ptychographic microscopy," *Opt. Exp.*, vol. 24, no. 18, pp. 20724–20744, 2016.

[18] S. Jiang, K. Guo, J. Liao, and G. Zheng, "Solving Fourier ptychographic imaging problems via neural network modeling and TensorFlow," *Biomed. Opt. Exp.*, vol. 9, no. 7, pp. 3306–3319, 2018.

[19] A. Kappeler, S. Ghosh, J. Holloway, O. Cossairt, and A. Katsaggelos, "Ptychnet: CNN based Fourier ptychography," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 1712–1716.

[20] X. Wang, Y. Piao, J. Yu, J. Li, H. Sun, Y. Jin, L. Liu, and T. Xu, "Deep multi-feature transfer network for Fourier ptychographic microscopy imaging reconstruction," *Sensors*, vol. 22, no. 3, p. 1237, Feb. 2022.
[21] J. Zhang, T. Xu, Z. Shen, Y. Qiao, and Y. Zhang, "Fourier ptychographic microscopy reconstruction with multiscale deep residual network," *Opt. Exp.*, vol. 27, no. 6, pp. 8612–8625, 2019.
[22] M. Sun, L. Shao, Y. Zhu, Y. Zhang, S. Wang, Y. Wang, Z. Diao, D. Li, Q. Mu, and L. Xuan, "Double-flow convolutional neural network for rapid large field of view Fourier ptychographic reconstruction," *J. Biophotonics*, vol. 14, no. 6, Jun. 2021, Art. no. e202000444.
[23] X. Lu, M. Wang, H. Wu, and F. Hui, "Deep learning for fast image reconstruction of Fourier ptychographic microscopy with expanded frequency spectrum," *Proc. SPIE*, vol. 11781, Feb. 2021, Art. no. 117810M.
[24] V. Bianco, M. D. Priscoli, D. Pirone, G. Zanfardino, P. Memmolo, F. Bardozzo, L. Miccio, G. Ciaparrone, P. Ferraro, and R. Tagliaferri, "Deep learning-based, misalignment resilient, real-time Fourier ptychographic microscopy reconstruction of biological tissue slides," *IEEE J. Sel. Topics Quantum Electron.*, vol. 28, no. 4, pp. 1–10, Jul. 2022.
[25] T. Nguyen, S. Aslam, D. Bower, J. L. Eigenbrode, N. Gorius, T. Hewagama, L. Miko, and G. Nehmetallah, "Portable flow device using Fourier ptychography microscopy and deep learning for detection in biosignatures," *Proc. SPIE*, vol. 11401, Apr. 2020, Art. no. 114010H.
[26] F. Shamshad, F. Abbas, and A. Ahmed, "Deep Ptych: Subsampled Fourier ptychography using generative priors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 7720–7724.
[27] R. Wang, P. Song, S. Jiang, C. Yan, J. Zhu, C. Guo, Z. Bian, T. Wang, and G. Zheng, "Virtual brightfield and fluorescence staining for Fourier ptychography via unsupervised deep learning," *Opt. Lett.*, vol. 45, no. 19, pp. 5405–5408, 2020.
[28] J. Zhang, X. Tao, L. Yang, C. Wang, C. Tao, J. Hu, R. Wu, and Z. Zheng, "The integration of neural network and physical reconstruction model for Fourier ptychographic microscopy," *Opt. Commun.*, vol. 504, Feb. 2022, Art. no. 127470.
[29] A. Abbott, "Sequence analysis: New methods for old ideas," *Annu. Rev. Sociol.*, vol. 21, no. 1, pp. 93–113, Aug. 1995.
[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[31] J. Zhang, T. Xu, J. Zhang, Y. Chen, and J. Li, "Cross-level channel attention network for Fourier ptychographic microscopy reconstruction," *IEEE Photon. J.*, vol. 14, no. 1, pp. 1–8, Feb. 2022.
[32] C. Yican, W. Xia, L. Zhi, Y. Huidong, and H. Bo, "Fourier stack microscopic imaging based on depth learning," *Prog. Laser Optoelectron.*, vol. 57, no. 22, 2020, Art. no. 221106.
[33] J. Zhang, T. Xu, J. Li, Y. Zhang, S. Jiang, Y. Chen, and J. Zhang, "Physics-based learning with channel attention for Fourier ptychographic microscopy," *J. Biophoton.*, vol. 15, no. 3, Mar. 2022, Art. no. e202100296.
[34] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5197–5206.
[35] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.
[36] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
[37] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[38] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin Transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.
[39] B. Li, E. Ouyang, W. Hu, G. Zhang, L. Zhao, and J. Wu, "Multi-granularity vision transformer via semantic token for hyperspectral image classification," *Int. J. Remote Sens.*, vol. 43, no. 17, pp. 6538–6560, Sep. 2022.
[40] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.
[41] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 47–57, Mar. 2017.
[42] 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.

LIN ZHAO (Member, IEEE) received the M.S. degree in computer application technology from the Huazhong University of Science, in 2007, and the Ph.D. degree in control science and engineering from Central South University, in 2017.

As an Association Professor, he is currently conducting research programs in optical imaging, hyperspectral image processing. His research interests include image processing, pattern recognition, and the modeling and optimal control of complex industrial process.

XUHUI ZHOU received the B.S. degree from the Hunan Institute of Science and Technology, Yueyang, China, in 2020, where he is currently pursuing the M.S. degree.

His research interests include deep learning, pattern recognition, image reconstruction, and Fourier ptychographic imaging technology.

XIN LU received the B.S. degree in opto-electronic information science and engineering from Shantou University, Shantou, Guangdong, in 2019, and the M.S. degree in optical engineering from Shenzhen University, Shenzhen, Guangdong, in 2022.

His research interest includes physical guide deep learning for computational super-resolution, such as Fourier ptychographic microscopy (FPM).

HAIPING TONG received the bachelor's degree from Fujian Normal University. He is currently pursuing the Graduate degree with Shenzhen University.

His current research interest includes Fourier ptychographic imaging technology.

HUI FANG received the bachelor's and master's degrees in physics from the University of Science and Technology of China, and the Ph.D. degree in physics from Boston University.

He is currently a Professor with the Nanophotonics Research Center, Institute of Microscale Optoelectronics, Shenzhen University. Previously, he was a Professor with Nankai University. Then, he was a Postdoctoral Fellow of the Harvard Medical School and Biomedical Engineering Department, Washington University in St. Louis. His major research interests include optical spectroscopic microscopy, photoacoustic sensing, atomic force microscopy, and structured light.