Opto-Electronic Advances

ISSN 2096-4579 CN 51-1781/TN

High performance “non-local” generic face reconstruction model using the lightweight Speckle-Transformer (SpT) UNet

Yangyundou Wang, Hao Wang and Min Gu

Citation: Wang YYD, Wang H, Gu M. High performance “non-local” generic face reconstruction model using the lightweight Speckle-Transformer (SpT) UNet. *Opto-Electron Adv*, **6**, 220049(2023).

https://doi.org/10.29026/oea.2023.220049

Received: 7 March 2022; Accepted: 25 May 2022; Published online: 8 October 2022

Opto-Electronic Advances

Article

February 2023, Vol. 6, No. 2

DOI: 10.29026/oea.2023.220049

High performance “non-local” generic face reconstruction model using the lightweight Speckle-Transformer (SpT) UNet

Yangyundou Wang[1,2]*, Hao Wang[3] and Min Gu[1,2]*

Significant progress has been made in computational imaging (CI), in which deep convolutional neural networks (CNNs) have demonstrated that sparse speckle patterns can be reconstructed. However, due to the limited “local” kernel size of the convolutional operator, for the spatially dense patterns, such as the generic face images, the performance of CNNs is limited. Here, we propose a “non-local” model, termed the Speckle-Transformer (SpT) UNet, for speckle feature extraction of generic face images. It is worth noting that the lightweight SpT UNet reveals a high efficiency and strong comparative performance with Pearson Correlation Coefficient (PCC), and structural similarity measure (SSIM) exceeding 0.989, and 0.950, respectively.

Keywords: speckle reconstruction; non-local model; generic face images; lightweight network

Wang YYD, Wang H, Gu M. High performance “non-local” generic face reconstruction model using the lightweight Speckle-Transformer (SpT) UNet. *Opto-Electron Adv* **6**, 220049 (2023).

Introduction

Imaging through scatters is a classical inverse problem[1]. As a direct forward modeling method, deep learning (DL) was recently implemented in computational imaging (CI), and it provided high-quality solutions to several CI problems[2]. Seminal works have demonstrated that deep convolutional neural networks (CNNs) can extract statistical features of speckle patterns[3–13]. Compared with the support vector regression (SVR)[14], deep convolutional UNet architectures demonstrated better performance on sparse feature extraction and certain generalization ability. The UNet architecture IDiffNet, first proposed by S. Li et al., realized speckle image reconstruction, especially for the sparse patterns[4]. Y. Li et al. demonstrated a network for scalable diffusers with various microstructures for different sparse pattern reconstructions[5]. The PDSNet was proposed by E. Guo et al. for sparse feature extraction. For the generic human face dataset, the network achieved far less accuracy with SSIM is about 0.75[6]. In other words, the performance of deep convolutional UNet on the spatially dense speckle feature extraction and reconstruction is limited.

Due to the limited size of the convolutional kernel, CNNs are a “local” model. As a “non-local” mechanism, the attention weighs the significance of each part of the input data and extracts long-term dependencies of the feature maps[15]. The generalization ability of the attention mechanism has revealed an excellent performance in speckle reconstructions of sparse patterns[16]. The

[1]Institute of Photonic Chips, University of Shanghai for Science and Technology, Shanghai 200093, China; [2]Centre for Artificial-Intelligence Nanophotonics, School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China; [3]School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China.
*Correspondence: YYD Wang, E-mail: ywang0606@usst.edu.cn; M Gu, E-mail: gumin@usst.edu.cn
Received: 7 March 2022; Accepted: 25 May 2022; Published online: 8 October 2022

transformers are modules that rely entirely on the attention mechanism and can be easily parallelized[17]. Moreover, the transformer assumes minimal prior knowledge about the structure of the problem as compared to its convolutional and recurrent counterparts in deep learning. In vision, transformers have been successfully used for image recognition[18,19], object detection, segmentation[20], image super-resolution[21], video understanding[22,23], image generation[24], text-image synthesis[25] and so on[26]. However, based on our knowledge, none of the investigations has explored the performance of the transformers in CI, such as speckle reconstruction.

Here, we propose a high-performance "non-local" generic feature extraction and reconstruction model - SpT UNet. The network is a UNet architecture including advanced transformer encoder and decoder blocks. For better feature reservation/extraction, we propose and demonstrate three key mechanisms, i.e., pre-batch normalization (pre-BN), and position encoding in multi-head attention/multi-head cross-attention (MHA/MHCA), and self-built up/down sampling pipelines. For the "scalable" data acquisition, four different grits of diffusers within the 40 mm detection range are considered. We further quantitatively evaluate the network performance with four scientific indicators, namely Pearson correlation coefficient (PCC), structural similarity measure (SSIM), Jaccard index (JI), and peak signal-to-noise ratio (PSNR). The SpT UNet shows less computational complexity and far better reconstruction and generalization ability compared with the other state-of-the-art transformer models in vision[18,27].

Method

SpT UNet implementation

The architecture of the SpT UNet

As shown in Fig. 1, the network is based on the UNet architecture. The skip connections are used to transfer information directly between blocks of the same size. Three blocks of the identical structure are included in both the encoder and decoder. In the encoder, the block contains two layers, i.e., the MHA mechanism and position-wise Feed-Forward Network (FFN). We employ a residual connection[28] around each of the two layers. The pre-batch normalization (Pre-BN), and residual connection are implemented for the stabilization in the training of the network.

Besides the two layers in each encoder block, the decoder block contains an extra MHCA layer which is used to aggregate the feature through the skip connections. The embedded label is fed to the MHCA layer as Q in the first block of the decoder, whereas the Q in the second and third blocks is a feature map that transmits through the skip connections. The activation function for the output layer is Sigmoid and the loss function is cross-entropy (CE).

Transformer module for the SpT UNet

Transformer adopts attention mechanism[15] with Query-Key-Value (QKV) module. To be specific, we define the attention function for the SpT UNet as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\mathrm{T}}}{d_k}\right)V\,, \qquad (1)$$

where the spaces for queries, keys, and values are $Q \in \mathbb{R}^{d_{\text{model}} \times d_k \times d_k}$, $K \in \mathbb{R}^{d_{\text{model}} \times d_k \times d_k}$, and $V \in \mathbb{R}^{d_{\text{model}} \times d_v \times d_v}$, separately. Here, d_{model} denotes the number of channels, i.e., the number of feature maps for Q, K, and V, respectively. And $d_k \times d_k$, and $d_v \times d_v$ represent the dimension of the feature maps for keys (or queries) and values, respectively. Moreover, the softmax is applied in a row-wise manner to get the attention matrix between each pixel, and d_k is a scaling factor that is implemented to alleviate gradient vanishing.

The core of SpT UNet encoder blocks is the multi-head attention (MHA) mechanism for joint information extraction from different representation subspaces at different positions. The MHA based model can be expressed as:

$$\begin{aligned} &\text{MHA}(Q, K, V) = \text{Concat}(\text{head}_1, \cdots, \text{head}_H)W^{\mathrm{O}}\,, \\ &\text{where} \quad \text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)\,. \end{aligned} \qquad (2)$$

Here, the projections are parameter tensors W_i^Q, W_i^K, and W_i^V which are the linear transformations on Q, K, and V, respectively. What's more, W^{O} is the projection for the output of all heads. In this work, the head number H varied in both the encoder and decoder blocks.

Besides the MHA mechanism, we propose the multi-head cross-attention (MHCA) mechanism for SpT UNet decoder blocks which can be expressed as:

$$\begin{aligned} &\text{MHCA}(Q,K,V) = \text{Concat}(\text{head}_1, \cdots, \text{head}_H)W^{\mathrm{O}}\,, \\ &\text{where} \quad \text{head}_i = \text{Attention}(f(Q^*)W_i^Q, KW_i^K, VW_i^V)\,. \end{aligned} \qquad (3)$$

Here, $f(Q^*)$ is the function for the feature embedding with Q^* representing the embedded label or the feature maps that transmit through the skip connections.

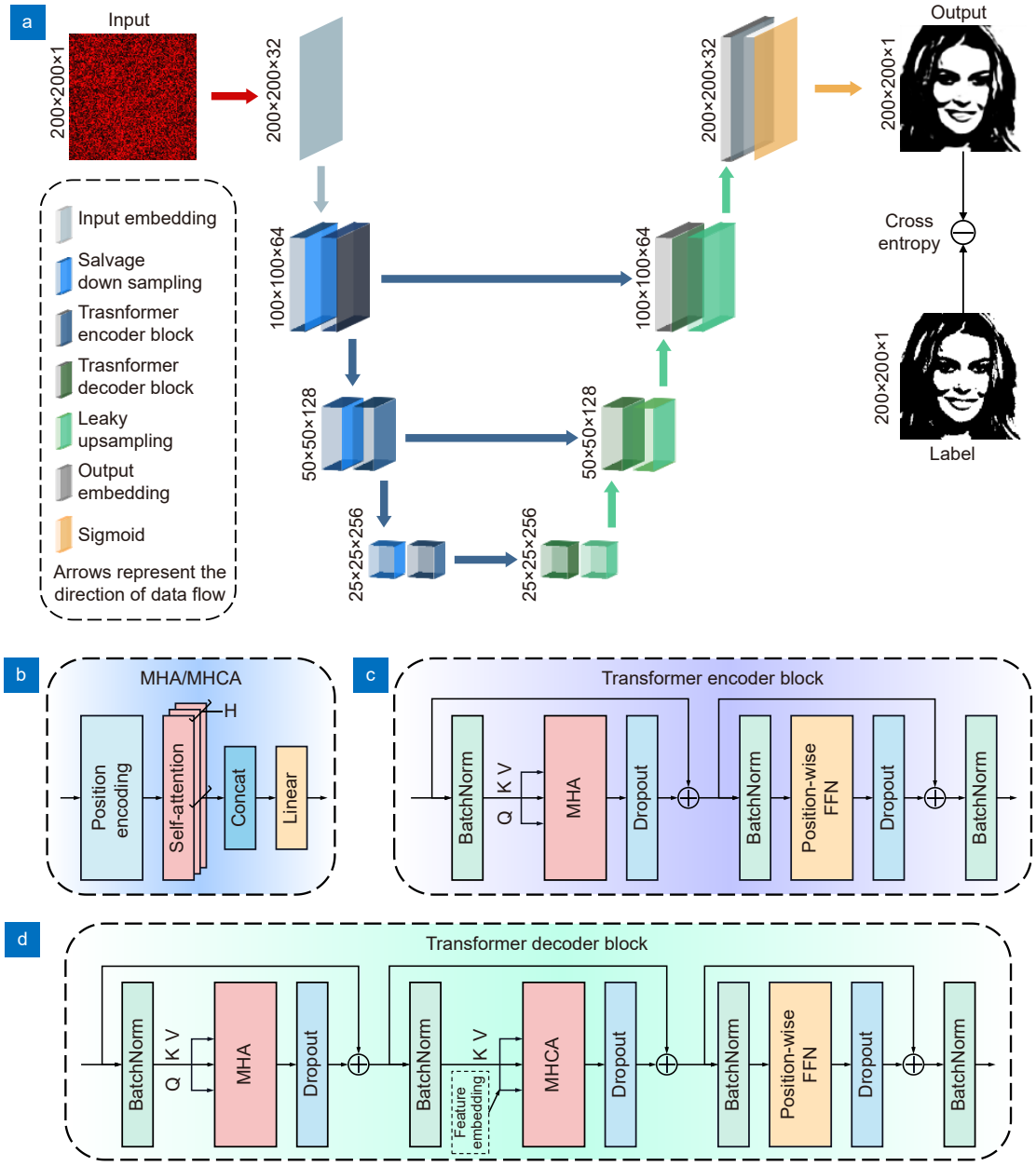

Fig. 1 | SpT UNet architecture for spatially dense feature reconstruction (**a**) with the multi-head attention (or cross attention) module (**b**) included transformer encoder block (**c**) and decoder block (**d**).

Module-level modification of SpT UNet

Pre-batch normalization (Pre-BN)

For better training of the network, pre-normalization in each block is implemented for stabilized convergence of the gradients. As each batch contains more varied cross-speckle features compared with the features between channels, we further upgrade the pre-layer normalization to the pre-batch normalization. As shown in Fig. 1(b) and Fig. 1(c), batch normalization (BN), along with residual connection, is considered as a mechanism for stabilizing training of the network (e.g., alleviating ill-posed gradients and model degeneration). Therefore, Pre-BN included transformer encoder and decoder blocks can better extract the cross-speckle features of the spatially dense images, i.e., generic face images with high training efficiency and less sensitivity to the initialization of the network.

Position encoding in MHA/MHCA

The SpT UNet backbone is a transformer module using MHA and MHCA as the core. MHA and MHCA are coped with three-dimensional position encoding, i.e., the inductive deviation. The purpose of the position

encoding is for stable and efficient feature extraction. In specific, we designed a three-dimensional absolute sinusoidal position encoding for feature maps. For the index t in feature maps, we define a vector pair $\boldsymbol{P}_t = (\boldsymbol{P}_{x,t}; \boldsymbol{P}_{y,t})$ with $\boldsymbol{P}_t \in \mathbb{R}^{d_{\text{model}}}$. And $\boldsymbol{P}_{x,t}$ and $\boldsymbol{P}_{y,t}$ are sinusoidal functions, i.e., $f(t,k)$ in x and y coordinate with pre-defined frequency, as shown below:

$$f(t,k) = \begin{cases} \sin(t/10000^k), \text{if } k \text{ is even} \\ \cos(t/10000^k), \text{if } k \text{ is odd} \end{cases}, \tag{4}$$

where k is the index of the vector, i.e., $k = [0,1, \cdots, d_{\text{model}}-1]$.

Puffed downsampling and leaky upsampling

For better speckle feature extraction, we invent two sampling methods, i.e., puffed downsampling with sandwich-like autoencoder structure, and leaky upsampling with a bottleneck structure inspired by compressed sensing.

As shown in Fig. 2, the first two layers of convolutional (Conv), BN, and ReLU expand channels of the feature map four times. The max pooling layer is sandwiched in between the five-layer autoencoder, and the last two layers of Conv, BN, and ReLU compress channels of the feature map to two times smaller as the input for "salvage" speckle feature information.

As shown in Fig. 3, the bilinear interpolation layer is positioned in the middle of the bottleneck structure. Compared with the conventional upsampling methods in UNet, leaky upsampling retains the most valuable features, and discards the less valuable features with high computational efficiency.

Optical imaging system and data acquisition

As shown in Fig. 4, the central 800×800 pixels of the SLM are illuminated by the filtered and collimated CW laser at 632.8 nm. The spatial light modulator (SLM, Thorlabs EXULUS-HD2 pixel size 8 μm, 1920×1200) uploads the phase patterns of binary generic face images. Glass diffusers with four grit types (Thorlabs DG10-120-MD, DG10-220-MD, DG10-600-MD, DG10-1500-MD) were positioned at the conjugated plane of the SLM sequentially. To match the pixel size of the CMOS camera (Thorlabs DCC1645C, pixel size 3.6 μm, 1280×1024) with that of the SLM, we built a 4F system using two lenses L1 (f = 300 mm) and L2 (f = 125 mm). For the training, testing, and validation dataset collection, the CMOS camera was placed within a distance of 40 mm from the focal plane of lens 4.

To collect the training and testing datasets, 1500 Faces-LFW face images[29] were considered. We

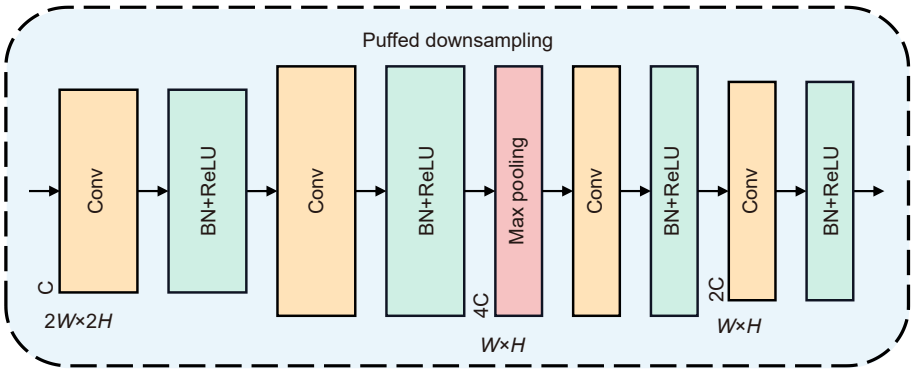

Fig. 2 | The puffed downsampling - module architecture.

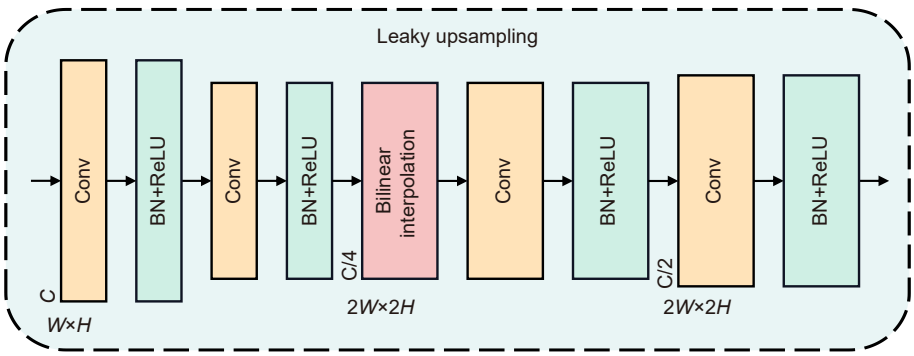

Fig. 3 | The leaky upsampling - module architecture.

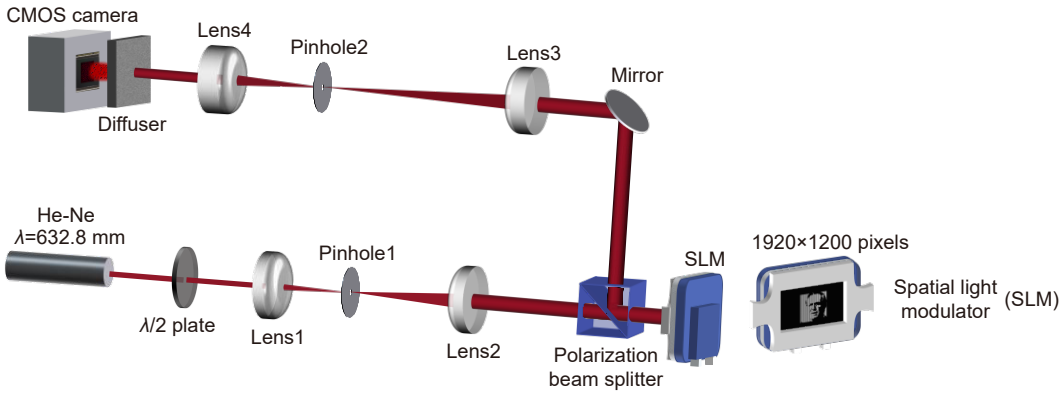

Fig. 4 | Experiment set-up.

experimentally achieved a space-bandwidth product of 800×800 pixels using up to 12000 training/testing pairs. As shown in Fig. 5, it includes four cases, as follows:

Case 1: Train/test the network with the speckles produced by a 120 grits diffuser at 0, and 20 mm away from the image plane.

Case 2: Train/test the network with the speckles produced by a 220 grits diffuser at 0, and 20 mm away from the image plane.

Case 3: Train/test the network with the speckles produced by a 600 grit diffuser at 0, and 20 mm away from the image plane.

Case 4: Train/test the network with the speckles produced by a 1500 grit diffuser at 0, and 20 mm away from the image plane.

To better evaluate the generalization ability of the network especially for varied depth of range, the validation dataset consists of 6000 pairs produced by four diffusers with 1500 seen Faces-LFW face images at 40 mm away from the focal plane.

Data processing

The speckle patterns were first normalized between 0 and 1, and the labels for the generic face images were binary values. To reduce the parameters of the network and the demand for GPU and training data, the input speckle patterns were first downsampled from 800 × 800 pixels to 200 × 200 pixels using the bilinear interpolation approach. And the network was implemented using Python version 3.8.5 and PyTorch framework version 1.7.1 (Facebook Inc.) and ran on NVIDIA GeForce RTX 3090. The network was trained with 200 epochs with a learning rate of 10^{-4} for the first 100 epochs, 10^{-5} for the next 50 epochs, and 10^{-6} for the final 50 epochs. The batch size in the training/testing process is 2. Moreover, the lightweight SpT UNet contains 6.6 million neurons. For

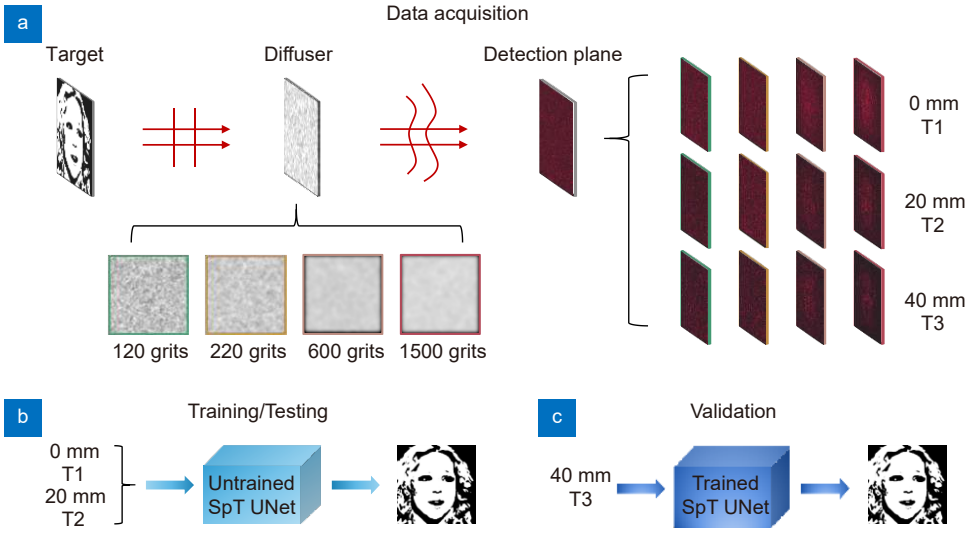

Fig. 5 | Overview of the data acquisition under various conditions and the training/testing/validation of the SpT UNet. (**a**) The training/testing data set is captured at T1 (0 mm), and T2 (20 mm). And the validation data set is captured at T3 (40 mm). The training/testing stage (**b**) and the validation stage (**c**) of the SpT UNet for the speckle reconstruction of the generic face images.

the network configurations, Adam optimizer, L2 norm regularizer, and the CE/NPCC as loss functions were chosen. Once the model was trained, each prediction was made within 16 ms.

Results and discussions

To intuitively visualize the JI score, the generic human faces, and related reconstructed pictures are shown in Fig. 6. The ground truth of the face images and their zoom-in structures are listed in the left column. As shown in the right column, the related predicted results are further broken down into the true positive (white), the false positive (green), and the false-negative (red). For the validation of the untrained detection position,

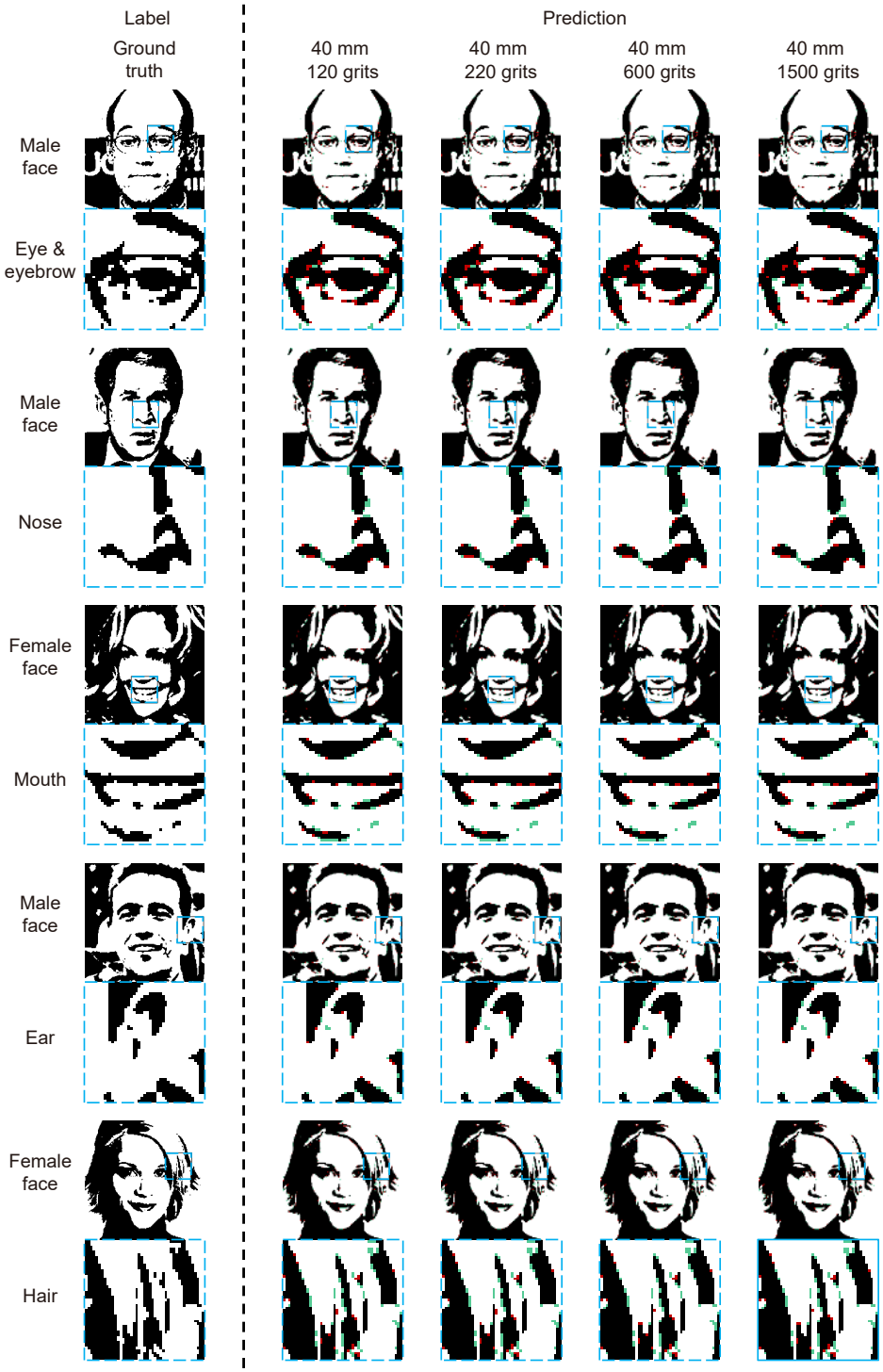

Fig. 6 | The ground truth (left column) and prediction (right column) of the trained SpT UNet with the camera placed at 40 mm away from the focal plane. The prediction results are overlaid with the true positive (white), false positive (green), and false negative (red).

the facial features (ears included) and facial contour (hair included) of both the males and females are well extracted and reconstructed.

We further quantitatively evaluate the performance of the network using PCC, JI, SSIM, and PSNR. The PCC is essentially a normalized measurement of covariance, with the value 1 representing perfect correlation. The SSIM evaluates the similarity between reconstructed patterns and related ground truth. It is a decimal value between 0 and 1, value 1 represents perfect structural similarity, and 0 indicates no structural similarity. Similar to the SSIM, the JI gauges the similarity and diversity between prediction and its ground truth. The PSNR is used to quantify the quality of the reconstruction: the higher PSNR, the better the reconstructed image. As shown in Table 1, for the four different diffusers, the values of the PCC, JI, and SSIM are all above 0.989, 0.976, and 0.950, respectively. Like the PCC, JI, and SSIM, the value of the PSNR increases slightly as the increase of the girt number of diffusers.

Moreover, to evaluate the loss and reconstruction accuracy of the SpT UNet. The plots of loss and accuracy for the trained SpT UNet as the function of the epoch are shown in Fig. 7.

We also quantitatively evaluate the complexity of the Spt UNet and its downsize version—SpT UNet-B. The performance of the SpT UNet and the SpT UNet-B is shown in Table 2. For the SpT UNet-B, we reduce the position-wise feed-forward network (FFN) parameters by 50%, and the number of the heads or parallel attention layers in MHA and MHCA is decreased by 50%. Moreover, we implement the bottleneck for the first and third convolution layers for the Leaky and Puffed sampling.

It is worth noting that, as a lightweight network, the SpT UNet and Spt UNet-B reveal less than one order of parameters compared with ViT[22], and SWIN transformer[23]. The comparison is shown in Table 3.

Conclusions

We have proposed a "non-local" spatially dense object

Table 1 | The validation performance of the trained SpT UNet.

Indicator	Diffuser/grit			
	120	220	600	1500
PCC	0.98986	0.98988	0.98990	0.98994
JI	0.97655	0.97658	0.97661	0.97666
SSIM	0.95001	0.95009	0.95024	0.95035
PSNR/dB	19.3826	19.3887	19.3954	19.4052

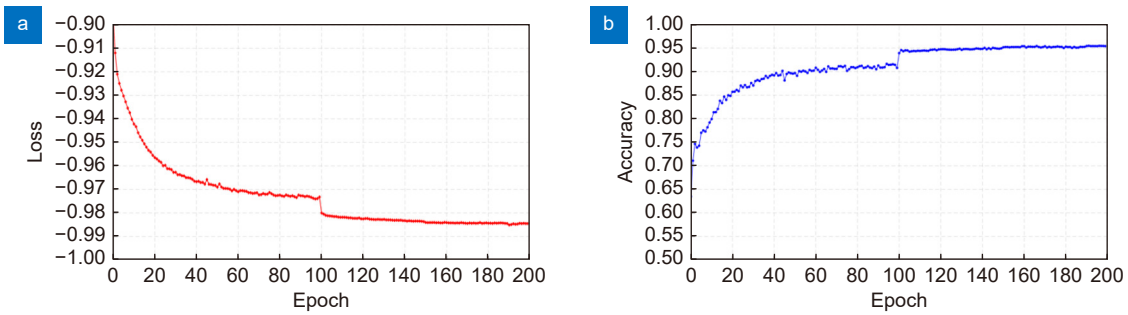

Fig. 7 | Quantitative analysis of the trained SpT UNet using NPCC as the loss function (**a**) and SSIM as the indicator for accuracy (**b**).

Table 2 | Performance of the SpT UNet.

Method	Image size	FLOPs/10^9	Parameters	Throughput (image/s)	Inference time (batch/ms)
SpT UNet	256×256	31.7	6.6 M	62.5	31.34
SpT UNet	224×224	24.3	6.6 M	83.3	24.02
SpT UNet	200×200	19.4	6.6 M	86.9	23.02
SpT UNet-B	256×256	19.6	4.0 M	78.5	25.46
SpT UNet-B	224×224	15.0	4.0 M	90.8	22.02
SpT UNet-B	200×200	12.0	4.0 M	95.1	21.02

Table 3 | The comparison of the SpT UNet, ViT, and SWIN transformer on parameter numbers.

Method	Parameters
SpT UNet	6.6 M
SpT UNet-B	4.0 M
SWIN transformer	197 M
SWIN transformer-B	88 M
ViT	303 M
ViT-B	86 M

feature extraction and reconstruction model, i.e., the lightweight SpT UNet. It shows an excellent performance with comparative values on the scientific indicators for generic face images through varied types of diffusers at different detection planes. Although we just consider the reconstruction of binary generic face images, the reconstruction of spatially dense images at grayscale using the SpT UNet can be considered in the future. For the biomedical imaging, we believe that the network can be further implemented in complex tissue imaging to boost the image contrast and depth of range. For photonic computing, as the paralleling processing model, the SpT UNet can be further implemented as an all-optical diffractive neural network with surpassing feature extraction ability, light speed and even lower energy consumption.

References

1. Goodman JW. *Speckle Phenomena in Optics: Theory and Applications* (Roberts and Company Publishers, Englewood, 2007).
2. Barbastathis G, Ozcan A, Situ GH. On the use of deep learning for computational imaging. *Optica* **6**, 921–943 (2019).
3. Li W, Xi TL, He SF, Liu LX, Liu JP et al. Single-shot imaging through scattering media under strong ambient light interference. *Opt Lett* **46**, 4538–4541 (2021).
4. Li S, Deng M, Lee J, Sinha A, Barbastathis G. Imaging through glass diffusers using densely connected convolutional networks. *Optica* **5**, 803–813 (2018).
5. Li YZ, Xue YJ, Tian L. Deep speckle correlation: a deep learning approach toward scalable imaging through scattering media. *Optica* **5**, 1181–1190 (2018).
6. Guo EL, Zhu S, Sun Y, Bai LF, Zuo C et al. Learning-based method to reconstruct complex targets through scattering medium beyond the memory effect. *Opt Express* **28**, 2433–2446 (2020).
7. Liao MH, Zheng SS, Pan SX, Lu DJ, He WQ et al. Deep-learning-based ciphertext-only attack on optical double random phase encryption. *Opto-Electron Adv* **4**, 200016 (2021).
8. Liao K, Chen Y, Yu ZC, Hu XY, Wang XY et al. All-optical computing based on convolutional neural networks. *Opto-Electron Adv* **4**, 200060 (2021).
9. Lei YS, Guo YH, Pu MB, He Q, Gao P et al. Multispectral scattering imaging based on metasurface diffuser and deep learning. *Phys Status Solidi Rapid Res Lett* **16**, 2100469 (2022).
10. Ma J, Huang YJ, Pu MB, Xu D, Luo J et al. Inverse design of broadband metasurface absorber based on convolutional autoencoder network and inverse design network. *J Phys D Appl Phys* **53**, 464002 (2020).
11. Wang JY, Tan XD, Qi PL, Wu CH, Huang L et al. Linear polarization holography. *Opto-Electron Sci* **1**, 210009 (2022).
12. Lin ZS, Wang YYD, Wang H et al. Expansion of depth-of-field of scattering imaging based on DenseNet. *Acta Optica Sinica* **42**, 0436001 (2022).
13. Wang YYD, Wang H et al. High-accuracy, direct aberration determination using self-attention-armed deep convolutional neural networks. *Journal of Microscopy* **286**, 13–21 (2022).
14. Horisaki R, Takagi R, Tanida J. Learning-based imaging through scattering media. *Opt Express* **24**, 13738–13743 (2016).
15. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L et al. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* 6000–6010 (ACM, 2017).
16. Wang YYD, Lin ZS, Wang H, Hu CF, Yang H et al. High-generalization deep sparse pattern reconstruction: feature extraction of speckles using self-attention armed convolutional neural networks. *Opt Express* **29**, 35702–35711 (2021).
17. Lin TY, Wang YX, Liu XY, Qiu XP. A survey of transformers. (2021); https://arxiv.org/abs/2106.04554.
18. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH et al. An image is worth 16x16 words: transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations* (ICLR, 2020).
19. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A et al. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning* 10347–10357 (PMLR, 2021).
20. Ye LW, Rochan M, Liu Z, Wang Y. Cross-modal self-attention network for referring image segmentation. In *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 10494–10503 (IEEE, 2019).
21. Yang FZ, Yang H, Fu JL, Lu HT, Guo BN. Learning texture transformer network for image super-resolution. In *Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 5790–5799 (IEEE, 2020).
22. Sun C, Myers A, Vondrick C, Murphy K, Schmid C. Videobert: a joint model for video and language representation learning. In *Proceedings of 2019 IEEE/CVF International Conference on Computer Vision* 7463–7472 (IEEE, 2019).
23. Girdhar R, Carreira JJ, Doersch C, Zisserman A. Video action transformer network. In *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 244–253

(IEEE, 2021).

24. Chen HT, Wang YH, Guo TY, Xu C, Deng YP et al. Pre-trained image processing transformer. In *Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 12294–12305 (IEEE, 2021); http://doi.org/10.1109/CVPR46437.2021.01212.
25. Ramesh A, Pavlov M, Goh G, Gray S, Voss C et al. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning* 8821–8831 (PMLR, 2021).
26. Khan S, Naseer M, Hayat M, Zamir SW, Khan FS et al. Transformers in vision: a survey. (2021); https://arxiv.org/abs/2101.01169.
27. Liu Z, Lin YT, Cao Y, Hu H, Wei YX et al. Swin transformer: hierarchical vision transformer using shifted windows. In *Proceedings of 2021 IEEE/CVF International Conference on Computer Vision* 9992–10002 (IEEE, 2021).
28. He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016); http://doi.org/10.1109/CVPR.2016.90.
29. Huang GB, Mattar M, Berg T, Learned-Miller E. Labeled faces in the wild: a database forstudying face recognition in unconstrained environments. In *Proceedings of Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition* (HAL, 2008).

Acknowledgements

M. Gu acknowledges the funding support from the Science and Technology Commission of Shanghai Municipality (Grant No. 21DZ1100500), the Shanghai Frontiers Science Center Program (2021-2025 No. 20), and the Zhangjiang National Innovation Demonstration Zone (Grant No. ZJ2019-ZD-005)). Y. Y. D. Wang is supported by a fellowship from the China Postdoctoral Science Foundation (2020M671169) and the International Postdoctoral Exchange Program from the Administrative Committee of Post-Doctoral Researchers of China ([2020]33).

Author contributions

M. Gu supervised the project. Y. Y. D. Wang proposed the original idea. Y. Y. D. Wang and H. Wang designed the network architecture, acquired the datasets and evaluated the reconstruction performance. Y. Y. D. Wang, H. Wang and M. Gu wrote the manuscript with input from all authors. All authors discussed the results and commented on the manuscript.

Competing interests

The authors declare no competing financial interests.