Restoration of Laser Interference Image Based on Large Scale Deep Learning

XIANGYU ZHOU, ZHONGJIE XU, XIANGAI CHENG, AND ZHONGYANG XING

College of Advanced Interdisciplinary Studies, National University of Defense Technology, Changsha 410073, China
Nanhu Laser Laboratory, National University of Defense Technology, Changsha 410073, China
State Key Laboratory of Pulsed Power Laser Technology, Changsha 410073, China

Corresponding author: Zhongyang Xing (sunnia_xing@126.com)

ABSTRACT
Although the deep learning method has achieved some results in the challenging task of repairing the missing areas of images over the years, there is no report on the images disturbed by the laser entering the field of view. We propose a restoration model that can restore the laser interference image, the corresponding countermeasure network deep learning model and a new model training method, and no additional manual marking work is required in the data training set. After a large number of experiments, the loss function can rapidly converge under this method, accurately give the reasonable reconstruction results of the image disturbed by laser, and significantly improve the scores of many common image quality evaluation methods: high-quality repair results are obtained on the laser interference composite dataset of face (CelebA), Stanford automobile, aircraft, buildings (Facade) and satellite images. The model used in paper has the characteristics of fast training speed, strong robustness, modular model design and wide application.

INDEX TERMS Laser beams, image processing, deep learning.

I. INTRODUCTION

Recent deep learning studies have shown that large-scale deep learning methods can repair and restore the image loss of various complex scenes, perform tasks including missing completion, denoising, improving resolution, and generate reasonable content in the damaged area. Its essence is to use a model to take the image information that has been converted into digital format as input, and use the model for reasoning, so that some of the images in digital format can be changed, and the semantic apriori and meaningful hidden representation can be reflected through end-to-end learning. Generally, the structure of the model that can generate new visual data to manipulate the existing visual data (i.e. images and videos) of various visual scenes is an autoregressive model. Different from the previous diffusion based methods, that is, adjacent information is propagated to these areas or the same patch is copied through the patch based method. Facing the damaged area which is independent of the unrepaired area, we directly input the whole image into the neural network as a separate module. Multiple codecs input in parallel at the same time, and finally input the output into the super-resolution reconstruction network. This task allocation mode will effectively reduce the training time, increase the total number of features that can be generated by the model, and allow the self encoder to generate a slightly blurred image, and then hand the blurred image to the super-resolution reconstruction network for resolution improvement.

We have tested our method on several synthetic data sets: fill the missing image content according to the overall scene, and maintain the structural continuity between the laser interference edge and the undisturbed part. A variety of image quality evaluation methods are used to evaluate the results, and the image features match the original image. The main contributions of the work are as follows:

A new framework of deep learning self encoder is proposed. As far as we know, it is used for the first time in different scene image restoration tasks of laser interference entering the field of view. Different from the pure blank restoration network model architecture, the model we designed will retain part of the image information that is not completely interfered by the laser, rather than simply adjust all the interfered image pixels to the blank missing state for restoration. Due to the characteristics of residual network, this
modular design architecture has simple scalability, and has proven excellent stability when expanding, and the restoration results can be accurate to the pixel level.

We combine the advantages of self encoder convolution network, residual network, super-resolution reconstruction network and anti training network, which greatly shortens the training time. The loss function of the restoration generator can be reduced from a few weeks’ training time to less than 2 days, and no additional manual marking is required. Only the original image is required for supervised learning.

II. RELATED WORK

As the previous semiconductor technology is in the development stage, the computing power of ordinary researchers is difficult to support the parameter training of large-scale neural networks. Some image restoration methods are based on the statistical information of low-level features (such as the mean square deviation feature of RGB images) to propagate from the adjacent pixels of the undisturbed area in the image to the target disturbed pixels [1], [2], [3], or patch matching based on the similarity between the background and the disturbed area [4], [5], [6] Image gradient [7] information is used for image restoration to spread the background area to the interference or missing area, which is suitable for static texture. These methods have limited effect on complex non-stationary information image restoration, and basically do not have the ability to restore large-area images disturbed by laser, and it is more difficult to accurately restore the inherent characteristics of target missing in the image.

The image restoration technology based on neural network has developed vigorously in recent years, and the image restoration technology based on convolutional neural network has been continuously improved. Yu et al. [8] proposed an end-to-end image restoration model to further ensure the color and texture consistency between the generated area and the surrounding environment by using the overlay generation network. In addition, in order to capture the long-term spatial dependence, a context attention module is proposed and integrated into the network to explicitly borrow information from distant spatial locations. However, this method is mainly trained on large rectangular masks and can not be well extended to free-form masks.

The network technology to deal with irregular mask adopts partial convolution [9] or gate convolution [10], and only effective pixels are used as input, which can deal with irregular shaped blank holes. However, the shape of blank holes needs to be manually specified during use, which is not in line with the restoration under the variable scene of laser interference. Or divide the spatial pixels into different regions and calculate the average value and variance of each region for normalization [11], normalize the pixels in the damaged and undamaged regions respectively based on the original repair mask to solve the problem of mean value and variance shift. However, it is not suitable for laser interference, which is a feature of partial complete interference. For example, the study [12] explored the influence of laser on the depth neural network, but did not make the laser enter the field of view and generate irregular pixel interference caused by laser diffraction in the image.

Recent research shows that Low light or blurred images without large area damage of images can be directly extracted from the input images for unsupervised training by the depth based learning method. Other recent research shows that on the premise of having expanded training data, the new model transformer [13], [14] can surpass the convolution residual network architecture of the classic state-of-the-art in image processing [15], and the latest image restoration and technology based on Transformer’s attention mechanism image restoration can restore the image in the case of almost complete loss of the original image [16], [17]. However, this method focuses on the imagination of the model, rather than the ability to accurately restore the original target features of the image in the case of partial laser interference.

III. METHOD

In this section, we will describe our deep learning neural network model from the network architecture. Firstly, the specific input / output interfaces and layer definitions of encoder,
decoder and super-resolution reconstruction network in the network architecture of image generator are introduced. Secondly, the antagonism network architecture of antagonism training with generator is introduced. The image generator is used for deep learning. The synthetic image after laser interference is restored to the original image containing target features. The countermeasure network is used to assist the generator training. The goal is to make it difficult to distinguish the network generated by the generator from the original image. We have verified that the large-scale deep learning model using only the traditional convolution network can suppress the local feature invalid or color difference, blur and obvious edge response caused by the disturbed pixels as described in [9]. It proposes to use mask and re-normalization steps to make convolution only depend on effective pixels for operation, as in

\[
O_{y,x} = \begin{cases} 
\sum \sum W \cdot \left( I \odot \frac{M}{\text{sum}(M)} \right), & \text{if } \text{sum}(M) > 0 \\
0, & \text{otherwise}
\end{cases}
\]  

(1)

where M is the corresponding binary mask, 1 indicates that the pixel in position \((y, x)\) is valid, 0 indicates that the pixel is invalid, and \(\odot\) indicates element multiplication. After each partial convolution operation, a mask update step is required to propagate the new \(m\) according to the following rules:

\[m'_{y,x} = 1, \text{iff } \text{sum}(M) > 0.\]

Or use gated convolution as shown in [10]

\[
\text{Gatting}_{y,x} = \sum \sum W_g \cdot I
\]  

(2)

\[
\text{Feature}_{y,x} = \sum \sum W_f \cdot I
\]  

(3)

\[
O_{y,x} = \varphi(\text{Feature}_{y,x}) \odot \sigma(\text{Gatting}_{y,x})
\]  

(4)

Among \(\sigma\) is a sigmoid function, so the output strobe value is between 0 and 1. \(\varphi\) Can be any activation function (for example, ReLU, ELU, and LeakyReLU). \(W_g\) and \(W_f\) are two different convolution filters.

However, these two methods classify all pixels as valid or invalid, and do not use them for laser scenes with partial interference, because the laser may only damage part of the disturbed area, causing color difference rather than pixel loss. If all the laser interference areas are set to be invalid directly, the target information will be lost. Even if the model is more advanced and the restoration result is clearer, the probability of accurate restoration of the location area based on the existing information will be greatly reduced in logic theory. Our method does not require the user to manually specify the range and size of missing pixels in the use stage, which enhances the automation of the whole image restoration process, and has advanced training efficiency of parallel residual convolution architecture. It can infer from three parts: the completely missing area of laser interference, the area containing laser interference noise and the area not disturbed by laser, so as to achieve accurate image restoration with the goal of approaching the original image.

We use a direct convolution network with convolution layers. The loss function of this method is the mean square error between the reconstructed generated image and the original image in the training data set. Many studies have confirmed the ability of convolutional network to encode image information effectively. For image inpainting task, convolution layer is mainly used to construct codec network.

We will start with the network architecture to describe our deep learning neural network model. Firstly, the specific input/output interfaces and layer definitions of encoder, decoder and super-resolution reconstruction network in the network architecture of image generator are introduced. Secondly, the antagonism network architecture of antagonism training with generator is introduced. The image generator is used for deep learning. It uses the synthetic image after laser interference to restore to the original image that saves the target features, that is, it makes four improvements to the neural network:

1. A residual layer structure similar to U-Net [18] and shift net [19] is added to improve the gradient vanishing problem and maximize semantic extraction. With the deepening of the number of layers, when the gradient propagates backward to the shallow network, it can hardly cause the disturbance of parameters, so the network can not be trained. And because many convolution operations are difficult to realize the basic operation of identity transformation, adding residual structure can effectively realize the identity transformation which is difficult to achieve in convolution layer.

2. We change the tail output structure of the decoder so that the network output is no longer limited to the missing part, especially for the irregular interference such as laser, which only has a complete missing effect on some pixels, and some original information is retained for some pixels. The decoder tail output structure is an image input structure. The size can make the coders and decoders connected together in series or in parallel. Combined with the residual structure, the content of each codec to be restored can be reduced, making it easier to be trained to generate hidden variable features. The model based on VAE (variable autoencoder) [20] uses an encoder to create an average vector and a standard deviation vector, which are combined to form a late vector to the decoder. Based on VQ-VAE [21], discrete hidden variables are generated. We use a parallel codec to generate latent variables, forming a transformer like multi head attention mechanism, so that each self encoder allows different recovery feature concerns.

3. As for the convolution layer, we generally increase the convolution core size, increase the number of layers, perform the convergence operation before the convolution layer, or use hole convolution. We superimpose multiple \(1^1\), \(3^3\) small convolutions behind each large convolution, so that the receptive field of each local convolution layer reaches \(3^7 = 2187\), far exceeding the width and height of the image, so that the convolution feature map can effectively restore the image after learning the global state of the image.
4. A super-resolution reconstruction network is added at the end of the codec. Even if the image output by the codec is slightly blurred, the image resolution can be improved through the super-resolution reconstruction network. The two networks cooperate with each other to make the restored image clearer.

In the encoding phase, the characteristic graph is updated by down sampling convolution with step size of 2, as in

$$\psi_{\text{fd}}(l) = f_{\text{down}}(\psi_f(l))$$  \tag{5}

In the decoding stage, the characteristic graph is updated by convolution up sampling with a step of 2, as in

$$\psi_{fu}(l) = f_{\text{up}}(\psi_f(l))$$  \tag{6}

Up-sampling function $f_{up}$ its parameters need to be adjusted in the training phase through device convolution. The decoding layer is very important. If the traditional nearest neighbor interpolation method is used, the performance of the model will be greatly reduced.

The AutoEncoder includes the encoding phase and the decoding phase, and the overall input is $\psi_{in}(l)$, as in

$$\psi_{in}(l) = f_{\text{concat}}(\psi_{in}(l), f_{\text{up}}(f_{\text{down}}(\psi_{in}(l))))$$  \tag{7}

Before super-resolution reconstruction of the model layer and by laser damaged image $\psi_{in}(l)$ parallel input to multiple codecs to get an output of $\psi_{out}(l)$. Transmit $\psi_{out}(l)$ and input to the current super-resolution reconstruction network to obtain $\phi(l)$, as in

$$\psi_{out}(l) = f_{\text{concat}}(\psi_{in}(l), \psi_{out}(l))$$  \tag{8}

$l$ indicates the number of codecs currently connected in parallel. $f_{\text{concat}}$ is an operation for connecting feature graphs. That is, the input of the current super-resolution reconstruction also includes the output of the previous super-resolution reconstruction network.

We use $f_{vm}$ represents an effective multidimensional learning reconstruction process, i.e. deep learning convolution mode. Then, the feature offset is learned through input feature mapping $\phi(l)$, as in

$$\psi_{vm}(l) = f_{\text{vm}}(\psi_f(l))$$  \tag{9}

$\psi_{vm}(l)$ represents the deformation characteristic diagram of $\psi_f(l)$.

After the combination of multiple codecs and super-resolution reconstruction network, the output characteristic diagram of the module coding phase is calculated as in

$$\hat{\psi}_{vm}(l) = f_{\text{concat}}(\psi_{vm}(l-1), f_{\text{concat}}(\psi_{in}(l), f_{\text{up}}(f_{\text{down}}(\psi_{in}(l)))))$$  \tag{10}

Invalid mask pixels encountered in general image restoration: after multiple convolution layers, the damaged areas gradually fuse, and it is difficult to obtain an accurate area mask from the original mask. For laser interference. The proposed model takes the synthetic laser interference image at random position as the input, and normalizes the parameters layer by layer at the same time, so that most of the inputs of the neural network layer are in an unsaturated state to avoid the problem of gradient disappearance, as in

$$\hat{x}(k) = \frac{x(k) - E[x(k)]}{\sqrt{\text{Var}[x(k)]} + \epsilon}$$  \tag{11}

$$y(k) = \gamma(k)\hat{x}(k) + \beta^{(k)}$$  \tag{12}

where $x^{(k)}$ represents the network input of layer $k$, $y^{(k)}$ indicates the normalized network layer output is used. $E[x^{(k)}]$ and $\text{Var}[x^{(k)}]$ refer to the expectation and variance of each dimension of $x^{(k)}$ on the whole training set. In this way, the optimized terrain of the neural network is smoother and the gradient is more stable, which allows the use of a larger learning rate and improves the convergence speed.
IV. EXPERIMENTS
We evaluated our proposed model on five data sets, including Celeba dataset [22], aircraft data set from the network, satellite dataset [23], facades building dataset [24], Stanford automobile dataset [25], and used corresponding synthetic laser interference data sets for model training and restoration test. Because the position and state of laser interference are uncertain, we have generated different kinds of laser interference composite images for each original image, forming the state of an original image corresponding to multiple interference images. Randomly select the interfering image and its original image for model training. It is meaningful to randomly generate interference images required for training, which increases the robustness of model parameters and the adaptability to laser interference in different scenes. The interference images used in model verification are also random interference images that are not available in the training set.

The countermeasure network or the discrimination network [26] is only useful during training, and may not appear
during model reasoning after training. Its function is to assist the training of the codec generator. During training, when our codec generator \( G(z; \theta) \) recovers the laser interference image, it can input the recovered image to the countermeasure network or the real image to the generator.

In the training, after the generator is used to repair the image, the repaired image and the real image are mixed into a mixed image data set containing the real image or the model generated image. The data set is handed over to the discriminator for judgment, and all the discrimination results are counted. The loss function can be defined as cross entropy discriminator for judgment, and all the discrimination results are counted. The loss function can be defined as cross entropy

\[
\mathcal{L}_{GAN}(G, D) = -(\mathbb{E}_x[y \log p(y = 1 \mid x) \\
+ (1 - y) \log p(y = 0 \mid x))
\]

The discriminator will be able to accurately determine the source of the image. At the same time, we constantly adjust the parameters \( \theta \) of the generator network, so that when the image \( G(z; \theta) \) generated by the model \( G \) is input to the discriminator \( D \), the statistical error of the final discriminator is the largest, which is recorded as:

\[
\max_{\theta} \mathbb{E}_{z \sim p(z)} [\log (D(x = G(z; \theta) ; \phi))] \\
\min_{\theta} \mathbb{E}_{z \sim p(z)} [\log (1 - D(x = G(z; \theta) ; \phi))]
\]

Then the whole training process of the image restoration model \( G \), that is, the parameter adjustment of the discriminator and generator, can be recorded as:

\[
G^* = \arg \min_{\theta} \max_{\phi} \mathcal{L}_{GAN}(G, D)
\]

For the loss function adjustment method, we use the adaptive method to accelerate its convergence speed in the early training stage: alternately train the generator and the discriminator, combines RMSprop algorithm, Momentum method and AdaDelta algorithm. The learning rate is adjusted by calculating the exponential decay average \( \Delta X_{t-1}^2 \) of the square of the gradient \( \Delta \theta_{t-1} \). At the same time, the exponential weighted average \( G_t \) of the square of the gradient \( g_t \) and the weighted average of the gradient \( M_t \) are calculated (similar to the momentum method) as in

\[
\Delta X_{t-1}^2 = \beta_1 \Delta X_{t-2}^2 + (1 - \beta_1) \Delta \theta_{t-1} \circ \Delta \theta_{t-1}
\]

\[
M_t = \beta_1 M_{t-1} + (1 - \beta_1) g_t
\]

\[
G_t = \beta_2 G_{t-1} + (1 - \beta_2) g_t \circ g_t
\]

Then correct the deviation as in

\[
\hat{G}_t = \frac{G_t}{1 - \beta_1^t}
\]

\[
\hat{M}_t = \frac{M_t}{1 - \beta_1^t}
\]

The final parameter update difference is:

\[
\Delta \theta_t = -\frac{\sqrt{\Delta X_{t-1}^2 + \epsilon}}{\hat{G}_t + \epsilon} \hat{M}_t
\]

This method not only has the acceleration effect of momentum method at the initial stage of iteration, but also has the deceleration effect of the oscillation of convergence value at the later stage of iteration. The initial learning rate \( \alpha \) is also changed to \( \sqrt{\Delta X_{t-1}^2} \) calculated dynamically to stabilize the fluctuation of learning rate. In the iterative process of model training, the learning rate of each parameter does not show a declining trend, which can be reduced or increased.

At the later stage of training, the more traditional small batch random gradient descent method is used to ensure that it accurately falls into the optimal solution. The learning rate

![FIGURE 3. We propose a new model training method, which loss function training records and PSNR results of various data sets. From the first column to the fifth column are: (a) Stanford automobile data set, (b) Satellite data set, (c) Facades building data set, (d) Airplanes dataset, and (e) Celeba dataset.](image-url)
is set to 0.0003, and the momentum parameter $\beta_1 = 0.5$, $\beta_2 = 0.999$. On a single Nvidia RTX3090 GPU (24GB), the 1000m size parameters of the complete neural network are trained, and the applicable input size is $512 \times 512$, all datasets were trained within 2 days, and their loss functions were recorded as Fig. 3. We visualize the joint output of the

**FIGURE 4.** Visualization of training process. The first to fifth lines show the learning process of laser interference caused by randomly selecting synthetic images to train the model in (a) Stanford automobile data set, (b) Satellite data set, (c) Facades building data set, (d) Airplanes data set, and (e) Celeba dataset.

**FIGURE 5.** Comparison of results of airplanes dataset. From the first row to the third row are: (a) the laser interference synthetic image, (b) the reconstruction and restoration results of our model, and (c) the original image.
FIGURE 6. Comparison of results on facades building dataset. From the first row to the third row are: (a) The laser interference synthetic image, (b) The reconstruction and restoration results of our model, and (c) The original image.

FIGURE 7. Comparison of results on the LEVIR-CD satellite photo data set. From the first row to the third row are: (a) the laser interference synthetic image, (b) the reconstruction and restoration results of our model, and (c) the original image.

codec and the super-resolution reconstruction network during the training process to analyze our method. In Fig. 4, it can be shown that in the initial stage of training, the existence of laser spot interferes with almost every pixel of the image, and the image has damaged areas globally.

There are not only color differences and artifacts, but also some features that are difficult to describe subjectively for human beings, which do not exist in the original image and the interference image. As the number of iterations increases, the color difference and artifact gradually weaken, the laser edge part gradually keeps consistent with the real pixel value, part of the laser interference center is perceived by the model, and the fusion effect through the convolution layer is constantly changing. The model continuously performs pixel conversion, the damaged area is reconstructed, and the undamaged area is preserved. The final restoration results make the features of the laser spot almost disappear, and the target features become more and more visible.

Accurate restoration of the randomly occluded part of the image and simple uniform interference area [27] are a difficulty in image restoration. Some restoration results based on deep learning [28], [29] do not use the full reference image.
quality evaluation method to evaluate the restoration results, and it indicates that [8] does not have a good numerical measurement to evaluate the image restoration results. Although the restoration results are very realistic using subjective evaluation, however, some of them are different from the original images and are not accurate enough. A few, such as [30], used pixel level indicators (including multi-scale structure similarity (ms-ssim) and mean absolute error (MAE)) to measure the reconstruction results. In Fig. 5-10, our reconstruction results are given. Under various non-stationary complex scenes,
there is no obvious color difference between the original image and the original image, no major restoration failure occurs, and the semantics are consistent with the original image. However, some reconstruction results have slight artifacts and edge responses around the laser interference, such as [31]. Our model produces reasonable restoration results in repeated structure and random interference in various complex scenes, and is suitable for different types of laser spots. Even if the laser spots cause irregular crosstalk in the image, they cause irregular and nonstationary interference in the image.

We used a series of evaluation indicators to evaluate the quality of reconstructed images, including SSIM [32], MSE and PNSR. Results as shown in the table, all the full reference image evaluation methods are close to the full score, and all the features of the reconstructed image are very close to the original image. In terms of subjective evaluation method, the restoration result of the image disturbed by the laser is very good, that is, it is realistic enough to completely restore the characteristics of different types of targets, and we can hardly see the interference of the original laser spot in the reconstructed image, as shown in Fig. 5. The pixel level reconstruction and restoration results we adopted can accurately restore the effect to the specific value of each pixel, resulting in a variety of high-frequency details. The full reference image restoration method uses the value of the pixel as the basic parameter to evaluate the image quality, so as to evaluate whether the reconstruction result is close to the original image. The evaluation results show that the reconstruction results are close to the original image at the pixel level. And if the MSE evaluation method is used to get an evaluation result with small difference, any other evaluation method will get a better result, because MSE directly counts the difference between the reconstructed image and the original image.

According to the current viewpoint [30], due to the instability of convolution features in the network learning and reasoning stage, it is very difficult to reconstruct directly based on convolution. Standard convolution is not suitable for image repair tasks, which may lead to failure, including defects, structure distortion and texture blur, especially in the case of complex background or huge missing areas. However, our direct convolution model plus super-resolution reconstruction network has overcome these difficulties in the case of laser interference through experimental verification, and only slight distortion and blurring results have appeared, showing a strong ability to restore semantic content and high-quality details in the case of laser interference.

### V. CONCLUSION

We propose a laser interference image restoration model system that combines and sends from the encoder and the resolution reconstruction network. After training with the countermeasure network, the image restoration task of laser interference in the field of view can be carried out. The extendible concurrent module is designed to solve the problem of rapid and stable learning of image features in convolution stage. The resolution reconstruction module helps to solve the fuzzy output problem of the self encoder. The experimental results show that without pre training, our large-scale network architecture has the characteristics of rapid convergence, and achieves excellent restored image quality evaluation results with different light spot interference. In the future, we will explore the parallel VQ-VAE and 3D attention transformer mechanism for a wide range of optical image processing.

### REFERENCES

[1] M. Bertalmio, “Image inpainting,” in Proc. 27th Annu. Conf. Comput. Graph. Interact. Techn., 2000, pp. 417–424.
X. Zhou et al.: Restoration of Laser Interference Image Based on Large Scale Deep Learning

[2] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera, “Filling-in by joint interpolation of vector fields and gray levels,” IEEE Trans. Image Process., vol. 10, no. 10, pp. 1200–1211, Aug. 2001.

[3] Y. Zheng, Q. Shen, M. Wang, M. Yang, J. Huang, and C. Feng, “Semantic segmentation sample augmentation based on simulated scene generation: Case study on dock extraction from high spatial resolution imagery,” Int. J. Remote Sens., vol. 42, no. 13, pp. 4961–4984, Jul. 2021.

[4] A. A. Efros and W. T. Freeman, “Image quilting for texture synthesis and transfer,” in Proc. 28th Annu. Conf. Comput. Graph. Interact. Techn., 2001, pp. 341–346.

[5] A. A. Efros and T. K. Leung, “Texture synthesis by non-parametric sampling,” in Proc. IEEE Int. Conf. Comput. Vis., vol. 2, Sep. 1999, pp. 1033–1038.

[6] C. Barnes, “The generalized patchmatch correspondence algorithm,” in Proc. Eur. Conf. Comput. Vis. Berlin, Germany: Springer, 2010, pp. 29–43.

[7] D. Mahajan, “Exploring the limits of weakly supervised pretraining,” in Proc. Adv. Neural Inf. Process. Syst. Berlin, Germany: Springer, 2018, pp. 1438–1447.

[8] 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.

[9] G. Liu, “Image inpainting for irregular holes using partial convolutions,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 4471–4480.

[10] T. Yu, “Region normalization for image inpainting,” in Proc. AAAI Conf. Artif. Intell., 2020, vol. 34, no. 7, pp. 12733–12740.

[11] R. Duan, X. Mao, A. A. Qin, Y. Chen, S. Ye, Y. He, and Y. Yang, “Adversarial laser beam: Effective physical-world attack to DNNs in a blink,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 16062–16071.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhou, 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.

[13] A. Vaswani, “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 1–11.

[14] D. Mahajan, “Exploring the limits of weakly supervised pretraining,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 181–196.

[15] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, and R. Girshick, “Masked autoencoders are scalable vision learners,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 16000–16009.

[16] C. Wu, J. Liang, L. Ji, F. Yang, D. Jiang, and N. Duan, “NUWA: Visual synthesis pre-training for neural visual world creation,” 2021, arXiv:2111.12417.

[17] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. Cham, Switzerland: Springer, 2015, pp. 234–241.

[18] Z. Yan, “Shift-Net: Image inpainting via deep feature rearrangement,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 1–17.

[19] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” 2013, arXiv:1312.6114.

[20] A. Van Den Oord and O. Vinyals, “Neural discrete representation learning,” in Proc. Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 1–10.

[21] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 3730–3738.

[22] H. Chen and Z. Shi, “A spatial–temporal attention-based method and a new dataset for remote sensing image change detection,” Remote Sens., vol. 12, no. 10, p. 1662, May 2020.

[23] Y. Zheng, Q. Shen, M. Wang, M. Yang, J. Huang, and C. Feng, “Semantic segmentation sample augmentation based on simulated scene generation: Case study on dock extraction from high spatial resolution imagery,” Int. J. Remote Sens., vol. 42, no. 13, pp. 4961–4984, Jul. 2021.

[24] I. Goodfellow, “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst., vol. 27, 2014, pp. 1–15.

[25] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, “3D object representations for fine-grained categorization,” in Proc. IEEE Int. Conf. Comput. Vis. Workshops, Dec. 2013, pp. 554–561.

[26] S. Iizuka, E. Simo-Serra, and H. Ishikawa, “Globally and locally consistent image completion,” ACM Trans. Graph., vol. 36, no. 4, pp. 1–14, Jul. 2017.

[27] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1125–1134.

[28] C. Zheng, T.-J. Cham, and J. Cai, “Pluralistic image completion,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 1438–1447.

[29] N. Wang, Y. Zhang, and L. Zhang, “Dynamic selection network for image inpainting,” IEEE Trans. Image Process., vol. 30, pp. 1784–1798, 2021.

Xiangyu Zhou received the B.S. degree in astronomy from the University of Science and Technology of China, Anhui, China, in 2018. He is currently pursuing the M.S. degree with the College of Advanced Interdisciplinary Studies, National University of Defense Technology, under the supervision of Prof. Xiangai Cheng. His research interests include deep learning and computer vision.

Zhongjie Xu received the Ph.D. degree in optical engineering from Tsinghua University. He is currently an Associate Professor with the College of Advanced Interdisciplinary Studies, National University of Defense Technology. His current research interest includes light-matter interaction.

Xiangai Cheng received the B.S. and Ph.D. degrees from the National University of Defense Technology. She is currently a Professor in optical engineering with the College of Advanced Interdisciplinary Studies, National University of Defense Technology. Her current research interest includes light-matter interaction.

Zhongyang Xing received the B.S. degree in optical engineering from the National University of Defense Technology, China, and the Ph.D. degree in physics from the University of Cambridge, U.K., in 2014 and 2019, respectively. She is currently an Associate Professor with the College of Advanced Interdisciplinary Studies, National University of Defense Technology, China. Her current research interests include light-matter interaction, laser technology, and image information processing.

* * *