Research Article

ADVANCED PHOTONICS NEXUS

Teacher-student learning of generative adversarial network-guided diffractive neural networks for visual tracking and imaging

Hang Su,[a,b,†] Yanping He,[a,b,†] Baoli Li,[a,b] Haitao Luan,[a,b] Min Gu,[a,b] and Xinyuan Fang[a,b,*]
[a]University of Shanghai for Science and Technology, School of Artificial Intelligence Science and Technology, Shanghai, China
[b]University of Shanghai for Science and Technology, Institute of Photonic Chips, Shanghai, China

Abstract. Efficiently tracking and imaging interested moving targets is crucial across various applications, from autonomous systems to surveillance. However, persistent challenges remain in various fields, including environmental intricacies, limitations in perceptual technologies, and privacy considerations. We present a teacher-student learning model, the generative adversarial network (GAN)-guided diffractive neural network (DNN), which performs visual tracking and imaging of the interested moving target. The GAN, as a teacher model, empowers efficient acquisition of the skill to differentiate the specific target of interest in the domains of visual tracking and imaging. The DNN-based student model learns to master the skill to differentiate the interested target from the GAN. The process of obtaining a GAN-guided DNN starts with capturing moving objects effectively using an event camera with high temporal resolution and low latency. Then, the generative power of GAN is utilized to generate data with position-tracking capability for the interested moving target, subsequently serving as labels to the training of the DNN. The DNN learns to image the target during training while retaining the target's positional information. Our experimental demonstration highlights the efficacy of the GAN-guided DNN in visual tracking and imaging of the interested moving target. We expect the GAN-guided DNN can significantly enhance autonomous systems and surveillance.

Keywords: visual tracking; diffractive neural network; generative adversarial network; teacher-student learning; event-based camera; optical machine learning.

Received Sep. 16, 2024; accepted for publication Oct. 17, 2024; published online Nov. 8, 2024.

[DOI: 10.1117/1.APN.3.6.066010]

1 Introduction

Visual tracking, as a pivotal focus within the domain of computer vision, has found extensive applications across diverse fields, such as surveillance,[1] autonomous driving,[2] and human-computer interaction.[3,4] The goal of visual tracking is to continuously estimate the position, shape, and possibly other attributes of the target over time when it moves through a scene.[5,6] Therefore, visual tracking is a challenging task in computer vision, requiring robust algorithms capable of addressing complexities associated with object appearance variations, object position changes,[7] dynamic environmental conditions, etc. However, large amounts of data and increasingly complex algorithms require high power consumption and significant processing time.[8–10]

Event cameras, also known as neuromorphic or dynamic vision sensors, are responsive to changes in brightness rather than capturing frames at fixed intervals compared with traditional cameras.[11] The sparse data output of event cameras[12] contributes to efficient data transmission and storage. Moreover, their temporal sensitivity allows for real-time responses to dynamic scenes, making them particularly suitable for applications requiring quick reactions, such as robotics and autonomous vehicles.[13] In addition, the high dynamic range and robustness of motion blur enable effective performance in diverse lighting conditions and fast-paced environments.[14] The sparse data output of event cameras[12] contributes to efficient data transmission and less data storage. The integration of event cameras in visual tracking systems signifies a promising direction for advancing

[*]Address all correspondence to Xinyuan Fang, xinyuan.fang@usst.edu.cn

[†]These authors contributed equally to this work.

energy-efficient and high-performance visual perception technologies in various fields.[15]

Currently, in addressing intricate tasks, such as object detection and motion tracking, data processing technology predominantly relies on machine learning algorithms, especially convolutional neural networks (CNNs).[16–19] However, conventional machine learning algorithms typically demand substantial labeled data and encounter challenges in preserving the positional information of dynamic targets.[20,21] In contrast, operating within an unsupervised learning paradigm, generative adversarial networks (GANs) adeptly leverage unlabeled data, effectively mitigating issues associated with constrained labeled datasets.[22,23] GANs demonstrate exceptional proficiency in generating authentic data samples across diverse domains, encompassing tasks such as image generation,[24–26] style transfer,[27,28] and data augmentation.[29–32] This inherent capability positions GANs as a promising solution for tasks requiring nuanced handling of dynamic and unlabeled datasets, thereby contributing to the advancement of machine learning methodologies.

Moreover, conventional electronic computing is constrained by slower information transmission, connectivity complexities, and energy consumption concerns. In recent developments, a novel diffractive neural network (DNN) architecture has been introduced, which can accomplish a diverse array of complex functions that can be achieved by computer-based neural networks in an all-optical way.[33–37] Hence, this groundbreaking framework demonstrates the capacity to perform machine learning operations at the speed of light while maintaining low energy consumption, which has found extensive applications in diverse domains, including but not limited to handwritten digit recognition,[38,39] action recognition,[40–43] imaging,[44–48] and data encryption.[49–51] In addition, there is a notable progression in the development of integrated and scalable DNN-based chips characterized by ultra-compact dimensions and reduced energy consumption.[52–56]

In this work, we demonstrate a GAN-guided DNN, which performs visual tracking and imaging of the interested moving target. The main contributions and innovative aspects of our work are as follows: (1) Leveraging an event camera for effective moving object capture. Event cameras for effective moving object capture offer significant advantages, including high-speed, low-latency performance, efficiency, and robustness to varying lighting conditions. These attributes make the event camera a valuable tool for visual tracking tasks, particularly in dynamic or challenging environments; (2) using GANs to generate labels for DNN training; and (3) pioneering the application of DNNs in visual tracking and imaging. We further harness the generative power of the GAN-based teacher model to produce data endowed with position-tracking capabilities for the interested moving target. This generated data is subsequently employed to provide labels for the DNN-based student model, which learns to selectively image the interested moving target during training while retaining the target's positional information. Traditionally, GANs have been employed primarily as output terminals for generating style or shape images. Here, using GANs to guide and generate labels for DNN training presents a novel and efficient solution to the problem of insufficient DNN training datasets. Notably, to the best of our knowledge, this is the first instance of utilizing DNNs in visual tracking and imaging. This approach effectively combines the strengths of both GANs and DNNs, offering a powerful and innovative methodology. Then, we apply the GAN-guided DNN to track and selectively image only a car or pedestrian in a complex scene containing a car, a pedestrian, and a tree. Our results show the successful imaging and tracking of only the interested moving target within this dynamic context. In addition, we extend our investigation to the experimental demonstration, which highlights the efficacy of GAN-guided DNN in visual tracking and imaging only the airplane in a scenario involving airplanes and missiles. Our work has also proven effective in challenging environments where traditional methods often struggle, such as in low-light conditions or complex backgrounds with inferring objects. Furthermore, the results obtained with GAN-guided DNNs, compared to those without GAN guidance, clearly show that integrating GANs significantly enhances the performance of DNNs in visual tracking and imaging.

2 Theory and Results

2.1 Overview of GAN-Guided DNNs

Figure 1(a) presents the entire pipeline of a GAN-guided DNN, with a time-series inputs derived from event-based video acquired by an event camera. The GAN-guided DNN is trained to selectively image the morphology and track the trajectory of the interested moving target within the input data. Here, we use the example of cars and pedestrians in a traffic scenario. As shown in Fig. 1(a), the inputs are images of moving cars (the interested target) and pedestrians, and the output of this GAN-guided DNN portrays only the selectively imaged representation of the car, enriched with positional information. This approach to selectively imaging and tracking the desired target not only enhances the efficiency of image storage but also alleviates the burden on transmission resources.

The training process of the GAN-guided DNN is demonstrated in Fig. 1(b). We first train the GAN-based teacher model to have the capability of selectively keeping the morphology and trajectory of the interested target, which takes the car as an example in Fig. 1(b). The inputs to the GAN-based teacher model consist of time-series data capturing the movement of cars and pedestrians, obtained from event-based videos recorded by an event camera, while the sample of the GAN-based teacher model is the target car without positional information produced by numerical simulations. After training on a limited dataset, the GAN-based teacher model demonstrates the capability to generate the output target car with associated positional information for previously unseen inputs. Then, the GAN-based teacher model's output is utilized as the labels for the subsequent training of the DNN-based student model. Incorporating corresponding time-series inputs, the trained DNN-based student model has the ability to selectively image and track the target.

2.2 Acquisition of Training Dataset and Testing Set

To obtain the training and test datasets for the GAN-guided DNN, the event-based videos with different dynamic scenarios are encoded into still images for the selective imaging task as shown in Fig. 2(a). An event-based video is a stream of events composed of numerous events, an event is a vector, which can be expressed as[14,57–61]

$$\vec{\varepsilon}_i = (x_i, y_i, t_i, p_i), \tag{1}$$

where t_i is the time of occurrence, (x_i, y_i) is the pixel coordinate at the time of the event, p_i represents the polarity, and there are

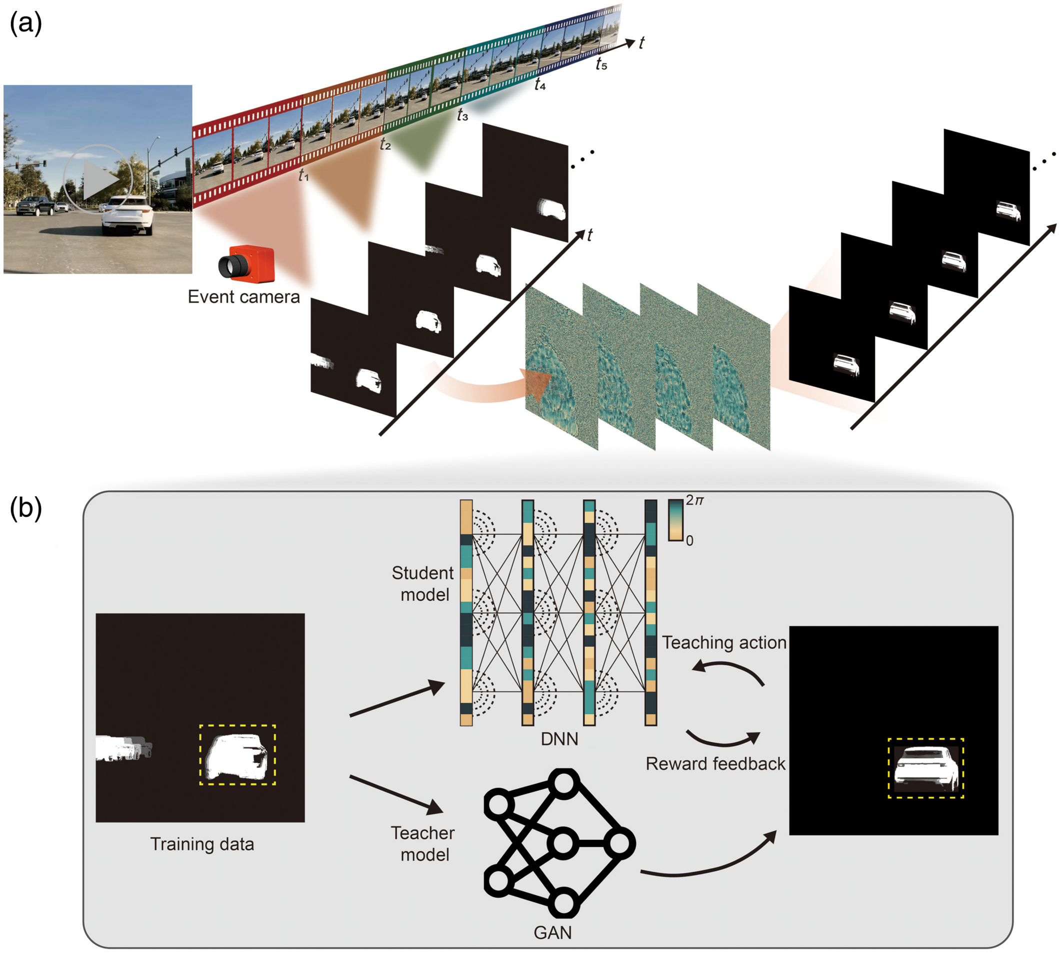

Fig. 1 The overall working principle of the GAN-guided DNN. (a) GAN-guided DNN for visual tracking and imaging of the interested moving target. (b) The training process of the GAN-guided DNN.

three values of p_i: $p_i = +1$ (increase in brightness), $p_i = -1$ (decrease in brightness), and $p_i = 0$ (no change in brightness). The vector can be described as follows: the pixel (x_i, y_i) has changed in some way at the time of the event t_i, and as soon as a pixel (x_i, y_i) perceives a change in brightness, an event is returned. Thus, throughout the field of view of the camera, whenever a pixel perceives a change in brightness, an event is returned, and since the times are all different, these events all occur asynchronously.

Since the optical diffraction neural network cannot process the event-based video information captured by the event camera, the information should be converted into a two-dimensional gray-scale image using dimensionality compression. In our work, we encode the polarity information of the event vectors according to Eq. (2), and then the increase and decrease of pixel brightness are represented as two different gray levels on the grayscale image:

$$\varepsilon_i(x_i, y_i, t_i, p_i) = \begin{cases} 255, & p_i = +1 \\ 0, & p_i = -1 \text{ or } p_i = 0 \end{cases}. \tag{2}$$

By stacking all the events over a period of time, we obtain an image that contains information about the motion of the target object over that period of time, which can be represented as an image:

$$E_i(x_i, y_i) = \sum_{t_i} \varepsilon_i(x_i, y_i, t_i). \tag{3}$$

Based on the previous core steps of acquiring the training and test datasets, when an event camera is used to capture the dynamic action, a noise filter is set on the appropriate software for denoising, and then the event-based video is converted into an image using the stacking-based on-time method. More details about the event camera can be found in Note 1 in the Supplementary Material.

Using this approach, we then obtained the dataset for training and testing the GAN-guided DNN. A total of three scenarios are captured for training and simulation testing of our GAN-guided DNN on the computer, including 600 images for training and 120 images for testing. In addition, there is another scenario used to test our DNN through experimental light paths, which consists of 480 and 120 images for training and testing sets, respectively. All images are 512 pixel × 512 pixel arrays with a pixel pitch of 12.5 μm.

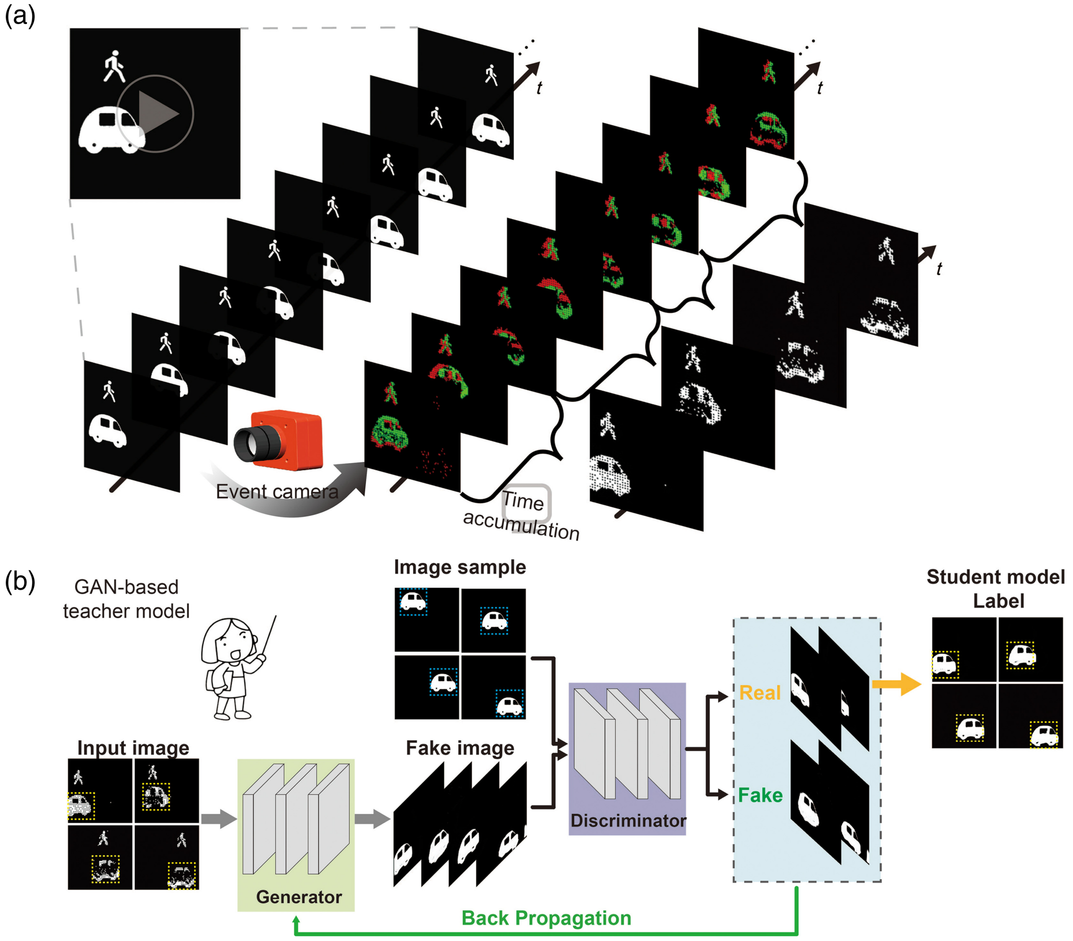

Fig. 2 The process of training the GAN-based teacher model, which involves acquiring datasets and optimizing models. (a) The principle of input dataset acquisition using the event-based camera. (b) The architecture of the GAN-based teacher model.

2.3 Principle and Architecture of the GAN-Based Teacher Model

Position information plays a vital role in the domain of visual tracking. However, conventional neural networks applied for target imaging frequently neglect such crucial positional details, which mainly arises from the inherent difficulty associated with obtaining precise positioning information when imaging a target with an uncertain location. Hence, owing to the absence of precise location information, conventional DNNs are trained without incorporating target labels that contain positional information. In our work, a GAN-based teacher model is used to generate the target labels with positional information. Figure 2(b) shows the training process of the GAN-based teacher model framework, which contains two fundamental components, a generator and a discriminator. The GAN-based teacher model constructs a mapping function between two domains by learning the feature information from two sets of images and attempts to convert one set of images into another using a generator.[62–66] As shown in Fig. 2(b), by allowing the GAN-based teacher model to learn the features of the input dataset (images involving pedestrians and cars taken by an event camera) against another image sample dataset (computer-generated images of cars), the GAN-based teacher model is able to produce an image of a fake target (a car) at the corresponding position for a given input image.

The architecture of the generator implemented in the GAN-based teacher model consists of inputs and outputs with multiple convolutional layers. To reduce the size of the input information and extract shallow low-level features, downsampling is performed on the input images. The presence of the nonlinear activation function ReLU causes irreversibility in both input and output, resulting in information loss in the model. Moreover, the deeper layers of the network use an increased number of ReLU functions, leading to further information loss. This makes it difficult to preserve the shallow features with the propagation of the antecedent terms. To address this issue, a residual block is introduced in the intermediate convolutional layers. This block is directly incorporated into the original network with a cross-layer connection to preserve information integrity and simplify the learning objective and difficulty. The convolution is then used to up-sample and recover the size of the input information while extracting deeper advanced features. The final output image is then derived from this process.

The discriminator model in the GAN-based teacher model aims to differentiate between authentic and generated images.

The generated images are inputted into the discriminator, undergoing compression into a discriminant matrix through multiple convolutional layers. The discriminant matrix determines the quality of the generator's image, aiding the optimization process (see Sec. 4 for details). Through iterative training, the generator refines its output to become increasingly indistinguishable from real data, while the discriminator enhances its ability to make accurate distinctions.

2.4 Imaging and Tracking Interested Moving Targets Using GAN-Guided DNNs

We first present a simulated numerical demonstration of the GAN-guided DNN within scenarios involving a pedestrian and a car in typical encounters in traffic, with a specific emphasis on car imaging and tracking. As illustrated in Fig. 3(b), a four-layer DNN-based student model with phase-only modulation is trained, which is learned as a result of this data-driven training approach. The phase modulation range for each diffractive layer spans from 0 to 2π. Each diffractive layer contains 512×512 neurons, leading to a total number of trainable neurons $N = 4 \cdot Ni = 1{,}048{,}576$, each with a size of 12.5 μm. The axial distance between any two consecutive planes is set as 150 mm. The trainable neurons are subject to optimization through the iterative process of minimizing the mean square error loss (see Sec. 4 for details), which aims to enhance the performance of GAN-guided DNN to image and track the interested moving target.

Following training, the GAN-guided DNN model undergoes numerical testing utilizing grayscale images involving a pedestrian and a car, which are represented in amplitude within the range of 0 to 1. Figure 3(a) presents a subset of the examples utilized for testing, which includes inputs, labels, and outputs. The output images are depicted as light-intensity maps following normalization. It is crucial to highlight that the deficiency of information in images due to rapid motion can be effectively addressed through the implementation of the GAN-guided DNN. Figure 3(c) illustrates the quantitative evaluation results of the GAN-guided DNN model, wherein structural similarity index (SSIM) values and peak signal-to-noise ratio (PSNR) values are computed through a comparison between the outputs and the labels across 40 frame. The SSIM values exhibit a fluctuation hovering around 0.9, all of which surpass the threshold of 85. So the output images can be considered very similar to the labels from the perspective of structural similarity (closer to human eye recognition), which includes brightness, contrast, and structure. Meanwhile, PSNR values display a similar pattern of fluctuation, with an average value circling 26, which underscores the excellence in signal-to-noise ratio, affirming the quality and precision of the GAN-guided DNN output. The resultant

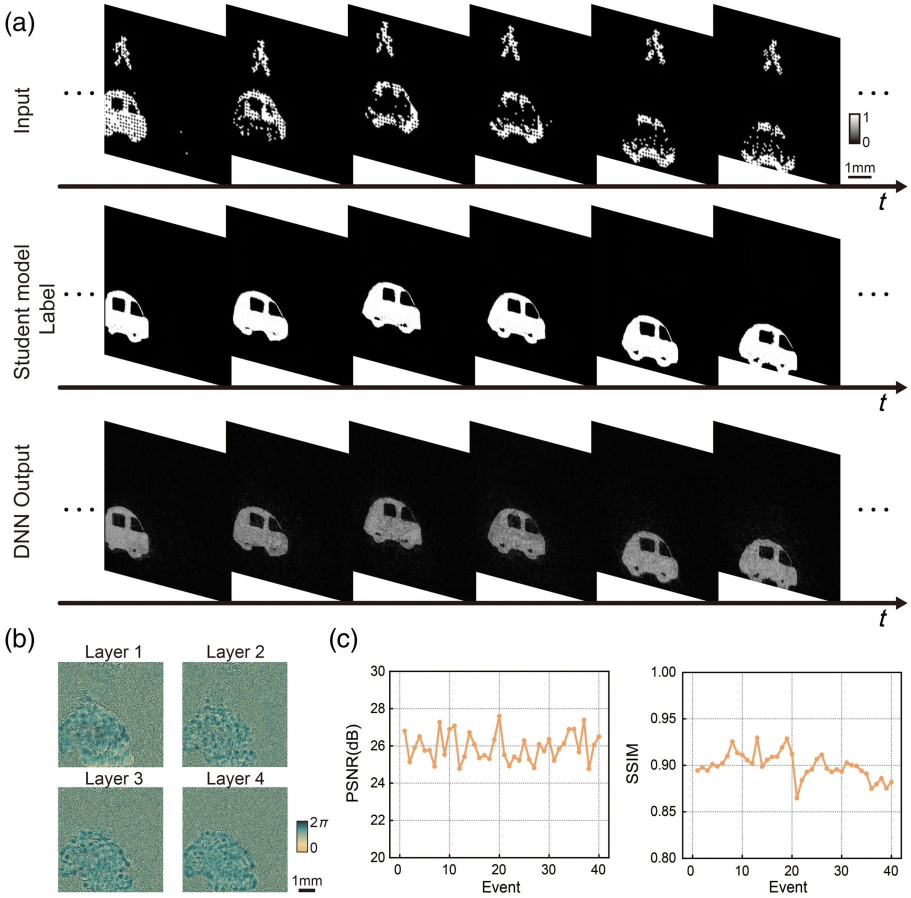

Fig. 3 Simulation results of GAN-guided DNN. (a) Examples of training results for the visual tracking and imaging of the target car. (b) The phase profiles of diffractive layers after deep learning-based optimization. (c) The PSNR and SSIM values with different input images.

output showcases the successful application of the GAN-guided DNN in isolating and tracking the designated moving car, offering a streamlined and resource-efficient solution for processing event-based video data.

To demonstrate the stability of our model, we evaluated its performance in a more complex scenario involving an interfering car with the same speed, profile, and size as the target car. As shown in Fig. S1 in the Supplementary Material, the results confirm that our method remains effective under these challenging conditions. Furthermore, we also evaluate our model under varying lighting conditions to assess its robustness in more general scenarios. As demonstrated in Fig. S2 in the Supplementary Material, the model continues to deliver excellent performance even in difficult environments, enhancing its relevance to real-world applications.

2.5 Quantification of the Performance of the GAN-Guided DNN with Different Numbers of Layers

To quantitatively evaluate the performance of different numbers of layers on the accuracy of imaging and tracking the interested moving target, we systematically trained three GAN-guided DNN models with different layer numbers: $L = 2$, 4, and 6. As shown in Fig. 4, we showcase a scenario with a car, a pedestrian, and trees, where only the target pedestrian is moving, and others are static. Our results demonstrate that all three models effectively and exclusively image and track the target pedestrian.

Then, we conducted a comprehensive comparison of the three models, examining not only their visualizations but also considering metrics, such as SSIM values and PSNR values,

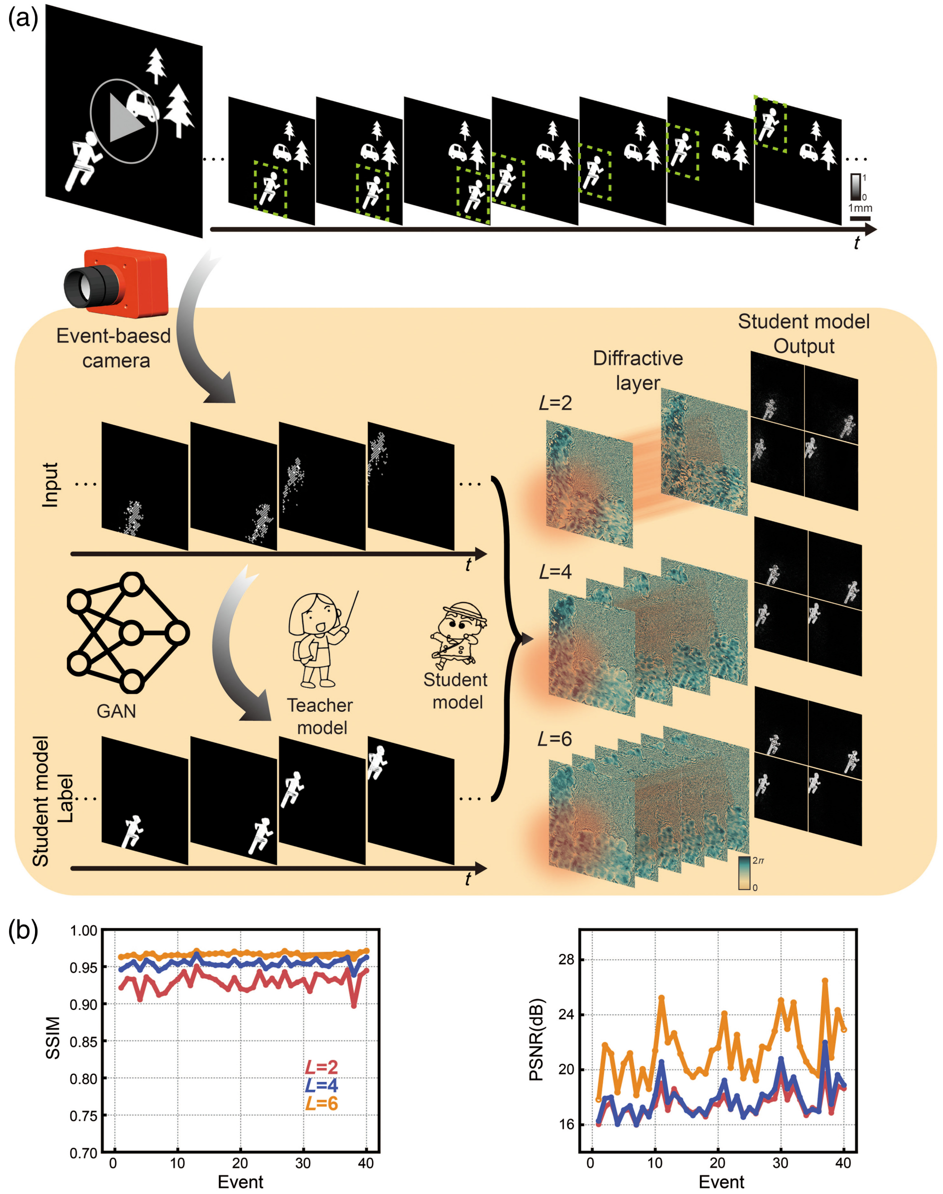

Fig. 4 The GAN-guided DNN trained and tested with different numbers of diffractive layers. (a) The performance of the GAN-guided DNN with different numbers of diffractive layers (*L*) for the scenario with just a pedestrian and some static objects. (b) The SSIM and PSNR values with different input images.

for the outputs. In Fig. 4(b), we can see that SSIM values for the three models are all above 0.9, which signifies a notable level of structural similarity between the compared images. The PSNR values are all larger than 16 dB, which means the quality of output images is very similar. Despite encountering some fluctuations, it is imperative to emphasize that the overall outcomes remain favorable. Our analysis underscores that the two-layer GAN-guided DNN exhibits comparatively lower accuracy in both imaging and tracking. Importantly, our results elucidate a notable enhancement in the performance of the GAN-guided DNN with a substantial increase in the number of diffractive layers.

2.6 Experimental Demonstration

We further experimentally demonstrated the feasibility of the GAN-guided DNN with the experimental setup shown in Fig. 5(a). The incident continuous wave at a wavelength of 632.8 nm is generated by a He-Ne laser (CW, HNL210L, Thorlabs) with a power output of 14.4 mW. A half-wave plate (HWP) and a polarization beam splitter (PBS) are integrated to continuously adjust both the power and polarization of the laser beam. The laser beam is then spatially magnified using the $4f$ system, comprising Lens 1 (OLD1430-T2M, JCOPTIX, China) with a focal length of 30 mm and Lens 2 (OLD2474-T2M, JCOPTIX, China) with a focal length of 500 mm. Concurrently, a 30 μm pinhole is positioned at the Fourier plane of Lens 1 for spatial filtering. To load the amplitude information of the input image on the spatial light modulator (SLM), two quarter-wave plates (QWPs) with orthogonal fast axes are mounted on the front and back of the SLM1 (X13138, Hamamatsu). The next two SLMs (SLM 2 and SLM 3, X13138, Hamamatsu) with a spacing distance of 150 mm are utilized to construct DNNs for visual tracking and imaging of the interested moving target. Finally, the output image was captured by the CCD camera (acA2040-90uc, Basler). Images and details of the laboratory optical setup are provided in the Supplementary Material (see Note 4 and Fig. S3 in the Supplementary Material).

For the experimental validation, we constructed a GAN-guided DNN with two diffractive layers denoted as layer 1 and layer 2, as shown in Fig. 5(b). Each layer is comprised of

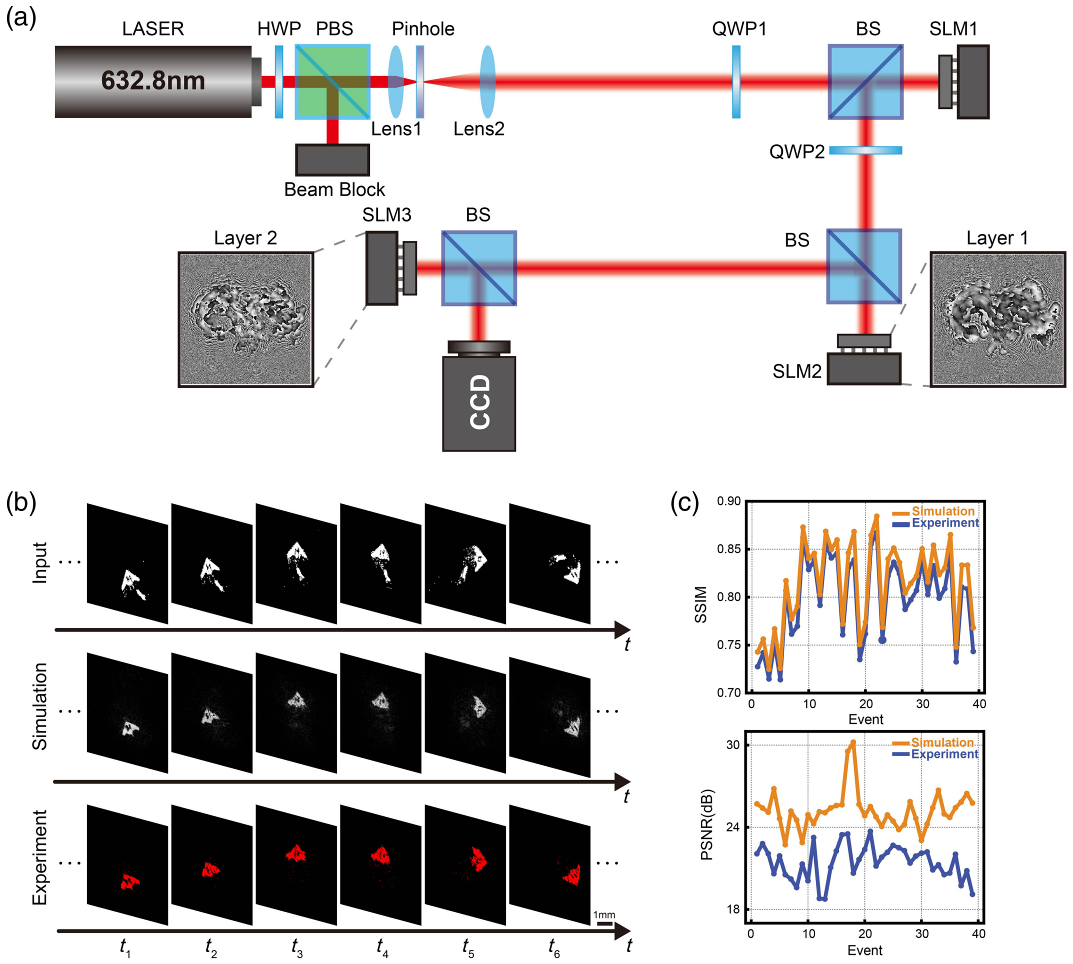

Fig. 5 Experimental demonstration of the visual tracking using a GAN-guided DNN. (a) Schematic diagram of the experimental setup and phase used in the experiment (layer 1 and layer 2 are loaded on SLM 1 and SLM 2, respectively). HWP, half-wave plate; PBS, polarization beam splitter; QWP, quarter-wave plate; BS, beam splitter; SLM, spatial light modulator. (b) Simulation and experimental results of visual tracking and imaging of the target airplane in a scenario involving airplanes and missiles. (c) The SSIM and PSNR values of the simulation and experimental results with different input images.

512×512 diffractive neurons aligned in the forward direction. The distance between the input layer, diffractive layers, and output layer is consistently set at 150 mm. The network is trained to perform the task of visual tracking and imaging of the target airplane in a scenario involving airplanes and missiles.

To test the GAN-guided DNN, 40 different images containing airplanes and missiles are used as the inputs to the network. A comparison of the obtained experimental results with the simulation results is presented in Fig. 5(c). The results show that the experimental measurements are in good agreement with the numerical simulations. The GAN-guided DNN faithfully represents the target (airplane) in the corresponding position while effectively inhibiting the formation of another object (missile) in precise agreement with our numerical simulations. Figure 5(d) shows the SSIM and PSNR of the experimental test results versus the simulated results, which reveals that the results of visual tracking and imaging of the interested moving target in the actual test exhibit a slightly lower performance than expected. Nonetheless, the results still demonstrate a commendable capability to effectively execute tasks of visual tracking and imaging of the interested moving target with a high level of quality.

3 Discussion and Conclusion

We have demonstrated a GAN-guided DNN that performs real-time visual tracking and imaging of the interested moving target. There are three key components in our method: (i) the event camera, acclaimed for its remarkable responsiveness to changes in brightness, facilitates the real-time, energy-efficient, and high-performance capture of dynamic scenes. (ii) The GAN-based teacher model, operating within an unsupervised learning paradigm, enables the generation of labels with position-tracking capability of the interested moving target for DNN-based student model. To demonstrate the necessity of the GAN-based teacher model, we present a comparative analysis between the GAN-guided DNN and a DNN without GAN guidance in the Supplementary Material (see Note 5 and Fig. S4 in the Supplementary Material). (iii) The DNN-based student model, computing at the speed of light, performs real-time visual tracking and imaging of the interested moving target with low energy consumption. Our work demonstrated exceptional performance in tracking fast-moving objects, particularly due to the integration of the event camera's high temporal resolution. The asynchronous data provided by the event camera allowed the model to maintain accurate tracking even when traditional frame-based methods struggled with motion blur or missed detections. Our method also performed robustly in low-light conditions, where traditional cameras typically suffer from noise and poor contrast. The event camera's resilience to varying lighting conditions contributed to stable tracking performance, which indicates the potential for real-world applications where lighting conditions are unpredictable. When tested in environments with different complex backgrounds, our methods showed improved generalization capabilities. However, it may still exhibit overfitting when exposed to environments that are significantly different from those in the training set. In addition, the model may struggle to differentiate between the target and similar objects, leading to potential tracking inaccuracies. To address these challenges, future work could explore domain adaptation or transfer learning techniques to improve the model's robustness across diverse and unseen environments. Techniques such as contrastive learning or data augmentation could further improve the model's ability to recognize and differentiate between similar objects. Incorporating these approaches would help mitigate the risk of overfitting and enhance the model's ability to distinguish between similar objects, ultimately improving its performance in complex and dynamic real-world applications.

Notably, compared to traditional manual methods for obtaining training labels, GANs offer significant advantages in terms of speed, convenience, and improved modeling of data distribution, resulting in sharper and clearer images. Unlike other image generation models, GANs do not require Markov chains for repeated sampling, do not necessitate inference during the learning process, and avoid complex variational lower bounds, thereby sidestepping the challenges of probabilistic approximation computation. In addition, the interplay between the generator and discriminator in GANs is more straightforward to integrate with other networks. The comparison of GAN-guided DNNs and existing visual tracking and imaging solutions has been discussed in Note 6 in the Supplementary Material. Our GAN-guided DNN has been successfully applied to diverse autonomous driving scenarios, which demonstrates an excellent performance with SSIM values of greater than 0.85 and PSNR values of over 16. Furthermore, a notable enhancement has been achieved when we increased the number of diffractive layers. We also apply the GAN-guided DNN for visual tracking and imaging of the ultrahigh-speed moving airplane. Both the simulation and experimental results indicate that our method has good performance in visual tracking and imaging of high-speed moving objects. Moreover, from the results we can see that the lack of information in images caused by high moving speed can be compensated by the GAN-guided DNN.

In summary, we have proposed a GAN-guided DNN for visual tracking and imaging of the dynamic target with high energy efficiency, low power consumption, and ultrahigh speed. The GAN-guided DNN eliminates unnecessary background information and enhances information utilization while reducing information storage costs. It is worth noting that the diffractive layers can be fabricated by two-photon polymerization technology and then assembled together to form a micrometer-scale optical chip. The optical chip operates without consuming any power other than the illumination source, which has the advantages of the speed of light, almost no power consumption, reduced weight and size, as well as robustness to diverse lighting conditions. In addition, if the DNN is used instead of the event camera to collect electrical data, resulting in all-optical computing in memory, it will greatly reduce energy consumption and significantly improve computing efficiency, making it highly applicable in fields such as artificial intelligence and autonomous driving.[67] Our method is anticipated to have a significant impact on various industries, offering a versatile and efficient solution for a wide range of optical data processing applications.

4 Methods

4.1 Numerical Forward Model of the GAN-Guided DNN

The GAN-guided DNN used for visual tracking and imaging consists of L diffraction layers, each followed by propagation through the free space. When the optical wave is incident on a diffractive layer, the transmission coefficient of the trainable diffractive neuron at the position (x, y) of the layer k, t^k, can be expressed as[68–70]

$$t^k(x,y) = a^k(x,y)\cdot \exp[j\phi^k(x,y)], \tag{4}$$

where $j=\sqrt{-1}$ and $a^k(x,y)$ represent the amplitude modulation of diffractive neuron for the diffraction layer modulating only the phase, the amplitude is 1, and $\exp[j\phi^k(x,y)]$ denotes the phase modulation of a diffractive neuron. Any two diffractive layers, including the last layer and the output plane, are connected by free space propagation. The propagation of the optical field is modeled using the angular spectrum method, according to which an optical field after propagation through an axial distance d can be calculated as

$$u(x,y,z+d) = F^{-1}\{F[u(x,y,z)]\cdot H(f_x,f_y;d)\}, \tag{5}$$

where F and F^{-1} represent the 2D Fourier transform and 2D inverse Fourier transform (IFFT) operations, respectively. f_x and f_y represent the spatial frequencies along the x and y directions, respectively. $H(f_x,f_y;d)$ is the free-space transfer function for propagation over an axial distance d, which can be defined as[50,71]

$$H(f_x,f_y;d) = \begin{cases} \exp\left(j2\pi d\sqrt{\left(\frac{1}{\lambda}\right)^2 - f_x^2 - f_y^2}\right), & f_x^2+f_y^2<\frac{1}{\lambda^2} \\ 0, & \text{otherwise} \end{cases}, \tag{6}$$

where λ is the illumination wavelength. In our numerical analyses, $\lambda = 632.8$ nm. The Fourier and inverse Fourier transforms (IFFT) are implemented using the fast Fourier transform and IFFT algorithms.

The forward propagation model of the GAN-guided DNN is modeled by alternating applying the operations of free-space propagation in Eq. (4) and phase modulation of the diffraction layer in Eq. (5). For a given input optical field, the complex amplitude field of the output result is obtained in the output plane of the diffraction network.

4.2 Digital Implementation and Training Details

In order to train the network to achieve visual tracking and imaging of the interested moving target, the parameters of the GAN-guided DNN are optimized by minimizing the loss function. The loss function is MAE, which is calculated using the intensities of the labels and network outputs,[49]

$$\text{Loss} = \frac{1}{n\times n}\sum_{y=1}^{n}\sum_{x=1}^{n}[G(x,y)-O(x,y)]^2, \tag{7}$$

where n represents the number of neurons of the diffractive network along the x and y axials, and the G and O represent the intensities of labels and network outputs, respectively, defined as

$$G(x,y) = |g(x,y)|^2, \tag{8}$$

$$O(x,y) = |o(x,y)|^2, \tag{9}$$

where $g(x,y)$ and $o(x,y)$ are the optical field of labels and network outputs, respectively.

For the experimented scenarios in Fig. 5, we present the GAN-guided DNN with two diffractive layers to perform the task of visual tracking and imaging of the fast-moving airplane. Each diffraction layer contains 512 × 512 diffractive pixels with a size of 12.5 μm, so the model contains a total of 524,288 trainable neurons, and the distance between any two adjacent planes is 150 mm.

4.3 Training Process of the GAN-Based Teacher Model

For GAN-based teacher model training, our goal is to learn the mapping function of the domains X and Y for the given two training samples $\{x_i\}_{i=1}^N \in X$ and $\{y_i\}_{i=1}^N \in Y$. As shown in Fig. 2, training samples x are images containing moving pedestrians and cars, while y are computer-generated images of cars with indeterminate locations. There are two mapping relations, $G{:}X\to Y$ and $F{:}Y\to X$. Therefore, we also introduce two adversarial discriminators, D_X and D_Y, where D_X aims to discriminate between the image $\{x\}$ and the translated image $\{F(y)\}$; similarly, D_Y aims to discriminate between $\{y\}$ and $\{G(x)\}$. The adversarial loss is obtained by comparing the generated images with the objects in the target domain; for the mapping function $G{:}X\to Y$ and its discriminator D_Y, the adversarial loss can be expressed as[65]

$$L_{\text{GAN}}(G,D_Y,X,Y) = E_{y\sim p\,\text{data}(y)}[\log D_Y(y)] + E_{x\sim p\,\text{data}(x)}\{\log\{1-D_Y[G(x)]\}\}, \tag{10}$$

where G tries to generate an image $G(x)$ similar to the image in the domain Y, and D_Y tries to distinguish between the translated sample $G(x)$ and the real sample y. Thus, for the mapping function $F{:}Y\to X$ and its discriminator D_X, the other loss function can be expressed as $L_{\text{GAN}}(F,D_X,Y,X)$.

Ideally, the GAN-based teacher model should learn cycle-consistent transformation functions $F{:}Y\to X$ and $G{:}X\to Y$. This means that, given an input x, the output x' after undergoing the transformations $F[G(x)] = x'$ should be exactly equal to the original input x. Theoretically, this is possible because function G can transform input x into a value within domain Y, and function F can then transform that value back into domain X.

In practice, however, these transformations do achieve the goal. It is important to note that x may not be exactly the same as x' because the two generators may have applied different changes to the data. Thus, a circular consistency loss is introduced, which is defined as the difference between the input value x and the forward transform $F[G(x)]$ and the input value y and the forward transform $G[F(y)]$. The prediction becomes further from the original input as the difference increases. The network aims to minimize this loss by making opposite changes to the input data in F and G. For instance, the input x produces $G(x)$ through G, and then $F[G(x)] = x'$ through F, making the two as equal as possible. This loss term can be expressed as

$$L_{\text{Cycle}} = E_{x\sim p\,\text{data}(x)}|F(G(x)-x)| + E_{y\sim p\,\text{data}(y)}|G(F(y)-y)|. \tag{11}$$

The function of the complete GAN-based teacher model loss used to train the network is defined as the sum of the two GAN-based teacher model losses and the cycle consistency loss. Therefore, the loss function of our model can be expressed as

$$L(G,F,D_X,D_Y) = L_{\text{GAN}}(G,D_Y,X,Y) + L_{\text{GAN}}(F,D_X,Y,X) + \lambda L_{\text{Cycle}}. \tag{12}$$

Our goal is to minimize the difference between real and generated samples, while maximizing the ability of the discriminators D_X and D_Y to discriminate whether the samples are from real or generated data as far as possible:

$$G^*, F^* = \arg \min_{G,F} \max_{D_X,D_Y} L_{\mathrm{GAN}}(F, D_X, Y, X). \tag{13}$$

Further details on the construction and training of the GAN are provided in Note 7 in the Supplementary Material.

Code and Data Availability

Data will be made available on request.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62422509 and 62405188), the Shanghai Natural Science Foundation (Grant No. 23ZR1443700), the Shuguang Program of Shanghai Education Development Foundation and Shanghai Municipal Education Commission (Grant No. 23SG41), the Young Elite Scientist Sponsorship Program by CAST (Grant No. 20220042), the Science and Technology Commission of Shanghai Municipality (Grant No. 21DZ1100500), and the Shanghai Municipal Science and Technology Major Project, and the Shanghai Frontiers Science Center Program (2021–2025 No. 20).

References

1. X. Wang, “Intelligent multi-camera video surveillance: a review,” *Pattern Recognit. Lett.* **34**, 3–19 (2013).
2. J. Janai et al., “Computer vision for autonomous vehicles: problems, datasets and state of the art,” *Found. Trends Comput. Graph. Vis.* **12**, 1–308 (2020).
3. Y. Shu et al., “Interactive design of intelligent machine vision based on human–computer interaction mode,” *Microprocess Microsyst.* **75**, 103059 (2020).
4. J. Li et al., “Moving target detection and tracking algorithm based on context information,” *IEEE Access* **7**, 70966–70974 (2019).
5. X. Chen et al., “Transformer tracking,” in *Proc. IEEE/CVF Int. Conf. Comput. Vision Pattern Recognit.*, pp. 8126–8135 (2021)
6. Z. Xu et al., “A multichannel optical computing architecture for advanced machine vision,” *Light Sci. Appl.* **11**, 255 (2022).
7. X. Feng et al., “Computer vision algorithms and hardware implementations: a survey,” *Integration* **69**, 309–320 (2019).
8. G. Wetzstein et al., “Inference in artificial intelligence with deep optics and photonics,” *Nature* **588**, 39–47 (2020).
9. W. Shi et al., “LOEN: lensless opto-electronic neural network empowered machine vision,” *Light Sci. Appl.* **11**, 121 (2022).
10. Y. Du et al., “Object-adaptive LSTM network for real-time visual tracking with adversarial data augmentation,” *Neurocomputing* **384**, 67–83 (2020).
11. H. Rebecq et al., “High speed and high dynamic range video with an event camera,” *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 1964–1980 (2019).
12. N. Messikommer et al., “Event-based asynchronous sparse convolutional networks,” in *Comput. Vision–Proc. Eur. Conf. Comput. Vision 2020*, Springer, pp. 415–431 (2020).
13. G. Gallego et al., “Event-based vision: a survey,” *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 154–180 (2020).
14. L. Wang et al., “Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks,” in *Proc. IEEE/CVF Int. Conf. Comput. Vision and Pattern Recognit.*, pp. 10081–10090 (2019).
15. G. Chen et al., “Event-based neuromorphic vision for autonomous driving: a paradigm shift for bio-inspired visual sensing and perception,” *IEEE Signal Process. Mag.* **37**, 34–49 (2020).
16. R. W. Baldwin et al., “Time-ordered recent event (TORE) volumes for event cameras,” *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 2519–2532 (2022).
17. M. Gehrig and D. Scaramuzza, “Recurrent vision transformers for object detection with event cameras,” in *Proc. IEEE/CVF Int. Conf. Comput. Vision Pattern Recognit.*, pp. 13884–13893 (2023)
18. X. Wang et al., “Visevent: reliable object tracking via collaboration of frame and event flows,” *IEEE Trans. Cybern.* (2023).
19. H. Luan et al., “768-ary Laguerre-Gaussian-mode shift keying free-space optical communication based on convolutional neural networks,” *Opt. Express* **29**, 19807–19818 (2021).
20. Y. Han et al., “Robust visual tracking based on adversarial unlabeled instance generation with label smoothing loss regularization,” *Pattern Recognit.* **97**, 107027 (2020).
21. H. Yu et al., “Conditional GAN based individual and global motion fusion for multiple object tracking in UAV videos,” *Pattern Recognit. Lett.* **131**, 219–226 (2020).
22. J. T. Springenberg, “Unsupervised and semi-supervised learning with categorical generative adversarial networks,” *arXiv:1511.06390* (2015).
23. J. Gui et al., “A review on generative adversarial networks: algorithms, theory, and applications,” *IEEE Trans. Knowl. Data Eng.* **35**, 3313–3332 (2021).
24. T. Karras et al., “A style-based generator architecture for generative adversarial networks,” in *Proc. IEEE/CVF Int. Conf. Comput. Vision Pattern Recognit.*, pp. 4401–4410 (2019).
25. T. Karras et al., “Alias-free generative adversarial networks,” *Adv. Neural Inform. Process. Syst.* **34**, 852–863 (2021).
26. C. Liu et al., “Intelligent coding metasurface holograms by physics-assisted unsupervised generative adversarial network,” *Photonics Res.* **9**, B159–B167 (2021).
27. W. Xu et al., “DRB-GAN: a dynamic resblock generative adversarial network for artistic style transfer,” in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, pp. 6383–6392 (2021).
28. Y. H. Kim et al., “GRA-GAN: generative adversarial network for image style transfer of gender, race, and age,” *Expert Syst. Appl.* **198**, 116792 (2022).
29. F. Fahimi et al., “Generative adversarial networks-based data augmentation for brain–computer interface,” *IEEE Trans. Neural Netw. Learn. Syst.* **32**, 4039–4051 (2020).
30. B. Bosquet et al., “A full data augmentation pipeline for small object detection based on generative adversarial networks,” *Pattern Recogn.* **133**, 108998 (2023).
31. B. Li et al., “Ultralow-power spiking neural networks for 1024-ary orbital angular momentum shift keying free-space optical communication,” *J. Opt.* **25**, 074001 (2023).
32. W. Peebles et al., “GAN-supervised dense visual alignment,” in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, pp. 13470–13481 (2022).
33. X. Lin et al., “All-optical machine learning using diffractive deep neural networks,” *Science* **361**, 1004–1008 (2018).
34. X. Yang et al., “Complex-valued universal linear transformations and image encryption using spatially incoherent diffractive networks,” *Adv. Photonics Nexus* **3**, 016010 (2024).
35. K. Zhang et al., “Advanced all-optical classification using orbital-angular-momentum-encoded diffractive networks,” *Adv. Photonics Nexus* **2**, 066006 (2023).
36. Y. Huang et al., “Sophisticated deep learning with on-chip optical diffractive tensor processing,” *Photonics Res.* **11**, 1125–1138 (2023).

37. Ç. Işıl et al., "All-optical image denoising using a diffractive visual processor," *Light Sci. Appl.* **13**, 43 (2024).
38. X. Luo et al., "Metasurface-enabled on-chip multiplexed diffractive neural networks in the visible," *Light Sci. Appl.* **11**, 158 (2022).
39. T. Zhou et al., "*In situ* optical backpropagation training of diffractive optical neural networks," *Photonics Res.* **8**, 940–953 (2020).
40. T. Yan et al., "All-optical graph representation learning using integrated diffractive photonic computing units," *Sci. Adv.* **8**, eabn7630 (2022).
41. Z. Li et al., "Event-based diffractive neural network chip for dynamic action recognition," *Opt. Laser Technol.* **169**, 110136 (2024).
42. J. Li et al., "Class-specific differential detection in diffractive optical neural networks improves inference accuracy," *Adv. Photonics* **1**, 046001 (2019).
43. M. S. Sakib Rahman and A. Ozcan, "Integration of programmable diffraction with digital neural networks," *ACS Photonics* **11**, 2906–2922 (2024).
44. Y. Li et al., "Analysis of diffractive neural networks for seeing through random diffusers," 7*IEEE J. Sel. Top. Quantum Electron.* **29**, 1–17 (2022).
45. B. Bai et al., "To image, or not to image: class-specific diffractive cameras with all-optical erasure of undesired objects," *eLight* **2**, 1–20 (2022).
46. X. Chang et al., "Complex-domain-enhancing neural network for large-scale coherent imaging," *Adv. Photonics Nexus* **2**, 046006 (2023).
47. M. Gu et al., "Optically digitalized holography: a perspective for all-optical machine learning," *Engineering* **5**, 363–365 (2019).
48. X. Fang et al., "Orbital angular momentum-mediated machine learning for high-accuracy mode-feature encoding," *Light Sci. Appl.* **13**, 49 (2024).
49. M. S. S. Rahman et al., "Learning diffractive optical communication around arbitrary opaque occlusions," *Nat. Commun.* **14**, 6830 (2023).
50. Y. Li et al., "Optical information transfer through random unknown diffusers using electronic encoding and diffractive decoding," *Adv. Photonics* **5**, 046009 (2023).
51. C. Qian et al., "Performing optical logic operations by a diffractive neural network," *Light Sci. Appl.* **9**, 59 (2020).
52. H. Zhu et al., "Space-efficient optical computing with an integrated chip diffractive neural network," *Nat. Commun.* **13**, 1044 (2022).
53. T. Fu et al., "Photonic machine learning with on-chip diffractive optics," *Nat. Commun.* **14**, 70 (2023).
54. T. Zhou et al., "Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit," *Nat. Photonics* **15**, 367–373 (2021).
55. S. Zheng et al., "Incoherent imaging through highly nonstatic and optically thick turbid media based on neural network," *Photonics Res.* **9**, B220–B228 (2021).
56. X. Fang et al., "High-dimensional orbital angular momentum multiplexing nonlinear holography," *Adv. Photonics* **3**, 015001 (2021).
57. J. Chen et al., "Dynamic graph CNN for event-camera based gesture recognition," in *IEEE Int. Symp. Circuits Systems (ISCAS)*, IEEE, pp. 1–5 (2020).
58. Y. Deng et al., "A voxel graph CNN for object classification with event cameras," in *Proc. IEEE/CVF Int. Conf. Comput. Vision and Pattern Recognit.*, pp. 1172–1181 (2022).
59. P. Wzorek and T. Kryjak, "Traffic sign detection with event cameras and DCNN," in *Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, IEEE, pp. 86–91 (2022).
60. F. Becattini et al., "Understanding human reactions looking at facial microexpressions with an event camera," *IEEE Trans. Ind. Inf.*. **18**, 9112–9121 (2022).
61. G. Jing et al., "Neural network-based surrogate model for inverse design of metasurfaces," *Photonics Res.* **10**, 1462–1471 (2022).
62. I. Goodfellow et al., "Generative adversarial nets," *Adv. Neural Inform. Process. Sys.* **27** (2014).
63. E. Brophy et al., "Generative adversarial networks in time series: a systematic literature review," *ACM Comput. Surv.* **55**, 1–31 (2023).
64. D. Torbunov et al., "UVCGAN: UNet vision transformer cycle-consistent GAN for unpaired image-to-image translation," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, pp. 702–712 (2023).
65. J.-Y. Zhu et al., "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 2223–2232 (2017).
66. Y. Yuan et al., "Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vision and Pattern Recognit.*, pp. 701–710 (2018).
67. M. Zhao et al., "A 3D nanoscale optical disk memory with petabit capacity," *Nature* **626**, 772–778 (2024).
68. B. Bai et al., "Pyramid diffractive optical networks for unidirectional image magnification and demagnification," *Light Sci. Appl.* **13**, 178 (2024).
69. K. Li et al., "Multi-dimensional multiplexing optical secret sharing framework with cascaded liquid crystal holograms," *Opto-Electron. Adv.* **7**, 230121 (2024).
70. W. Meng et al., "100 Hertz frame-rate switching three-dimensional orbital angular momentum multiplexing holography via cross convolution," *Opto-Electron. Sci.* **1**, 220004–220004 (2022).
71. J. Li et al., "Unidirectional imaging using deep learning–designed materials," *Sci. Adv.* **9**, eadg1505 (2023).

Hang Su is currently a PhD student in the School of Artificial Intelligence Science and Technology at the University of Shanghai for Science and Technology. He received his BS degree in optical information science and technology from the East China university of Science and Technology in 2021. His current research focuses on orbital angular momentum holography and optical imaging.

Yanping He is currently an assistant research fellow in the School of Artificial Intelligence Science and Technology at the University of Shanghai for Science and Technology. She received her PhD from The Chinese University of Hong Kong. Her research focuses on optical imaging, image flow cytometry, and optical diffractive neural networks.

Xinyuan Fang is a professor at the University of Shanghai for Science and Technology. His research areas include multi-dimensional light field manipulation, optical neural networks, and holography. Until now, he has published more than 30 peer-reviewed papers as the co-first/co-corresponding author, such as *Science, Nature Photon., Nature Nanotech., Light Sci. Appl., Adv. Photon., and Nano Lett.* He is the recipient of the Excellent Young Scientists Fund of NSFC.

Biographies of the other authors are not available.