Event-based Synthetic Aperture Imaging with a Hybrid Network

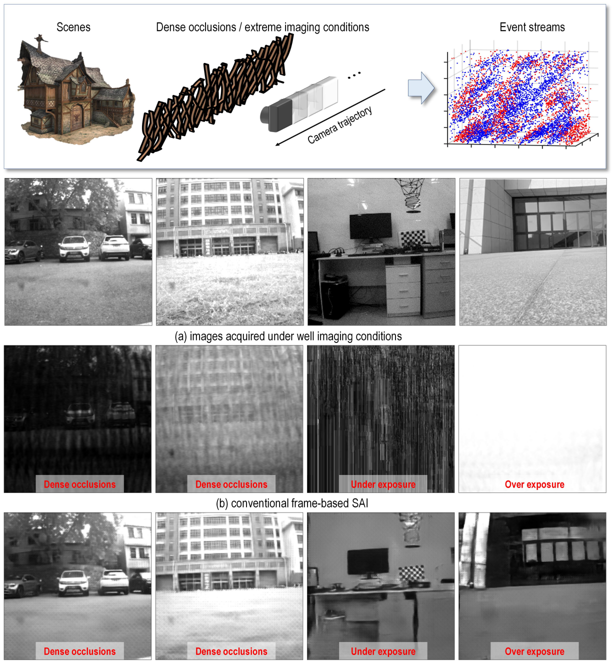
Xiang Zhang[*], Wei Liao[*], Lei Yu[†], Wen Yang, Gui-Song Xia
Wuhan University, Wuhan, China.
{xiangz, weiliao, ly.wd, yangwen, guisong.xia}@whu.edu.cn

Abstract

Synthetic aperture imaging (SAI) is able to achieve the ***see through*** *effect by blurring out the off-focus foreground occlusions and reconstructing the in-focus occluded targets from multi-view images. However, very dense occlusions and extreme lighting conditions may bring significant disturbances to the SAI based on conventional frame-based cameras, leading to performance degeneration. To address these problems, we propose a novel SAI system based on the event camera which can produce asynchronous events with extremely low latency and high dynamic range. Thus, it can eliminate the interference of dense occlusions by measuring with almost continuous views, and simultaneously tackle the over/under exposure problems. To reconstruct the occluded targets, we propose a hybrid encoder-decoder network composed of spiking neural networks (SNNs) and convolutional neural networks (CNNs). In the hybrid network, the spatio-temporal information of the collected events is first encoded by SNN layers, and then transformed to the visual image of the occluded targets by a style-transfer CNN decoder. Through experiments, the proposed method shows remarkable performance in dealing with very dense occlusions and extreme lighting conditions, and high quality visual images can be reconstructed using pure event data.*

1. Introduction

Harsh environments, *e.g.* with dense occlusions and extreme lighting conditions, often prohibit the efficient imaging of real scenes, due to the fact that the collected light information is very limited and moreover severely disturbed. Synthetic aperture imaging (SAI) tackles the problem of *seeing through* occlusions via multi-view exposures [18, 17], forming the light field [29] of the target scene under occlusions. The basic idea of SAI is to extract the light information of the occluded scenes while filter out foreground occlusions [11, 26]. However, very dense occlusions and extreme lighting scenes may bring severe disturbances, leading to serious degradation on imaging quality or even failure reconstructions (*e.g.*, Fig. 1).



(a) images acquired under well imaging conditions

(b) conventional frame-based SAI

(c) proposed event-based SAI

Figure 1: Prototype of the **event-based synthetic aperture imaging (E-SAI)** system and the illustrative examples of see through imaging under harsh environments via (b) the conventional frame-based SAI and (c) the event-based SAI. Under either very dense occlusions or extreme lighting scenes, the proposed E-SAI method can successfully generate high quality visual images for the occluded scenes.

- **Very dense occlusions:** With conventional frame-based cameras, the light cues are captured via brightness intensities. Very dense occlusions will greatly decrease the "signal", *i.e.* the light from target scenes, while increase the "noise", *i.e.* disturbances from fore-

[†] Corresponding author

[*] Equal contribution

The research was partially supported by the National Natural Science Foundation of China under Grants 61871297, the National Natural Science Foundation of China Enterprise Innovation Development Key Project under Grant U19B2004 and the Fundamental Research Funds for the Central University under Grant 2042020kf0019.

ground occlusions, leading to considerable reduction of the Light-SNR (ratio of “signal” to “noise”).

- **Extreme lighting scenes:** Due to the low dynamic range (*e.g*. about 60 dB), images from conventional frame-based cameras usually suffer from the over/under exposure problems under extreme lighting conditions. It will severely degrade the imaging quality and thus reduce the confidence of the light information from target scenes.

As a consequence, conventional frame-based SAI (F-SAI) often fails in these cases, and it is of great demand to develop new SAI methods to handle such harsh environments.

In this paper, we address the aforementioned problems by presenting a novel SAI method with event cameras [1]. Event cameras only measure the pixel-wise brightness changes of scenes in an asynchronous manner, leading to many outstanding properties including extremely low latency (in the order of μs), high dynamic range ($>$ 120 dB) and low power consumption [6, 1]. Instead of using frame-based intensity images, as shown in Fig. 1, event-based SAI (E-SAI) collects the light information from occluded targets via event streams, representing the brightness difference between the foreground occlusions and the occluded targets. This mechanism means that a higher density of occasions produces more events from occluded targets, *i.e*. more light information of targets can be recorded. With the low latency, event cameras can capture adequate information of the occluded object from almost continuous viewpoints. Thanks to the high dynamic range of event camera, E-SAI is able to collect confident light information from occluded targets even under extreme lighting conditions, making the reconstruction of scenes feasible (*e.g*., Fig. 1).

Although E-SAI can easily handle the aforementioned problems, we still have to answer the following question: *how to effectively process the event stream and reconstruct the high quality visual images of occluded targets?* Since the working mechanism of event camera differs radically from that of the frame-based one, conventional computer vision methods, *e.g*. convolutional neural networks (CNNs), cannot be directly applied to such asynchronous event streams, where the temporal and spatial information of events should be simultaneously considered [1].

The spiking neural network (SNN) [8, 7] serves as a perfect model for integrating spatio-temporal information. Unlike other artificial neural networks, spiking neurons do not respond to stimulus in a synchronous fashion. Instead, the membrane potential of spiking neurons updates over time, and a spike will be generated whenever the membrane potential exceeds a specific spiking threshold. Thus the spatio-temporal information is naturally encoded in the spike position and timing. Exploiting this, the influence of noise events can be further mitigated from the temporal dimension, leading to the improvement of Light-SNR. However, recent researches have observed the vanishing spike phenomenon [10] in deep spiking layers. Thus SNNs often suffer from performance degradation when the number of layers increases.

To tackle this, we propose a hybrid neural network that contains a SNN encoder and a CNN decoder. With initial spiking layers, the spatio-temporal information of events can be efficiently integrated and encoded. Then, the CNN is able to decode the rich output of SNN, and effectively reconstruct the visual image of occluded targets. Therefore, this architecture not only utilizes sufficient information of events, but also guarantees the overall performance of reconstruction.

In a nutshell, contributions of this paper are three-fold:

- We present a novel event-based SAI algorithm with systematic analysis, which can overcome the dilemma that the conventional F-SAI faces under very dense occlusions and extreme lighting conditions.
- We propose a hybrid SNN-CNN encoder-decoder network to reconstruct high quality visual images for E-SAI. By leveraging the merits of SNN and CNN, the spatio-temporal information of events can be well retained and utilized, and thus the occluded target can be effectively reconstructed.
- We construct an event-based SAI dataset to evaluate the proposed method, and make them available to the research community.

2. Related Work

Synthetic Aperture Imaging: How to see through the foreground occlusion has attracted considerable interest for decades [18, 27, 11, 26]. Through calibrating the images captured by camera arrays, a plane + parallax framework was proposed to solve the de-occlusion problem [18]. Since the output of camera arrays can be regarded as a virtual camera imaging with large-aperture lens, the foreground occlusion can be effectively blurred out when the background target is refocused on. But this method often results in blurry images because the information from both occlusions and targets will be indiscriminately used for reconstruction. To improve the performance, a variety of techniques have been exploited to filter the disturbance of occlusions, including depth-based approach [27], energy minimization [11] and k-means clustering [26]. By separating targets from foreground occlusions, a better de-occlusion effect can be achieved using only target information.

The principle of traditional F-SAI is to reconstruct the occluded target via multi-view images captured by a moving camera [29] or a camera array system [23, 21]. By projecting all images to the plane where targets are located, the light information of occluded target is aligned while the occlusion becomes out of focus. Afterward, reconstruction can be performed to achieve the see through effect. But due

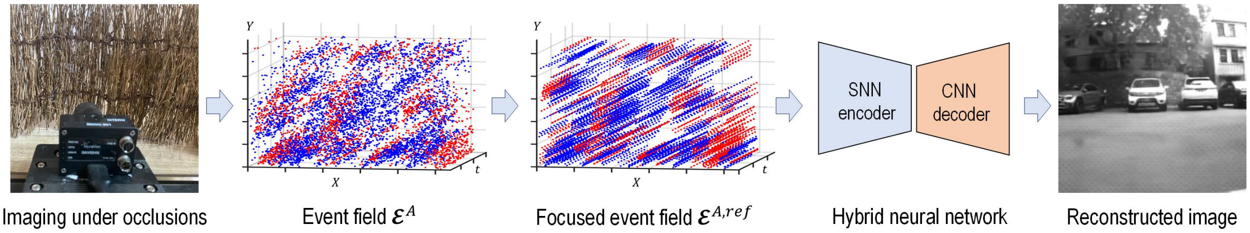

Imaging under occlusions
Y
X
t
Event field $\boldsymbol{\mathcal{E}}^A$
Y
X
t
Focused event field $\boldsymbol{\mathcal{E}}^{A,ref}$
SNN encoder
CNN decoder
Hybrid neural network
Reconstructed image

Figure 2: Overall pipeline of the proposed E-SAI. As moving the event camera, E-SAI collects event streams $\mathcal{E}_\theta^A$ with almost continuous viewpoints θ and forms the *event field* $\boldsymbol{\mathcal{E}}^A$. To reconstruct high quality images from $\boldsymbol{\mathcal{E}}^A$, we propose to employ the hybrid SNN-CNN network after the event refocusing process.

to the inherent mechanism of traditional camera, the Light-SNR of captured images is often severely reduced when encountering very dense occlusions or extreme light scenes, resulting in significant performance degradation.

Event Cameras: Instead of frame-based intensity images, event cameras generate asynchronous events [1], composed of pixel position, time stamp and polarity. Specifically, the i-th event $e_i = (p_i, x_i, t_i)$ is triggered at pixel position x_i and time t_i whenever the log-scale brightness change exceeds a pre-setting threshold η, *i.e*.

$$\log(I(x_i,t_i)) - \log(I(x_i,t_i-\Delta t_i)) \geq p_i \cdot \eta, \qquad (1)$$

where $I(\cdot)$ denotes the intensity of pixel; Δt_i indicates the time since the last event at position x_i; $p_i \in \{+1,-1\}$ is the polarity representing the sign of brightness change [6]. This paradigm shift in visual information acquisition leads to many outstanding properties like extremely low latency and high dynamic range, and promotes great potential in many computer vision tasks like optical flow estimation [20], high dynamic range (HDR) imaging [14] and simultaneous localization and mapping (SLAM) [19].

Similarly, event cameras pose great advantages in dealing with the see through tasks. Due to the low latency property, sufficient light information of occluded targets can be acquired by event cameras under disturbance of dense occlusions. On the other hand, the high dynamic range of event cameras provides the possibility of measuring light information under extreme lighting conditions. Thus it motivates us to exploit event cameras to tackle the problem of SAI under very dense occlusions and extreme lighting conditions, and propose the E-SAI.

3. Problem Statement

Suppose for some static unknown scene A with $I_\theta^A(u,v)$ representing the projected brightness intensity captured with the camera pose θ, where u,v are respectively horizontal / vertical coordinates. Then $\boldsymbol{I}^A \triangleq \{I_\theta^A\}_{\theta\in\mathcal{P}}$ forms a tensor of light field of A with $\mathcal{P}$ the set of camera poses. Analogically, the light field of occlusions O can be represented as $\boldsymbol{I}^O \triangleq \{I_\theta^O\}_{\theta\in\mathcal{P}}$ with $I_\theta^O(u,v)$ denoting the brightness intensity captured with the camera pose θ.

F-SAI: The task of F-SAI is to achieve the see through imaging from limited number of occluded observations, *i.e*., $\bar{\boldsymbol{I}}^A = \{\bar{I}_\theta^A\}_{\theta\in\mathcal{P}}$ with $|\mathcal{P}| < \infty$ and

$$\bar{I}_\theta^A = \mathcal{M}^O(I_\theta^A) + I_\theta^O + I^n \qquad (2)$$

where I^n denotes the measurement noises and $\mathcal{M}^O$ represents the masking operator for $\mathcal{M}^O(\cdot) = 0$ only when it is occluded by O. With very dense occlusions, $\bar{I}_\theta^A$ will be severely contaminated and the extreme lighting conditions make the observations completely saturated and incorrect, leading to failure reconstruction of visual images for A.

E-SAI: As illustrated in Fig. 1, events are induced by the brightness change as moving the event camera, then we can denote the collected events with camera pose θ as a set of stream $\mathcal{E}_\theta^A \triangleq \{e_i\}_{i=1}^M = \{(p_i,x_i,t_i)\}_{i=1}^M$ with $M = |\mathcal{E}_\theta^A|$. According to the generating process, we can divide $\mathcal{E}_\theta^A$ into two categories: (1) *Signal events*, denoted as $\mathcal{E}_\theta^{OA}$, are induced by the brightness difference between the scene A and the occlusion O. Then based on Eq. (1), the number of events emitted for $\mathcal{E}_\theta^{OA}$,

$$|\mathcal{E}_\theta^{OA}| \propto \left|\log(I_\theta^A) - \log(I_\theta^O)\right|. \qquad (3)$$

(2) *Noise events* include the physical noises $\mathcal{E}^n$ inherently from the event camera and the interference events induced by the brightness change (caused by textures) of occlusions $\mathcal{E}_\theta^{OO}$ and occluded targets $\mathcal{E}_\theta^{AA}$ as moving the event camera.

Due to the low latency property, E-SAI is able to collect events $\mathcal{E}_\theta^A$ from almost continuous viewpoints θ and form the *event field*, *i.e*. $\boldsymbol{\mathcal{E}}^A = \{\mathcal{E}_\theta^A\}_{\theta\in\mathcal{P}}$, with $|\mathcal{P}| \to \infty$ and

$$\mathcal{E}_\theta^A = \mathcal{E}_\theta^{OA} + \mathcal{E}_\theta^{OO} + \mathcal{E}_\theta^{AA} + \mathcal{E}^n \qquad (4)$$

where $\mathcal{E}_\theta^{OA}$ encodes the light information from the occluded targets and the other terms are considered as noises. Thus the main problem of E-SAI is to reconstruct the high quality visual images of the scene A from the event field $\boldsymbol{\mathcal{E}}^A$ which are severely disturbed by noise events.

4. Event-based SAI

Fig. 2 illustrates the overall pipeline of the proposed E-SAI algorithm, which aims at reconstructing the high quality visual images of occluded targets. It consists of two

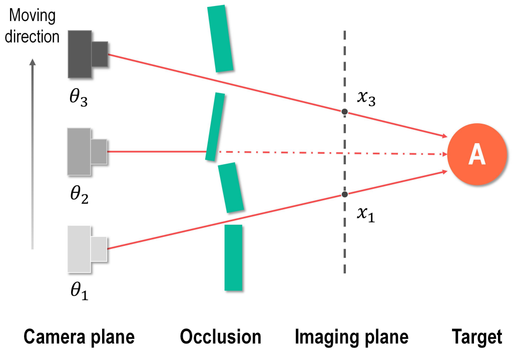


Figure 3: Diagram of event generating process in E-SAI system. As the event camera moves, signal events triggered with camera poses θ_1 and θ_3 are induced by the brightness difference between occlusions (green) and the target A (orange), while noise events triggered with camera pose θ_2 are induced by textures on occlusions.

main steps: *refocusing* and *reconstruction*. The purpose of refocusing is to align the signal events, while scatter out the noise events from both spatial and temporal dimensions. For the reconstruction, a hybrid SNN-CNN network is proposed to mitigate the disturbance of noise mentioned in Eq. (4). With spiking layers, the influence of scattered noise events can be eliminated from the temporal dimension, thus a clean visual result can be decoded by the CNN.

4.1. Event Refocusing

Previous works [2, 13] have presented the similar ideas of event refocusing. But in our case, the basic principles of the aforementioned techniques are violated due to the disturbance of dense occlusions and extreme lighting scenes. Thus we only consider a simple situation with linear camera motion and known target depth in this work. As displayed in Fig. 3, a moving camera is employed to collect events $\boldsymbol{\mathcal{E}}^A$ from multiple viewpoints. We assume that the event camera keeps staying on the camera plane and the optical axes of all camera poses are parallel. Since all the events are emitted asynchronously as the camera moves, a pixel-wise refocusing process needs to be performed for event alignment. Define θ^{ref} as the reference camera pose and X_i as the coordinate of pixel x_i in the camera coordinate system at pose θ_i. According to the multiple view geometry [3] and the pinhole imaging model [12], the refocusing equation can be formulated as [28]:

$$x_i^{ref} = KR_iK^{-1}x_i + \frac{KT_i}{d}, \tag{5}$$

where x_i^{ref} represents the target pixel position on the reference imaging plane; K is the intrinsic matrix of camera; R_i, T_i are the rotation and translation matrices between camera poses θ_i and θ^{ref}; target depth d is the distance between target A and the camera plane.

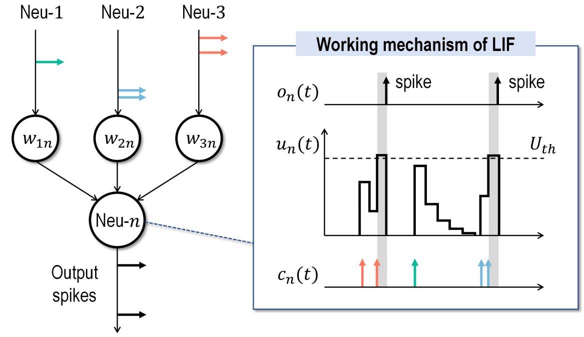


Figure 4: An illustrative example of the LIF neuron and its working mechanism. The spikes from pre-neurons are first weighted and then fed into the target neuron-n, charging the internal membrane potential $u_n(t)$. Spikes will be fired whenever $u_n(t) > U_{th}$. Due to the leakage mechanism, the LIF neuron is able to filter out the isolated spikes, *e.g.* the noise events scattered out in spatio-temporal dimensions.

Exploiting Eq. (5), the refocused event field can be obtained $\boldsymbol{\mathcal{E}}^{A,ref} = \{\mathcal{E}_\theta^{A,ref}\}_{\theta\in\mathcal{P}}$ where $\mathcal{E}_\theta^{A,ref} = \{e_i^{ref}\}_{i=1}^M = \{(p_i, x_i^{ref}, t_i)\}_{i=1}^M$ and all events are projected to the imaging plane of the reference camera at pose θ^{ref}. After refocusing, the events triggered by target A are successfully aligned, while others, *e.g.* the events generated by occlusions, are scattered out in both temporal and spatial dimensions, achieving a preliminary de-occlusion effect.

4.2. Reconstruction with a Hybrid Network

According to Eq. (3), the brightness intensity of the occluded scene is closely related to the number of events. Thus the visual image of the occluded scene can be recovered by event accumulation after the refocusing process without removal of noises. Even though CNN-based methods can be further exploited to alleviate the noise problem, the temporal information behind events cannot be effectively used. In view of this, we propose a hybrid neural network composed by a SNN encoder and a CNN decoder, where both spatial and temporal information of events can be efficiently considered and utilized.

SNN Encoder: Although the noise events are dispersed during the refocusing process, their presence still affects the quality of reconstruction. To deal with it, we implement the SNN encoder using the leaky integrate-and-fire (LIF) model [25]. As shown in Fig. 4, LIF neurons are usually activated when receiving more continuous spikes. If no new spikes are fed, the internal membrane potential will gradually leak over time. Recall that all signal events are successfully aligned during the refocusing process, *i.e.* they appear more continuously in the temporal dimension, while these noise events are scattered in both time and space. Thus, the leakage mechanism of LIF neuron is able to eliminate the

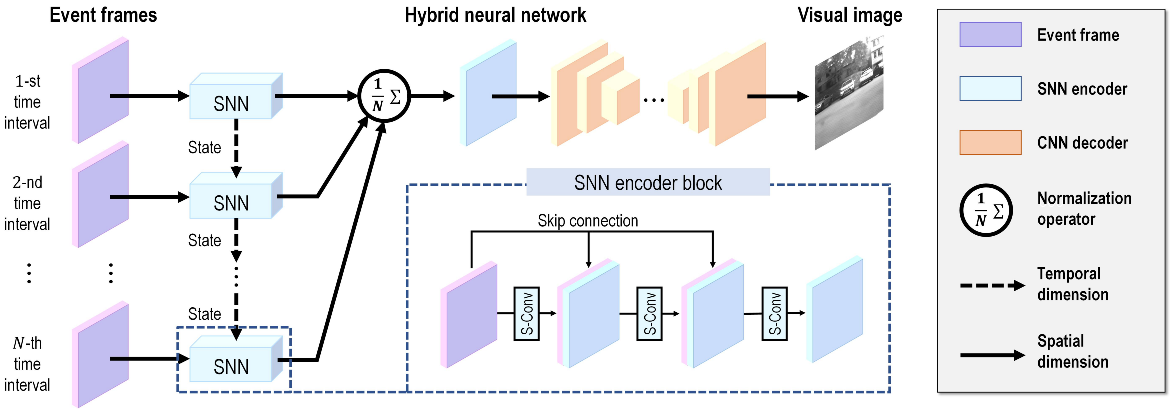


Figure 5: Structure of the hybrid encoder-decoder network. The spatio-temporal information of events is first encoded by SNN blocks, and then transformed to visual images by the CNN decoder. To reduce the information loss of events, we add skip connections between the event frame and the output of the 1-st, 2-nd spiking convolution (S-Conv) layers.

influence of noise events, meanwhile preserving the information of occluded targets.

LIF Neuron: Define $u_n^l(t)$ as the membrane potential of the neuron-n on the l-th layer at time t. The update of membrane potential can be described as

$$u_n^l(t) = \alpha u_n^l(t-1) + c_n^l(t), \tag{6}$$

where $\alpha \in [0,1]$ denotes the decay factor and $c_n^l(t)$ is the input current corresponding to neuron-n. Considering the convolution operation in spiking layers, Eq. (6) can be reformulated as:

$$u_n^l(t) = \alpha u_n^l(t-1) + \sum_m w_{mn} o_m^{l-1}(t-1), \tag{7}$$

where $o_m^{l-1}(t-1)$ represents the output spike of neuron-m on the $(l-1)$-th layer at time $t-1$, and w_{mn} denotes the synaptic weight between neuron-m and neuron-n. Further, we add the reset & fire mechanism into Eq. (7),

$$u_n^l(t) = \alpha u_n^l(t-1)(1-o_n^l(t-1)) + \sum_m w_{mn} o_m^{l-1}(t-1), \tag{8}$$

where the output spike $o_n^l(t)$ is defined by

$$o_n^l(t) = \begin{cases} 1, & \text{if } u_n^l(t) > U_{th}, \\ 0, & \text{otherwise,} \end{cases} \tag{9}$$

and U_{th} represents the spiking threshold. Eq. (8) indicates that the membrane potential of neuron-n is affected by both its own state and the input spikes. If no new spikes are fed, the membrane potential $u_n^l(t)$ will leak at a certain rate related to the factor α. In contrast, if the potential $u_n^l(t)$ is charged up to the spiking threshold U_{th}, the potential will be immediately reset to the resting potential $U_{rest} = 0$ and simultaneously a spike will be emitted to other neurons.

SNN Structure: As illustrated in Fig. 5, our SNN encoder consists of a three-layer structure composed of LIF neurons. To make a balance between computational complexity and information integrity, we present a spatio-temporal representation for events. Given a pre-setting number of event frames, *e.g.* N, the refocused event sequence are fairly divided into N time intervals. In each interval, an event frame can be generated by accumulating events over time, and each frame contains two channels (positive and negative events). Thus, every input group includes N event frames and the temporal relationship between event frames is retained. Over time, event frames sequentially pass through the spiking layers, and the membrane potential of spiking neurons updates between time intervals. Since noise events are scattered during refocusing, their influence can be gradually leaked out by the potential update of LIF neurons. Therefore, the noise issue is well alleviated, guaranteeing the reconstruction quality of occluded targets. To avoid the vanishing spike phenomenon in deep spiking layers [10], we instead implement the decoder with a deep CNN block.

CNN Decoder: Due to the inherent difference between the event feature map and the visual image, we regard the decoding process as a style-transfer task. Here, we adopt the decoder architecture from the generator network used in [30], which shows remarkable results in image style transferring, and adjust the kernel size of the output layer to fit the gray-scale images in our case. Benefiting from the hybrid structure, the spatio-temporal information of events can be fully utilized by the SNN encoder, and the occluded targets can be effectively reconstructed by the CNN decoder, guaranteeing the overall performance.

Training Hybrid Network: The synaptic weights in SNN can be trained in a supervised fashion via the spatio-temporal back propagation (STBP) technique [24, 25], where the gradient of each pixel can be derived based on time intervals. And CNNs can be trained via back propagation (BP). Thus the SNN and CNN in the proposed hybrid network can be jointly trained.

To guide the training, we first exploit the idea of perceptual loss [4] for high-level feature learning. With a pre-trained loss network ϕ, we denote $\phi_k(X)$ as the output of the k-th convolution layer when network ϕ processes image X. Assume that $\phi_k(X)$ has the shape $C_k \times H_k \times W_k$, we can formulate the perceptual loss $\mathcal{L}_{per}$ as:

$$\mathcal{L}_{per}(Y,\hat{Y}) = \sum_k \frac{\lambda_k}{C_k H_k W_k} \|\phi_k(Y) - \phi_k(\hat{Y})\|_2^2, \quad (10)$$

where Y represents the output of the hybrid network and $\hat{Y}$ is the corresponding ground truth; λ_k denotes the weight of the k-th feature map. Rather than encouraging the pixel-wise match between images, the perceptual loss encourages the network to learn the similarity between high-level features, leading to better visual results.

In the pixel level, we add the pixel loss $\mathcal{L}_{pix}$ to maintain the similarity in low-level features like shape and texture. We express the pixel loss as:

$$\mathcal{L}_{pix}(Y,\hat{Y}) = \frac{\|Y-\hat{Y}\|_1}{CHW}, \quad (11)$$

where $C \times H \times W$ represents the shape of Y and $\hat{Y}$. Besides, the total variance loss $\mathcal{L}_{tv}(Y)$ in [9] is exploited to encourage the spatial smoothness of reconstruction. Thus, the total loss can be summarized as follows.

$$\mathcal{L}(Y,\hat{Y}) = \beta_{per}\mathcal{L}_{per}(Y,\hat{Y}) + \beta_{pix}\mathcal{L}_{pix}(Y,\hat{Y}) + \beta_{tv}\mathcal{L}_{tv}(Y), \quad (12)$$

where β_{per}, β_{pix} and β_{tv} are the weights that control the importance of the corresponding loss function.

5. Experiments and Analysis

5.1. Experimental Settings

Event-based SAI Dataset: We build an event-based SAI dataset where the event streams are captured by a DAVIS346 camera [6] installed on a programmable sliding trail. A large variety of targets are considered including printed pictures, simple objects and real scenes. They are occluded by the wooden fence installed parallel to the sliding trail to imitate the very dense occlusions, as shown in Fig. 2. By linearly sliding the DAVIS346 camera, the events triggered by the brightness difference between the wooden fence and the occluded targets are collected. The dataset can be divided into two categories according to the shooting scenes: *indoor* and *outdoor*. The *indoor* dataset contains printed pictures, simple objects and real scenes, while the *outdoor* dataset only contains real complex scenes. The gray-scale images are captured simultaneously as collecting the events by DAVIS346 camera since it can output both events and APS (active pixel sensor) frames. Moreover, we collect the APS frames without occlusions and take them as the ground truth of the occluded targets. In summary, the event-based SAI dataset is built with 300 groups of data including 250 groups for *indoor* and 50 groups for *outdoor*, and each group contains a stream of events, a series of APS frames with occlusions and one APS frame without occlusions. For the extreme lighting scenes, there is no APS frame without occlusions due to the over/under exposure problem.

Training Details: We augment the event-based SAI dataset by flipping (horizontal, vertical, and horizontal-vertical) and rotating (random angles ranging from -10 to 10 degree). Finally, 216 groups (180 *indoor* groups and 36 *outdoor* groups) are augmented to 1296 groups for training, while the rest in dataset are left for the testing phase. All networks are trained on NVIDIA TITAN RTX GPUs with batch size 8 for around 500 epochs, and the Adam optimizer [5] is used, where the initial learning rate is set as 5×10^{-4} and the step decay learning rate schedule is applied after the 250 epochs. The 16-layer VGG network [16] pretrained on the ImageNet dataset [15] is employed as the loss network, where the perceptual loss is calculated at the 2-nd, 4-th, 7-th and 10-th convolution layers.

SAI Methods: For the frame-based SAI, the approach proposed in [18] (**F-SAI**) is employed, where 35 images are collected with frame-based cameras from different viewpoints. In addition, we design a learning-based F-SAI using CNNs (**F-SAI+CNN**) with the same 35 images. For the event-based SAI, we evaluate three different reconstruction methods, including accumulating method (**E-SAI+ACC**), pure CNN method (**E-SAI+CNN**) and the proposed hybrid network (**E-SAI+Hybrid**). In the E-SAI+CNN method, the refocused event streams are stacked as a $2N$-channel tensor (2 represents the polarity) for network input. To evaluate the effectiveness of our hybrid network, a pure CNN counterpart is designed by simply replacing the SNN encoder with a 3-layer CNN which has the same network setting as the SNN. By applying the pure CNN model to **F-SAI+CNN** and **E-SAI+CNN**, we can fairly compare these learning-based SAI methods.

5.2. Qualitative Analysis

As shown in Fig. 6, the reconstruction results of F-SAI methods are severely contaminated by dense occlusions where a lot of details are missing. For the extreme lighting scenes, the performance is even worse since the light from the occluded target cannot be correctly measured due to the over/under exposure problems encountered with frame-based cameras. On the contrary, E-SAI methods are able to produce results with better visual effects and retain more details. Due to the inherent high dynamic range of event camera, E-SAI methods do not suffer from over/under exposure problems, thus the occluded target can be effectively reconstructed under extreme lighting conditions.

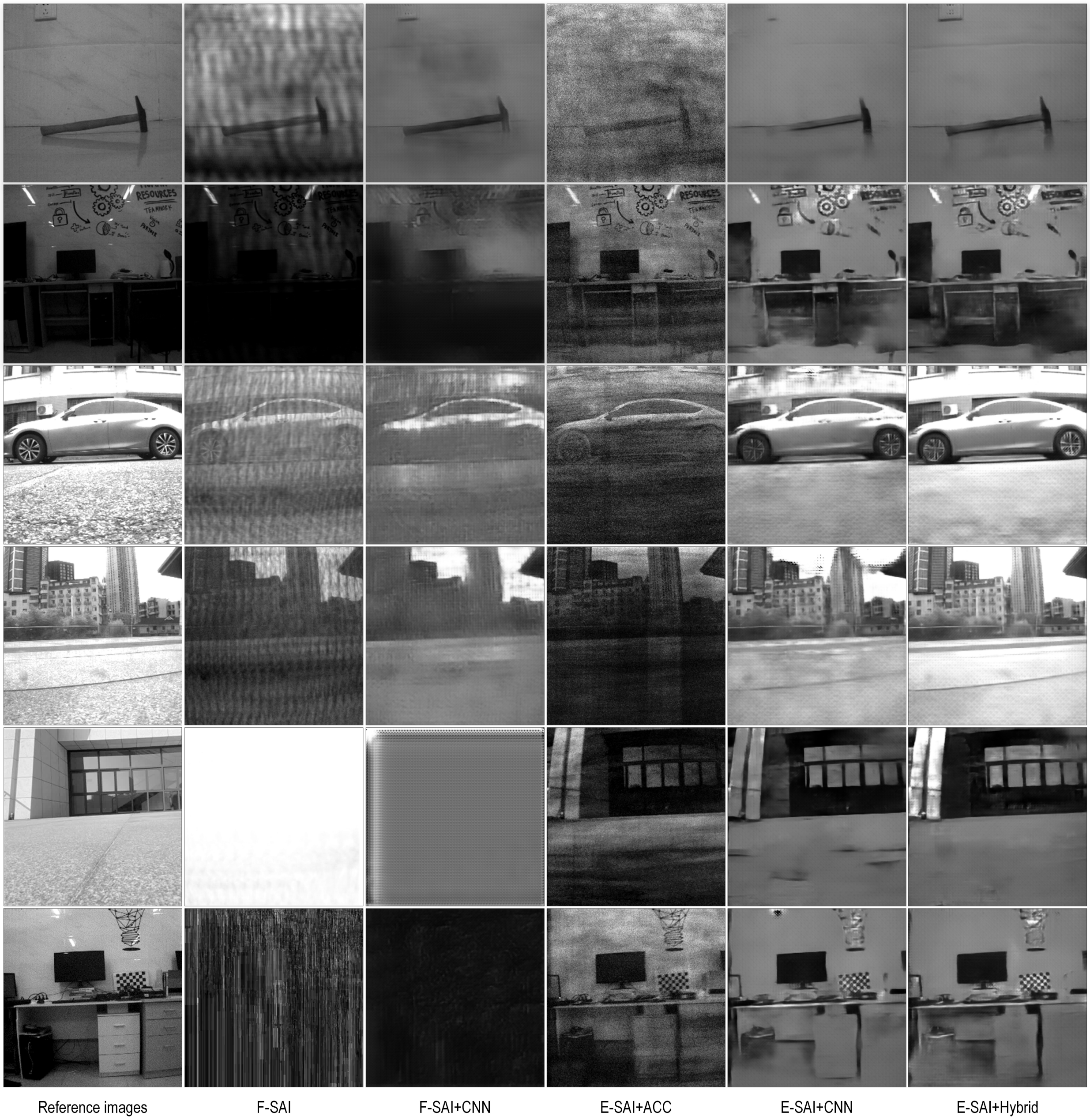


Figure 6: Qualitative comparisons between F-SAI and E-SAI algorithms under very dense occlusions (row 1-4) and extreme lighting conditions (row 5-6) for *indoor* (row 1,2 and 6) and *outdoor* (row 3-5) dataset.

To reveal the advantages of our hybrid network, we compare the visual results of E-SAI with different reconstruction techniques in Fig. 7. It is obvious that the result of E-SAI+ACC is often noisy because both signal and noise events are indiscriminately accumulated for reconstruction. In the learning-based approaches, the E-SAI+CNN fed directly with the stacked event frames cannot efficiently deal with the temporal information of asynchronous events, and thus degrading the visual quality with detail losses, artifacts and saturation. But these issues can be mitigated by the hybrid architecture, where the temporal information is utilized by the SNN encoder. Over time, LIF spiking neurons can efficiently leak out the influence of noise events which are either emitted randomly or scattered out after the refocusing process. Consequently, the proposed hybrid network generates images with the best visual quality.

5.3. Quantitative Analysis

In Table 1, we evaluate the quantitative results of the proposed system. In the dense occlusion experiment, the

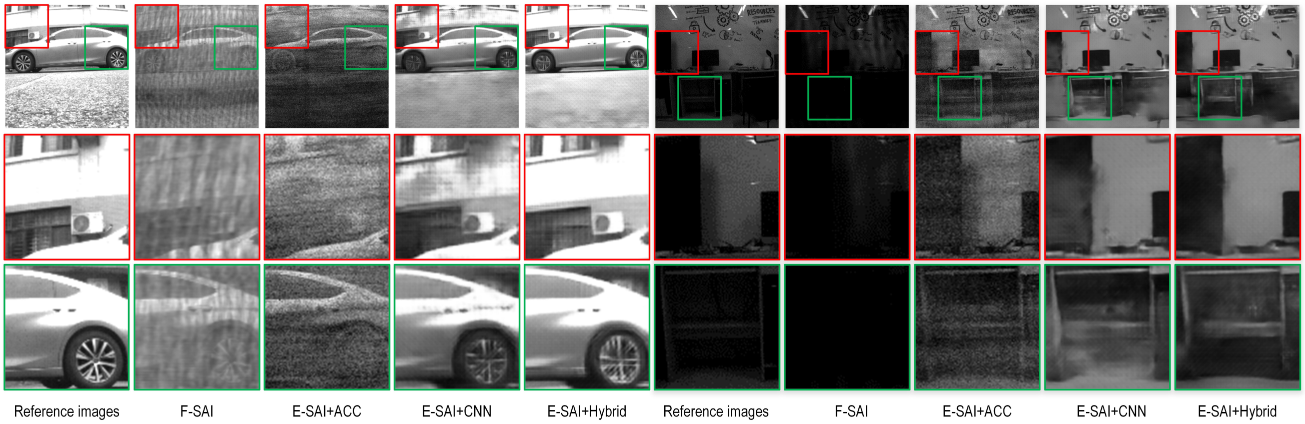


Figure 7: Comparisons of F-SAI and E-SAI with different reconstruction methods. Details are zoomed in for better view.

Method	Dense Occlusion				Extreme Exposure			
	Indoor		Outdoor		Over		Under	
	PSNR	SSIM	PSNR	SSIM	Entropy	STD	Entropy	STD
F-SAI [18]	13.89	0.4482	10.96	0.3124	4.855	16.86	5.022	26.06
F-SAI+CNN	26.44	0.8077	13.93	0.3744	5.684	20.77	4.165	4.264
E-SAI+ACC	14.71	0.2272	8.654	0.1887	5.706	41.16	5.314	**50.10**
E-SAI+CNN	31.27	0.8373	17.62	0.5237	6.921	56.16	6.105	33.84
E-SAI+Hybrid	**33.04**	**0.8429**	**24.02**	**0.6524**	**7.417**	**61.95**	**6.204**	34.87

Table 1: Quantitative comparisons between F-SAI and E-SAI algorithms. PSNR(dB) and SSIM are exploited for dense occlusion cases. No-reference metrics, *i.e*. 2D entropy and STD are exploited for extreme exposure cases due to the absence of the corresponding APS frames.

metrics PSNR and SSIM [22] are employed for quantitative comparison, where the aligned APS images captured by DAVIS346 are considered as the ground truth. In the extreme exposure part, the no-reference assessment metrics two-dimensional (2D) entropy [31] and standard deviation (STD) are exploited to evaluate the image quality. 2D entropy measures the amount of image information and higher value indicates more information. STD is used to assess the contrast of image and larger value means higher contrast.

Exploiting learning-based techniques, F-SAI+CNN is able to produce better results than F-SAI under dense occlusions. But both frame-based methods cannot deal with the over/under exposure problem due to the low dynamic range of traditional camera. On the contrary, event-based approaches can effectively reconstruct visual images with more information and better contrast. However, it is hard for E-SAI+ACC to produce satisfactory PSNR and SSIM results since the emission of events is based on the brightness change in the logarithmic domain, which differs from the intensity directly recorded in reference images. Through learning the mapping relationship between the event domain and the image domain, this problem can be well solved by the E-SAI+CNN and E-SAI+Hybrid. Regarding the learning-based E-SAI, the hybrid network excels its pure CNN counterpart over 6 dB in PSNR and 0.12 in SSIM under complex outdoor scenes. This demonstrates that the use of SNN encoder not only achieves the denoising purpose, but also maintains the integrity of overall structure. In summary, our E-SAI+Hybrid method largely outperforms other algorithms under dense occlusions, and can produce more natural visual results in extreme lighting environments.

6. Conclusion

In this work, we proposed a novel SAI algorithm based on event cameras. With extremely low latency and high dynamic range of event cameras, our method is able to handle the disturbance of dense occlusion and does not suffer from the over/under exposure problem. This greatly expands the usage of SAI algorithm, enabling the application under harsh conditions like daytime astronomical observation and nighttime penetrating imaging. Moreover, a hybrid SNN-CNN network is proposed to process the output of event camera. Benefiting from the combination of SNN and CNN, the spatio-temporal information of events is well utilized and the reconstruction quality of occluded targets is guaranteed. To test our method, we build an event-based SAI dataset including scenes under heavy occlusions and extreme lighting conditions. The result verifies that our approach is effective to these harsh environments and can reconstruct the occluded target with impressive visual effects.

References

[1] Guillermo Gallego, Tobi Delbruck, Garrick Michael Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Jorg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1–1, 2020. 2, 3

[2] Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2018. 4

[3] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 4

[4] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Eur. Conf. Comput. Vis.*, pages 694–711, 2016. 6

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[6] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 dB 15 μs Latency Asynchronous Temporal Contrast Vision Sensor. *IEEE J. Solid-State Circuits*, 43(2):566–576, 2008. 2, 3, 6

[7] Wolfgang Maass. Networks of Spiking Neurons: The Third Generation of Neural Network Models. *Neural Networks*, 10(9):1659–1671, 1997. 2

[8] Wolfgang Maass and Christopher Bishop. *Pulsed neural networks*. MIT Press, 1998. 2

[9] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5188–5196, 2015. 6

[10] Priyadarshini Panda, Sai Aparna Aketi, and Kaushik Roy. Toward Scalable, Efficient, and Accurate Deep Spiking Neural Networks With Backward Residual Connections, Stochastic Softmax, and Hybridization. *Frontiers in Neuroscience*, 14:653, 2020. 2, 5

[11] Zhao Pei, Yanning Zhang, Xida Chen, and Yee-Hong Yang. Synthetic aperture imaging using pixel labeling via energy minimization. *Pattern Recognition*, 46(1):174–187, 2013. 1, 2

[12] Zhao Pei, Yanning Zhang, Tao Yang, Xiuwei Zhang, and Yee-Hong Yang. A novel multi-object detection method in complex scene using synthetic aperture imaging. *Pattern Recognition*, 45(4):1637 – 1658, 2012. 4

[13] Henri Rebecq, Guillermo Gallego, Elias Mueggler, and Davide Scaramuzza. Emvs: Event-Based Multi-View Stereo—3D Reconstruction with an Event Camera in Real-Time. *Int. J. Comput. Vis.*, 126(12):1394–1414, 2018. 4

[14] Henri Rebecq, Rene Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-To-Video: Bringing Modern Computer Vision to Event Cameras. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3852–3861, 2019. 3

[15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and others. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015. 6

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[17] Vaibhav Vaish, Marc Levoy, Richard Szeliski, C Lawrence Zitnick, and Sing Bing Kang. Reconstructing Occluded Surfaces Using Synthetic Apertures: Stereo, Focus and Robust Measures. In *IEEE Conf. Comput. Vis. Pattern Recog.*, volume 2, pages 2331–2338, 2006. 1

[18] Vaibhav Vaish, Bennett Wilburn, Neel Joshi, and Marc Levoy. Using plane + parallax for calibrating dense camera arrays. In *IEEE Conf. Comput. Vis. Pattern Recog.*, volume 1, pages 2–9, 2004. 1, 2, 6, 8

[19] Antoni Rosinol Vidal, Henri Rebecq, Timo Horstschaefer, and Davide Scaramuzza. Ultimate SLAM? Combining Events, Images, and IMU for Robust Visual SLAM in HDR and High-Speed Scenarios. *IEEE Robot. Auto. Letters*, 3(2):994–1001, 2018. 3

[20] Bishan Wang, Jingwei He, Lei Yu, Gui-Song Xia, and Wen Yang. Event enhanced high-quality image recovery. In *Eur. Conf. Comput. Vis.*, 2020. 3

[21] Yingqian Wang, Tianhao Wu, Jungang Yang, Longguang Wang, Wei An, and Yulan Guo. DeOccNet: Learning to See Through Foreground Occlusions in Light Fields. In *IEEE Conf. Wint. Applic. Comput. Vis.*, pages 118–127, 2020. 2

[22] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *IEEE Asilomar Conf. Sign. Syst. Comput.*, volume 2, pages 1398–1402, 2003. 8

[23] Bennett Wilburn, Neel Joshi, Vaibhav Vaish, Eino-Ville Talvala, Emilio Antunez, Adam Barth, Andrew Adams, Mark Horowitz, and Marc Levoy. High performance imaging using large camera arrays. In *ACM SIGGRAPH*, pages 765–776, 2005. 2

[24] Yujie Wu, Lei Deng, Guoqi Li, Jun Zhu, and Luping Shi. Spatio-temporal backpropagation for training high-performance spiking neural networks. *Frontiers in Neuroscience*, 12:331, 2018. 5

[25] Yujie Wu, Lei Deng, Guoqi Li, Jun Zhu, Yuan Xie, and Luping Shi. Direct Training for Spiking Neural Networks: Faster, Larger, Better. In *AAAI Conf. Artif. Intell.*, volume 33, pages 1311–1318, 2019. 4, 5

[26] Zhaolin Xiao, Lipeng Si, and Guoqing Zhou. Seeing Beyond Foreground Occlusion: A Joint Framework for SAP-Based Scene Depth and Appearance Reconstruction. *IEEE J. Selected Topics in Signal Processing*, 11(7):979–991, 2017. 1, 2

[27] Tao Yang, Yanning Zhang, Xiaomin Tong, Xiaoqiang Zhang, and Rui Yu. Continuously tracking and see-through occlusion based on a new hybrid synthetic aperture imaging model. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3409–3416, 2011. 2

[28] Lei Yu, Wei Liao, You-Long Zhou, Wen Yang, and Gui-Song. Xia. Event camera based synthetic aperture imaging. *Acta Automatica Sinica*, 45(x):1–14, 2020. 4

[29] Xiaoqiang Zhang, Yanning Zhang, Tao Yang, and Yee-Hong Yang. Synthetic aperture photography using a moving camera-IMU system. *Pattern Recognition*, 62:175 – 188, 2017. 1, 2

[30] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Int. Conf. Comput. Vis.*, pages 2223–2232, 2017. 5

[31] Lin Zhu, Siwei Dong, Jianing Li, Tiejun Huang, and Yonghong Tian. Retina-like visual image reconstruction via spiking neural model. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1438–1446, 2020. 8