Learning to See Through With Events

Lei Yu, Xiang Zhang, Wei Liao, Wen Yang, and Gui-Song Xia

Abstract—Although synthetic aperture imaging (SAI) can achieve the seeing-through effect by blurring out off-focus foreground occlusions while recovering in-focus occluded scenes from multi-view images, its performance is often deteriorated by dense occlusions and extreme lighting conditions. To address the problem, this paper presents an Event-based SAI (E-SAI) method by relying on the asynchronous events with extremely low latency and high dynamic range acquired by an event camera. Specifically, the collected events are first refocused by a *Refocus-Net* module to align in-focus events while scattering out off-focus ones. Following that, a *hybrid network* composed of spiking neural networks (SNNs) and convolutional neural networks (CNNs) is proposed to encode the spatio-temporal information from the refocused events and reconstruct a visual image of the occluded targets. Extensive experiments demonstrate that our proposed E-SAI method can achieve remarkable performance in dealing with very dense occlusions and extreme lighting conditions and produce high-quality images from pure events. Codes and datasets are available at https://dvs-whu.cn/projects/esai/.

Index Terms—Synthetic aperture imaging, event camera, spiking neural network

1 Introduction

Harsh environment, e.g., dense occlusions or extreme lighting conditions, often makes it difficult to efficiently acquire images of real scenes, as the collected light information is usually very limited and severely disturbed. Among the methods attempting to achieve *seeing-through* effect, synthetic aperture imaging (SAI) tackles the problem via multi-view exposures [1], [2], forming the light field of the target scene under occlusions. The basic idea of SAI is to extract the light information of the occluded scenes while filtering out foreground occlusions [3], [4]. However, very dense occlusions and extreme lighting scenes may bring severe disturbances, leading to serious degradation of the imaging quality or even failure reconstructions, e.g., Fig. 1.

- *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 foreground occlusions, leading to a considerable reduction of the Light-SNR (ratio of "signal" to "noise").
- *Extreme lighting scenes:* Due to the low dynamic range (e.g., ≈ 60 dB), images from conventional frame-based cameras usually suffer from over/under exposure problems under extreme lighting conditions, severely degrading the imaging quality and the confidence of the light information from target scenes.

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

In this paper, we present a novel SAI method with event cameras to address the aforementioned problems. Event cameras measure the pixel-wise brightness changes of scenes asynchronously, leading to many promising properties, including extremely low latency (in the order of μs), high dynamic range (> 120 dB), and low power consumption [6], [7]. 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 the foreground occlusion will trigger more events for the occluded targets, i.e., more light information of targets can be recorded. With low latency, event cameras can capture adequate information of the occluded object from almost continuous viewpoints. Due to the high dynamic range of event cameras, E-SAI can collect confident light information from occluded targets even under extreme lighting conditions, leading to successful reconstructions.

Some preliminary results have shown the feasibility of seeing-through with events [8], but the E-SAI framework is still open, including the working mechanism and the reconstruction methodology. It is crucial to answer the following question: *how to effectively process the event stream and reconstruct the high-quality visual images of occluded targets?* The working mechanism of the 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

- *Lei Yu, Xiang Zhang, Wei Liao, and Wen Yang are with the School of Electronic Information, Wuhan University, Wuhan 430072, China. E-mail: {ly.wd, xiangz, wei.liao, yangwen}@whu.edu.cn.*
- *Gui-Song Xia is with the School of Computer Science, Wuhan University, Wuhan 430072, China. E-mail: guisong.xia@whu.edu.cn.*

Manuscript received 28 August 2021; revised 3 October 2022; accepted 4 December 2022. Date of publication 7 December 2022; date of current version 5 June 2023.

This work was supported in part by the National Natural Science Foundation of China under Grants 62271354, 61871297, 61922065, 41820104006, and 61871299, and in part by the Natural Science Foundation of Hubei Province, China under Grant 2021CFB467.

(Corresponding authors: Wen Yang and Gui-Song Xia.)

Recommended for acceptance by K. Nishino.

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2022.3227448, provided by the authors.

Digital Object Identifier no. 10.1109/TPAMI.2022.3227448

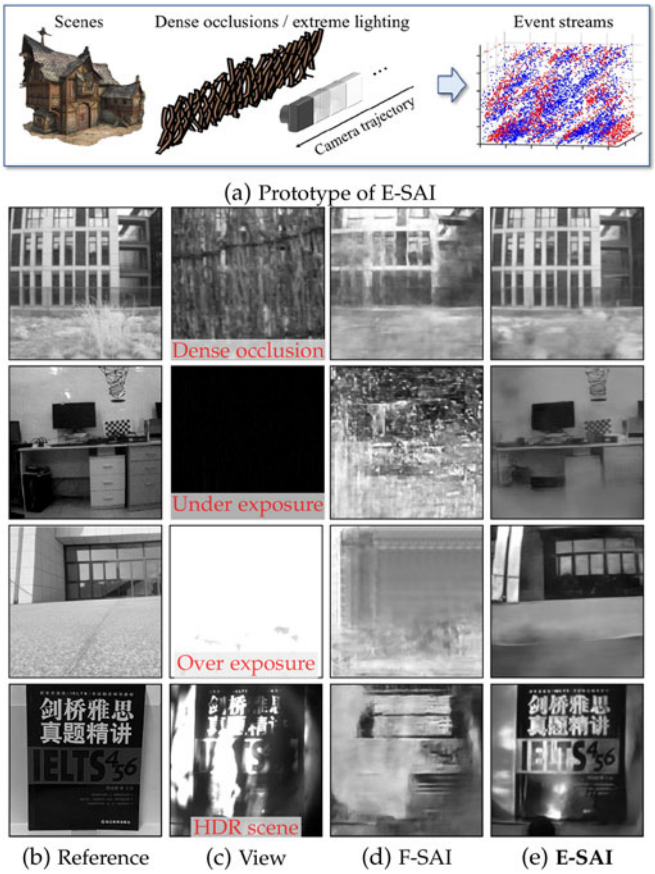


Fig. 1. Prototype of the **event-based synthetic aperture imaging (E-SAI)** system (a), the illustrative indoor and outdoor scenes (b) viewing under harsh environments (c), and the corresponding seeing-through results via the state-of-the-art F-SAI [5] (d) and the E-SAI (e). Under either very dense occlusions or extreme lighting scenes (c), the proposed E-SAI method can successfully generate high quality visual images for the occluded scenes.

event streams where the temporal and spatial information of events should be simultaneously considered [7]. The spiking neural network (SNN) [9], [10] serves as an optimal model for integrating spatio-temporal information. Unlike other artificial neural networks, spiking neurons do not respond to stimuli synchronously. 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, adopting pure SNNs for image reconstruction tasks often suffers from performance degradation. On the one hand, SNNs transmit information with sparse and binary spikes, which are not sufficient for high-quality image restoration tasks where high numerical precision is required to recover the accurate pixel intensities [11]. On the other hand, recent researches have observed the vanishing spike phenomenon [12] in deep spiking layers. To tackle both problems, we propose a hybrid neural network that contains an 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 can decode the rich output of SNN and effectively reconstruct the visual image of occluded targets. Therefore, this architecture utilizes sufficient information of events and guarantees the overall performance of reconstruction.

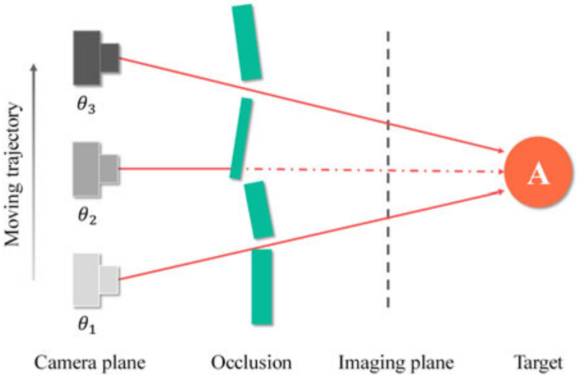


Fig. 2. Diagram of fronto-parallel uniform camera motion in our E-SAI system, where the event camera moves straightly with uniform velocity and fronto-parallel to occlusion and target planes. As the event camera moves, events triggered at camera viewpoint θ_1 are induced by the brightness difference between occlusions (green) and target A (orange), while events triggered at θ_2 and θ_3 are induced by high contrast textures of occlusions and targets, respectively.

The contributions of this paper are three-fold:

- We present a comprehensive analysis of the event-based SAI which can overcome the challenges of very dense occlusions and extreme lighting conditions.
- We propose a novel E-SAI algorithm to reconstruct visual images of occluded target scenes through refocusing-then-reconstructing implemented in a data-driven manner, where the Refocus-Net and the hybrid SNN-CNN network are proposed respectively for refocusing and reconstruction.
- We build a new SAI dataset containing both image frames and event streams to facilitate event-based SAI research. Extensive experiments on the SAI dataset demonstrate the superiority of E-SAI over F-SAI under very dense occlusions and extreme lighting scenes.

A preliminary version of this work was appeared in [13]. In contrast with [13], this paper provides more analysis of the E-SAI framework, including more details on the components of triggered events and the corresponding epipolar geometry. We also design a spatial transformer network, i.e., Refocus-Net, to automatically refocus the events collected by a moving event camera with fronto-parallel uniform motion as depicted in Fig. 2, relaxing the dependence on prior information such as camera velocity and target depth, and present a novel training strategy for the proposed Refocus-Net. Furthermore, we enlarge the SAI dataset from 300 groups to 588 groups, including 488 groups of *indoor* scenes and 100 groups of *outdoor* scenes. Based on the enlarged dataset, we finally provide an in-depth analysis of our proposed E-SAI method.

2 Related Work

2.1 Event Camera

As a bio-inspired vision sensor, the event camera poses a paradigm shift in visual information acquisition [7]. Instead of capturing full-intensity images at a fixed frame rate, event cameras only respond to the brightness change and emit asynchronous events composed of pixel position, time stamp, and polarity [6]. This mechanism offers many outstanding properties, e.g., low latency and high dynamic range [7], and previous researches have revealed the potential of events in a wide variety of computer vision and robotic applications such as feature tracking [14], optical flow estimation [15], and simultaneous localization and mapping (SLAM) [16].

For imaging tasks, event data also exhibits promising benefits, especially under harsh conditions such as high dynamic range (HDR) [17], [18], [19] and high-speed motion [20], [21]. Assuming the brightness constancy, one can reconstruct static scenes from events and simultaneously estimate the camera motion and the optical flow [22], [23], [24], [25], addressing the challenges raised by fast moving cameras. Furthermore, one can reconstruct the intensity images for dynamic scenes by directly integrating the triggered events that naturally encode the temporal differentiation of the brightness in the logarithmic domain [26], [27]. However, the collected events in real-world scenarios often contain a large amount of noise induced by background activity noise, false negatives, and temporal statistics [6], [7], which degrades the performance of event-based imaging methods. Recently, event-based imaging gains considerable progress by leveraging deep neural networks and learning to reconstruct high frame-rate and HDR videos of the target scenes from events with improved noise robustness [17], [19], [28], [29].

However, existing event-based imaging methods are devoted to occlusion-free imaging, and few of them can be directly applied to the task with dense occlusions.

2.2 Synthetic Aperture Imaging

How to see through the foreground occlusion has attracted considerable interest for decades [1], [3], [4], [5], [30], [31], [32], [33], [34]. The traditional F-SAI reconstructs the occluded target via multi-view images captured by a moving camera [33] or a camera array system [1], [5], [35]. Then, by projecting all images to the plane where targets are located, the light information of the occluded target is aligned while the occlusion becomes out of focus. Afterward, reconstruction can be performed to achieve the seeing-through effect. A plane + parallax framework has been proposed to solve the de-occlusion problem by calibrating the images captured by camera arrays [1]. Since the output of camera arrays can be regarded as a virtual camera imaging with a 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 are indiscriminately used for reconstruction. One can further improve the de-occlusion effect by filtering out the disturbance of occlusions using the depth-based approach [31], energy minimization [3], or k-means clustering [4]. Moreover, an all-in-focus SAI method based on image matting techniques [32] is developed to reconstruct target scenes with different depths. And a mobile camera-IMU system [33] is then designed for practical usage of SAI in real-world scenarios. Recently, deep learning based SAI methods have been proposed and achieved state-of-the-art performance [5]. With the seeing-through ability, SAI has been exploited in a wide range of computer vision tasks, e.g., multi-object detection [36], continuously tracking [30], [31], and 3D reconstruction of occluded objects [34].

However, the captured images of occluded targets are often severely contaminated when encountering very dense occlusions or extreme lighting conditions, making conventional F-SAI fail to achieve the seeing-through effect. In contrast, event cameras pose significant advantages in dealing with seeing-through tasks. On the one hand, sufficient light information of occluded targets can be acquired by event cameras even under the disturbance of dense occlusions due to the low latency property. On the other hand, the high dynamic range of event cameras enables the acquisition of light information even under extreme lighting conditions. Thus it motivates us to adopt event cameras to tackle the problem of SAI under very dense occlusions and extreme lighting conditions [13]. Previous work [8] has introduced a wide range of potential applications that can benefit from event-based vision, including low earth orbit satellite tracking, star mapping, and seeing through cloud gaps and bushes. Compared to [8], this paper not only presents a comprehensive analysis of the E-SAI task, but also proposes learning-based approaches respectively for refocusing and reconstructing which validate the superiority of E-SAI under very dense occlusions and extreme lighting conditions.

3 Problem Formulation and Analysis

Suppose for a static unknown scene A with I_θ^A representing the projected brightness intensity captured at the camera viewpoint θ. Then $\boldsymbol{I}^A \triangleq \{I_\theta^A\}_{\theta\in\mathcal{P}}$ forms a tensor of light field of A with $\mathcal{P}$ denoting the set of camera viewpoints. Analogically, the light field of occlusions O can be represented as $\boldsymbol{I}^O \triangleq \{I_\theta^O\}_{\theta\in\mathcal{P}}$ with I_θ^O denoting the brightness intensity captured at the camera viewpoint θ.

3.1 Frame-Based SAI (F-SAI)

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

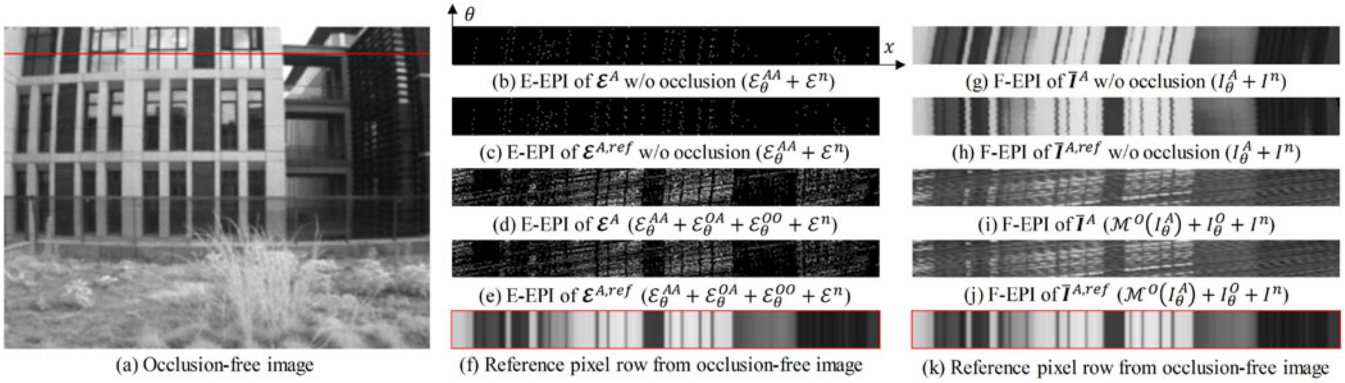
$$\bar{I}_\theta^A = \mathcal{M}^O(I_\theta^A) + I_\theta^O + I^n, \tag{1}$$

where I^n denotes the measurement noise and $\mathcal{M}^O$ represents the masking operator for $\mathcal{M}^O(\cdot) = 0$ only when it is occluded by O. Very dense occlusions will bring severe disturbances to $\bar{I}_\theta^A$ and the extreme lighting conditions often make the observations saturated, leading to failure reconstruction of visual images for A.

3.2 Event-Based SAI (E-SAI)

Instead of using frame-based cameras, we propose an event-based SAI system where the event camera is employed to collect the light field. Specifically, the i-th event $e_i = (p_i, \mathbf{x}_i, t_i)$ is triggered at pixel position $\mathbf{x}_i$ and time t_i whenever the log-scale brightness change exceeds the event threshold η, i.e.,

$$\tilde{I}(\mathbf{x}_i, t_i) - \tilde{I}(\mathbf{x}_i, t_i - \Delta t_i) = p_i \cdot \eta, \tag{2}$$



(a) Occlusion-free image

(b) E-EPI of $\boldsymbol{\mathcal{E}}^A$ w/o occlusion ($\mathcal{E}^{AA}_\theta + \mathcal{E}^n$)

(c) E-EPI of $\boldsymbol{\mathcal{E}}^{A,ref}$ w/o occlusion ($\mathcal{E}^{AA}_\theta + \mathcal{E}^n$)

(d) E-EPI of $\boldsymbol{\mathcal{E}}^A$ ($\mathcal{E}^{AA}_\theta + \mathcal{E}^{OA}_\theta + \mathcal{E}^{OO}_\theta + \mathcal{E}^n$)

(e) E-EPI of $\boldsymbol{\mathcal{E}}^{A,ref}$ ($\mathcal{E}^{AA}_\theta + \mathcal{E}^{OA}_\theta + \mathcal{E}^{OO}_\theta + \mathcal{E}^n$)

(f) Reference pixel row from occlusion-free image

(g) F-EPI of $\bar{\boldsymbol{I}}^A$ w/o occlusion ($I^A_\theta + I^n$)

(h) F-EPI of $\bar{\boldsymbol{I}}^{A,ref}$ w/o occlusion ($I^A_\theta + I^n$)

(i) F-EPI of $\bar{\boldsymbol{I}}^A$ ($\mathcal{M}^O(I^A_\theta) + I^O_\theta + I^n$)

(j) F-EPI of $\bar{\boldsymbol{I}}^{A,ref}$ ($\mathcal{M}^O(I^A_\theta) + I^O_\theta + I^n$)

(k) Reference pixel row from occlusion-free image

Fig. 3. Comparisons between frame-based EPIs (F-EPI) and event-based EPIs (E-EPIs) with and without (w/o) occlusions, where the red pixel row in the occlusion-free image (a) is selected for visualization. We compare the F-EPIs and E-EPIs under both (b, c, g, h) occlusion-free and (d, e, i, j) densely occluded scenes. E-EPIs (b, d) and F-EPIs (g, i) are generated with the collected event fields $\boldsymbol{\mathcal{E}}^A$ and light fields $\bar{\boldsymbol{I}}^A$, while E-EPIs (c, e) and F-EPIs (h, j) are generated with the refocused event fields $\boldsymbol{\mathcal{E}}^{A,ref}$ and light fields $\bar{\boldsymbol{I}}^{A,ref}$. Reference pixel row (f, k) from occlusion-free image is stretched for better comparison.

where $\tilde{I} = \log(I)$ denotes the log-scale pixel intensity, Δt_i indicates the time since the last event at position $\mathbf{x}_i$, $p_i \in \{+1, -1\}$ is the polarity representing the direction of brightness change, i.e., increase $(+1)$ or decrease (-1) [6].

As illustrated in Fig. 2, events are induced by the brightness change as moving the event camera, then we can denote the collected events at camera viewpoint θ as a set of stream $\mathcal{E}^A_\theta \triangleq \{e_i\}_{i=1}^M = \{(p_i, \mathbf{x}_i, t_i)\}_{i=1}^M$ with $M = |\mathcal{E}^A_\theta|$. According to the source of brightness change, we divide the collected events into four categories, i.e.,

$$\mathcal{E}^A_\theta = \mathcal{E}^{AA}_\theta + \mathcal{E}^{OA}_\theta + \mathcal{E}^{OO}_\theta + \mathcal{E}^n, \tag{3}$$

and respectively,

- $\mathcal{E}^{AA}_\theta$ and $\mathcal{E}^{OO}_\theta$ respectively denote the set of events triggered by *edges of targets A and occlusions O with high contrasts*. For $e_i = (p_i, \mathbf{x}_i, t_i) \in \mathcal{E}^{AA}_\theta$ or $\mathcal{E}^{OO}_\theta$, the left side of Eq. (2) can be written as $\tilde{I}(\mathbf{x}_i, t_i) - \tilde{I}(\mathbf{x}_i, t_i - \Delta t_i) = \Delta\tilde{I}_\theta(\mathbf{x}_i)$, i.e.,

$$\Delta\tilde{I}_\theta(\mathbf{x}_i) = p_i \cdot \eta.$$

By applying Taylor expansion and optical flow constraint on the brightness change on the left side, we have $\Delta\tilde{I}_\theta(\mathbf{x}_i) \approx -\nabla\tilde{I}_\theta(\mathbf{x}_i) \cdot \Delta\mathbf{x}_i$ with $\Delta\mathbf{x}_i$ denoting the pixel displacement during Δt_i [7]. Then,

$$-\nabla\tilde{I}_\theta(\mathbf{x}_i) \cdot \Delta\mathbf{x}_i \approx p_i \cdot \eta,$$

which indicates that events $\mathcal{E}^{AA}_\theta$ or $\mathcal{E}^{OO}_\theta$ are mainly induced by regions of A or O with large gradients $\nabla\tilde{I}_\theta(\mathbf{x}_i)$, i.e., high contrast edges. Thus, the number of events emitted for $\mathcal{E}^{AA}_\theta$ or $\mathcal{E}^{OO}_\theta$ is related to the edges of A or O, i.e.,

$$|\mathcal{E}^{AA}_\theta| \propto \left\|\nabla\tilde{I}^A_\theta(\mathbf{x}) \cdot \Delta\mathbf{x}\right\|, \text{ and } |\mathcal{E}^{OO}_\theta| \propto \left\|\nabla\tilde{I}^O_\theta(\mathbf{x}) \cdot \Delta\mathbf{x}\right\|, \tag{4}$$

with $|\cdot|$ and $\|\cdot\|$ respectively denoting the cardinality of the set and the norm / absolute value.

- $\mathcal{E}^{OA}_\theta$ denotes the set of events triggered by *brightness difference between occlusions O and targets A*. Specifically, for $e_i = (p_i, \mathbf{x}_i, t_i) \in \mathcal{E}^{OA}_\theta$ collected at the viewpoint θ, the left side of Eq. (2) denotes the brightness difference between A and O with $\tilde{I}(\mathbf{x}_i, t_i) = \tilde{I}^A_\theta(\mathbf{x}_i)$ and $\tilde{I}(\mathbf{x}_i, t_i - \Delta t_i) = \tilde{I}^O_\theta(\mathbf{x}_i)$, i.e.,

$$\tilde{I}^A_\theta(\mathbf{x}_i) - \tilde{I}^O_\theta(\mathbf{x}_i) = p_i \cdot \eta.$$

Thus the number of events emitted for $\mathcal{E}^{OA}_\theta$ is related to the brightness difference between A and O, i.e.,

$$|\mathcal{E}^{OA}_\theta| \propto \left\|\tilde{I}^A_\theta - \tilde{I}^O_\theta\right\|. \tag{5}$$

- $\mathcal{E}^n$ denotes noise events due to *physical imperfections of intrinsic circuits and ambient lights*.

We can conclude that for a given viewpoint θ, events $\mathcal{E}^{AA}_\theta$ and $\mathcal{E}^{OO}_\theta$ respectively contain edge information of targets A and occlusions O, while events $\mathcal{E}^{OA}_\theta$ provide the texture information of targets A relative to occlusions O. Thanks to the low latency property, E-SAI is able to collect events $\mathcal{E}^A_\theta$ from almost continuous viewpoints θ and form the event field $\boldsymbol{\mathcal{E}}^A = \{\mathcal{E}^A_\theta\}_{\theta\in\mathcal{P}}$ with $|\mathcal{P}| \to \infty$.

Epipolar Analysis. In our setup of fronto-parallel horizontal camera motion (Fig. 2), the collected event field $\boldsymbol{\mathcal{E}}^A$ can be parameterized by $(\mathbf{x}, \theta)$, with $\mathbf{x} \triangleq (x, y)$ representing the event coordinates and θ denoting the horizontal viewpoint. Similar to the epipolar plane images (EPIs) in light field [37], we fix y coordinate and generate the event-based EPI (E-EPI) of event field $\boldsymbol{\mathcal{E}}^A$ in the x-θ plane as shown in Figs. 3b and 3d for scenarios without and with occlusions. Following the frame refocusing procedure in F-SAI [1], an event refocusing process can be performed for event alignment from the collected event field $\boldsymbol{\mathcal{E}}^A$ to the refocused event field at the reference image plane $\boldsymbol{\mathcal{E}}^{A,ref} \triangleq \{\mathcal{E}^{A,ref}_\theta\}_{\theta\in\mathcal{P}}$ with $\mathcal{E}^{A,ref}_\theta \triangleq \{e^{ref}_i\}_{i=1}^M = \{(p_i, \mathbf{x}^{ref}_i, t_i)\}_{i=1}^M$. According to the multiple view geometry [38] and the pinhole imaging model [36], the event refocusing can be formulated as

$$\tilde{\mathbf{x}}^{ref}_i = KR_iK^{-1}\tilde{\mathbf{x}}_i + \frac{KT_i}{d}, \tag{6}$$

where $\tilde{\mathbf{x}}_i$, $\tilde{\mathbf{x}}^{ref}_i$ correspond to the homogeneous coordinates of $\mathbf{x}_i$, $\mathbf{x}^{ref}_i$; K is the intrinsic matrix of camera; R_i, T_i are the rotation and translation matrices between camera viewpoint θ_i and the reference one θ^{ref}; target depth d is the distance between target A and the camera plane. Similarly, we depict the E-EPI of the refocused event field $\mathcal{E}^{A,ref}$ without and with occlusions in Figs. 3c and 3e. Compared to the reference pixel row of the occlusion-free image in Fig. 3f, the E-EPIs in Figs. 3c and 3e respectively correspond to edges and textures of the target A, consistent with Eqs. (4) and (5).

Therefore, when refocusing on the target plane, events $\mathcal{E}_\theta^{AA}$ and $\mathcal{E}_\theta^{OA}$ are aligned and regarded as *signal* in the E-SAI system, where $\mathcal{E}_\theta^{AA}, \mathcal{E}_\theta^{OA}$ respectively contain the target information of high contrast edges and the scene texture, and the misaligned events $\mathcal{E}_\theta^{OO}$ and $\mathcal{E}^n$ are treated as *noise* to be filtered out. Although the signal events $\mathcal{E}_\theta^{AA}$ will reduce when the target scenes are densely occluded, the signal events $\mathcal{E}_\theta^{OA}$ triggered by occlusions can compensate for the lost edge information in $\mathcal{E}_\theta^{AA}$ and provide more scene texture for reconstruction. We will present detailed discussion of the refocusing and reconstructing methods in Section 4.

3.3 E-SAI Versus F-SAI

E-SAI is better behaved than F-SAI when encountering very dense occlusions and extreme lighting scenes.

- *Very Dense Occlusions:* we present the frame-based EPIs (F-EPIs) for light fields $\bar{I}^A$ and refocused light fields $\bar{I}^{A,ref}$ of F-SAI with or without occlusions in Figs. 3g, 3h, 3i and 3j. It is shown that *foreground occlusions bring severe disturbances to the captured frames but trigger additional signal events $\mathcal{E}_\theta^{OA}$ for reconstruction.* Comparing Figs. 3g and 3i, the F-EPI in occlusion-free scene highly matches the reference pixel row, but the one under occlusions is heavily contaminated by the light from foreground occlusions, i.e., I_θ^O. By contrast, E-EPI in occlusion-free scenes mainly contains events $\mathcal{E}_\theta^{AA}$ that respond to high contrast regions, e.g., edges as shown in Fig. 3b. By exploiting the brightness contrast between occlusions and targets, E-EPI in Fig. 3d contains abundant events $\mathcal{E}_\theta^{OA}$ that provide texture information, enabling reconstruction of densely occluded scenes.
- *Extreme lighting scenes:* the light fields captured in F-SAI system will be severely degraded due to the over/under exposure problems when encountering extreme lighting conditions. By contrast, this issue can be largely alleviated thanks to the high dynamic range property of event cameras in E-SAI system.

4 Methodology

The goal of E-SAI is to reconstruct the occlusion-free image from the collected event field. Fig. 4 illustrates the overall pipeline of the proposed E-SAI algorithm, which consists of two main steps: *refocusing* and *reconstruction*. The purpose of refocusing is to align the signal events and scatter out the noise events, and we will address it in Section 4.1. For the reconstruction, a hybrid SNN-CNN network is proposed to mitigate the disturbance of noise mentioned in Eq. (3) and reconstruct the occluded scenes from the refocused event streams. The detailed description of the reconstruction network can be found in Section 4.2.

4.1 Event Refocusing

Although the problem of event refocusing has been recently investigated [39], [40], [41], [42], [43], [44], previous methods are mainly designed for sparsely occluded or occlusion-free scenes, and the dense foreground occlusion in our case brings new challenges to correctly refocusing on the background scenes. For example, EMVS [39] can estimate the depth of scene points by locating the high-density regions where several viewing rays back-projected from events intersect. Thus, EMVS[1] can easily locate the position of target points in occlusion-free scenes as shown in Figs. 5a and 5c. However, under dense occlusions, EMVS tends to detect the structure of foreground occlusions instead of the background targets (Fig. 5c), since the occluded scene can only be sparsely and inconsistently observed and the rays back-projected from the acquired events mostly intersect on the occlusion plane as depicted in Fig. 5b. To deal with this issue, this section presents an auto refocus method to adaptively align signal events under dense occlusions.

4.1.1 *Auto Event Refocusing via Spatial Transformers*

We first define the warping process from the collected event field $\mathcal{E}^A$ to the refocused event field $\mathcal{E}^{A,ref}$ as

$$\mathcal{E}^{A,ref} = \mathcal{W}(\mathcal{E}^A, \boldsymbol{\psi}), \tag{7}$$

which is achieved by spatial projection for each event from $\mathbf{x}_i$ to $\mathbf{x}_i^{ref}$ parameterized by parameter $\boldsymbol{\psi}$. Since the event camera is moving straightly with uniform velocity in our case, the refocusing formulation Eq. (6) can be simplified to

$$\mathbf{x}_i^{ref} = \mathbf{x}_i + \boldsymbol{\psi} \cdot (t_i - t^{ref}), \tag{8}$$

where $\boldsymbol{\psi} = \frac{1}{d}[f_x \nu_x,\ f_y \nu_y]^\top$ is the coupled warping parameter with f_x, f_y denoting the pixel focal length and ν_x, ν_y indicating the camera speed in horizontal and vertical directions, respectively; t_i is the timestamp of the i-th event; t^{ref} represents the timestamp when the camera is at reference viewpoint θ^{ref}. After refocusing with the provided warping parameter $\boldsymbol{\psi}$, 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.

However, the refocusing process Eq. (8) heavily depends on the prior knowledge of the coupled warping parameter $\boldsymbol{\psi}$ associated with the camera motions and the depth of target scenes, which are difficult to be given in practice. In addition, the refocusing results are also sensitive to the accuracy of the target depth and camera motion, especially for close-view imaging (see Fig. 15). As a result, it is not easy to directly apply Eq. (8) for event refocusing in practice.

Fortunately, in our setup with 1D uniform camera motion, we can facilitate event refocusing by analyzing E-EPIs as shown in Fig. 6a, where spatially aligned events are parallel to the viewpoint dimension θ and appear vertical in E-EPI. Since a long baseline is often required in SAI systems to imitate the camera with a large aperture, which results in a very shallow depth-of-field, we only consider targets with

[1] We use codes from https://github.com/uzh-rpg/rpg_emvs.

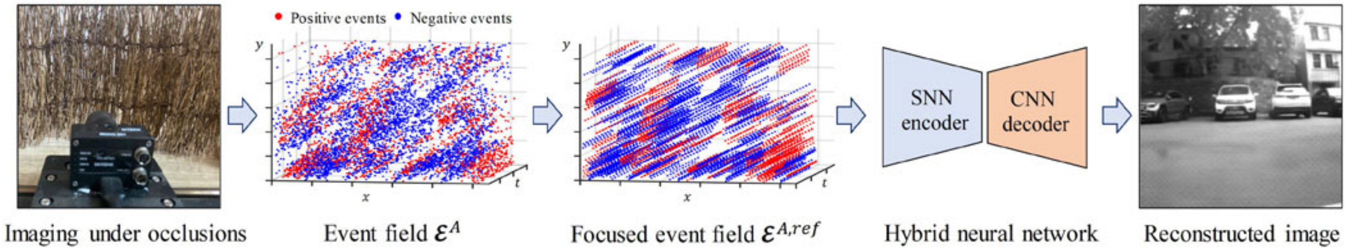


Fig. 4. Overall pipeline of the proposed E-SAI. As moving the event camera, E-SAI collects event streams $\mathcal{E}_\theta^A$ with almost continuous viewpoints θ and forms the *event field* $\mathcal{E}^A$. To reconstruct high quality images from $\mathcal{E}^A$, we propose to employ the hybrid SNN-CNN network after the event refocusing process.

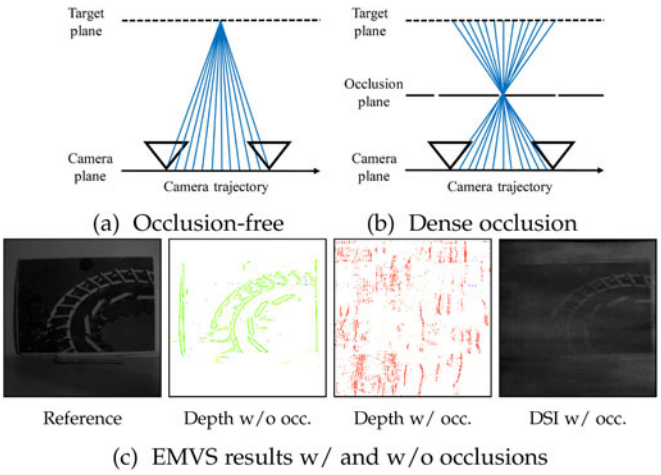
small depth variation. Therefore, events can be spatially aligned when refocusing on either the occlusion plane or the target plane using the corresponding warping parameter ψ. Furthermore, according to Eqs. (4) and (5), events will be respectively aligned on edges and scene textures when refocusing on the occlusion plane and the target plane, leading to different spatial density in E-EPIs as shown in Fig. 6a. Thus, the warping parameter ψ can be uniquely determined based on *event alignment* and *spatial density* in the E-EPI domain for refocusing on the target plane.

Based on the above analysis, we propose to address event refocusing via learning-based methods. Previous works of spatial transformer networks (STNs) [45], [46] have achieved spatial manipulation of data within networks, and the idea behind it can be also exploited for event refocusing. Following the methodology of STN [45], we design a Refocus-Net to predict the coupled warping parameter ψ from the collected event field, i.e.,

$$\psi = \mathrm{Refocus-Net}\left(\mathcal{E}^A\right). \tag{9}$$

Then, one can achieve event refocusing by warping the collected event field with the predicted parameter ψ. And finally, the warping process Eq. (7) can be relaxed

$$\mathcal{E}^{A,ref} = \mathcal{W}\left(\mathcal{E}^A, \mathrm{Refocus-Net}\left(\mathcal{E}^A\right)\right). \tag{10}$$



(a) Occlusion-free

(b) Dense occlusion

(c) EMVS results w/ and w/o occlusions

Fig. 5. Performance of EMVS [39] in occlusion-free and densely occluded scenes. (a) In occlusion-free scenes, rays back-projected from the events corresponding to the same target point successfully intersect on the target plane. (b) In densely occluded scene, rays back-projected from the events tend to intersect on the occlusion plane. (c) Results of EMVS in the scenes with (w/ occ.) and without (w/o occ.) dense occlusions, and the disparity space image (DSI) slice at the target depth is also depicted.

Comparing Figs. 6b and 6c, applying Refocus-Net to event refocusing reduces reliance on prior information. Therefore, it is convenient when the target depth is hard to estimate or the camera motion cannot be acquired accurately, enhancing the practicality of E-SAI.

4.1.2 Training Refocus-Net

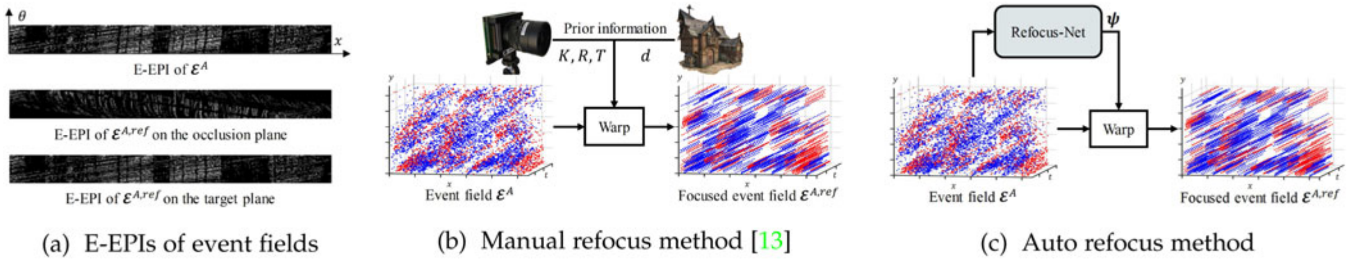
Directly training the Refocus-Net with the ground truth warping parameter is usually difficult to converge due to the severe noise issue caused by occlusions. Instead, we propose to train the Refocus-Net with the aid of the reconstruction module. We first train a reconstruction network over the event streams refocused by Eq. (8) with the ground truth ψ. By fixing the weights of the reconstruction network, the Refocus-Net can then be trained with the supervision of the reconstruction module and the reference image. One can find more training details in Section 6.1.2.

4.2 Reconstruction With a Hybrid Network

According to Eq. (5), the brightness intensity of the occluded scene is closely related to the number of events. Thus the image of the occluded scene can be recovered by accumulating events after the refocusing process or counting the rays back-projected from events at the target depth, e.g., the disparity space image (DSI) of EMVS [39] shown in Fig. 5c. However, the DSI slice is often noisy under dense occlusions as the rays back-projected from noise events $\mathcal{E}_\theta^{OO}$ are also counted without filtering. Even though CNN-based methods can be further exploited to alleviate the noise problem, the temporal information inside events cannot be effectively used. Because of this, we propose a hybrid neural network composed of an SNN encoder and a CNN decoder, where both spatial and temporal information of events can be efficiently considered and utilized, as depicted in Fig. 7.

4.2.1 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 [47], where *spike firing* and *leakage* mechanisms contribute to noise suppression. Specifically, the membrane potential of LIF neurons is constantly leaking but occasionally charging when


θ
x
E-EPI of $\mathcal{E}^A$
E-EPI of $\mathcal{E}^{A,ref}$ on the occlusion plane
E-EPI of $\mathcal{E}^{A,ref}$ on the target plane
Prior information
K, R, T
d
Warp
Event field $\mathcal{E}^A$
Focused event field $\mathcal{E}^{A,ref}$
Refocus-Net
ψ
Warp
Event field $\mathcal{E}^A$
Focused event field $\mathcal{E}^{A,ref}$

(a) E-EPIs of event fields (b) Manual refocus method [13] (c) Auto refocus method

Fig. 6. Illustration of event refocusing. (a) Based on the example in Fig. 3, we draw two E-EPIs of the event fields $\mathcal{E}^{A,ref}$ refocused on the occlusion and target planes. Although some lines in the E-EPI on occlusion plane are not strictly vertical due to the non-uniform surface of our wooden fence occlusions, events are successfully aligned on both occlusion and target planes. However, E-EPI on occlusion plane is spatially sparse since events are aligned on occlusion edges, while E-EPI on target plane is spatially dense as events encode the texture information of occluded scenes. (b) Manual refocus via Eq. (6) strongly relies on prior information of camera motions and target depth, while (c) auto refocus via spatial transformer (Refocus-Net) can adaptively align signal events, which largely facilitates the E-SAI for real-world scenarios.

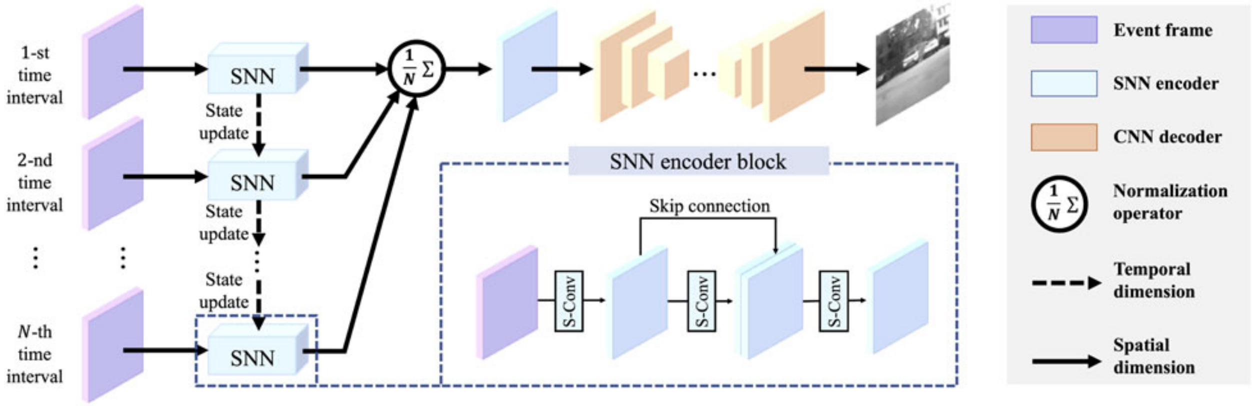


Fig. 7. Structure of the hybrid SNN-CNN network. The spatio-temporal information of events is first encoded by SNN blocks, and then transformed to visual images by the CNN decoder. To reduce the information loss of events, we add skip connections between the outputs of the 1^{st} and 2^{nd} spiking convolution (S-Conv) layers.

feeding with events, and spikes are fired as long as the membrane potential exceeds the spiking threshold. Thus temporally dense spikes are more likely to activate the LIF neurons than the isolated ones as shown in Fig. 8, which enables LIF neurons to suppress dispersed noise events but preserve aligned signal events.

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

$$u_n^l(t) = \alpha u_n^l(t-1) + c_n^l(t), \tag{11}$$

where $\alpha \in [0, 1]$ denotes the decay factor and $c_n^l(t)$ is the input current to neuron-n. Considering the convolution operation in spiking layers, Eq. (11) can be reformulated as

$$u_n^l(t) = \alpha u_n^l(t-1) + \sum_m w_{mn} o_m^{l-1}(t-1), \tag{12}$$

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. (12)

$$u_n^l(t) = \alpha u_n^l(t-1)(1 - o_n^l(t-1)) + \sum_m w_{mn} o_m^{l-1}(t-1), \tag{13}$$

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

$$o_n^l(t) = \begin{cases} 1, & \text{if } u_n^l(t) > U_{th}, \\ 0, & \text{otherwise,} \end{cases} \tag{14}$$

and U_{th} represents the spiking threshold. Eq. (13) 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,

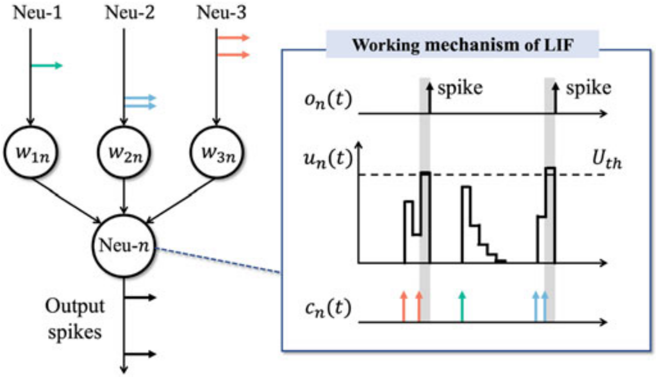


Fig. 8. An illustrative example of the LIF neuron and its working mechanism. The spikes from pre-synaptic neurons are first weighted and then fed into the target neuron-n, charging the internal membrane potential $u_n(t)$. Spikes will be fired whenever $u_n(t) > U_{th}$. Thanks to the spike firing and leakage mechanisms, the LIF neuron is able to filter out the isolated spikes, e.g., the noise events scattered out in spatial and temporal dimensions.

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. Although spiking neurons are able to directly process asynchronous event data when implemented on neuromorphic hardwares, many supervised learning methods are based on the discretized frame representation by stacking events to facilitate SNN training [11], [47], [48]. In this work, we mainly follow previous works [47], [48] to build our SNN encoder. As illustrated in Fig. 7, our SNN encoder is implemented with 3 neural layers composed of LIF neurons. To make a balance between computational complexity and information integrity, we present a spatio-temporal representation for events. The refocused event sequence is evenly divided into a pre-defined number of intervals N. In each interval, an event frame is generated by accumulating events over time, and each event 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. After that, we generate the output of SNN encoder $\mathcal{O}_s$ by normalizing the output spike tensor over time

$$\mathcal{O}_s = \frac{1}{N}\sum_{t=1}^{N} \mathbf{o}(t), \tag{15}$$

where $\mathbf{o}(t)$ denotes the output spike tensor of the time interval t. 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 SNNs [12], we instead implement the decoder with CNNs.

4.2.2 CNN Decoder

Features extracted from the SNN encoder are then fed into a style-transfer network to reconstruct visual images. Here, we adopt the decoder architecture from the generator network used in [49] 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.

4.2.3 Training Hybrid Network

The synaptic weights in SNN can be trained in a supervised fashion via the spatio-temporal backpropagation (STBP) technique [47], [48], where the gradient of each pixel can be derived based on time intervals. And CNN can be trained via backpropagation (BP). Thus the SNN and CNN in the proposed hybrid network can be jointly trained.

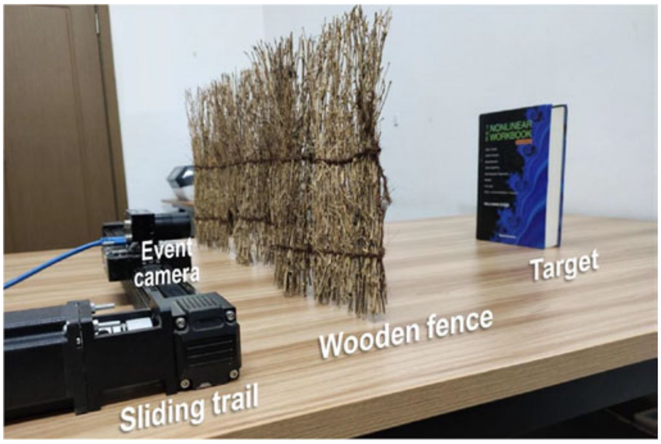


Fig. 9. An example of our experimental setup: occlusions (the wooden fence), targets (the book) and an event camera installed on a programmable sliding trail.

To guide the training, we first exploit the idea of perceptual loss [50] 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

$$\mathcal{L}_{per}(Y, \hat{Y}) = \sum_k \frac{\lambda_k}{C_k H_k W_k} \|\phi_k(Y) - \phi_k(\hat{Y})\|_2^2, \tag{16}$$

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

$$\mathcal{L}_{pix}(Y, \hat{Y}) = \frac{\|Y - \hat{Y}\|_1}{CHW}, \tag{17}$$

where $C \times H \times W$ represents the shape of Y and $\hat{Y}$. Besides, the total variance loss $\mathcal{L}_{tv}(Y)$ in [51] is exploited to encourage the spatial smoothness of reconstruction. Thus, the total loss can be summarized as follows:

$$\mathcal{L}(Y, \hat{Y}) = \beta_{per}\mathcal{L}_{per}(Y, \hat{Y}) + \beta_{pix}\mathcal{L}_{pix}(Y, \hat{Y}) + \beta_{tv}\mathcal{L}_{tv}(Y), \tag{18}$$

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

5 SAI Dataset

Due to the lack of available datasets for comparing F-SAI and E-SAI methods, we build a new SAI dataset containing both image frames and event streams. A DAVIS346 camera [6] is employed for data collection since it can output both events and gray-scale active pixel sensor (APS) frames.

As displayed in Fig. 9, we install the event camera on a programmable sliding trail and employ a wooden fence as dense occlusions. When the camera moves linearly on the sliding trail, the triggered events can be collected from different viewpoints, and APS frames are captured simultaneously by the DAVIS346 camera. We also collect the APS frames without occlusions (occlusion-free APS frames) for reference. Then event sequences and APS frames with occlusions are spatially matched with occlusion-free APS frames. To achieve this, we first generate an event frame by accumulating refocused events over time and then choose the occlusion-free image with the highest structural similarity (SSIM) [52] as the ground truth. All APS frames are collected under a fixed range of exposure time (25.0-38.4 ms) and each event stream experiences around 0.7 seconds under constant camera moving speed (17.7 cm/s).

In the SAI dataset, a large variety of targets are considered, including 2D printed pictures and 3D objects in simple and complex real-world scenarios. They are occluded by the wooden fence to imitate the very dense occlusions, as shown in Fig. 9, and some of them are captured under extreme lighting conditions, i.e., under/over exposure scenes. For clarity, we divide the SAI dataset into two main categories according to the shooting scenes: *indoor* and *outdoor*. The *indoor* dataset contains printed pictures and simple objects, while the *outdoor* dataset only contains real complex scenes. For under exposure scenes, we first acquire the occluded frames and events in *indoor* environments with the lights off, and then turn on the lights to capture the reference image. Regarding over exposure scenes, we collect the occluded frames and events using DAVIS346 in sunny *outdoor* environments and capture the reference images with an iPhone 11 Pro in HDR mode. Thus, there is no spatially matched occlusion-free APS frame for the extreme lighting scenes due to the over/under exposure problem.

In summary, the SAI dataset is built with 588 groups of data, including 488 groups for *indoor* and 100 groups for *outdoor*. To quantify the occlusion density in our SAI dataset, we first generate occlusion mask by subtracting the occlusion-free images from the corresponding occluded frames, and then calculate the proportion of occlusion pixels in the mask as occlusion density at the reference viewpoint. Overall, the occlusion density in our SAI dataset ranges from 73.5% to 99.8%, with the average value equal to 90.8%. More details of our SAI dataset can be found in the supplementary material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2022.3227448. And our SAI dataset is released at https://dvs-whu.cn/projects/esai/.

6 Experiments and Analysis

This section evaluates and analyzes the proposed E-SAI method. In Section 6.1, we first present the experimental settings, including the evaluation prototype and implementation details. The performance of different F-SAI and E-SAI methods are then compared in Section 6.2, under dense occlusions and extreme lighting conditions. After that, we analyze the effectiveness of refocusing and reconstruction modules of our proposed E-SAI method respectively in Secs. 6.3 and 6.4. Finally, we discuss the limitations of our E-SAI method and the related future work in Section 6.5.

6.1 Experimental Settings

6.1.1 Evaluation Prototype

For the frame-based SAI, we employ a representative traditional method [1] (*F-SAI+ACC*) and the recent state-of-the-art learning-based method [5] (*F-SAI+CNN*) for comparison. We use 30 APS frames with occlusions for reconstruction in F-SAI+ACC and stack them in chronological order as the input of F-SAI+CNN. For the event-based SAI, we compare three different reconstruction methods, including the accumulation method (*E-SAI+ACC*), CNN-based method (*E-SAI+CNN*), and the proposed hybrid network (*E-SAI+Hybrid*). We denote E-SAI+Hybrid with manual refocus and auto refocus methods as E-SAI+Hybrid (M) and E-SAI+Hybrid (A), respectively. The detailed network architecture is provided in the supplementary material, available online.

For E-SAI+ACC, we directly accumulate both positive and negative events along the time dimension after refocusing on the target plane. Specifically, for the refocused event field $\mathcal{E}^{A,ref}$, we define $\mathcal{E}^{A,ref}(\mathbf{x}) \triangleq \{e_i | e_i \in \mathcal{E}^{A,ref}, \mathbf{x}_i = \mathbf{x}\}$ as the set of events at pixel $\mathbf{x}$ and generate the event frame $E(\mathbf{x}) \triangleq |\mathcal{E}^{A,ref}(\mathbf{x})|$. Due to the lack of event threshold η, we further apply a min-max normalization to obtain the E-SAI +ACC results

$$E_{\text{ACC}} = \frac{E - \min(E)}{\max(E) - \min(E)}.$$

Regarding the E-SAI+CNN method, the refocused event frames are stacked as a $2N$-channel tensor (corresponding to 2 polarities) for network input. For the sake of fairness, we also build a pure CNN counterpart by simply replacing the SNN encoder with a 3-layer CNN, with the same number of network parameters as the SNN encoder. Therefore, we can evaluate the effectiveness of the SNN encoder by comparing E-SAI+Hybrid with E-SAI+CNN.

6.1.2 Training Details

Networks implemented in this paper are all trained on NVIDIA GeForce RTX 2080 Ti GPUs with batch size 12 for around 500 epochs, and the initializer in [53] is applied. The Adam optimizer [54] is used with the initial learning rate setting to 5×10^{-4} and the SGDR schedule [55] by setting $T_{max} = 64$ (reset the learning rate every 64 epochs).

Data Augmentation. We choose 90% scenes of the SAI dataset for training and leave the rest for the testing phase. In training reconstruction networks, we apply flipping (horizontal, vertical, and horizontal-vertical) and random rotating (random angles ranging from -15 to 15 degrees) to ground truth images and the refocused event frames for data augmentation. Meanwhile, the reflection padding technique is applied to reduce boundary artifacts. For the training of Refocus-Net, we only perform flipping for augmentation since rotation will break the parallax structure. And the horizontal flipping is performed together with the temporal order flipping of event streams to maintain the original parallax structure of event field.

F-SAI+CNN. The pre-trained models of F-SAI+CNN in [5] are oriented to a two-dimensional camera array. Sliding the camera in our experimental setting corresponds to a 1×30

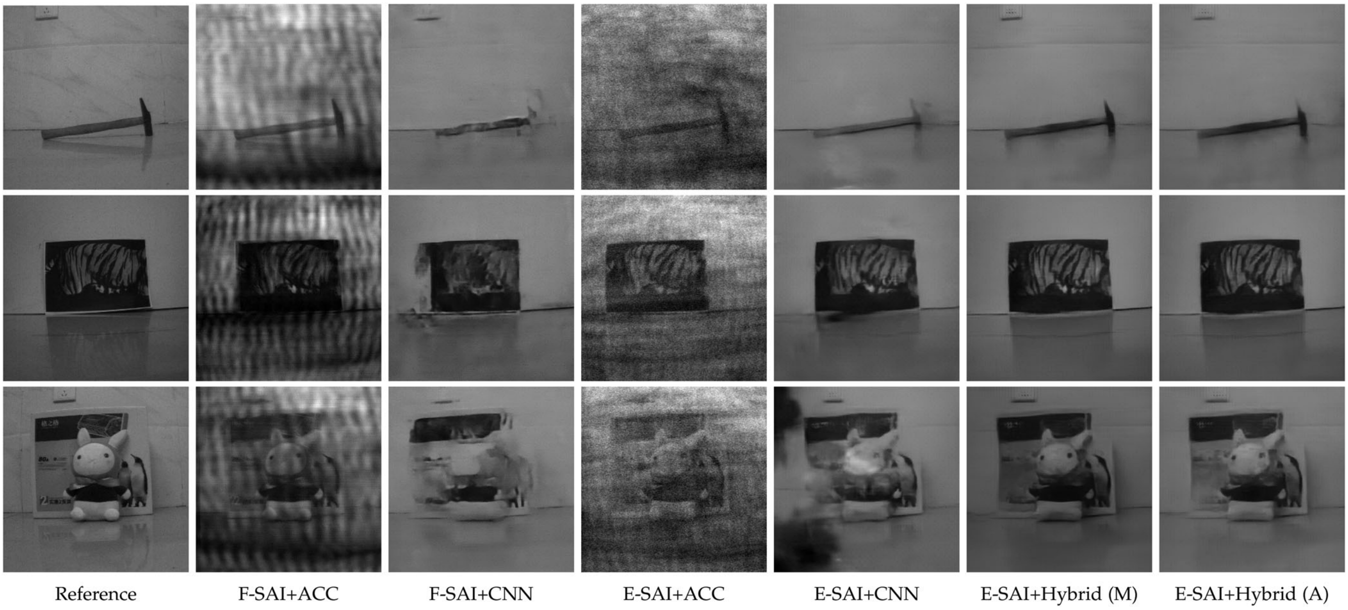


Fig. 10. Qualitative comparisons between F-SAI and E-SAI algorithms under very dense occlusions for *indoor* dataset.

camera array. For fair comparisons, we re-train F-SAI+CNN over our SAI dataset with the official codes[2].

E-SAI Networks. To train E-SAI+CNN and E-SAI+Hybrid, we divide each event sequence into 30 slices with equal time intervals, i.e., $N=30$, and set the loss weights as $[\beta_{per}, \beta_{pix}, \beta_{tv}] = [1, 32, 2\times10^{-4}]$. The 16-layer VGG network [56] pre-trained on the ImageNet dataset [57] is employed to calculate the perceptual loss based on the 2-nd, 4-th, 7-th, and 10th convolution layers and the corresponding loss weights are respectively $[\lambda_2, \lambda_4, \lambda_7, \lambda_{10}] = [1\times10^{-1}, 1/21, 10/21, 10/21]$.

Refocus-Net. The Refocus-Net is fed with a $2N$-channel tensor composed of $2N$ event frames from the unfocused event stream and outputs the warping parameter $\boldsymbol{\psi}$. We first pre-train the E-SAI+Hybrid network with manually refocused event streams using ground truth warping parameters $\boldsymbol{\psi}$. After that, fixing the weights of the pre-trained E-SAI+Hybrid network and replacing the manual refocus module with the Refocus-Net, we train the Refocus-Net individually by backpropagating the image reconstruction loss. The Refocus-Net is finally trained for 200 epochs using the same training strategy (Adam optimizer, batch size 12 and SGDR schedule) and loss function as the hybrid network, except we set the initial learning rate to 3×10^{-4}.

6.2 Comparisons Between F-SAI and E-SAI

In this subsection, we first evaluate the performance of SAI methods respectively over the indoor and outdoor datasets with dense occlusions. Comparisons of the proposed E-SAI +Hybrid networks with manual and auto refocusing methods to the state-of-the-art methods are made qualitatively and quantitatively. After that, the superiority of E-SAI to F-SAI under extreme lighting conditions is also validated.

6.2.1 Results With Dense Occlusions

The qualitative results on indoor and outdoor scenes are presented in Figs. 10 and 11, respectively. In the indoor experiments, we mainly test F-SAI and E-SAI methods with simple objects. As displayed in Fig. 10, the reconstruction results of F-SAI methods are severely disturbed by dense occlusions where a lot of details are missing or blurred. This is because the signal information for F-SAI, i.e., I^A_θ in Eq. (1), is heavily contaminated by occlusions, and a lot of redundant information like the texture of occlusions is recorded in each input frames, as shown in Fig. 1c. Thus, although F-SAI +CNN can achieve a better de-occlusion effect than F-SAI +ACC via deep learning-based methods, the results still suffer from detail losses and artifacts. Compared with F-SAI, E-SAI methods can retain more details and produce results with better visual effects since the major signal events for E-SAI, i.e., $\mathcal{E}^{OA}_\theta$ in Eq. (3), can be effectively triggered by dense occlusions. To reveal the advantages of our hybrid network, we compare the results of E-SAI with different reconstruction techniques in Fig. 12 and Table 1. It is difficult for E-SAI+ACC to produce satisfactory quantitative results since the emission of events is based on the log-scale brightness change, 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 resolved by E-SAI+CNN and E-SAI+Hybrid. However, E-SAI+CNN fed directly with the stacked event frames cannot efficiently deal with the temporal information of asynchronous events, thus degrading the visual quality with detail losses, artifacts, and saturation. These issues can be well mitigated by the hybrid architecture where the SNN encoder utilizes temporal information. Furthermore, LIF neurons in SNN can efficiently leak out the influence of noise events which are either emitted randomly or scattered after the refocusing process. Consequently, the proposed hybrid network generates images with more uniform structure and realistic details.

For the outdoor dataset, we consider more general targets, including cars, fields, and buildings. Compared with the indoor scenes, outdoor lighting conditions are much more complicated, making it harder for F-SAI methods to generate sharp results, as shown in Fig. 11. Similarly, complex lighting

[2] Official codes at https://github.com/YingqianWang/DeOccNet.

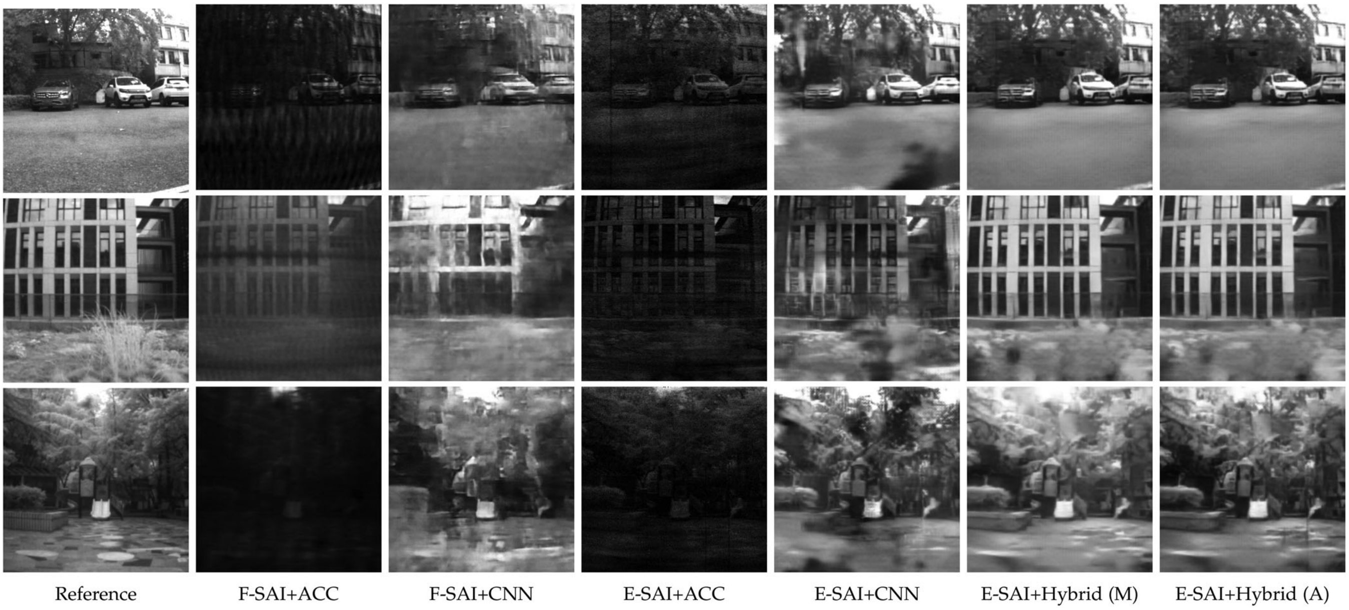


Fig. 11. Qualitative comparisons between F-SAI and E-SAI algorithms under very dense occlusions for *outdoor* dataset.

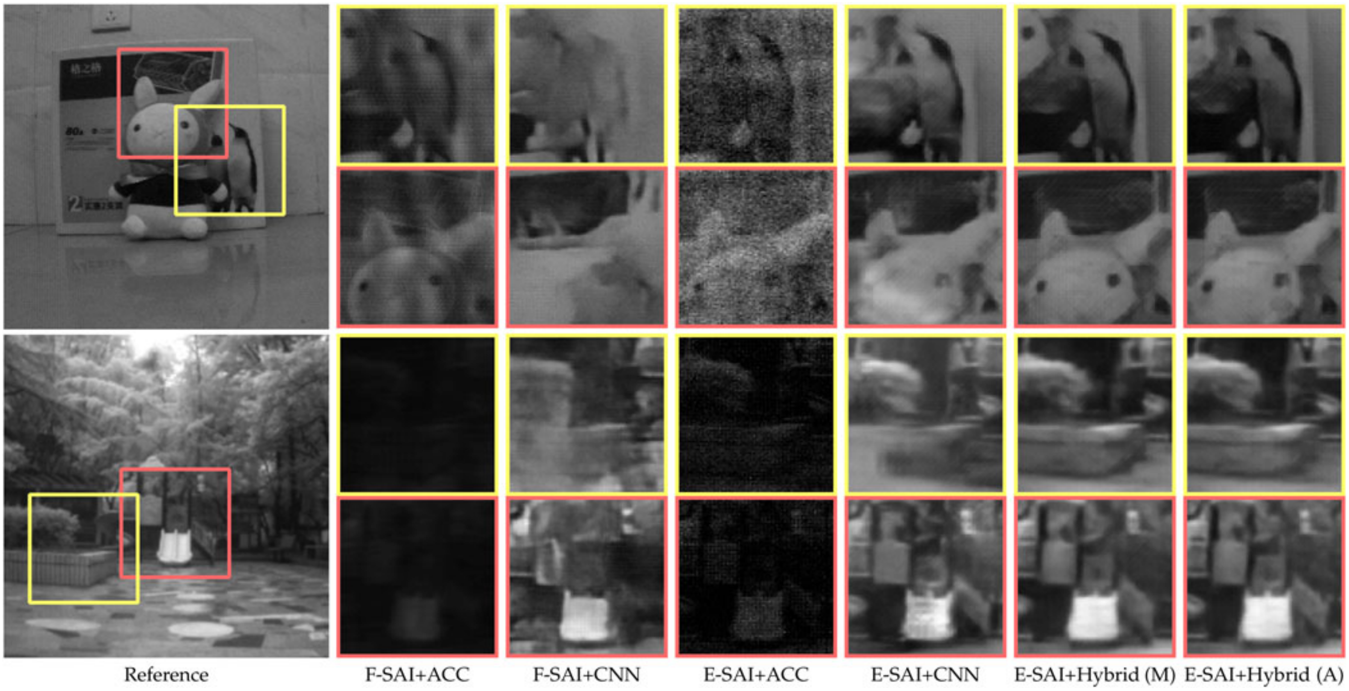


Fig. 12. Comparisons of F-SAI and E-SAI with different reconstruction methods. Details are zoomed in for better view.

conditions also degrade the performance of E-SAI due to the increased noise events, e.g., $\mathcal{E}_\theta^{OO}$. The rising number of noise events makes the target indistinguishable in the results of E-SAI+ACC and brings more disturbances to E-SAI+CNN, deteriorating the reconstruction quality with severe saturation problems. Thanks to the hybrid SNN-CNN architecture, noise events can be alleviated from the temporal dimension. In Table 1, our E-SAI+Hybrid method excels its pure CNN counterpart with a 4 dB increase in PSNR, a 24% improvement in SSIM, and a 44% decrease in LPIPS. This shows that the use of an SNN encoder achieves a better denoising effect and improves the overall reconstruction performance, which is consistent with the qualitative results shown in Figs. 11 and 12.

To further verify the superiority of E-SAI over F-SAI, we investigate the performance of F-SAI+ACC with more angular sampling. We collect up to 200 frames for 10 indoor scenes during camera movement. Fig. 13 shows that the reconstruction performance of F-SAI+ACC becomes better when the input rises from 20 to 120 frames since more target information, i.e., I_θ^A in Eq. (1), is acquired from multiple viewpoints. With sufficient angular sampling, the effect of $\mathcal{M}^O$ is alleviated and the occluded scenes can be fully observed as shown in Fig. 13b with F-SAI (100). However, further increasing the number of observations, e.g., 120-200 views, hardly results in better reconstruction quality of F-SAI since the disturbance from dense occlusions, i.e., I_θ^O,

TABLE 1
Quantitative Comparisons of F-SAI and E-SAI

Method	Indoor			Outdoor		
	PSNR(dB) ↑	SSIM ↑	LPIPS ↓	PSNR(dB) ↑	SSIM ↑	LPIPS ↓
F-SAI+ACC [1]	14.22	0.3484	0.2955	11.86	0.3925	0.2821
F-SAI+CNN [5]	25.01	0.7933	0.1245	17.39	0.5616	0.1347
E-SAI+ACC	14.96	0.2608	0.3219	10.26	0.2956	0.2875
E-SAI+CNN	26.53	0.7653	0.0871	15.98	0.5240	0.1838
E-SAI+Hybrid (M) [13]	**30.71**	**0.8311**	**0.0374**	**20.39**	**0.7037**	**0.0981**
E-SAI+Hybrid (A)	29.55	0.8086	0.0546	**20.39**	0.6961	0.1013

Results are the average over all test sequences of the corresponding datasets. Our E-SAI+Hybrid outperforms the state-of-the-art SAI method (F-SAI+CNN) with a 3-5 dB improvement in peak signal to noise ratio (PSNR) and a 24%-56% decrease in perceptual distance (LPIPS) [58] on our SAI dataset.

exists at each view and simultaneously increases with more input frames. Therefore, despite more angular sampling brings more target information for reconstruction, the performance of F-SAI is limited by dense occlusions. Compared with F-SAI, our E-SAI method produces visual images with better contrast and less noise, validating its superiority over F-SAI under dense occlusions.

6.2.2 Results With Extreme Lighting Conditions

Extreme lighting conditions often degrade the reconstruction quality of F-SAI methods or even lead to failure reconstruction, as shown in Fig. 14. This is because the light from the occluded target cannot be correctly measured due to the over/under exposure problems encountered with frame-based cameras, as displayed in Fig. 1c. By contrast, E-SAI methods do not suffer from over/under exposure problems thanks to the high dynamic range of event cameras. Therefore, the light information of occluded targets can be reliably captured by event cameras and effectively reconstructed via our E-SAI methods under extreme lighting conditions.

We compare the performance of E-SAI methods in extreme lighting scenes. Although the refocusing process scatters out the noise events, their presence still degrades the image quality of E-SAI+ACC since all the events are directly accumulated in the results. For the learning-based E-SAI methods, the saturation issue becomes more severe in E-SAI+CNN because the distribution of the collected event field in extreme lighting scenes differs from that in standard lighting scenes. By exploiting the SNN encoder to utilize the spatio-temporal information inside events, this issue can be mitigated by E-SAI+Hybrid, thus generating results with more natural textures and better contrast.

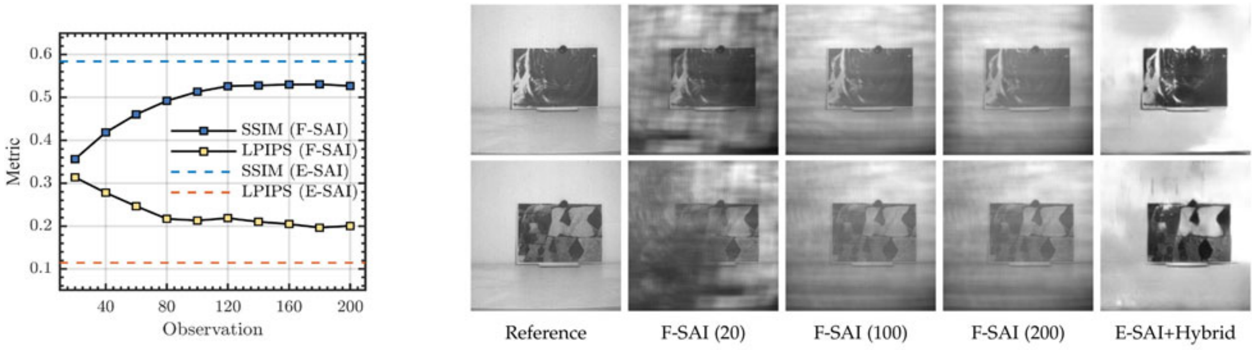
6.3 Analysis of Refocus Module

An in-depth analysis to the refocus module will be given in this subsection, including the influence of the refocusing accuracy to the reconstruction performance, the comparison of auto and manual refocus methods, and the effectiveness of the proposed Refocus-Net.

6.3.1 Influence of Warping Parameter Error

In this subsection, we investigate the robustness of E-SAI +Hybrid to the estimation error of the camera poses R, T and the target depth d. Since our data is mainly recorded under the fronto-parallel camera motions as depicted in Fig. 2, we only need to consider the warping parameter $\boldsymbol{\psi}$ in the horizontal direction (ψ represents the horizontal warping parameter hereinafter). Denoting the ground truth ψ as $\hat{\psi}$, we apply the warping projection Eq. (8) to two random pairs of data respectively selected from indoor and outdoor datasets with the parameter ratio $\psi/\hat{\psi}$ varying from 0.4 to 1.6, and reconstruct the corresponding visual images with our hybrid network.

As illustrated in Figs. 15a and 15b, the reconstruction quality of indoor data is severely degraded if the warping parameter is under/over estimated. Since targets of the indoor dataset are usually close to the camera (i.e., close-view targets), even a small error of the warping parameter will cause a significant pixel shift of events on the imaging plane. As a consequence, it makes the failure alignments of the signal events and thus leads to severe blurs and missing



(a) Quantitative results

(b) Qualitative results

Fig. 13. Quantitative (a) and Qualitative (b) results of F-SAI+ACC under different numbers of observations, where F-SAI (20/100/200) represents the F-SAI+ACC result with 20/100/200 views and results of E-SAI are provided for comparison.

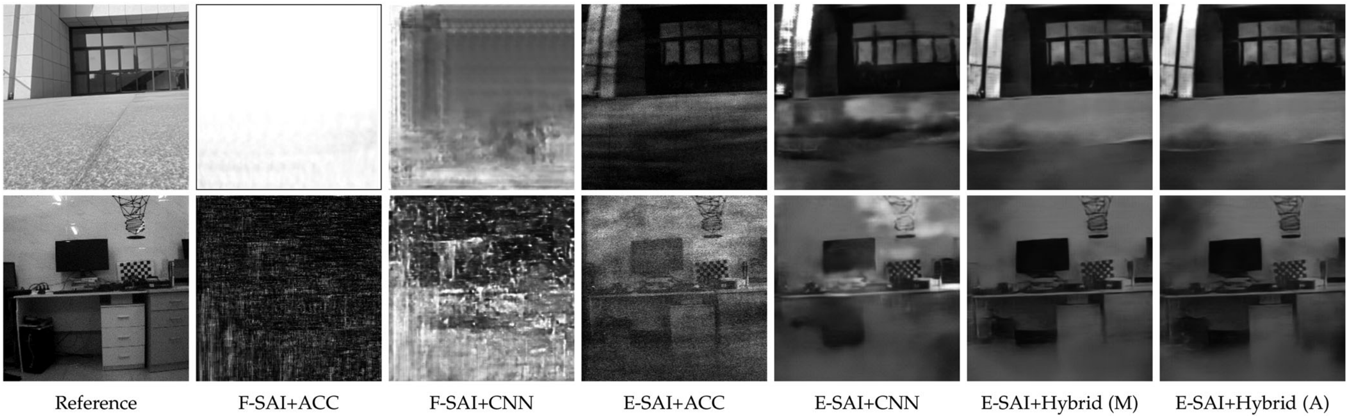


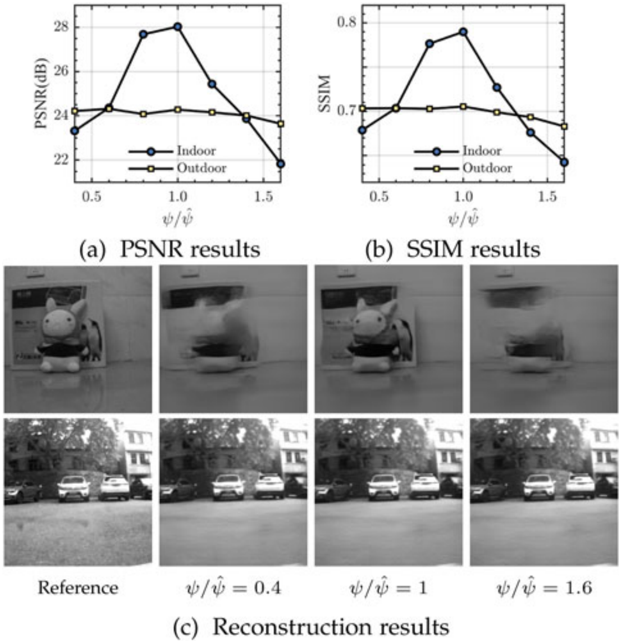
Fig. 14. Qualitative comparisons between F-SAI and E-SAI algorithms under extreme lighting conditions, where the reference images are captured under well lighting conditions.

details in the final results as shown in Fig. 15c. On the other hand, E-SAI on the outdoor dataset mainly contains far-view targets. Therefore, the corresponding E-SAI performance is less sensitive to the estimation error of the warping parameter than that of the indoor dataset.

6.3.2 *Auto Versus Manual Refocus Module*

According to above discussions, the prior information of the camera velocity and target depth is essential for the refocus module, especially when encountering close-view target scenes. To investigate the performance of the proposed auto refocus method, we respectively employ Eq. (6) (manual refocus) and a pre-trained Refocus-Net (auto refocus) to generate the refocused events and apply the same hybrid network for reconstruction. Thus, we can evaluate the performance of Refocus-Net based on both reconstruction quality and event refocusing accuracy. To quantitatively assess the refocusing accuracy, we introduce the average pixel shift error (APSE)

$$\mathrm{APSE} \triangleq \frac{1}{N_e}\sum_{i=1}^{N_e}\left\|(\boldsymbol{\psi}_i - \hat{\boldsymbol{\psi}}_i)\Delta t_i\right\|,$$



(a) PSNR results

(b) SSIM results

(c) Reconstruction results

Fig. 15. Influence of the warping parameter ψ: (a-b) Quantitative and (c) qualitative results of indoor and outdoor scenes with different $\psi/\hat{\psi}$.

TABLE 2
APSE of Refocus-Net Over All *Indoor* and *Outdoor* Test Sequences and the Corresponding Performance Drops of E-SAI+Hybrid From (M) [13] to (A)

Scenes	APSE	PSNR Drop	SSIM Drop	LPIPS Drop
Indoor	0.722	1.16	0.0225	0.0172
Outdoor	1.047	0	0.0076	0.0032

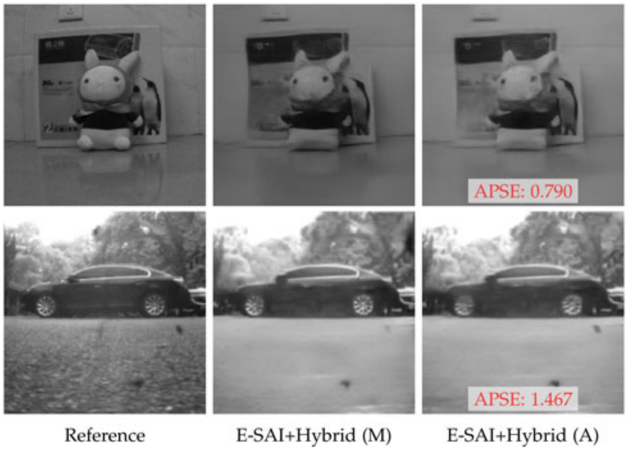
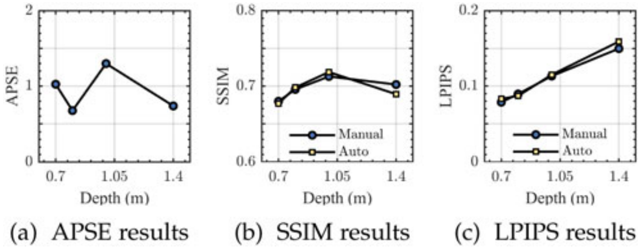


Fig. 16. Qualitative results generated by the E-SAI+Hybrid with manual refocus (M) and auto refocus (A).

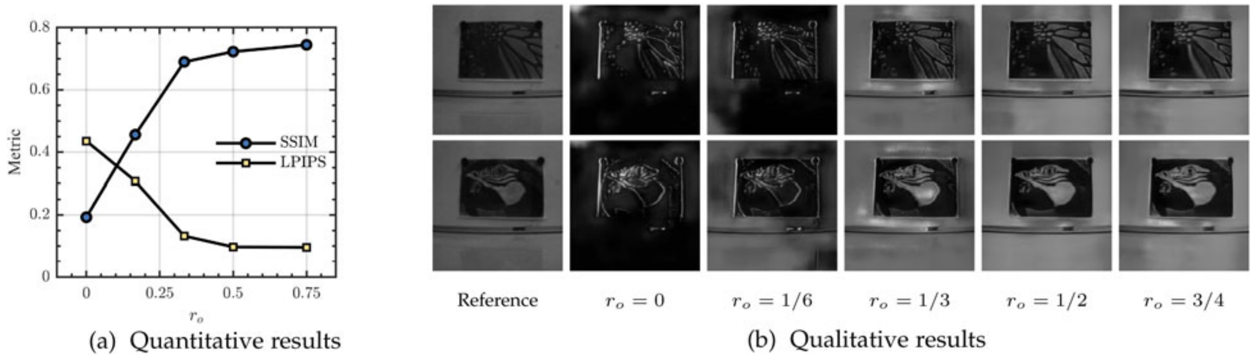


(a) APSE results

(b) SSIM results

(c) LPIPS results

Fig. 17. Quantitative comparisons under different target depths: (a) Refocus error (APSE) of Refocs-Net; (b-c) Reconstruction performance of E-SAI+Hybrid (A/M).



(a) Quantitative results

(b) Qualitative results

Fig. 18. Performance of E-SAI+Hybrid under different occlusion densities (r_o indicates the ratio of the occluded area).

where $\Delta t_i = |t_i - t^{ref}|$, N_e denotes the total number of events input to the Refocus-Net, and $\boldsymbol{\psi}_i, \hat{\boldsymbol{\psi}}_i$ indicate the predicted and ground truth warping parameters, respectively. APSE measures the pixel alignment error caused by auto refocus method compared to the manual refocus one, and lower values indicate more accurate pixel alignment. Note we only compute the APSE of horizontal warping parameter here since our experiments only includes horizontal motion.

In Table 2, applying the auto refocus method results in an average APSE of 0.7-1 pixels, which is generally acceptable from the qualitative perspective as depicted in Fig. 16. Although the APSE of the outdoor dataset is relatively high, the corresponding performance degradation is still tolerable since outdoor data is less sensitive to the error of ψ (strongly related to APSE) as discussed in Section 6.3.1. Thus, under fronto-parallel camera motions shown in Fig. 2, our auto refocus method can largely relax the dependence on camera motions and target depth with only slight performance degradation and thus facilitate practical usage.

6.3.3 *Effectiveness of the Refocus-Net*

We further validate the effectiveness of the Refocus-Net on the same targets at different depths d varying from 0.7 to 1.4 meters. It can also be regarded as testing Refocus-Net with different camera speeds or camera intrinsic parameters since they are coupled to the warping parameter ψ. The proposed auto refocus method, i.e., Refocus-Net, can successfully estimate the warping parameter ψ and achieve promising performance under different target depths, as shown in Fig. 17a. Thus E-SAI +Hybrid method with Refocus-Net can achieve comparable performance to that with the manual refocus approach, as shown in Figs. 17b and 17c. And the Refocus-Net can even boost the performance of E-SAI+Hybrid and achieve better results than the manual refocusing method. This is because the spatial matching method employed in our dataset (detailed in Section 5) is based on the SSIM value calculated from the noisy

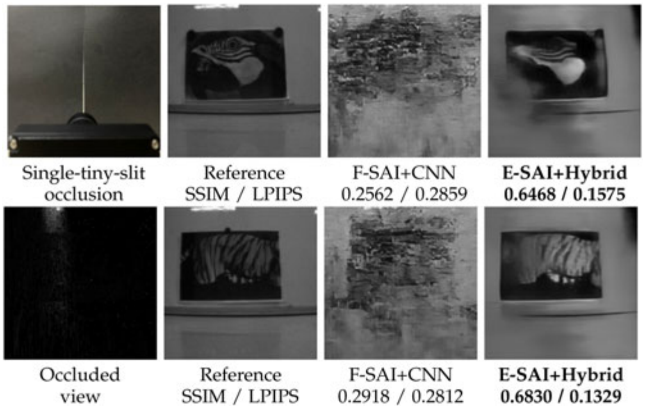
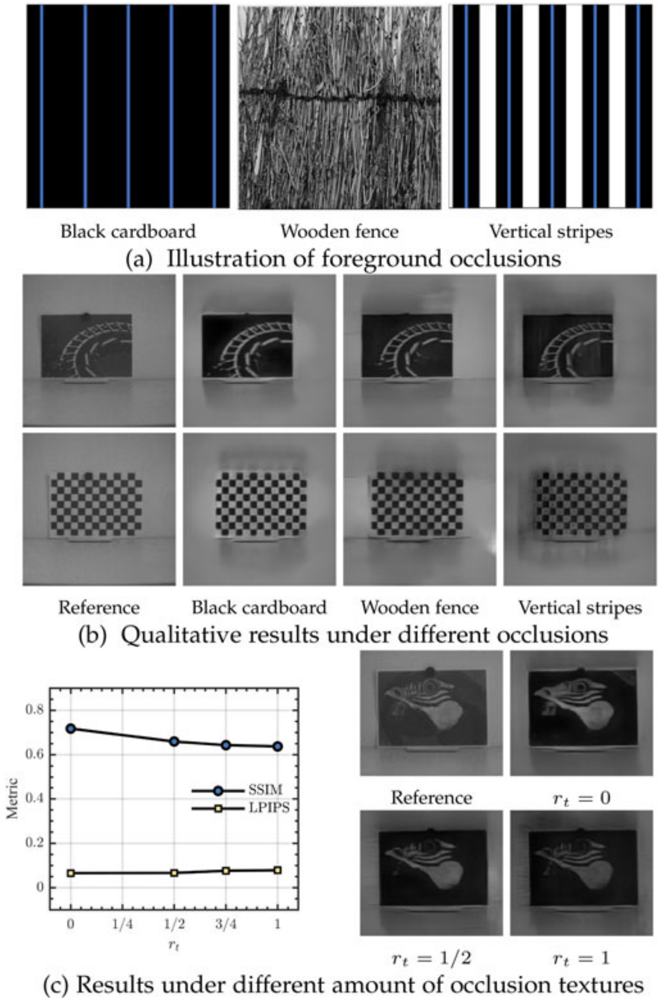


Fig. 19. Comparisons between the state-of-the-art SAI method (F-SAI +CNN) and the proposed E-SAI+Hybrid to see through a single-tiny-slit.



(a) Illustration of foreground occlusions

(b) Qualitative results under different occlusions

(c) Results under different amount of occlusion textures

Fig. 20. Influence of occlusion textures. (a) Different types of occlusions with the blue lines indicating the cropped slits where the event camera can see the occluded scene. (b) Reconstruction results under different foreground occlusions. (c) We employ the high contrast vertical stripes in (a) with different r_t as occlusions and provide the corresponding results, where r_t denotes the ratio of the number of occlusion edges for $\mathcal{E}_\theta^{OO}$ (boundaries of stripes) to that for $\mathcal{E}_\theta^{OA}$ (boundaries of cropped slits).

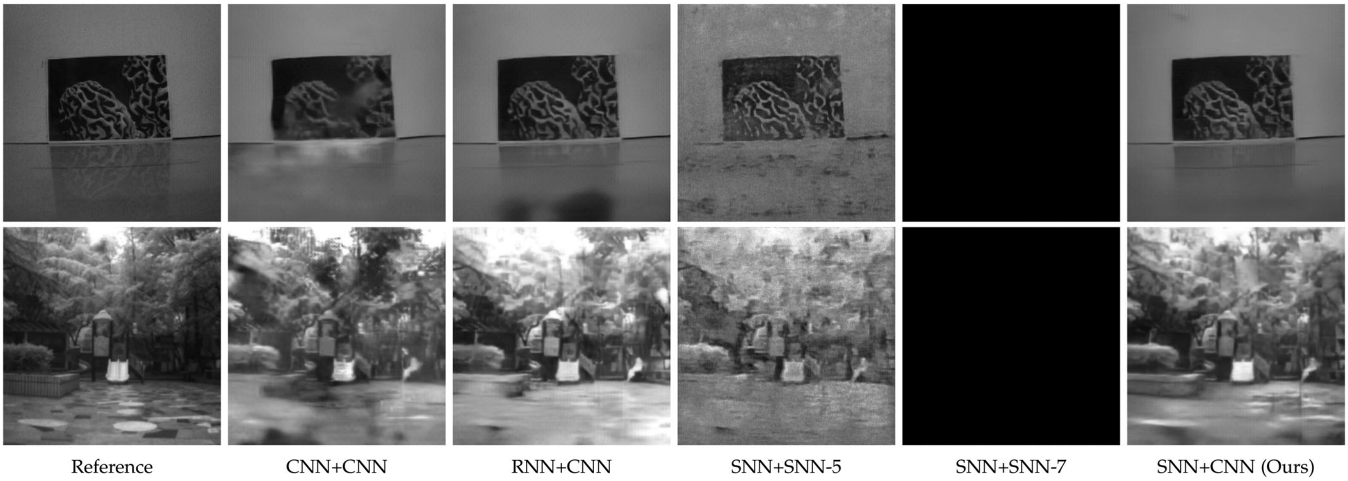


Fig. 21. Qualitative results of E-SAI with different reconstruction networks.

E-SAI+ACC results and the corresponding occlusion-free APS frames, and thus may occasionally cause a mismatch. By contrast, the Refocus-Net is supervised directly by the image reconstruction error, as discussed in Section 4.1.2.

6.4 Analysis of Reconstruction Module

The effectiveness of the reconstruction module is evaluated in this subsection, including the influences of occlusion density and occlusion texture, the ablation study on the proposed E-SAI+Hybrid, and the analysis of hyper-parameters.

6.4.1 *Influence of Occlusion Density*

Based on the results in Section 6.2.1, the proposed E-SAI +Hybrid method achieves promising results on our SAI dataset with dense occlusions, but how well does it perform under more general or sparse occlusions? To investigate this point, we replace the wooden fence with cuttable cardboards as the occlusion for quantitative evaluation and separately analyze the performance of the reconstruction module under the different occlusion densities. To facilitate the quantification of occlusion density, we introduce the metric $r_o \triangleq A_o/A_m$ where A_o, A_m respectively denote the area of occlusions and the total observation area during camera moving. Then we conduct experiments with r_o ranging from 0 (occlusion-free) to 3/4.

As shown in Fig. 18, the proposed E-SAI+Hybrid performs better in both qualitative and quantitative perspectives when the occlusion density increases. In occlusion-free scenes, the collected events are mainly caused by edges of target scenes with high-contrast regions, i.e., $\mathcal{E}_\theta^{AA}$ in Eq. (3), which will be aligned on target edges after the refocusing process, leading to edge-like reconstructions as shown in Fig. 18b with $r_o = 0$. Thus for the occlusion-free scenes, one can directly reconstruct the visual image using event to video translation methods, e.g., E2VID [19]. When occlusion ratio increases, more signal events can be collected, i.e., $\mathcal{E}_\theta^{OA}$, which encodes the overall scene texture of targets since it responds to both low and high contrast target regions by exploiting foreground occlusions as brightness reference. As depicted in Fig. 18, the overall texture of target scene can be restored when $r_o \geq 1/3$, and the performance of E-SAI becomes better as the occlusion ratio increases.

TABLE 3
Quantitative Results Averaged Over All Test Sequences With the Same Refocus-Net and Different Reconstruction Networks (Recon-Nets), Where SNN-5 and SNN-7 Indicate 5-Layer and 7-Layer Spiking Neural Networks, Respectively

Recon-Net	Indoor	Outdoor
Enc.+Dec.	PSNR / SSIM / LPIPS	PSNR / SSIM / LPIPS
CNN+CNN	26.53 / 0.7653 / 0.0871	15.98 / 0.5240 / 0.1838
RNN+CNN	28.38 / 0.7993 / 0.0571	18.26 / 0.6477 / 0.1244
SNN+SNN-5	23.34 / 0.5015 / 0.1298	15.61 / 0.4334 / 0.1964
SNN+SNN-7	7.33 / 0.0008 / 0.7173	6.36 / 0.0006 / 0.8610
SNN+CNN	**29.55** / **0.8086** / **0.0546**	**20.39** / **0.6961** / **0.1013**

TABLE 4
Quantitative Results of E-SAI With Different Numbers of Time Intervals N, Averaged Over All Test Sequences

N	Indoor	Outdoor
	PSNR / SSIM / LPIPS	PSNR / SSIM / LPIPS
1	29.30 / 0.8073 / 0.0491	17.80 / 0.5714 / 0.1635
3	29.89 / 0.8156 / 0.0421	18.99 / 0.6067 / 0.1412
6	30.22 / 0.8176 / 0.0398	19.47 / 0.6139 / 0.1377
15	30.60 / 0.8218 / 0.0377	20.31 / 0.6586 / 0.1081
30	**30.71** / **0.8311** / **0.0374**	**20.39** / **0.7037** / **0.0981**

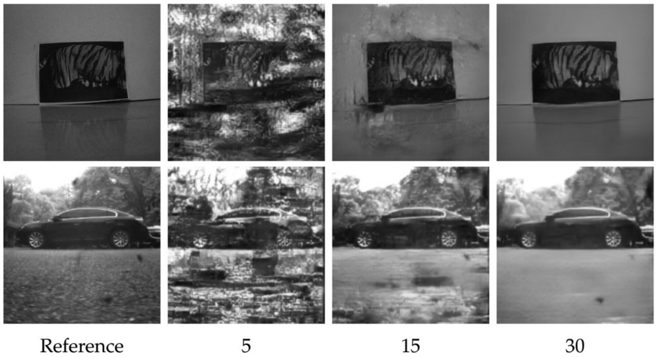


Fig. 22. Qualitative results of E-SAI with events of the first 5, 15, and 30 time intervals.

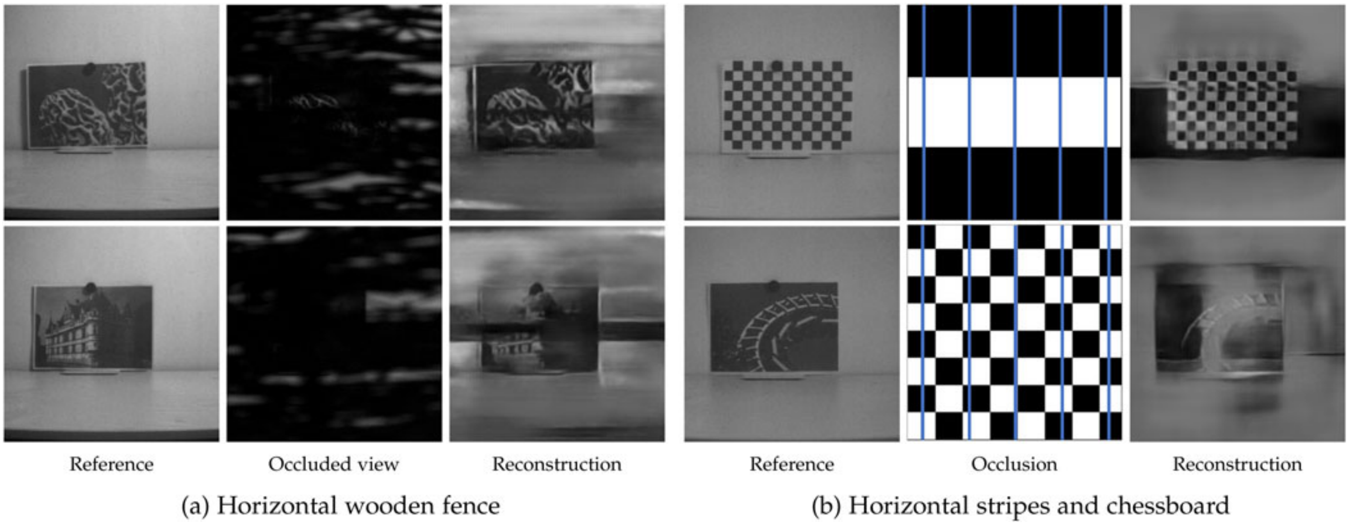


Fig. 23. Examples of failure cases. (a) Reconstruction results under horizontal dense wooden fence. (b) Reconstruction results under the occlusions of horizontal stripes and chessboard, where the blue lines indicate the cropped slits where the event camera can see the occluded scene.

We also considered an extremely occluded scene where the camera can only observe the target scene through a single tiny slit on the cardboard, as shown in Fig. 19. Compared to the state-of-the-art F-SAI+CNN method, the proposed E-SAI+Hybrid gains a 142% increase in SSIM and a 48% decrease in LPIPS. Qualitatively, F-SAI+CNN fails to reconstruct the occluded scenes while E-SAI+Hybrid can produce visually acceptable results with clear shapes and natural textures. Thanks to the low temporal latency, the event camera can “scan” the target scene through the tiny slit, acquiring enough information to guarantee the performance of E-SAI+Hybrid.

6.4.2 Influence of Occlusion Texture

The noise events $\mathcal{E}_\theta^{OO}$ triggered by the high contrast occlusion edges will interfere with reconstruction. To study the influence of occlusion texture, we perform experiments with two additional occlusions as depicted in Fig. 20a, where black cardboard triggers few noise events with homogeneous texture and vertical stripes emit a large number of noise events $\mathcal{E}_\theta^{OO}$ with high contrast stripe boundaries.

As shown in Fig. 20b, the reconstruction performance is related to the number of noise events, where our E-SAI method generates the most reliable texture and brightness under the occlusion of black cardboard while the results of vertical stripes are disturbed by heavy noise events $\mathcal{E}_\theta^{OO}$. To further qualify the influence of occlusion texture on our E-SAI method, we employ multiple vertical stripes and denote r_t as the ratio of the number of occlusion edges for $\mathcal{E}_\theta^{OO}$ (boundaries of stripes) to that for $\mathcal{E}_\theta^{OA}$ (boundaries of cropped slits) with higher r_t indicating more foreground texture to emit $\mathcal{E}_\theta^{OO}$. As shown in Fig. 20c, although noise events $\mathcal{E}_\theta^{OO}$ will bring disturbances to the reconstruction results, the overall texture of occluded scenes is still successfully restored by our E-SAI method even under severe noise disturbances, e.g., $r_t = 1$.

6.4.3 Ablation Studies of the Hybrid Network

In this section, we analyze the hybrid architecture of our reconstruction network. For encoders, we implement two counterpart networks by replacing our 3-layer SNN encoder with a 3-layer CNN, i.e., E-SAI+CNN, and a 3-layer RNN (using ConvLSTM as [19]) for comparison. The results in Table 3 demonstrate that the RNN encoder surpasses the CNN one by efficiently utilizing the temporal information of events in a recurrent manner. Compared with RNN, our SNN encoder not only exploits spatio-temporal information in events but also alleviates the influence of noise events via the leakage and spike firing mechanisms in spiking neurons, achieving the best performance.

Although SNN shows promising results in processing spatio-temporal events, we opt for the hybrid architecture instead of pure SNNs since (i) *vanishing spike phenomenon*: it is commonly observed in deep SNNs that spike activities dramatically reduce as the network depth grows [12], which often causes information loss in the results; (ii) *low numerical precision*: SNN outputs sparse and binary spike signals (in input and hidden layers), which are usually not sufficient for high-quality image restoration tasks. In our experiments, we design two decoders with 5-layer SNN and 7-layer SNN, denoted by SNN-5 and SNN-7, respectively. As depicted in Fig. 21, SNN-5 cannot restore high quality details of target scenes due to the low numerical precision issue, and SNN-7 totally fails in reconstruction since spikes vanish in the output layer. Apparently, our proposed hybrid architecture not only alleviates the disturbances from noise events but also prevents the vanishing spike phenomenon and low numerical precision problems of SNN, and thus achieves the best reconstruction performance.

6.4.4 Analysis of Hyper-Parameters

The number of time intervals N controls the granularity of temporal information in input events. In Table 4, we compare the performance of E-SAI+Hybrid with different N. When $N = 1$, the temporal information in events is discarded and only the spike firing mechanism in Eq. (14) functions, which benefits noise suppression since the spiking threshold acts as a blocker for noise events. As N increases, the reconstruction performance becomes better since both leakage and spike firing mechanisms contribute to noise suppression. In addition, the leakage mechanism gains more performance improvements with larger N since finer temporal granularity helps spiking neurons to leak out the influence of noise events. Thus, we set $N = 30$ to balance reconstruction performance and computational efficiency.

We also study the performance of E-SAI+Hybrid with partial input events under fixed $N = 30$. Specifically, we first record the SNN outputs with events at the first 5, 15, and 30 time intervals, and then feed them to the CNN decoder to obtain the intermediate reconstruction results. As shown in Fig. 22, the reconstruction suffers from noise and loss of details if only with events of the first 5 time intervals. When more events are fed to the network, e.g., the first 15 or 30 time intervals, more signal information in events is utilized and the leakage mechanism in spiking neurons better functions to alleviate noise, leading to visual results with smoother texture and less noise.

6.5 Limitations and Future Works

The principle of SAI is to see through occlusions from multi-view observations where the targets should be partially observed, and the same is true for our E-SAI. For the areas occluded from all viewpoints, e.g., horizontal camera motions with horizontal occlusions shown in Fig. 23a, the proposed E-SAI cannot achieve successful reconstructions due to the lack of signal information. This issue may be addressed by developing E-SAI methods with a 4D event field using more flexible camera motions, e.g., both horizontal and vertical motions, to acquire more effective observations of the occluded scenes.

The proposed E-SAI system can handle a large set of natural occluded scenes where the brightness of occlusion boundaries is almost uniform with limited variations. For some special cases where the boundaries of occlusions have significant varying intensities and provide different offsets, our E-SAI system may fail to reconstruct occluded targets or give reversed intensities. As shown in Fig. 23b, the reconstruction results are not consistent with the target scenes under the occlusion of horizontal stripes, where intensities vary in occlusion edges and offer different brightness offsets for different regions. Chessboard not only disturbs the reconstruction via varying brightness offsets but also emits a massive number of noise events $\mathcal{E}_\theta^{OO}$, resulting in failure reconstruction. This problem could be tackled by combining frames and events to form a multi-modal SAI method, such as [59], where the occlusion intensities contained in frames could provide reconstruction guidance for events.

Moreover, our method is tailored to restore the occluded targets with small depth variation as discussed in Section 4.1.1. For multi-depth scenes, multiple refocusing procedures might be required to align events triggered at different depths. To fulfill all-in-focus reconstruction, several techniques could be incorporated such as image matting [32] and depth estimation [31], and we leave it for future work.

7 Conclusion

In this work, we propose a novel SAI method based on event cameras (E-SAI). Benefiting from the extremely low latency and the high dynamic range of event cameras, E-SAI can achieve the seeing-through effects under very dense occlusions and extreme lighting conditions. Specifically, a Refocus-Net and a hybrid SNN-CNN network are proposed to effectively reconstruct the visual images of occluded scenes from collected events. 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. We also construct a new SAI dataset containing both events and frames for a variety of indoor and outdoor scenes. Quantitative and qualitative results over the SAI dataset show the effectiveness of our proposed E-SAI method and validate its superiority to the frame-based counterpart, i.e., F-SAI, under dense occlusions and extreme lighting scenes.

References

[1] V. Vaish, B. Wilburn, N. Joshi, and M. Levoy, “Using plane + parallax for calibrating dense camera arrays,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2004, pp. 2–9.

[2] V. Vaish, M. Levoy, R. Szeliski, C. L. Zitnick, and S. B. Kang, “Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 2331–2338.

[3] Z. Pei, Y. Zhang, X. Chen, and Y.-H. Yang, “Synthetic aperture imaging using pixel labeling via energy minimization,” *Pattern Recognit.*, vol. 46, no. 1, pp. 174–187, 2013.

[4] Z. Xiao, L. Si, and G. Zhou, “Seeing beyond foreground occlusion: A joint framework for SAP-based scene depth and appearance reconstruction,” *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 979–991, Oct. 2017.

[5] Y. Wang, T. Wu, J. Yang, L. Wang, W. An, and Y. Guo, “DeOccNet: Learning to see through foreground occlusions in light fields,” in *Proc. IEEE Conf. Winter Appl. Comput. Vis.*, 2020, pp. 118–127.

[6] P. Lichtsteiner, C. Posch, and T. Delbruck, “A 128×128 120 dB 15 μs latency asynchronous temporal contrast vision sensor,” *IEEE J. Solid-Stat. Circuits*, vol. 43, no. 2, pp. 566–576, Feb. 2008.

[7] G. Gallego et al., “Event-based vision: A survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 154–180, Jan. 2020.

[8] G. Cohen, “Active sensing and its application to neuromorphic space imaging [talk],” ICONS, 2019. [Online]. Available: https://youtu.be/mnfQwngwW78

[9] W. Maass and C. Bishop, *Pulsed Neural Networks*. Cambridge, MA, USA: MIT Press, 1998.

[10] W. Maass, “Networks of spiking neurons: The third generation of neural network models,” *Neural Netw.*, vol. 10, no. 9, pp. 1659–1671, 1997.

[11] C. Lee, A. K. Kosta, A. Z. Zhu, K. Chaney, K. Daniilidis, and K. Roy, “Spike-flownet: Event-based optical flow estimation with energy-efficient hybrid neural networks,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 366–382.

[12] P. Panda, S. A. Aketi, and K. Roy, “Toward scalable, efficient, and accurate deep spiking neural networks with backward residual connections, stochastic softmax, and hybridization,” *Front. Neurosci.*, vol. 14, 2020, Art. no. 653.

[13] X. Zhang, W. Liao, L. Yu, W. Yang, and G.-S. Xia, “Event-based synthetic aperture imaging with a hybrid network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14230–14239.

[14] D. Gehrig, H. Rebecq, G. Gallego, and D. Scaramuzza, “EKLT: Asynchronous photometric feature tracking using events and frames,” *Int. J. Comput. Vis.*, vol. 128, no. 3, pp. 601–618, 2020.

[15] J. Hagenaars, F. Paredes-Vallés, and G. De Croon, “Self-supervised learning of event-based optical flow with spiking neural networks,” *Adv. Neural Inform. Process. Syst.*, vol. 34, pp. 7167–7179, 2021.

[16] A. R. Vidal, H. Rebecq, T. Horstschaefer, and D. Scaramuzza, “Ultimate SLAM? Combining events, images, and IMU for robust visual SLAM in HDR and high-speed scenarios,” *IEEE Trans. Robot. Autom.*, vol. 3, no. 2, pp. 994–1001, Apr. 2018.

[17] M. Mostafavi, L. Wang, and K.-J. Yoon, “Learning to reconstruct HDR images from events, with applications to depth and flow prediction,” *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 900–920, 2021.

[18] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "Events-to-video: Bringing modern computer vision to event cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3852–3861.

[19] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "High speed and high dynamic range video with an event camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 1964–1980, Jun. 2021.

[20] B. Wang, J. He, L. Yu, G.-S. Xia, and W. Yang, "Event enhanced high-quality image recovery," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 155–171.

[21] L. Pan, R. Hartley, C. Scheerlinck, M. Liu, X. Yu, and Y. Dai, "High frame rate video reconstruction based on an event camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2519–2533, May 2020.

[22] M. Cook, L. Gugelmann, F. Jug, C. Krautz, and A. Steger, "Interacting maps for fast visual interpretation," in *Proc. Int. Joint Conf. Neural Netw.*, 2011, pp. 770–776.

[23] H. Kim, A. Handa, R. Benosman, S.-H. Ieng, and A. Davison, "Simultaneous mosaicing and tracking with an event camera," in *Proc. Brit. Mach. Vis. Conf.*, 2014, pp. 1–12.

[24] H. Kim, S. Leutenegger, and A. J. Davison, "Real-time 3D reconstruction and 6-DoF tracking with an event camera," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 349–364.

[25] P. Bardow, A. J. Davison, and S. Leutenegger, "Simultaneous optical flow and intensity estimation from an event camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 884–892.

[26] C. Scheerlinck, N. Barnes, and R. Mahony, "Continuous-time intensity estimation using event cameras," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 308–324.

[27] G. Munda, C. Reinbacher, and T. Pock, "Real-time intensity-image reconstruction for event cameras using manifold regularisation," *Int. J. Comput. Vis.*, vol. 126, no. 12, pp. 1381–1393, 2018.

[28] L. Wang, S. M. M. I. Y.-S. Ho, and K.-J. Yoon, "Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10081–10090.

[29] Y. Zou, Y. Zheng, T. Takatani, and Y. Fu, "Learning to reconstruct high speed and high dynamic range videos from events," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2024–2033.

[30] N. Joshi, S. Avidan, W. Matusik, and D. J. Kriegman, "Synthetic aperture tracking: Tracking through occlusions," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[31] T. Yang, Y. Zhang, X. Tong, X. Zhang, and R. Yu, "Continuously tracking and see-through occlusion based on a new hybrid synthetic aperture imaging model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 3409–3416.

[32] Z. Pei, X. Chen, and Y.-H. Yang, "All-in-focus synthetic aperture imaging using image matting," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 28, no. 2, pp. 288–301, Feb. 2018.

[33] X. Zhang, Y. Zhang, T. Yang, and Y.-H. Yang, "Synthetic aperture photography using a moving camera-imu system," *Pattern Recognit.*, vol. 62, pp. 175–188, 2017.

[34] Z. Pei et al., "Occluded-object 3D reconstruction using camera array synthetic aperture imaging," *Sensors*, vol. 19, no. 3, 2019, Art. no. 607.

[35] B. Wilburn et al., "High performance imaging using large camera arrays," in *Proc. ACM SIGGRAPH*, 2005, pp. 765–776.

[36] Z. Pei, Y. Zhang, T. Yang, X. Zhang, and Y.-H. Yang, "A novel multi-object detection method in complex scene using synthetic aperture imaging," *Pattern Recognit.*, vol. 45, no. 4, pp. 1637–1658, 2012.

[37] G. Wu et al., "Light field image processing: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 926–954, Oct. 2017.

[38] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[39] H. Rebecq, G. Gallego, E. Mueggler, and D. Scaramuzza, "EMVS: Event-based multi-view stereo—3D reconstruction with an event camera in real-time," *Int. J. Comput. Vis.*, vol. 126, no. 12, pp. 1394–1414, 2018.

[40] G. Gallego, H. Rebecq, and D. Scaramuzza, "A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3867–3876.

[41] G. Gallego, M. Gehrig, and D. Scaramuzza, "Focus is all you need: Loss functions for event-based vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12280–12289.

[42] T. Stoffregen and L. Kleeman, "Event cameras, contrast maximization and reward functions: An analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12300–12308.

[43] U. M. Nunes and Y. Demiris, "Entropy minimisation framework for event-based vision model estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 161–176.

[44] J. Xu, M. Jiang, L. Yu, W. Yang, and W. Wang, "Robust motion compensation for event cameras with smooth constraint," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 604–614, 2020.

[45] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inform. Process. Syst.*, 2015, pp. 2017–2025.

[46] C.-H. Lin and S. Lucey, "Inverse compositional spatial transformer networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2252–2260.

[47] Y. Wu, L. Deng, G. Li, J. Zhu, Y. Xie, and L. Shi, "Direct training for spiking neural networks: Faster, larger, better," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 1311–1318.

[48] Y. Wu, L. Deng, G. Li, J. Zhu, and L. Shi, "Spatio-temporal backpropagation for training high-performance spiking neural networks," *Front. Neurosci.*, vol. 12, 2018, Art. no. 331.

[49] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.

[50] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.

[51] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5188–5196.

[52] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *IEEE Asilomar Conf. Sign. Syst. Comput.*, vol. 2, 2003, pp. 1398–1402.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[55] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–13.

[56] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[57] O. Russakovsky et al., "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

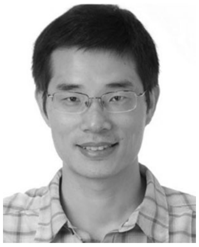
[58] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.

[59] W. Liao, X. Zhang, L. Yu, S. Lin, W. Yang, and N. Qiao, "Synthetic aperture imaging with events and frames," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17735–17744.

Lei Yu received the BS and PhD degrees in signal processing from Wuhan University, Wuhan, China, in 2006 and 2012, respectively. From 2013 to 2014, he has been a postdoc researcher with the VisAGeS Group, Institut National de Recherche en Informatique et en Automatique (INRIA) for one and half years. He is currently working as an associate professor with the School of Electronics and Information, Wuhan University. From 2016 to 2017, he has also been a visiting professor with Duke University for one year. He has been working as a guest professor with the École Nationale Supérieure de l'Électronique et de ses Applications (ENSEA), Cergy, France, for one month, in 2018. His research interests include event-based vision, neuromorphic computation, and signal processing.

Xiang Zhang received the BE degree in communication engineering from Wuhan University, Wuhan, China, in 2020. He is currently working toward the MS degree in information and communication engineering with the electronic information school, Wuhan University. His research interests include computer vision and neuromorphic computation.

Wei Liao received the BE degree in communication engineering from Wuhan University, Wuhan, China, in 2020. He is currently working toward the MS degree in information and communication engineering with the electronic information school, Wuhan University. His research interests include image processing and signal processing.



Wen Yang received the BS degree in electronic apparatus and surveying technology, the MS degree in computer application technology, and the PhD degree in communication and information system from Wuhan University, Wuhan, China, in 1998, 2001, and 2004, respectively. From 2008 to 2009, he worked as a visiting scholar with the Apprentissage et Interfaces (AI) Team, Laboratoire Jean Kuntzmann, Grenoble, France. From 2010 to 2013, he worked as a postdoctoral researcher with the State Key Laboratory of Information Engineering, Surveying, Mapping and Remote Sensing, Wuhan University. Since then, he has been a full professor with the School of Electronic Information, Wuhan University. He is also a guest professor of the Future Lab AI4EO in Technical University of Munich. His research interests include object detection and recognition, multisensor information fusion, and remote sensing image processing.

Gui-Song Xia received the PhD degree in image processing and computer vision from CNRS LTCI, Télécom ParisTech, Paris, France, in 2011. From 2011 to 2012, he has been a post-doctoral researcher with the Centre de Recherche en Mathématiques de la Decision, CNRS, Paris-Dauphine University, Paris, for one and a half years. He is currently working as a full professor in computer vision and photogrammetry with Wuhan University. He has also been working as visiting scholar with DMA, École Normale Supérieure (ENS-Paris) for two months, in 2018. He is also a guest professor of the Future Lab AI4EO in Technical University of Munich (TUM). His current research interests include mathematical modeling of images and videos, structure from motion, perceptual grouping, and remote sensing image understanding. He serves on the Editorial Boards of several journals, including *ISPRS Journal of Photogrammetry and Remote Sensing, Pattern Recognition, Signal Processing: Image Communications, EURASIP Journal on Image & Video Processing, Journal of Remote Sensing, and Frontiers in Computer Science: Computer Vision*.