Synthetic Aperture Imaging with Events and Frames

Wei Liao[1]*, Xiang Zhang[1]*, Lei Yu[1]†, Shijie Lin[2], Wen Yang[1]†, Ning Qiao[3]
[1]Wuhan University, Wuhan, China
[2]The University of Hong Kong, Hong Kong, China
[3]Chengdu SynSense Tech. Co. Ltd., Chengdu, China
{weiliao,xiangz,ly.wd,yangwen}@whu.edu.cn, lsj2048@connect.hku.hk, ning.qiao@synsense.ai

Abstract

The Event-based Synthetic Aperture Imaging (E-SAI) has recently been proposed to see through extremely dense occlusions. However, the performance of E-SAI is not consistent under sparse occlusions due to the dramatic decrease of signal events. This paper addresses this problem by leveraging the merits of both events and frames, leading to a fusion-based SAI (EF-SAI) that performs consistently under the different densities of occlusions. In particular, we first extract the feature from events and frames via multi-modal feature encoders and then apply a multi-stage fusion network for cross-modal enhancement and density-aware feature selection. Finally, a CNN decoder is employed to generate occlusion-free visual images from selected features. Extensive experiments show that our method effectively tackles varying densities of occlusions and achieves superior performance to the state-of-the-art SAI methods. Codes and datasets are available at https://github.com/smjsc/EF-SAI

1. Introduction

The Event-based Synthetic Aperture Imaging (E-SAI) [34, 35] has been recently proposed for occlusion removal, benefiting from the low latency and the high dynamic range of events. It shows promising performance, especially when facing extremely dense occlusions, as shown in Fig. 1. Unlike the frame-based SAI (F-SAI) collecting the light information from conventional frame-based cameras [8, 15, 16, 22, 33], the E-SAI collects the light information with events asynchronously triggered by brightness contrast [34, 35] in almost continuous viewpoints. Existing E-SAI approaches have shown the superiority of restoring clear images from the occluded light field with extremely dense occlusions. However, their performance degrades dramatically when occlusions become sparse (see Fig. 1).

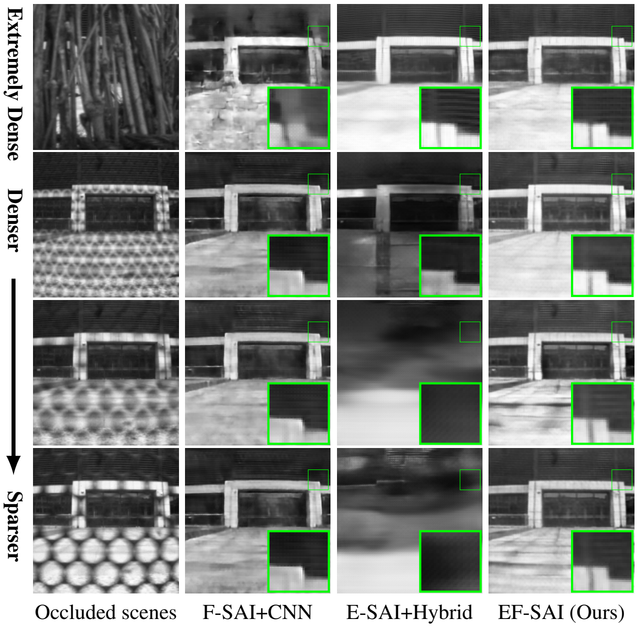

Figure 1. Qualitative comparisons of F-SAI+CNN [27], E-SAI+Hybrid [35] and our proposed EF-SAI under different densities of occlusions. The EF-SAI can reconstruct high-quality occlusion-free images under either sparse or dense occlusions by exploiting the information from both events and frames.

The principle of existing E-SAI is to predict the light field of occluded targets via accumulating events triggered by occlusion-to-target contrast, which is proportional to the intensity of occluded targets [35]. Thus, E-SAI achieves high performance in occluded scenes with dense occlusions, where the signal events caused by target-to-occlusion contrast is dominant. However, when occlusions become sparse, the performance of existing E-SAI will inevitably deteriorate as signal events decrease.

*Equal contribution

†Corresponding author

The research was partially supported by the National Natural Science Foundation of China under Grants 61871297, the Natural Science Foundation of Hubei Province, China under Grant 2021CFB467, the Fundamental Research Funds for the Central University under Grant 2042020kf0019, and the National Natural Science Foundation of China Enterprise Innovation Development Key Project under Grant U19B2004.

To deal with the SAI problem under sparse occlusions, one can alternatively employ the prior event-based imaging models [2, 17, 25] to firstly recover intensity frames from events triggered by targets and then utilize the F-SAI pipeline for occlusion removal. However, event cameras only respond to intensity contrasts theoretically until the brightness change reaches the triggering threshold [20]. Thus, the collected events may only contain light information from high contrast textures (mostly around sharp edges), leading to the missing of low contrast textures in the final SAI reconstructions if only based on events.

Motivated by the high performance of F-SAI in dealing with sparse occlusions, we propose a novel fusion-based SAI method (EF-SAI). The EF-SAI exploits intensity frames to address above problems of E-SAI, achieving high-quality reconstruction performance invariant to the density of occlusions. However, F-SAI may suffer from performance deterioration when facing dense occlusions [27, 34, 35]. Both E-SAI and F-SAI exhibit inconsistent performance under varying occlusion densities in the real-world scenarios. Thus, the main challenge of EF-SAI is straightforward, *i.e*., *how to take the merits of both events and frames to bridge the gap between E-SAI and F-SAI, and finally achieve the see-through effects with performance invariant to occlusion densities?*

To this end, we propose a novel EF-SAI-Net for high-quality reconstruction based on the fusion of events and frames. The network is based on the Encoder-Decoder architecture to encode and fuse the multi-modal signals and then decode an intensity frame without occlusions. Specifically, a two-stage fusion mechanism in our network is proposed for feature enhancement and adaptive fusion, composed of a cross-modal enhancement module and a density-aware fusion module. The cross-modal enhancement module aims to suppress event noises and occlusion disturbances by mutual compensations between features of different modalities, *i.e*., events and frames. And we employ a cross-attention-based swin transformer to achieve this end. On the other hand, events and frames have their advantages under specific conditions, *e.g*., different occlusion densities or lighting conditions. Thus, the density-aware fusion module aims to adaptively select the features with high confidence according to the occlusion densities. To this end, the multi-modal inputs are directly fed into a channel attention block to provide the information about occlusion densities, serving as a guidance for the feature selection procedure.

The main contributions of this paper are three-fold.

- We propose a novel fusion-based SAI, *i.e*., EF-SAI, which takes the merits of both E-SAI and F-SAI, to achieve high-quality imaging performance invariant to occlusion densities.
- We propose a deep neural network, *i.e*., EF-SAI-Net, to reconstruct the image of targets from the occluded events and frames, where a cross-attention module is introduced to suppress noises and disturbances, and a density-aware feature selection module is proposed for adaptive fusions.
- We train our proposed EF-SAI-Net on a new EF-SAI dataset which contains various targets under occlusions with different densities. Extensive evaluations show that our method is superior to existing SAI approaches and exhibits consistent performance with respect to different occlusion densities.

2. Related Work

2.1. Event-based Image Reconstruction

Instead of capturing the whole frame at a fixed rate, event cameras report asynchronous events which respond to per-pixel brightness changes [1, 20]. This paradigm shift in visual perception leads to many outstanding advantages, *e.g*., high dynamic range (HDR) and extremely low latency, showing great potential for practical applications like HDR imaging [13, 18, 25] and high-frame-rate videoing [12, 17, 21].

Imaging with pure events: Restoring intensity images from events is a challenging problem due to the lack of absolute brightness information inside events. Previous attempts tackle this issue by exploiting techniques like sparse coding [29] and manifold regularisation [11], but their performance often degrades in real-world scenarios due to the heavy noises induced by false negatives or temporal instability [3]. To mitigate this, learning-based approaches have been proposed to fit event distributions and directly synthesize intensity images from noisy event sequences [17, 26], achieving better imaging performance. However, existing event-driven imaging methods are generally designed for occlusion-free scenes, limiting their performance under occlusions.

Imaging with the fusion of events and frames: Due to the extremely high temporal resolution, the event camera is able to observe target scenes continuously. Several works extract motion and texture information from events and fuse them with intensity frames for motion deblurring [6,9,13,32], video interpolation [12,21,25], and super resolution [24]. However, the task of EF-SAI couples fusion and de-occlusion together, posing more challenges compared to previous works.

2.2. Synthetic Aperture Imaging

Synthetic aperture imaging tackles the disturbances of occlusions via multi-view measurement and image synthesis. The basic idea of SAI is to form the light field [5] of occluded scenes from multi-view exposures and project the lights onto a virtual focal plane, which is equivalent to imaging with a large aperture and shallow depth of field [8].

In this case, the objects on the focal plane remain sharp while the others outside the plane appear blurry, leading to the "seeing through" effect. The pioneering work [23] proposes a plane+parallax framework to warp multi-view images onto a focal plane for SAI, but the reconstructed images are often noisy and blurry due to the blending of lights from the occlusions. Several techniques have been developed to improve the reconstruction quality, including multiple cost functions [22], background subtraction [16], and occlusion labeling via energy minimization [15]. Recent work of [27] approaches SAI with the learning-based methods and proposes DeOccNet, which can effectively remove occlusions and reconstruct high-quality images. However, frame-based methods often suffer from performance degradation in densely occluded scenes due to the limited light information and severe disturbances from occlusions.

With the extremely low latency of events, event camera poses advantages in dealing with the SAI task under dense occlusions. Exploiting the brightness contrast between target scenes and dense occlusions, event-based SAI (E-SAI) methods [34, 35] are able to collect sufficient signal information of the occluded scenes and produce occlusion-free visual images from pure event data. However, their performance is inconsistent when encountering sparse occlusions due to the dramatic decrease of signal events, which often results in failure reconstruction as shown in Fig. 1. Thus, it is difficult for the existing F-SAI or E-SAI methods to deal with different densities of occlusions, which motivates us to take the merits of both events and frames and propose the EF-SAI.

3. Problem Statement

Synthetic Aperture Imaging (SAI) aims to remove occlusions by collecting the light information from multi-view measurements, *i.e.*,

$$I^A = \mathrm{SAI}(\mathcal{I}^O_{\mathcal{P}}(A), \mathcal{P}),$$

where $\mathcal{I}^O_{\mathcal{P}} \triangleq \{\mathbf{L}^O_p\}_{p\in\mathcal{P}}$ is a set of projected light information $\mathbf{L}^O_p(A)$ of the scene A captured at the camera pose $p \in \mathcal{P}$ under occlusions O which is generally unknown in practice, and I^A denotes the reconstructed image of the scene A via the SAI method. Generally, sufficient light information is required to ensure correct imaging results.

Frame-based SAI directly measures the projected intensities $I_p(A)$ of A. If A is occluded by O, the captured light information would be

$$\mathbf{L}^O_p(A) = \{I^O_p(A) \triangleq \mathcal{M}^O(I_p(A)) + I_p(O) + I^n\},$$

where $\mathcal{M}^O(\cdot)$ returns the occlusion-free pixels of $I_p(A)$ under occlusions O, $I_p(O)$ is the projected intensities from occlusions O, and I^n denotes noises. The task of F-SAI is to identify target regions $\mathcal{M}^O(I_p(A))$ from occluded observations and then remove disturbances from occlusions. Generally, F-SAI can provide reliable SAI results when occlusions O is sparse; on the contrary, when occlusions O become denser, the performance of F-SAI deteriorates as the light information from the target A becomes insufficient.

Event-based SAI is an approach that can tackle the problem of dense occlusion removal [35], where the light information is captured in the form of brightness change through an event camera. The collected light information are represented as a set of event points $E^O_p(A)$ triggered during camera movement. Typically, $E^O_p(A)$ contains four different subsets, *i.e.*,

$$\mathbf{L}^O_p(A) = \{E^O_p(A) \triangleq E^{AA}_p + E^{OA}_p + E^{OO}_p + E^n\},$$

where E^{AA}_p, E^{OA}_p, and E^{OO}_p are respectively induced by brightness changes from target textures, occlusion-to-target contrast, and occlusion textures, and E^n denotes event noises. Since events triggered by occlusion-to-target contrast E^{OA}_p is proportional to the brightness of the target A [35], the task of E-SAI is to eliminate noisy events while enhancing the signal information, *i.e.*, E^{OA}_p. Due to the low latency and high dynamic range, E-SAI is able to collect events from almost continuous viewpoint and thus can provide sufficient light information under extremely dense occlusions and poor lighting conditions [34, 35]. However, when facing sparse occlusions, E^{OA}_p decreases significantly, making it insufficient for correct reconstructions.

It is difficult for the existing E-SAI or F-SAI methods to deal with different densities of occlusions, which motivates us to utilize both frames and events as the input of SAI to compensate such inconsistency.

EF-SAI leverages the merits of both E-SAI and F-SAI, where the light information of targets A is extracted from both events and frames, *i.e.*,

$$\mathbf{L}^O_p(A) = \{E^O_p(A), I^O_p(A)\}. \tag{1}$$

The target of the EF-SAI is to recover the occlusion-free image of the target A from $\mathbf{L}^O_p(A), p \in \mathcal{P}$, *i.e.*,

$$I^A = \text{EF-SAI}(\{\mathbf{L}^O_p(A)\}_{p\in\mathcal{P}}, \mathcal{P}). \tag{2}$$

To fulfill this, two main obstacles exist for EF-SAI:

(1) Disturbances and Noises. The light information from both events and frames are contaminated by noises and disturbances of foreground occlusions, which should be necessarily suppressed. Generally, $I^O_p(A)$ is less noisy but contains more disturbances than $E^O_p(A)$, since frame-based cameras are of low frame rate and only a limited number of observations are captured. By contrast, $E^O_p(A)$ contains more structure information benefiting from the continuous observation, but it suffers from heavy noises and lack of low-contrast textures [3]. Both $I^O_p(A)$ and $E^O_p(A)$ are collected with the same target and thus share the common latent structures embedded in the light information. Thus, it is straightforward to utilize latent consistency for the mutual compensation between frames and events, promoting the performance of EF-SAI.

(2) Performance Inconsistency to Occlusion Densities. The light information and disturbance in frames and events varies with the density of occlusions. When facing dense occlusions, the information from frames becomes a burden due to the increased disturbances of $I_p(O)$ and should be suppressed during the fusion process. On the contrary, when encountering sparse occlusions, the number of E_p^{OA} decreases and E_p^{AA} becomes dominant. Thus, E_p^{AA} should be also utilized for fusion since it reflects the light information from the target, especially for scenes under extreme lighting conditions. Based on the above discussion, how to adaptively select features according to occlusion densities is another essential problem to the EF-SAI.

4. Method

4.1. Multi-modal Signals

Due to the multi-view observation, the parallax exists in both intensity pixels and events from the target A. Thus we need to first warp frames and events to refocus on the plane where target A is located. Given the depth d of target A, we implement the refocusing procedure similar as [35], and the refocused frames and events are denoted respectively as $\mathrm{I}^{ref}, \mathrm{E}^{ref}$, *i.e.*,

$$\begin{aligned} \mathrm{I}^{ref} &= \mathrm{Refocus}(I_p^O(A)), \\ \mathrm{E}^{ref} &= \mathrm{Refocus}(E_p^O(A)), \end{aligned} \tag{3}$$

where $\mathrm{Refocus}(\cdot)$ represents the refocusing operator defined in [35]. Apart from I^{ref} and E^{ref}, E_p^{AA} also contains the light information from targets and plays an important role in compensating the brightness information when encountering poor lighting conditions, where I^{ref} is often severely disturbed. To make full use of E_p^{AA}, we first reconstruct the intensity frames from $E_p^O(A)$ via the pre-trained E2VID [17] followed by the refocusing process, *i.e.*,

$$\mathrm{I}_{E\to F}^{ref} = \mathrm{Refocus}(\mathrm{E2VID}(E_p^O(A))), \tag{4}$$

and treat $\mathrm{I}_{E\to F}^{ref}$ as another information source, which complements brightness information by exploiting the pre-learned event-to-image features in E2VID. In summary, we call $\mathrm{I}^{ref}, \mathrm{E}^{ref}$, and $\mathrm{I}_{E\to F}^{ref}$ as a set of multi-modal signals, *i.e.*,

$$\mathrm{I}^{EF} \triangleq \{\mathrm{I}^{ref}, \mathrm{E}^{ref}, \mathrm{I}_{E\to F}^{ref}\} \tag{5}$$

And the effectiveness of each component of I^{EF} varies with respect to specific conditions, as shown in Tab. 1.

Table 1. Effectiveness of different signals for SAI performance. ✓denotes positive impact under specific conditions.

Signal	Dense	Sparse	Poor Lighting
I^{ref}	-	✓	-
E^{ref}	✓	-	✓
$\mathrm{I}_{E\to F}^{ref}$	-	✓	✓

4.2. The EF-SAI Network

According to Tab. 1, it is crucial for EF-SAI to leverage the merits of multi-modal signals under different densities of occlusions. To this end, we propose a novel EF-SAI-Net to adaptively fuse the light information from events and frames according to the occluded scenes, and achieve consistent performance under occlusions or poor lighting conditions. As shown in Fig. 2, our proposed EF-SAI-Net is composed of a Multi-modal Feature (MF) encoder, a Cross-modal Enhancement module (CME), a Density-Aware Fusion module (DAF), and a Multi-modal Feature Decoder. The EF-SAI-Net firstly transforms the multi-modal signals into the feature domain by the MF encoder composed of three sub-encoders. For intensity frames I^{ref} and event frames $\mathrm{I}_{E\to F}^{ref}$, two CNN-based sub-encoders are employed to separately extract shallow features $\mathrm{f}_{F,0}^{A}$ and $\mathrm{f}_{E,0}^{AA}$. For events E^{ref}, an SNN-based sub-encoder [35] is used to process the events with leaky integrate-and-fire (LIF) [31] neurons, which extract the features $\mathrm{f}_{E,0}^{OA}$ while filtering the event noises scattered by the refocusing process. Then, the outputs of the MF encoder $\mathrm{f}_0^{EF} \triangleq \{\mathrm{f}_{F,0}^{A}, \mathrm{f}_{E,0}^{OA}, \mathrm{f}_{E,0}^{AA}\}$ are fed to a two-stage fusion network composed of a CME module and a DAF module for multi-modal feature enhancement and adaptive feature selection.

Cross-Modal Enhancement. Based on the assumption that frames and events share the common structure, *e.g.*, edges and textures, of the same targets, we use a CME module to learn the coarse latent structure and achieve cross-modal enhancement. In the CME module, N Cross-attention based Swin Transformer Layers (Cross-STLs) [10] are firstly employed to mutually enhance the multi-modal features, as shown in Fig. 2. In the i-th Cross-STL, the input multi-modal features f_{i-1}^{EF} are firstly partitioned into K non-overlapping windows, denoted as $\mathrm{f}^1, \mathrm{f}^2, ..., \mathrm{f}^K$. And for every feature window $\mathrm{f}^k (k = 1, \ldots, K)$, we estimate an independent query q^j, key k^j and value matrices v^j $(j = 1, \ldots, 3)$ for each feature branch in f^k, and the self-attention weight w^j is calculated by $\mathrm{w}^j = \mathrm{q}^j {\mathrm{k}^j}^{\mathrm{T}} / \sqrt{d}$. Then, the fused self-attention weight w can be obtained by

$$\mathrm{w} = \mathrm{w}^1 + \mathrm{w}^2 + \mathrm{w}^3. \tag{6}$$

In this manner, we can mutually compensate the common latent structures of occluded scenes and suppress the noise in each feature branch, achieving cross-modal feature enhancement. Following that, we apply the fused self-attention weight w separately to $\mathrm{v}^1, \mathrm{v}^2, \mathrm{v}^3$ and generate the self-attentions by $\mathrm{Attention}\left(\mathrm{v}^j; \mathrm{w}\right) = \mathrm{softmax}(\mathrm{w})\mathrm{v}^j$. We then connect the self-attentions with their inputs in f^k and generate the enhanced feature per window via an MLP

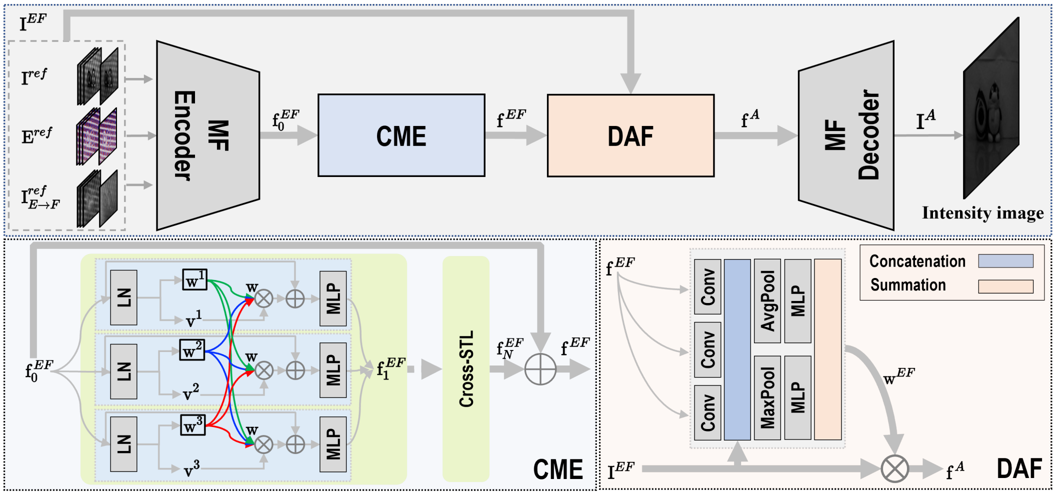

Figure 2. Architecture of the proposed EF-SAI-Net. A multi-modal feature (MF) encoder firstly extracts features $\mathrm{f}_0^{\mathrm{EF}}$ from multi-modal signals I^{EF}. Then, a two-stage fusion mechanism is employed to suppress noises and disturbances by a Cross-Modal Enhancement (CME) module and a Density-Aware Fusion (DAF) module. Finally, an MF decoder is used to decode I^A from the fused features f^A.

layer. After traversing all the feature windows, the i-th Cross-STL outputs the i-th enhanced feature f_i^{EF}, which will be fed to the next Cross-STL. Finally, the CME module produces the enhanced features f^{EF} by residually connecting the original input f_0^{EF} with f_N^{EF} output from the N-th Cross-STL.

Density-Aware Feature Selection. The DAF module is designed to predict the confidence of features and adaptively select the signal information that varies with occlusion densities. To facilitate the feature selection, we provide extra information for DAF via concatenating the multi-modal features f^{EF} with the original input signals I^{EF} since the input information, *e.g.*, I^{ref}, is often closely related to the occlusion density. After that, we employ a Channel Attention Block (CAB) [30] to predict the importance of each feature branch and produce the fused feature f^A by

$$\mathrm{f}^A = \mathrm{w}^{EF} * \mathrm{f}^{EF}, \tag{7}$$

where w^{EF} denotes the weighting vector output by DAF. Finally, an MF decoder is used to reconstruct the occlusion-free visual images from the multi-modal feature f^A.

5. The EF-SAI Dataset

We construct an EF-SAI dataset using the same manner as [35], where the events and frames are collected simultaneously by a DAVIS346 camera installed in a programmable sliding trail. The occlusion-free intensity frames are also captured as the ground truth and camera poses (translation) are determined by the constant velocity of the slider. To validate the performance of SAI under different occluded scenes, we employ *random thorn fences* and *regular wooden grids* to imitate the occlusions with different densities ranging from extremely dense to very sparse, as shown in Fig. 3. Considering the lighting variance, our EF-SAI dataset also contains *indoor* and *outdoor* scenes with different targets, including toys, desks, paintings for indoor scenes and cars, motorcycles, buildings, play yards for outdoor scenes.

Table 2. Overview of the proposed EF-SAI dataset.

Usage	Dense Occlusion		Sparse Occlusion		Summary
	Indoor	Outdoor	Indoor	Outdoor	
Train	220	138	495	27	880
Test	25	10	55	18	108
Total	245	148	550	45	988

Tab. 2 summarizes the proposed EF-SAI dataset. We totally record 988 groups of paired dataset which contains 30 APS frames and the concurrent events collected with occlusions, and 1 occlusion-free APS frame as the ground truth. Our proposed EF-SAI dataset differs from [35] with respect to both size and varieties, where the occlusion density is varying from sparse to extremely dense. Thus the EF-SAI dataset can be employed for training and evaluation of both frame-based and event-based SAI algorithms under occlusions with different densities. Our EF-SAI dataset is available at https://github.com/smjsc/EF-SAI.

6. Experiments and Analysis

The proposed EF-SAI-Net is implemented using Pytorch [14] and trained on 2 NVIDIA TITAN RTX3090 GPUs with a batch size of 4. For the training phase, we adopt the ℓ_1 loss to supervise the pixel-wise low-level features (denoted as $\mathcal{L}_{pix}$), the perceptual loss $\mathcal{L}_{per}$ to encourage network to learn the similarity between high-level visual features, and the total variation loss $\mathcal{L}_{tv}$ for noise suppression. The total loss function $\mathcal{L}$ can be formulated as

$$\mathcal{L} = \alpha\mathcal{L}_{pix} + \beta\mathcal{L}_{per} + \gamma\mathcal{L}_{tv}, \tag{8}$$

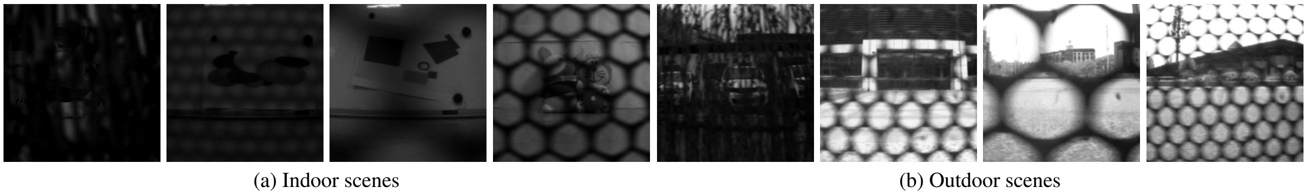

Figure 3. Occluded views of indoor (left four) and outdoor (right four) scenes under occlusions with different densities.

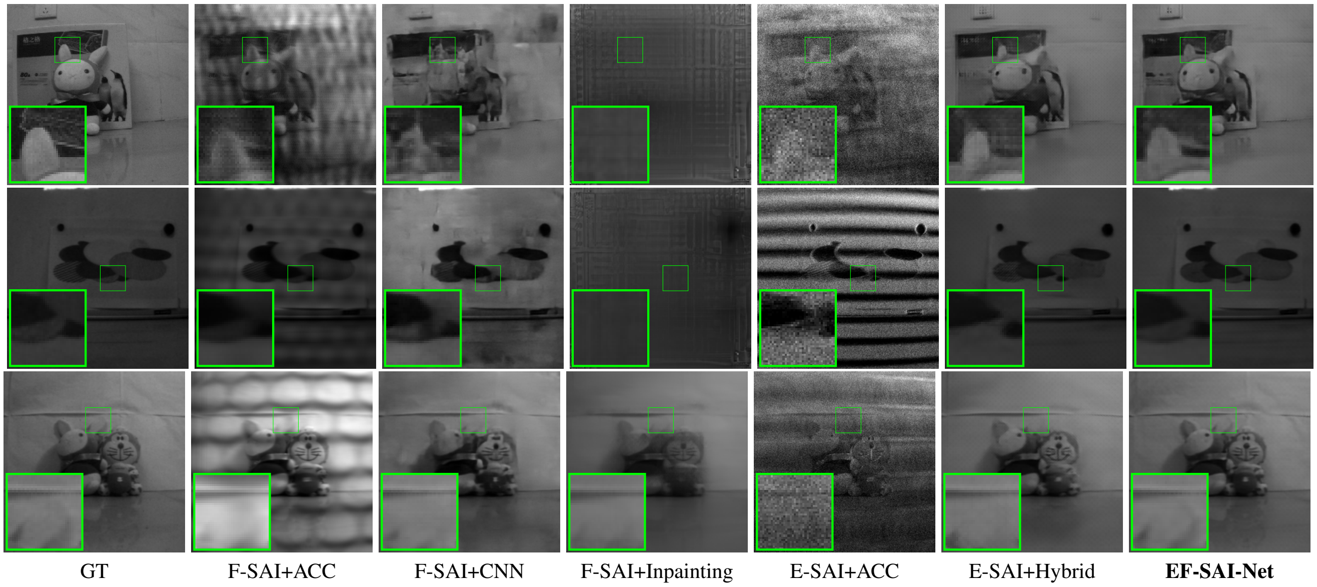

Figure 4. Qualitative comparisons on the indoor scenes under dense occlusions (1st row) and sparse occlusions (2nd to 3th rows).

with $[\alpha, \beta, \gamma] = [1, 32, 2 \times 10^{-4}]$ denoting the balancing factors. We train the EF-SAI-Net for 500 epochs with the Adam optimizer [7] and a cosine annealing strategy where the learning rate decays from 6×10^{-4} to 10^{-7} for every 64 epochs. To enhance the robustness of our network, we augment the EF-SAI dataset by flipping (horizontal, vertical, and horizontal-vertical) and rotating (random angles ranging from -10 to 10 degree) and obtain 2355 pairs of data for training. We adopt the pre-trained E2VID network [17] to generate intensity frames from event sequences, *i.e.*, $\mathrm{I}_{E \to F}^{ref}$, and implement the MF decoder with a U-net [19].

Evaluation. We compare our method against several SAI methods, including F-SAI with accumulation (**F-SAI+ACC**) [23], F-SAI with DeOccNet (**F-SAI+CNN**) [27], F-SAI with inpainting (**F-SAI+Inpainting** [4]), E-SAI with accumulation (**E-SAI+ACC**) [34], and E-SAI with the hybrid network (**E-SAI+Hybrid**) [35]. We use 30 frames as the input of F-SAI methods, *i.e.*, F-SAI+ACC, F-SAI+CNN and F-SAI+Inpainting. The metrics Peak Signal to Noise Ratio (PSNR, higher is better) and Structural SIMilarity (SSIM, higher is better) [28] are employed for quantitative evaluation.

6.1. Qualitative Analysis

From Figs. 4 and 5, the results of F-SAI+ACC are often noisy and blurry under dense occlusions since it directly accumulates all the information without distinguishing signals and noises. Exploiting the learning-based approaches, F-SAI+CNN effectively alleviates the noise issue as shown in Fig. 5, but the visual details are still heavily contaminated by the disturbances from occlusions, see Fig. 4. For densely occluded scenes, F-SAI+Inpainting cannot recover the targets due to the limited signal information. For sparsely occluded scenes, it also suffers from the blur due to the inconsistent inpainted patches on different frames. For E-SAI methods, it is difficult for E-SAI+ACC to generate satisfying visual results due to the heavy event noises. To mitigate this, E-SAI+Hybrid employs an SNN encoder to filter noises from the temporal dimension, achieving the most competing results under dense occlusions. However, the overwhelming performance of E-SAI+Hybrid is not consistent in the case of spare occlusions as shown in Fig. 5, where the SAI results suffer from severe artifacts and missing details. This is because events are only triggered when the log-scale brightness change exceeds the event threshold, thus often leading to the ignorance of low-contrast textures. By compensating the full texture information from frames, our EF-SAI method can reconstruct rich and natural details under sparse occlusions. Meanwhile, our method does not suffer from the disturbances of dense occlusions, showing the best and consistent performance under different occlusion densities.

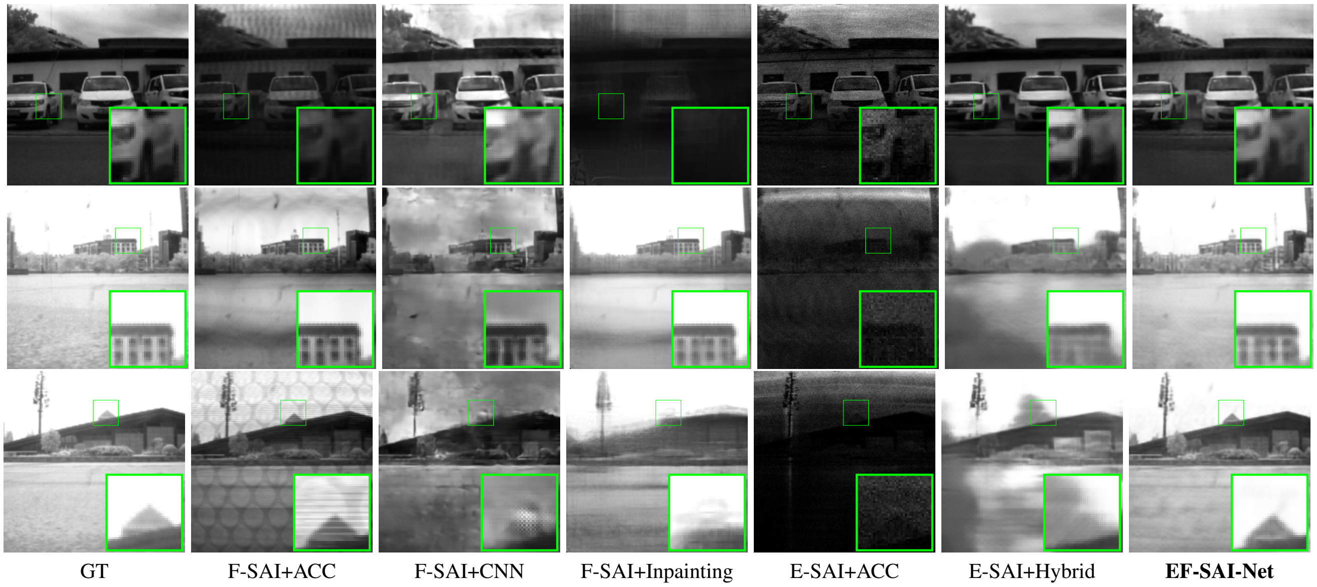

Figure 5. Qualitative comparisons on the outdoor scenes under dense occlusions (1st row) and sparse occlusions (2nd to 3th rows).

Table 3. Quantitative comparisons of EF-SAI-Net to the state-of-the-art SAI methods on indoor and outdoor scenes under different occlusion densities. The networks re-trained on our EF-SAI dataset are marked with the symbol *.

Method	Dense Occlusions				Sparse Occlusions			
	Indoor		Outdoor		Indoor		Outdoor	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
F-SAI+ACC [23]	13.3832	0.3379	11.7946	0.4056	16.8071	0.6262	12.0371	0.5695
F-SAI+CNN* [27]	24.7108	0.7805	18.1457	0.5985	22.5317	0.7991	13.3662	0.5906
F-SAI+Inpainting [4]	15.2886	0.4723	11.1015	0.2831	23.7813	0.7565	15.3313	0.5675
E-SAI+ACC [34]	14.4509	0.2202	10.4116	0.2702	15.6526	0.2438	4.4836	0.0756
E-SAI+Hybrid [35]	**31.0715**	**0.8277**	**20.2579**	**0.6879**	15.7629	0.5163	6.9290	0.2469
E-SAI+Hybrid* [35]	29.6905	0.8003	19.2968	0.6450	31.0529	0.8926	12.1227	0.4064
EF-SAI-Net (ours)	30.5387	0.8273	19.7834	0.6631	**35.0089**	**0.9279**	**21.0394**	**0.7065**

6.2. Quantitative Analysis

The quantitative results are summarized in Tab. 3. For accumulation methods, F-SAI+ACC suffers from the disturbances of occlusions while E-SAI+ACC often meets brightness inconsistency with the ground truth images, leading to poor PSNR and SSIM results. Employing the learning-based techniques, F-SAI+CNN gains a general improvement under both dense and sparse occlusions, but its performance is often limited by the number of observations. For inpainting methods, the performance of F-SAI+Inpainting is better than F-SAI+CNN when occlusions are sparse but significantly drops as the occlusion information becomes dominant. Despite E-SAI+Hybrid achieves the best PSNR and SSIM results in densely occluded scenes, its performance drops dramatically when facing sparse occlusions since the signal events E_p^{OA} becomes minority. After retraining on the EF-SAI dataset, E-SAI+Hybrid can learn to approach occlusion removal under sparse occlusions, but it also pays performance losses to balance the different distributions of varying occlusion densities. Compared to the above F-SAI and E-SAI methods, the proposed EF-SAI shows the most robust performance by taking the complementary advantage of events and frames, achieving remarkable performance under either sparsely or densely occluded scenes.

6.3. Ablation study

In this section, we study the contributions of the network modules CME and DAF, and the importance of multi-modal signals, *i.e.*, $\mathrm{I}^{ref}, \mathrm{E}^{ref}$, $\mathrm{I}^{ref}_{E\rightarrow F}$ in our EF-SAI-Net. From Tab. 4 and Fig. 6, we can draw the following conclusions:

Importance of CME and DAF. As demonstrated in Tab. 4, the CME module gains a general performance improvement under both sparse and dense occlusions via enhancing cross-modal signals while suppressing noises. For the DAF module, it plays an important role in dealing with varying occlusion densities or even extreme lighting conditions. In Fig. 6, the network without DAF module suffers from brightness inconsistency in the over-exposure scene

Table 4. Ablation study of our EF-SAI-Net w/o the network modules CME, DAF, and the multi-modal signals I^{ref}, E^{ref}, and $\mathrm{I}^{ref}_{E\to F}$.

Method	Dense Occlusions				Sparse Occlusions			
	Indoor		Outdoor		Indoor		Outdoor	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
w/o all	29.7052	0.8046	19.2030	0.6443	34.0899	0.9237	19.5192	0.6542
w/o CME	29.9534	0.8216	19.2921	0.6557	34.4912	0.9251	20.4002	**0.7138**
w/o DAF	29.9625	0.8150	19.4271	0.6615	34.9402	0.9266	19.4667	0.6641
w I^{ref}	27.9170	0.7739	18.5470	0.6081	33.0011	0.9111	20.5511	0.7034
w E^{ref}	29.1839	0.7974	18.8486	0.5949	29.6165	0.8676	12.8511	0.3715
w $\mathrm{I}^{ref}_{E\to F}$	23.9406	0.7032	14.5684	0.3660	25.1331	0.7589	12.9840	0.3558
w/o I^{ref}	29.7329	0.8114	19.4327	0.5289	29.8331	0.8751	14.2442	0.3708
w/o E^{ref}	29.5823	0.8167	19.7604	0.6483	34.8559	0.9268	20.7289	0.7012
w/o $\mathrm{I}^{ref}_{E\to F}$	30.2764	0.8240	19.2886	0.6561	34.9962	0.9208	**21.3782**	0.7057
w/ all	**30.5387**	**0.8273**	**19.7834**	**0.6631**	**35.0089**	**0.9279**	21.0394	0.7065

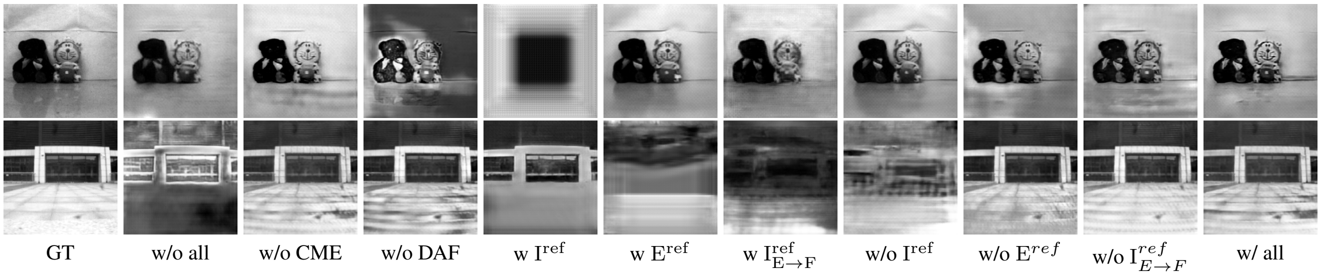
GT
w/o all
w/o CME
w/o DAF
w $\mathrm{I}^{\mathrm{ref}}$
w $\mathrm{E}^{\mathrm{ref}}$
w $\mathrm{I}^{\mathrm{ref}}_{\mathrm{E\to F}}$
w/o $\mathrm{I}^{\mathrm{ref}}$
w/o E^{ref}
w/o $\mathrm{I}^{ref}_{E\to F}$
w/ all

Figure 6. Qualitative ablation study under extreme lighting conditions (1st row) and sparse occlusions (2nd row).

and can barely reconstruct the target in the under-exposure case. This is because the network mistakenly treats the severely disturbed frames I^{ref} as signal information, showing the importance of DAF in our EF-SAI-Net.

Complementarity of Multi-Modal Signals. From the results of Tab. 4, the frame I^{ref} is essential for the SAI under sparse occlusions as it directly provides the brightness information of target scenes. Under dense occlusions, the information from events E^{ref} is dominant since it collects abundant signal information from the rich occlusion-to-target contrast. As for the event frames $\mathrm{I}^{ref}_{E\to F}$, it is helpful in dealing with extremely lighting conditions under sparse or occlusion-free scenes because it provides intensity information when the frames I^{ref} is not reliable, as shown in Fig. 6. Thus, by taking the merits of the above multi-modal signals, our EF-SAI-Net achieves consistent performance under varying densities of occlusions and does not suffer from the over- or under-exposure problems.

6.4. Computational Complexity

Tab. 5 shows the comparisons of learning-based SAI methods when inferring 256 × 256 images. Although F-SAI+Inpainting requires minimal computational resources, its performance in SAI is not satisfying according to Tab. 3. Compared to F-SAI+CNN and E-SAI+Hybrid, the proposed EF-SAI-Net can cope with varying densities of occlusions while maintaining overall efficiency with the smallest model size and the comparable computational costs. However, due to a large amount of the element-wise operations and the storage of intermediate variables in transformer-based CME module, the EF-SAI-Net requires a longer inference time, and we leave optimization for future work.

Table 5. Comparisons of computational complexity with learning-based SAI methods.

Method	FLOPs	#Param.	Infer. time
F-SAI+CNN [27]	188.71G	39.04M	29.89ms
F-SAI+Inpainting [4]	**17.67G**	52.15M	**28.35ms**
E-SAI+Hybrid [35]	167.82G	18.59M	34.74ms
EF-SAI	173.44G	**12.04M**	123.12ms

7. Conclusions

This paper introduces a novel EF-SAI method that utilizes events and frames to reconstruct high-quality occlusion-free images under various densities of occlusions. Specifically, we propose a cross-modal enhancement transformer CME to enhance the signal information while suppressing noises according to the learned latent structure of occluded targets. Following that, a density-aware feature selection DAF module is employed to judge the signal confidence and guarantee the performance consistency under different occluded scenes. To evaluate our method, we construct an EF-SAI dataset composed of both indoor and outdoor scenes under various occlusion densities. Experiments show that our method is effective in dealing with multi-modal signals and achieves robust performance under different occluded situations and extreme lighting conditions.

References

[1] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240 × 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *IJSC*, 49(10):2333–2341, 2014. 2

[2] Hadar Cohen Duwek, Albert Shalumov, and Elishai Ezra Tsur. Image reconstruction from neuromorphic event cameras using laplacian-prediction and poisson integration with spiking and artificial neural networks. In *CVPR*, pages 1333–1341, 2021. 2

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

[4] Xiefan Guo, Hongyu Yang, and Di Huang. Image inpainting via conditional texture and structure dual generation. In *ICCV*, pages 14134–14143, 2021. 6, 7, 8

[5] Aaron Isaksen, Leonard McMillan, and Steven J Gortler. Dynamically reparameterized light fields. In *PACMCGIT*, pages 297–306, 2000. 2

[6] Zhe Jiang, Yu Zhang, Dongqing Zou, Jimmy Ren, Jiancheng Lv, and Yebin Liu. Learning event-based motion deblurring. In *CVPR*, pages 3320–3329, 2020. 2

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

[8] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 31–42, 1996. 1, 2

[9] Songnan Lin, Jiawei Zhang, Jinshan Pan, Zhe Jiang, Dongqing Zou, Yongtian Wang, Jing Chen, and Jimmy Ren. Learning event-driven video deblurring and interpolation. In *European Conference on Computer Vision*, pages 695–710. Springer, 2020. 2

[10] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–1110022, 2021. 4

[11] Gottfried Munda, Christian Reinbacher, and Thomas Pock. Real-time intensity-image reconstruction for event cameras using manifold regularisation. *IJCV*, 126(12):1381–1393, 2018. 2

[12] Genady Paikin, Yotam Ater, Roy Shaul, and Evgeny Soloveichik. Efi-net: Video frame interpolation from fusion of events and frames. In *CVPR*, pages 1291–1301, 2021. 2

[13] Liyuan Pan, Richard Hartley, Cedric Scheerlinck, Miaomiao Liu, Xin Yu, and Yuchao Dai. High frame rate video reconstruction based on an event camera. *IEEE TPAMI*, 2020. 2

[14] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NIPS*, 32:8026–8037, 2019. 5

[15] Zhao Pei, Yanning Zhang, Xida Chen, and Yee-Hong Yang. Synthetic aperture imaging using pixel labeling via energy minimization. *PR*, 46(1):174–187, 2013. 1, 3

[16] Zhao Pei, Yanning Zhang, Tao Yang, Xiuwei Zhang, and Yee-Hong Yang. A novel multi-object detection method in complex scene using synthetic aperture imaging. *PR*, 45(4):1637–1658, 2012. 1, 3

[17] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *CVPR*, pages 3857–3866, 2019. 2, 4, 6

[18] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE TPAMI*, 2019. 2

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MCCAI*, pages 234–241. Springer, 2015. 6

[20] Teresa Serrano-Gotarredona and Bernabé Linares-Barranco. A 128 × 128 1.5% contrast sensitivity 0.9% FPN 3 µs latency 4 mW asynchronous frame-free dynamic vision sensor using transimpedance preamplifiers. *JSSC*, 48(3):827–838, 2013. 2

[21] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *CVPR*, pages 16155–16164, 2021. 2

[22] Vaibhav Vaish, Marc Levoy, Richard Szeliski, C Lawrence Zitnick, and Sing Bing Kang. Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures. In *CVPR*, pages 2331–2338, 2006. 1, 3

[23] Vaibhav Vaish, Bennett Wilburn, Neel Joshi, and Marc Levoy. Using plane+ parallax for calibrating dense camera arrays. In *CVPR*, 2004. 3, 6, 7

[24] Bishan Wang, Jingwei He, Lei Yu, Gui-Song Xia, and Wen Yang. Event enhanced high-quality image recovery. In *ECCV*, pages 155–171. Springer, 2020. 2

[25] Lin Wang, Yo-Sung Ho, Kuk-Jin Yoon, et al. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *CVPR*, pages 10081–10090, 2019. 2

[26] Lin Wang, Tae-Kyun Kim, and Kuk-Jin Yoon. Eventsr: From asynchronous events to image reconstruction, restoration, and super-resolution via end-to-end adversarial learning. In *CVPR*, pages 8315–8325, 2020. 2

[27] Yingqian Wang, Tianhao Wu, Jungang Yang, Longguang Wang, Wei An, and Yulan Guo. Deoccnet: Learning to see through foreground occlusions in light fields. In *WACV*, pages 118–127, 2020. 1, 2, 3, 6, 7, 8

[28] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 6

[29] Yijing Watkins, Austin Thresher, David Mascarenas, and Garrett T Kenyon. Sparse coding enables the reconstruction of high-fidelity images and video from retinal spike trains. In *Proceedings of the International Conference on Neuromorphic Systems*, pages 1–5, 2018. 2

[30] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018. 5

[31] Yujie Wu, Lei Deng, Guoqi Li, Jun Zhu, Yuan Xie, and Luping Shi. Direct training for spiking neural networks: Faster, larger, better. In *AAAI*, pages 1311–1318, 2019. 4

[32] Fang Xu, Lei Yu, Bishan Wang, Wen Yang, Gui-Song Xia, Xu Jia, Zhendong Qiao, and Jianzhuang Liu. Motion deblurring with real events. In *ICCV*, pages 2583–2592, 2021. 2

[33] Tao Yang, Yanning Zhang, Jingyi Yu, Jing Li, Wenguang Ma, Xiaomin Tong, Rui Yu, and Lingyan Ran. All-in-focus synthetic aperture imaging. In *ECCV*, pages 1–15. Springer, 2014. 1

[34] Lei Yu, Wei Liao, You-Long Zhou, Wen Yang, and Gui-Song. Xia. Event camera based synthetic aperture imaging. *Acta Automatica Sinica*, 45(x):1–14, 2020. 1, 2, 3, 6, 7

[35] Xiang Zhang, Wei Liao, Lei Yu, Wen Yang, and Gui-Song Xia. Event-based synthetic aperture imaging with a hybrid network. In *CVPR*, pages 14235–14244, 2021. 1, 2, 3, 4, 5, 6, 7, 8