Event-based Synthetic Aperture Imaging with a Hybrid Network
- Supplementary Material -

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

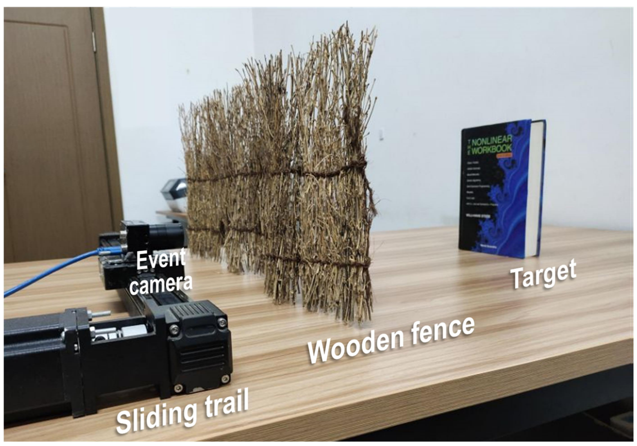


Figure 1: An example of our experimental scenes. It consists of occlusions (wooden fence), targets (book) and an event camera installed on a programmable sliding trail.

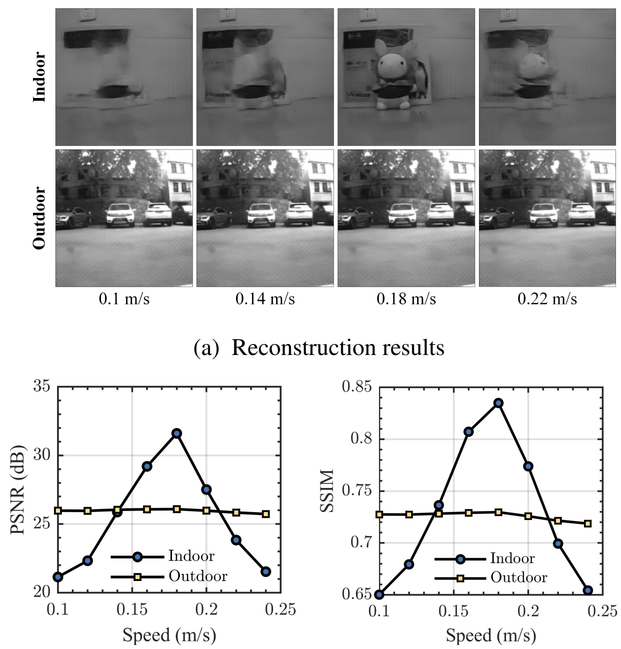
1. Experimental Scenes

As displayed in Fig. 1, we install the event camera on a programmable sliding trail and employ a wooden fence to simulate the densely occluded scenes. When the camera moves linearly on the sliding trail, the events triggered by the brightness difference between occlusions and targets can be collected from different viewpoints.

2. Extra Experimental Results

2.1. Influence of Camera Motion Accuracy

Our experiments are mainly based on one-dimensional uniform camera motion, thus the estimation of the camera speed is directly related to the refocusing module and the overall performance of our proposed method. To investigate this, we choose two pairs of data (from indoor and outdoor datasets each) and apply E-SAI+Hybrid to them with camera speeds varying from $0.1\ m/s$ to $0.24\ m/s$. Note that the actual camera speed is $0.177\ m/s$.



(a) Reconstruction results

(b) PSNR results

(c) SSIM results

Figure 2: Influence of camera motion accuracy. (a) Reconstruction results of indoor and outdoor data under different estimates of camera speed. (b), (c) Quantitative results. The real camera speed is $0.177\ m/s$.

As illustrated in Fig. 2, the reconstruction quality of indoor data is severely degraded when the speed estimation error is large. Since the targets in our indoor dataset are often closer to the camera (*i.e.* close-view targets), estimation error will cause a significant shift of targets on the imaging plane. Thus these signal events cannot be reliably aligned during refocusing and may be treated as noise during reconstruction, leading to serious blur and missing details in final results. On the contrary, our outdoor dataset mainly contains far-view targets, and thus is less sensitive to the

†Corresponding author
*Equal contribution

Figure 3: Qualitative comparisons between F-SAI and E-SAI algorithms on the *indoor* dataset.

estimation error.

2.2. Analysis of Indoor and Outdoor Results

In the experiments on indoor dataset, we mainly test F-SAI and E-SAI methods with simple objects. As displayed in Fig. 3, the results of F-SAI and E-SAI+ACC are often blurry and noisy since the light information of both occlusions and targets are equally treated during reconstruction. For the learning-based SAI, F-SAI+CNN is able to recover the shape of targets and achieve a better de-occlusion effect than F-SAI, but the result still suffers from the issues of detail losses and artifacts. On the other hand, it is hard for E-SAI+CNN to simultaneously handle spatial and temporal information inside events, thus the reconstruction quality is often degraded by the disturbance of noise events.

For the outdoor dataset, we consider more general targets including cars, fields and buildings. Compared with the indoor scenes, outdoor lighting conditions are much

Figure 4: Qualitative comparisons between F-SAI and E-SAI algorithms on the *outdoor* dataset.

more complicated, making it harder for frame-based SAI methods (F-SAI and F-SAI+CNN) to generate clean results, as shown in Fig. 4. Similarly, complex lighting conditions also degrade the performance of E-SAI due to the increase of noise events, *e.g*. the events triggered by the brightness change of occlusions $\mathcal{E}_\theta^{OO}$ and occluded scenes $\mathcal{E}_\theta^{AA}$. The rising number of noise events not only makes the target indistinguishable in the results of E-SAI+ACC, but also brings more disturbances to E-SAI+CNN, deteriorating the reconstruction quality with serious saturation problem. Thanks to the hybrid SNN-CNN architecture, the issue of noise events can be alleviated from the temporal dimension. Therefore, our E-SAI+Hybrid is more robust to complex lighting conditions compared to other SAI methods, and can achieve the best visual effects on both indoor and outdoor datasets.

3. Extra Information

3.1. Implementation Details

Each data sequence lasts about 0.7 seconds in our datasets. In our experiments, we divide each sequence into 30 time intervals, *i.e.* $N = 30$, for E-SAI+CNN and E-SAI+Hybrid to make the input information equal. In network training, we set the loss weights as $[\beta_{per}, \beta_{pix}, \beta_{tv}] = [1, 32, 2e{-}4]$. For the perceptual loss, we set the weights $[\lambda_2, \lambda_4, \lambda_7, \lambda_{10}] = [1e{-}1, 1/21, 10/21, 10/21]$.

3.2. Network Architectures

Let $\mathrm{cCsS\text{-}K}$ denotes a $\mathrm{C} \times \mathrm{C}$ Convolution-BatchNorm-ReLU layer with stride S and K kernels. $\mathrm{r\text{-}K}$ denotes a residual block composed by a $\mathrm{c3s1\text{-}K}$ layer and a 3×3 Convolution-BatchNorm layer with K kernels and stride 1. $\mathrm{convCpP\text{-}K}$ denotes a $\mathrm{C} \times \mathrm{C}$ Convolution layer with stride 1, padding P and K kernels. $\mathrm{deconv\text{-}K}$ denotes a 3×3 fractional-strided-Convolution-BatchNorm-ReLU layer with K kernels and stride $1/2$.

Then, the CNN decoder consists of: $\mathrm{c7s1\text{-}64}$, $\mathrm{c3s2\text{-}128}$, $\mathrm{c3s2\text{-}256}$, $\mathrm{r\text{-}256}$, $\mathrm{r\text{-}256}$, $\mathrm{r\text{-}256}$, $\mathrm{r\text{-}256}$, $\mathrm{r\text{-}256}$, $\mathrm{r\text{-}256}$, $\mathrm{r\text{-}256}$, $\mathrm{r\text{-}256}$, $\mathrm{r\text{-}256}$, $\mathrm{deconv\text{-}128}$, $\mathrm{deconv\text{-}64}$, $\mathrm{c7s1\text{-}1}$. Note that in the output layer $\mathrm{c7s1\text{-}1}$, we replace ReLU function with Tanh function to normalize the output and do not use batch normalization. With the same CNN decoder, these learning-based SAI methods in our experiments can be described as:

- **F-SAI+CNN:** $\mathrm{conv3p1\text{-}16}$, $\mathrm{conv1p0\text{-}16}$, $\mathrm{conv1p0\text{-}32}$ followed by the CNN decoder.
- **E-SAI+CNN:** $\mathrm{conv3p1\text{-}16}$, $\mathrm{conv1p0\text{-}16}$, $\mathrm{conv1p0\text{-}32}$ followed by the CNN decoder.
- **E-SAI+Hybrid:** $\mathrm{Sconv3p1\text{-}16}$, $\mathrm{Sconv1p0\text{-}16}$, $\mathrm{Sconv1p0\text{-}32}$ followed by the CNN decoder, where $\mathrm{SconvCpP\text{-}K}$ denotes a $\mathrm{C} \times \mathrm{C}$ spiking-Convolution layer with stride 1, padding P and K kernels.

For fair comparison, we also add skip connections between the input tensor and the output of the 1-st, 2-nd Convolution layers in **F-SAI+CNN** and **E-SAI+CNN**.