SCIENCE CHINA
Information Sciences

• Supplementary File •

Event-Enhanced Synthetic Aperture Imaging

Siqi Li[1,2], Shaoyi Du[3], Junhai Yong[1] & Yue Gao[1,2*]

[1]*BNRist, School of Software, Tsinghua University, Beijing , China;*
[2]*THUIBCS, KLISS, BLBCI, Tsinghua University, Beijing* 100084*, China;*
[3]*National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, National Engineering Research Center for Visual Information and Applications, Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an* 710049*, China*

Appendix A Network Architecture

Here we provide the detailed network architecture of our proposed attention-based event encoder and the event-frame decoder. In practice, the architecture of the event encoder and decoder is shown in Fig. A1, in which the input channel of the event encoder is set to 36 corresponding to the event voxel, and output channel of the event-frame decoder is set to 3 corresponding to RGB frame. As shown in the figure, a convolutional layer and a channel attention module is contained in each down-sample layer. In each up-sample layer, the input feature is first up-sampled using bilinear interpolation, and is then forwarded into 4 parallel convolutional layers with convolutional kernel sizes of 1, 3, 5 and 7 to extract multi-scale features. The multi-scale features are fused using a convolutional layer with the kernel size of 1. Then, it is concatenated with the frame feature and extracted by a convolutional layer with the kernel size of 3.

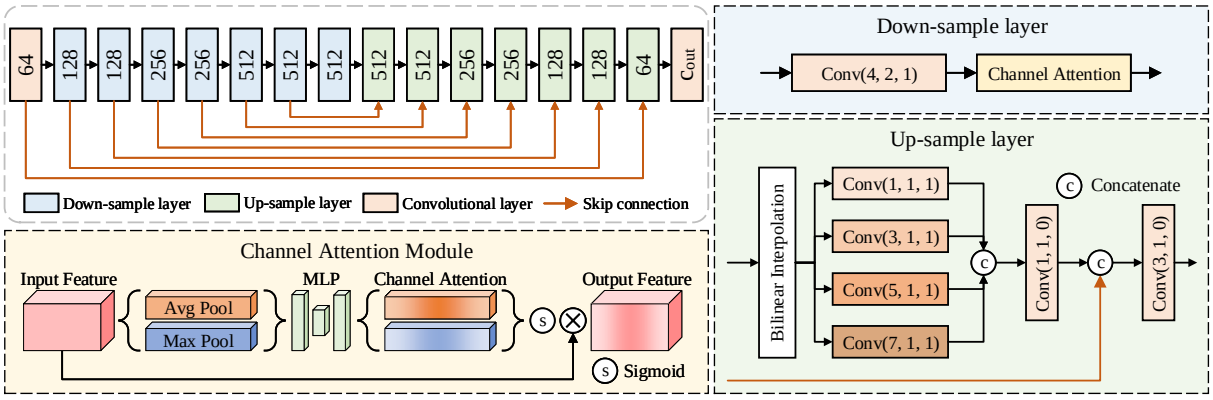

Figure A1 The network architecture of the encoder and the decoder. The number labeled in each layer in the top indicates the output channel number of that layer, and the tuple of numbers labeled in each block below denotes the kernel size, stride, and padding respectively.

Appendix B THU$^{\text{ERGB-SAI}}$ Dataset

Table B1 Comparison of our THU$^{\text{ERGB-SAI}}$ dataset with existing event-based SAI dataset, including E-SAI [2] dataset and EF-SAI [3] dataset. Our THU$^{\text{ERGB-SAI}}$ dataset has a larger scale, more occlusion types, and contains RGB information.

Dataset	Events	Frames	RGB	Type of Occlusions	Dataset Scale
E-SAI [2]	✓	✗	✗	1	588
EF-SAI [3]	✓	✓	✗	3	988
THU$^{\text{ERGB-SAI}}$ (Ours)	✓	✓	✓	4	2560

We construct a visual acquisition system and collect a large-scale **E**vent-enhanced **RGB** **S**ynthetic **A**perture **I**maging dataset, named THU$^{\text{ERGB-SAI}}$ dataset. As shown in Fig. B1 (a), our visual acquisition system contains a DAVIS346 Color camera placed on a linear slide moving straightly from one side to another side. The occlusion is placed between the scene and the slide. In our data collection, we use four different occlusions, including baffle, grids, fence, and grille, as shown in Fig. B1 (b). In practice, the occlusion is placed parallel to the slide and vertically to the optical axis of the event camera. Our dataset is collected under both

* Corresponding author (email: kevin.gaoy@gmail.com)

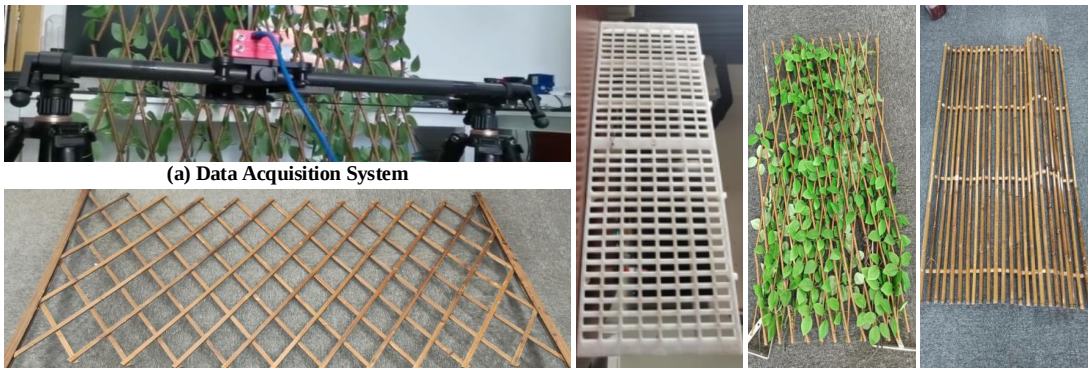

Figure B1 (a) Our data acquisition system. (b-c) Four different occlusions used to capture our THU$^{\text{ERGB-SAI}}$ dataset.

indoor and outdoor scenarios. Specifically, The indoor parts of our dataset are collected in a laboratory environment and contain common objects such as cabinets, chairs, books, toys, *etc.*, and are occluded by all four types of occlusions. The outdoor parts of our dataset mainly contain relatively large objects such as buildings, bicycles, cars, *etc.*, and are occluded by the fence. For each sample, a series of occluded frames and the corresponding event stream are collected by the DAVIS346 camera simultaneously during the camera moving, and the ground truth clear scene image without occlusions is obtained when the occlusion is removed. Using our visual acquisition system, we totally collect 2560 samples indoors and outdoors in our THU$^{\text{ERGB-SAI}}$ dataset, which is over 2× larger than the existing EF-SAI dataset. In our dataset, 720 samples are occluded by the sparse occlusion (baffle), and the remaining 1840 samples are occluded by dense occlusions (grids, fence, and grille).

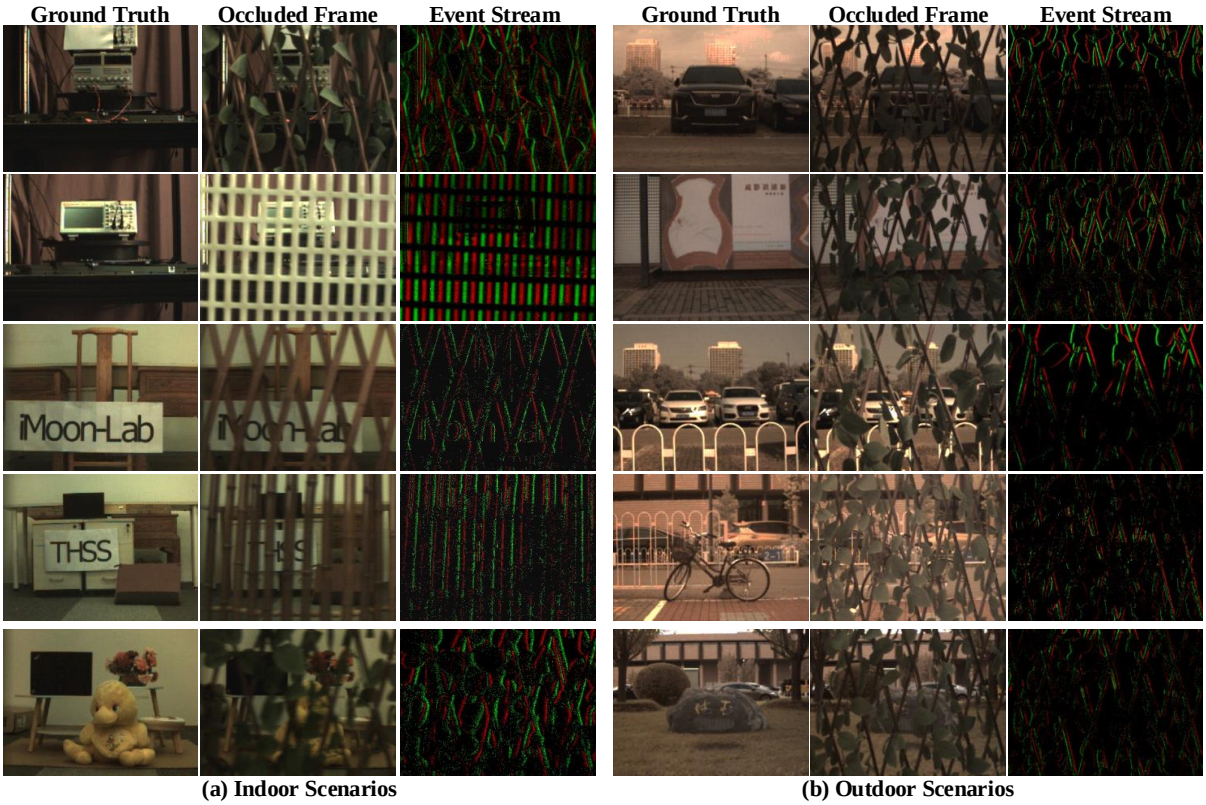

Figure B2 Some example samples of our THU$^{\text{ERGB-SAI}}$ dataset. (a) Indoor scenarios that are occluded by all four types of occlusions. (b) Outdoor scenarios that are occluded by the fence. For each sample, the ground truth clear scene image without occlusion, the occluded frame, and the event stream are visualized.

Table B1 shows the comparison of our THU$^{\text{ERGB-SAI}}$ dataset with existing event-based synthetic aperture imaging datasets, including E-SAI dataset [2] and EF-SAI dataset [3]. From the table, we can observe that compared with existing datasets, our THU$^{\text{ERGB-SAI}}$ dataset has a larger scale, *i.e.*, over 2× larger than EF-SAI dataset. Meanwhile, more varied occlusions are used in our dataset, which could better simulate the occluded situations that may occur in real-world applications. In addition, compared with the EF-SAI dataset, RGB occluded frames and ground truth scene images are contained in our dataset, which could be more advantageous for computational photography applications.

Figure B2 shows some example samples of our THU$^{\text{ERGB-SAI}}$ dataset. We choose 5 indoor and 5 outdoor typical scenarios for demonstration. The occluded frames and the corresponding event streams as well as the ground truth clear scene images are visualized. It should be noted that since our task, *i.e.*, event-enhanced synthetic aperture imaging, aims to reconstruct clear images

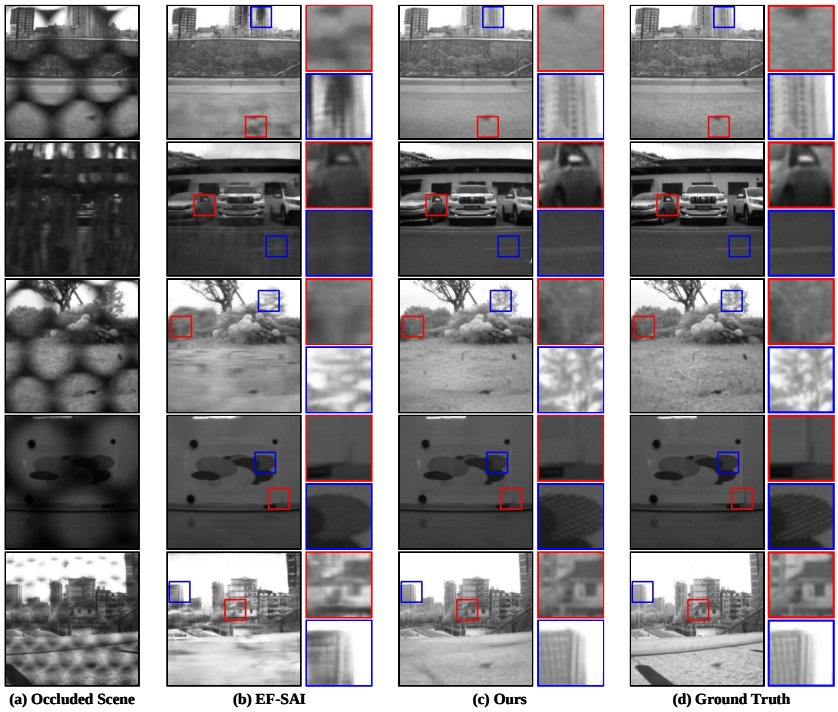

Figure C1 The qualitative results on the EF-SAI dataset [3]. From left to right: the occluded scenarios, the results generated by EF-SAI [3] and by our proposed method, and the ground truth images. Details are zoomed in for better comparison.

with detailed textures, we deliberately added some printed letters in our designed indoor scenarios, as shown in Fig. B2 (a), which could better demonstrate the capability of the method to reconstruct detailed clear images.

Appendix C Qualitative Results

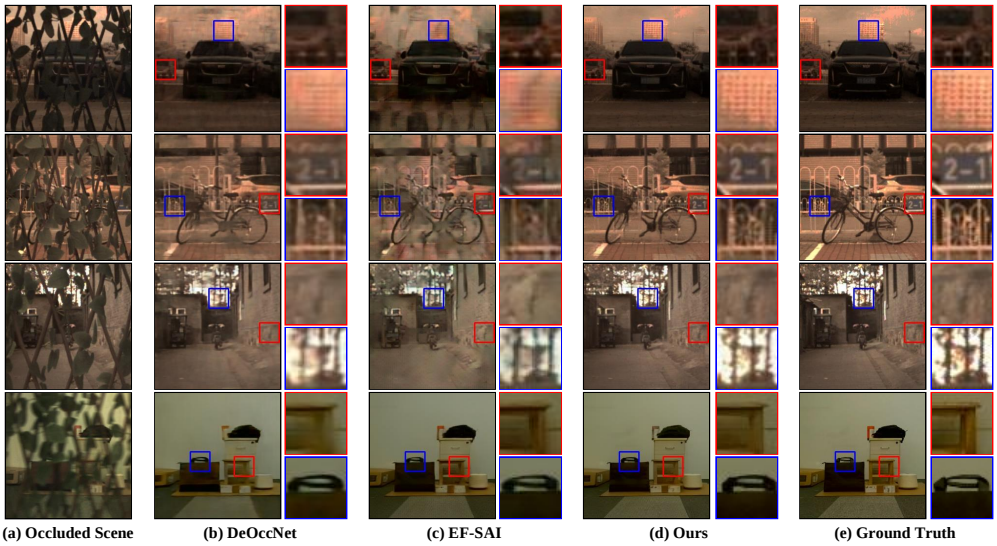

Figure C2 More visualization results on our THU$^{\text{ERGB-SAI}}$ dataset. From left to right: the occluded scenarios, the results generated by DeOccNet [1], EF-SAI [3] and by our proposed method, and the ground truth images. Details are zoomed in for better comparison.

Figure C1 shows the qualitative results on the existing EF-SAI dataset [3]. We choose 5 scenarios for comparison, and our proposed method is compared with the state-of-the-art event- and frame-based method EF-SAI [3]. From the visualization results, we could observe that our proposed method could generate clearer images with more detailed structures. As shown in the blue box in row 1, our proposed method could clearly reconstruct detailed structures, *e.g.*, the windows of the distant house. In contrast, the result generated by EF-SAI is blurry. Meanwhile, we could also observe from the figure that our proposed method could generate clear scene images without artifacts. As shown in column (b) in row 2, the result generated by EF-SAI contains artifacts due to the occlusions in the input frames. In contrast, our proposed method could generate clear scene images without artifacts due to our event-enhanced foreground segmentation module, as shown in column (c) in row 2. Another typical example is shown in the blue

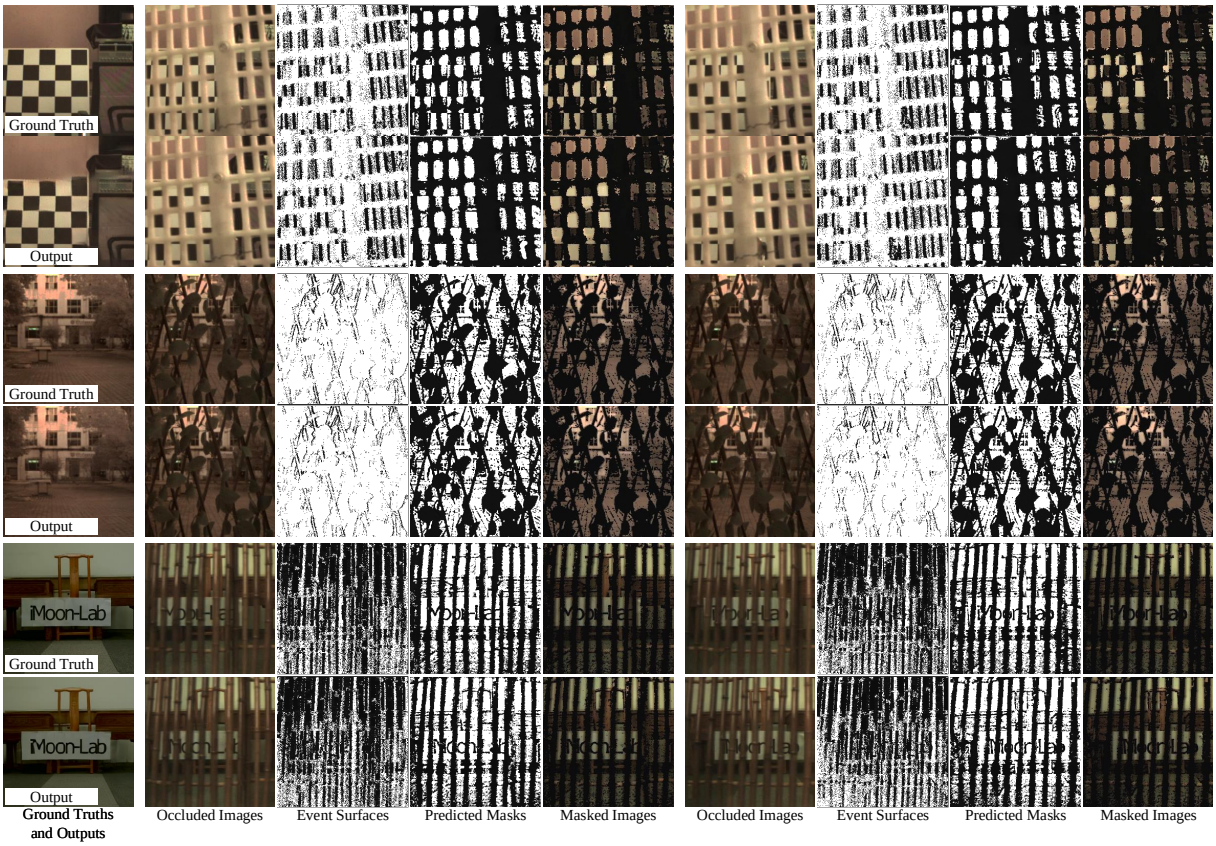

Figure D1 The qualitative results of the event-enhanced image matting module. We choose 3 different scenarios for demonstration. For each scenario, we visualize four occluded images, and the corresponding event surfaces, the predicted masks, and the masked images. The ground truth images and the outputs of our method are also provided for reference.

box in row 4, compared with EF-SAI, our proposed method could reconstruct the very detailed grid texture in the circle, which could demonstrate the effectiveness of our proposed method.

Figure C2 shows more visualization results on our THU$^{\text{ERGB-SAI}}$ dataset. The occluded scenarios, the results generated by DeOccNet [1], EF-SAI [3] and by our proposed method, and the ground truth images are visualized from left to right. To demonstrate the capability of our proposed method to tackle irregular occlusions, we have chosen 4 scenarios that are occluded by irregular fences (middle one of Fig. B1 (c)). From the figure, we could observe that serious artefacts are appeared in the results generated by both DeOccNet and EF-SAI when facing dense irregular occlusions, *e.g.*, the first row and the second row. In contrast, our proposed method could effectively remove foreground irregular occlusions and reconstruct unobstructed scene images, *e.g.*, the detailed numbers in the red box of the second row. This is due to the fact that our proposed method leverages an event-guided occlusion segmentation module, which could effectively distinguish the foreground occlusions with the background target scene and further remove the invalid occlusions. In contrast, all existing methods directly use the occluded frames as input, which will lead to the appearance of artifacts.

Appendix D Ablation Experiments

To demonstrate the effectiveness of our proposed event-enhanced frame-based SAI branch and the event- and frame-based SAI branch, we test and analyze the performance of each branch, respectively. The quantitative results are shown in Tab. D1. From the table, we could observe that our proposed event-enhanced frame-based SAI branch (denoted as EE-F-SAI Branch) performs better when facing spare occlusions since sufficient valid visual information is contained in the input occluded frames and could be effectively extracted by our event-enhanced frame-based SAI branch. The event- and frame-based SAI branch (denoted as EF-SAI Branch) is more suitable for dense occluded scenarios and could achieve better performance under such scenarios since extra visual information could be provided by the input event stream when facing dense occlusions.

Table D1 Ablation experiments on our THU$^{\text{ERGB-SAI}}$ dataset.

Method	Sparse Occlusions		Dense Occlusions		Total	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
EE-F-SAI Branch	<u>26.81</u>	<u>0.808</u>	26.03	0.774	26.29	0.785
EF-SAI Branch	26.47	0.786	<u>26.56</u>	<u>0.793</u>	<u>26.53</u>	<u>0.791</u>
Full Model	**27.39**	**0.824**	**27.18**	**0.817**	**27.25**	**0.819**

To further demonstrate the effectiveness of our proposed event-enhanced foreground segmentation module, we visualize the input and the output of the foreground segmentation module, as shown in Fig. D1. Here we select 3 different scenarios for illustration. The event count map E^s used as the guidance, the mask predicted by the event-enhanced foreground segmentation module, and the masked image are shown from left to right. The ground truth scene images without occlusion and the output de-occlusion results are also provided as references. The masked image is obtained by multiplying the input occluded image with the predicted mask, which is the same as the operation in partial convolution. The black pixels in the masked image are the invalid parts that are filtered out and are not performed convolution. As shown in Fig. D1, we can observe that our proposed event-enhanced foreground segmentation module can predict the mask of the occlusion accurately from the occluded image with the guidance of the event count map, which can strengthen the partial convolutional layer to extract valid information of the occluded scene, *i.e.*, the pixels of the occlusions are filtered out precisely, and only the feature of valid part is extracted.

References

1 Wang Y., Wu T., Yang J., *et al.* DeOccNet: Learning to See Through Foreground Occlusions in Light Fields. In: Proceedings of the IEEE WACV, 2020, 118-127.
2 Yu L., Zhang X., Liao W., *et al.* Learning to See Through with Events. IEEE T-PAMI, 2023, 45(7): 8660-8678.
3 Liao W., Zhang X., Yu L., *et al.* Synthetic Aperture Imaging with Events and Frames. In: Proceedings of the IEEE CVPR, 2022, 17735-17744.