SCIENCE CHINA
Information Sciences

• MOOP •

March 2025, Vol. 68, Iss. 3, 134101:1–134101:3
https://doi.org/10.1007/s11432-023-4298-8

Event-enhanced synthetic aperture imaging

Siqi LI[1,2], Shaoyi DU[3], Jun-Hai YONG[1] & Yue GAO[1,2*]

[1]*Beijing National Research Center for Information Science and Technology, School of Software, Tsinghua University, Beijing 100084, China*
[2]*Institute for Brain and Cognitive Sciences, Beijing Laboratory of Brain and Cognitive Intelligence, 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*

Received 13 June 2023/Revised 19 January 2024/Accepted 25 September 2024/Published online 13 February 2025

Citation Li S Q, Du S Y, Yong J-H, et al. Event-enhanced synthetic aperture imaging. Sci China Inf Sci, 2025, 68(3): 134101, https://doi.org/10.1007/s11432-023-4298-8

Synthetic aperture imaging (SAI) methods aim to see through dense occlusions and reconstruct the target scene behind occlusions. Traditional frame-based SAI methods, e.g., DeOccNet [1], take the occluded light field images captured by a camera array as input, and fuse them to achieve image de-occlusion. However, when facing dense occlusions, the valid information of the scene available to the camera array with only a certain number of viewpoints is limited.

Event cameras are bio-inspired vision sensors that respond to pixel-wise brightness changes asynchronously. Specifically, an event is triggered whenever the change of the logarithm of the brightness at a pixel exceeds a certain threshold, i.e., $|\Delta \log(I(x,y,t))| > C$, where $I(x,y,t)$ is the brightness at pixel (x,y) and timestamp t, and C is the threshold. The output event is defined as $e = (x,y,t,p)$, where $p \in \{+1,-1\}$ is the polarity, indicating the increase or decrease of the brightness. Event cameras own outstanding advantages due to the specific working principle, e.g., high dynamic range (up to 140 dB) and high temporal resolution (about 1 ms). Therefore, event cameras could record extra visual information of the occluded scene.

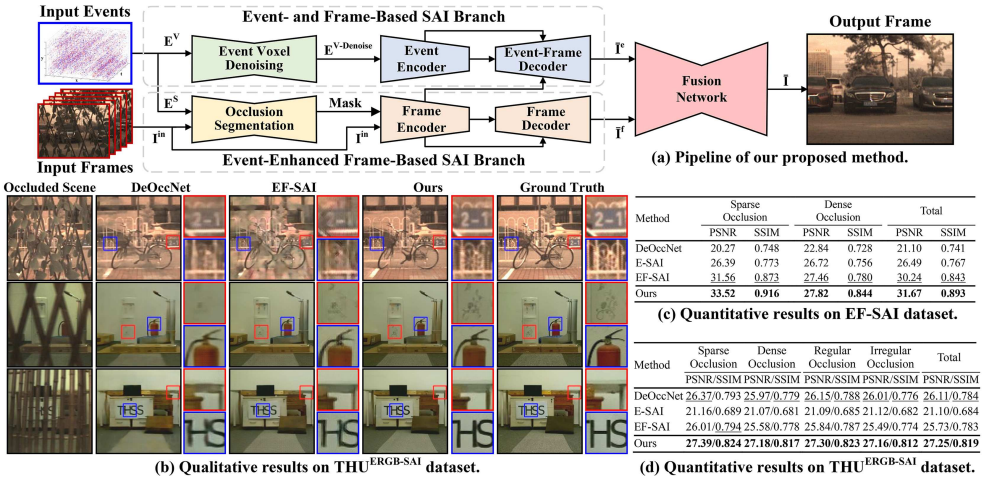
Since event cameras own ultra high temporal resolution, the event stream obtained by a moving event camera could record extra visual information of the scene behind the occlusions, which could greatly facilitate the image de-occlusion task under complex occlusion scenarios. Zhang et al. [2] proposed the first event-based SAI method, which can synthesize clear scene images from input event streams captured by a moving event camera. However, generating images solely from input event streams is an ill-posed problem due to the fact that the event stream could only provide the records of the brightness changes at each pixel while the initial brightness value is unknown, which may lead to fatal errors. Meanwhile, since the event camera is heavily affected by noise and its spatial resolution is insufficient, it is tough to generate image details, e.g., textures, from pure event streams. To tackle this problem, Liao et al. [3] and Li et al. [4] further proposed event- and frame-based hybrid models, which take the occluded frames and the event streams as input to generate clear scene images. However, although these event-based methods exploit the high temporal resolution advantage of the event stream, they fail to take full use of the visual information contained in occluded frames, which may lead to performance decline when facing sparse occlusions. Meanwhile, the input occluded frames contain plenty of invalid occlusion visual features, simply extracting all features of the input occluded frames to generate clear scene images may raise severe artifacts. All existing methods directly use the occluded frames with interference as input, which leads to performance restriction.

In this study, we propose a two-branch event-enhanced SAI method that makes full use of both the event stream and the occluded frames to achieve satisfactory image de-occlusion under both sparse and dense occlusions. Compared with existing methods [2–4] that directly use original occluded frames as input, our proposed method for the first time leverages the event surface as guidance to explicitly predict the occlusion mask of each occluded frame and then filters out invalid occlusions. Then, the event-enhanced frame-based SAI result could be obtained from valid portions of occluded frames based on partial convolution, which could achieve satisfying performance under sparse occlusions. Meanwhile, the event stream is leveraged to provide extra information and synthesize the event- and frame-based SAI results, which could generate clearer images when facing dense occlusions. Finally, the results obtained by both branches are fused to generate the final output image.

In addition, we construct a large-scale event-enhanced RGB synthetic aperture imaging ($\text{THU}^{\text{ERGB-SAI}}$) dataset. Compared with the existing largest event-based SAI dataset, i.e., EF-SAI [3] dataset, our dataset has a larger scale and further contains RGB information, which is more advantageous for computational photography applications.

Method. Figure 1(a) shows the pipeline of our proposed

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



Method	Sparse Occlusion		Dense Occlusion		Total	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DeOccNet	20.27	0.748	22.84	0.728	21.10	0.741
E-SAI	26.39	0.773	26.72	0.756	26.49	0.767
EF-SAI	<u>31.56</u>	<u>0.873</u>	<u>27.46</u>	<u>0.780</u>	<u>30.24</u>	<u>0.843</u>
Ours	**33.52**	**0.916**	**27.82**	**0.844**	**31.67**	**0.893**

(c) Quantitative results on EF-SAI dataset.

Method	Sparse Occlusion	Dense Occlusion	Regular Occlusion	Irregular Occlusion	Total
	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
DeOccNet	<u>26.37</u>/0.793	<u>25.97</u>/0.779	<u>26.15</u>/0.788	<u>26.01</u>/0.776	<u>26.11</u>/0.784
E-SAI	21.16/0.689	21.07/0.681	21.09/0.685	21.12/0.682	21.10/0.684
EF-SAI	26.01/<u>0.794</u>	25.58/0.778	25.84/0.787	25.49/0.774	25.73/0.783
Ours	**27.39/0.824**	**27.18/0.817**	**27.30/0.823**	**27.16/0.812**	**27.25/0.819**

(d) Quantitative results on THU^ERGB-SAI dataset.

Figure 1 (Color online) (a) Pipeline of our proposed method and (b)–(d) comparison results. The best results are in bold. The second best results are underlined. Zoom in for a better view. Code and dataset: https://github.com/lisiqi19971013/THU-ERGB-SAI.

method, which could reconstruct the 2D appearance of the target scene behind foreground occlusions from visual data captured by a moving event camera. Specifically, our proposed method takes k consecutive occluded frames, defined as $I^{\text{in}} = \{I_1, \ldots, I_k\}$ and the corresponding event stream $\mathcal{E} = \{e_i = (x_i, y_i, t_i, p_i) | T_1 < t_i < T_k\}$ as input to reconstruct the unobstructed scene image at the middle position, where T_i is the timestamp of the occluded frame I_i. In practice, we select $k = 17$.

(1) Event-enhanced frame-based SAI. This branch mainly focuses on relatively sparse occluded portions, which could generate unobstructed images from the valid features extracted from the occluded frames. Since the input occluded frames contain plenty of invalid occlusions, we propose an occlusion segmentation module to predict the occlusion masks with the guidance of the event surface. Our proposed method for the first time explicitly tackles foreground occlusion and could significantly improve SAI performance. In practice, the guided event surface is formulated as

$$E_i^s = \sum_{i \in \{i | T_{i-1} < t_i < T_{i+1}\}} \delta(x - x_i, y - y_i), \tag{1}$$

where $\delta(\cdot)$ is the Dirac function. Thus, each occluded frame is concatenated with its corresponding event surface and forwarded into a lightweight U-Net like occlusion segmentation module to predict occlusion masks. The segmentation module contains 5 down-sample layers and 5 up-sample layers.

After the occlusion masks are predicted, it is combined with the input occluded frames and forwarded into the frame encoder and decoder, which have similar architecture as the segmentation module but are based on partial convolutional layers. Let $\boldsymbol{W}$ and b represent the convolution filter weight and bias for a partial convolution layer, $\boldsymbol{F}$ represent the feature at the current convolution window, and $\boldsymbol{M}$ denote the occlusion mask. The partial convolution is defined as

$$x_{\text{out}} = \begin{cases} \boldsymbol{W}^\top (\boldsymbol{F} \odot \boldsymbol{M}) \frac{\text{sum}(\mathbf{1})}{\text{sum}(\boldsymbol{M})} + b, & \text{sum}(\boldsymbol{M}) > 0, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where $\odot$ is the point-wise multiplication. Using partial convolution, features of valid parts of the scene in the occluded frame could be effectively extracted, and the parts of invalid occlusions could be filtered out by the mask. Then, using the occlusion segmentation module and frame encoder-decoder, the SAI result $\bar{I}^f$ could be obtained, which could achieve satisfactory performance for sparse occlusions.

(2) Event- and frame-based SAI. When facing extremely dense occlusions, the valid visual information provided by the input occluded frames is insufficient. Therefore, we additionally leverage the event stream $\mathcal{E}$ to provide extra visual information of the target scene and generate the event- and frame-based SAI result $\bar{I}^e$.

The asynchronous input event stream is first converted into a grid-based event voxel E^V using an event representation method proposed in [5]. In practice, the event stream is discretized into N_b time bins among temporal dimension, i.e., $t_i' = \lfloor (N_b - 1) \frac{t_i - T_{m-k}}{T_{m+k} - T_{m-k}} \rfloor$. Then, E^V could be obtained by $E^V(x, y, t) = \sum_i p_i \delta(x - x_i) \delta(y - y_i) \delta(t - t_i')$, where $\delta(\cdot)$ is the Dirac function. In practice, the number of the time bins N_b is set to 36. Considering that the raw event stream is heavily affected by noise, we first leverage an event voxel denoising module to denoise the event voxel, whose architecture is similar to the segmentation module.

After the denoised event voxel $E^{\text{V-Denoise}}$ is obtained, it is forwarded into the attention-based event encoder module to extract multi-scale features. Then, the event features are multiplied with the frame features generated from the partial convolution-based frame encoder and are forwarded into the event-frame decoder. In practice, the event encoder contains 7 down-sample layers, each of which is followed by a channel attention module. The event-frame decoder contains 7 up-sample layers, each of which uses 4 parallel convolutional layers with convolutional kernel sizes of 1, 3, 5, and 7 to extract multi-scale features. Detailed network architecture is provided in the supplementary material. With the event- and frame-based SAI branch, the unobstructed image $\bar{I}^e$ could be reconstructed when facing dense occlusions.

(3) Fusion network. After $\bar{I}^f$ and $\bar{I}^e$ are obtained, they are concatenated and then forwarded into a fusion network to generate the final image de-occlusion result $\hat{I}$. The architecture of the fusion network is similar to that of the segmentation module, except for input and output channels.

(4) Loss functions. Our proposed model is trained with the supervision of $\mathcal{L} = \mathcal{L}_{\text{pix}} + \lambda \mathcal{L}_{\text{per}}$, where λ is the hyper-

parameters. $\mathcal{L}_{\text{pix}} = \|\bar{I} - I\|_1$ is the pixel-level Manhattan distance between the output image $\bar{I}$ and ground truth image I. $\mathcal{L}_{\text{per}}$ is the perceptual loss, which could maintain the high-level vision features.

THU$^{ERGB\text{-}SAI}$ dataset. We construct a large-scale event-enhanced RGB synthetic aperture imaging dataset, named THU$^{\text{ERGB-SAI}}$. In our data collection, we use four different occlusions, including baffle, grids, fence, and grille. Our dataset is collected under both indoor and outdoor scenarios. For each sample, occluded frames and event stream are collected by a moving DAVIS346 event camera simultaneously, and the ground truth unobstructed scene image is obtained after the occlusion is removed. We totally collect 2560 samples in our THU$^{\text{ERGB-SAI}}$ dataset, which is over 2× larger than the existing EF-SAI dataset [3] (988 samples). 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. In our dataset, 720 samples are occluded by the sparse occlusion, and the remaining 1840 samples are occluded by dense occlusions.

Experiments. To validate the performance of our proposed method, we conduct experiments on the existing EF-SAI [3] dataset and our THU$^{\text{ERGB-SAI}}$ dataset. The peak signal to noise ratio (PSNR, higher is better) and structural similarity (SSIM, higher is better) are used as the evaluation metrics. In practice, we first train the event-enhanced frame-based SAI branch and the event- and frame-based SAI branch separately, and then the entire model is trained jointly. The two sub-modules are trained for 400 epochs with a batch size of 8, and the entire model is trained for 1000 epochs with a batch size of 16. The optimization method is Adam. For the loss function, the weight λ is set to 0.1.

(1) Quantitative comparisons. From Figure 1(c), we could observe that compared with the second-best method EF-SAI, our proposed method could achieve improvements of 1.43 dB and 0.050 in terms of PSNR and SSIM, respectively. Compared with dense occluded scenarios, we could observe that our proposed method achieves greater advantages compared to EF-SAI when facing sparse occlusions, i.e., 1.96 dB on PSNR (vs. 0.36 dB for dense occlusions). This is due to the fact that occluded frames may be more useful when facing sparse occlusions, and our proposed method leverages the event surface as guidance to predict the occlusion mask and filters out the invalid part of the input occluded frames, which could better extract the valid visual information from the occlusion frames.

Figure 1(d) shows the quantitative results on our THU$^{\text{ERGB-SAI}}$ dataset. All compared methods are trained on our dataset from scratch for fair comparison. From Figure 1(d), we could observe that compared with the second-best method DeOccNet [1], our proposed method could achieve improvements of 1.14 dB and 0.035 in terms of PSNR and SSIM, respectively. We could also observe that when facing dense occlusions, our proposed method could achieve greater advantages compared to DeOccNet. This is due to the fact that compared with the frame-based method, our proposed method further leverages the input event stream to provide extra visual information of the target scene behind dense occlusions. In addition, we could observe that our method could achieve better performance compared with the second-best method DeOccNet when facing both regular and irregular occlusions due to the proposed two-branch architecture.

(2) Qualitative comparisons. Figure 1(b) shows the qualitative results on our THU$^{\text{ERGB-SAI}}$ dataset. Our proposed method is compared with the frame-based method DeOccNet [1] and the event- and frame-based method EF-SAI [3]. Some details are zoomed in for better comparison. From Figure 1(b), we could observe that when facing dense occlusions, our proposed method could reconstruct clear scene images with detailed structures compared with the frame-based method DeOccNet, e.g., the numeric characters in the first row and the letters "HS" in the last row. When facing sparse occlusions, our proposed method could generate clearer details, e.g., the pattern on the drinking fountain shown in the red box in the second row, owing to the full use of the valid information in the input occlusion frames. We could also observe that artifacts appear in the results generated by DeOccNet and EF-SAI when facing dense occlusions, as shown in the first row, and our proposed method could generate images without artifacts due to our proposed event-enhanced foreground segmentation module.

Conclusion. In this work, we propose a two-branch event-enhanced synthetic aperture imaging method for the image de-occlusion task, which contains an event-enhanced frame-based SAI branch, an event- and frame-based SAI branch, and a fusion module. The event-enhanced frame-based SAI branch takes the occluded frames as input and generates frame-based SAI results with the guidance of the occlusion mask predicted by our proposed event-enhanced foreground segmentation module, which is suitable for sparse occluded scenes. The event- and frame-based SAI branch additionally leverages event streams to provide extra visual information and could achieve better performance under dense occlusions. Finally, the fusion network fuses the outputs of the two branches to generate clear output images without occlusion. Our proposed method for the first time achieves explicit handling of invalid occlusions, which could effectively improve the SAI performance. In addition, we collect and construct the THU$^{\text{ERGB-SAI}}$ dataset containing RGB occluded frames and event streams together with ground truth clear scene images, which is over 2× larger than the existing largest dataset. Experimental results on our THU$^{\text{ERGB-SAI}}$ dataset and the existing EF-SAI dataset show that our proposed method achieves state-of-the-art performance.

Acknowledgements This work was supported by Beijing Natural Science Foundation (Grant No. L242167), National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Xi'an Jiaotong University (Grant No. HMHAI-202412), National Natural Science Foundation of China (Grant No. U24A20252), and Key Research and Development Program of Shaanxi Province of China (Grant No. 2024PT-ZCK-66).

Supporting information Videos and other supplemental documents. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

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 Trans Pattern Anal Mach Intell, 2023, 45: 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
4 Li S Q, Gao Y, Dai Q H. Image de-occlusion via event-enhanced multi-modal fusion hybrid network. Mach Intell Res, 2022, 19: 307–318
5 Zhu A Z, Yuan L, Chaney K, et al. Unsupervised event-based optical flow using motion compensation. In: Proceedings of the ECCVW, 2018