EvDiG: Event-guided Direct and Global Components Separation

Supplementary Material

Xinyu Zhou[1] Peiqi Duan[2,3] Boyu Li[2,3] Chu Zhou[1] Chao Xu[1] Boxin Shi[*2,3]

[1]National Key Lab of General AI, School of Intelligence Science and Technology, Peking University

[2]National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University

[3]National Engineering Research Center of Visual Technology, School of Computer Science, Peking University

{zhouxiny, duanqi0001, liboyu, zhou_chu, shiboxin}@pku.edu.cn xuchao@cis.pku.edu.cn

6. Analysis on network design

To validate the effectiveness and necessity of each part of our method, we conduct several ablation studies and show results in Tab. 3. To verify the necessity of our two-stage network design, we introduce a vanilla version of EvDiG which is a simple U-Net architecture network (denoted as "EvDiG-vanilla"). The impact of our proposed Separation Correction Block (SCB) and Color Correction Block (CCB) is evaluated by substituting them with standard convolution layers (denoted as "EvDiG w/o SCB" and "EvDiG w/o CCB" respectively). To evaluate impact of the coarse separation (CS) results of the direct and global components obtained using event accumulation, we remove the coarse separation results from the inputs of EvSepNet (denoted as "EvDiG w/o CS"). As indicated in Tab. 3, our complete model demonstrates superior performance, which highlights the contribution of each component in our methodology.

Table 3. Ablation study of the proposed network design on our collected dataset. ↑(↓) indicates the higher (lower), the better throughout this paper. The best performances are highlighted in **bold**. The content in each cell refers to the results for direct and global components respectively.

	PSNR ↑	SSIM ↑	LPIPS ↓
EvDiG-vanilla	29.05/30.54	0.862/0.829	0.097/0.139
EvDiG w/o SCB	29.55/31.33	0.878/0.840	0.085/0.129
EvDiG w/o CCB	29.54/31.22	0.876/0.839	0.086/0.131
EvDiG w/o CS	28.90/30.40	0.872/0.834	0.083/0.126
Ours	**30.01/31.66**	**0.883/0.846**	**0.077/0.117**

7. More details on data collection

Examples of captured scenes are presented in Fig. 9. The data capture setup involves one projector, several types of source occluders, and a hybrid camera system consisting of a machine vision camera (HIKVISION MV-CA050-12UC) and an event camera (PROPHESEE GEN4.1), which are co-aligned using a beam splitter (Thorlabs CCM1-BS013). We utilize the hybrid camera system to capture RGB images and events simultaneously. For geometric calibration, we use a checkerboard to deal with homography transformation and radial distortion between two views. For temporal synchronization, we use an independent signal generator to synchronize two cameras.

Figure 9. Examples of scenes in our captured real-world dataset.

In our experiment, we employ a Panasonic PT-WX3201 LCD projector for indoor scenes. DLP projectors are not recommended because their spinning color wheel mechanism can trigger redundant events that interfere with the separation accuracy.

In our indoor-scene dataset, we experiment with various types of occluders, including line and mesh types. Mesh occluders, such as a 2D grid with circular holes, are traditionally favored for their efficiency in image-based methods. However, the high temporal resolution of event cameras allows for equally efficient capture with a simpler line occluder. While mesh occluders may present slight efficiency gains in the data capture phase, they also notably increase the size of event data. This increase in data size complicates and hinders efficient data processing. Given these considerations, we predominantly utilize line occluders for achieving event-guided direct and global component separation. This choice is informed by a balance between data capture efficiency and the practicality of data processing.

8. Analysis of the moving speed of occluders

We attach the stick occluder to a motor to precisely control its moving speed and evaluate the performance of our method under different moving speed of source occluders. We capture 16 scenes at 6 varied speeds in total. The duration for the occluder to traverse the scene spans a range

* Corresponding author

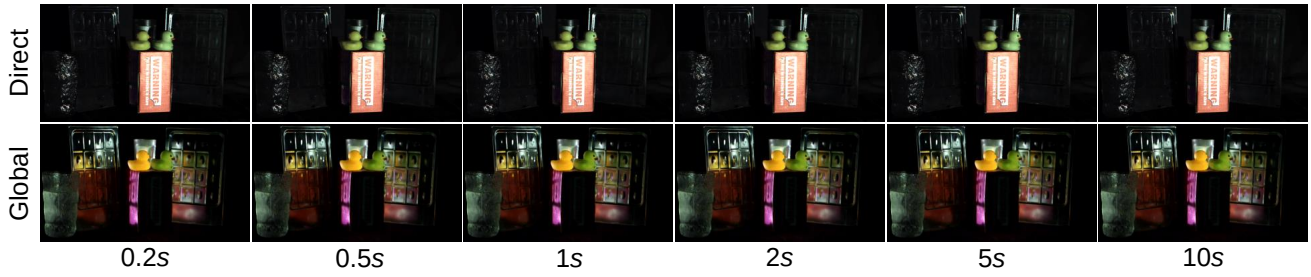

Figure 10. A challenging example of our method under varied moving speed of source occluders. The time below represents the duration for the occluder to traverse the scene.

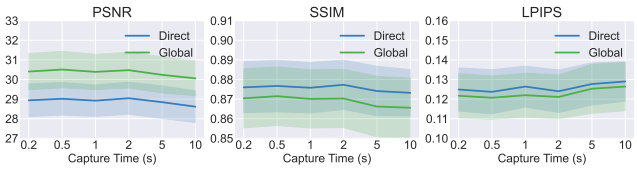

Figure 11. Quantitative results of our method under 6 varied moving speeds of source occluders. The capture time denotes the duration for the occluder to sweep across the scene.

from $0.2s$ to $10s$. The quantitative results are shown in Fig. 11. Our method achieves stable performance across all speeds, which demonstrates that our method is capable of handling scenarios without fast-moving occluders. The slight performance decline at lower speeds is attributable to the increased capture of noise events. We show a challenging scene with strong interreflections in Fig. 10. This example demonstrates that our method is able to accurately separate the direct and global components under different moving speed of occluders.

9. Determination of the threshold

In our experiment, we meticulously capture dynamic scene videos across a range of specific camera settings and derive the appropriate contrast threshold θ for the corresponding camera parameter. This is achieved through event accumulation between adjacent frames, allowing us to accurately obtain the threshold to trigger an event. Subsequently, the data gathered from this process leads to the creation of a look-up table for θ, serving as a valuable resource for quickly identifying the optimal contrast threshold based on the specific parameters of the hybrid camera system in use.

10. More qualitative comparisons

In this section, we provide additional qualitative comparison results on our real-world indoor scenes in Fig. 12. Both SF-pattern-classic [3] and SF-pattern-deep [1] derive separation results from a single pattern image. As observed in Fig. 12, These methods tend to produce blurry separation results with checkerboard-like artifacts. SF-scene-deep [4] predicts the direct and global components from a single scene image without physical cues. The comparison results in Fig. 12 demonstrate that the absence of physical cues renders the performance of SF-scene-deep highly dependent on its training dataset, thereby constraining its ability to generalize to unseen scenes. If some regions are not recorded within the umbra of the shadow in the captured image sequence, MF-shadow-classic [3] will produce serious artifacts in the separation results. Event cameras, with their ability to detect scene brightness changes with microsecond precision, continuously record the brightness changes over time, which enables the proposed EvDiG to achieve effective and efficient direct and global components separation.

11. Results on real-world dynamic scenes

In Sec. 4.2, we demonstrate how the high temporal resolution of event cameras substantially reduces the capture time close to that of single-frame methods. This advancement renders our approach suitable for dynamic scenes. We conduct the comparative experiment against SF-scene-deep [4] on real-world dynamic scenes. The comparison results are illustrated in the supplementary video which can be found on our project page.

For the data acquisition of dynamic scenes, we meticulously regulate the velocity of the source occluder to surpass the capture frequency of the RGB camera. Specifically, for every frame obtained, we utilize the events recorded within a $30ms$ window after the frame's exposure as the input for our method. Both EvDiG and SF-scene-deep process video sequences in a frame-by-frame manner. Note that the application of multi-frame-based methods to dynamic scenes is impractical, due to their long data capture time. For the purposes of visualization, we apply the same post-processing technique [2] with identical parameter settings for both the results from our method and those obtained using SF-scene-deep [4] to improve temporal consistency. The comparison results reveal that EvDiG achieves more accurate separation of direct and global components in various scenarios, including those with camera ego-motion and object-motion.

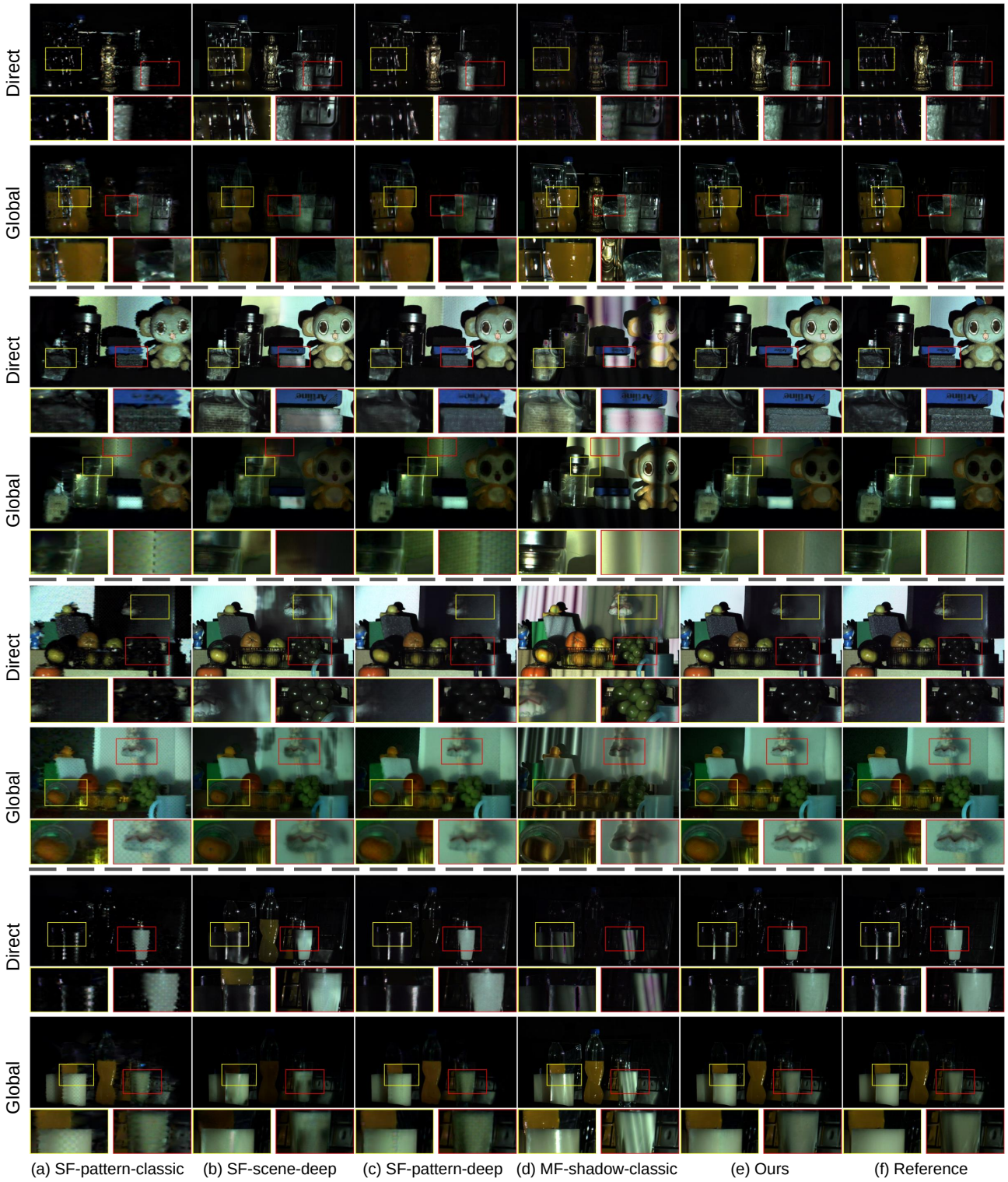

Figure 12. Direct and global components separation results on our real-world indoor scenes. (a)-(f) Separation results of SF-pattern-classic [3], SF-scene-deep [4], SF-pattern-deep [1], MF-shadow-classic [3], Ours, and Reference [3].

References

[1] Zhaoliang Duan, James Bieron, and Pieter Peers. Deep separation of direct and global components from a single photograph under structured lighting. *Computer Graphics Forum*, 39(7):459–470, 2020.

[2] Chenyang Lei, Yazhou Xing, and Qifeng Chen. Blind video temporal consistency via deep video prior. In *Advances in Neural Information Processing Systems*, 2020.

[3] Shree K Nayar, Gurunandan Krishnan, Michael D Grossberg, and Ramesh Raskar. Fast separation of direct and global components of a scene using high frequency illumination. *ACM Transactions on Graphics*, pages 935–944, 2006.

[4] Shijie Nie, Lin Gu, Art Subpa-Asa, Ilyes Kacher, Ko Nishino, and Imari Sato. A data-driven approach for direct and global component separation from a single image. In *Proc. of Asian Conference on Computer Vision*, 2018.