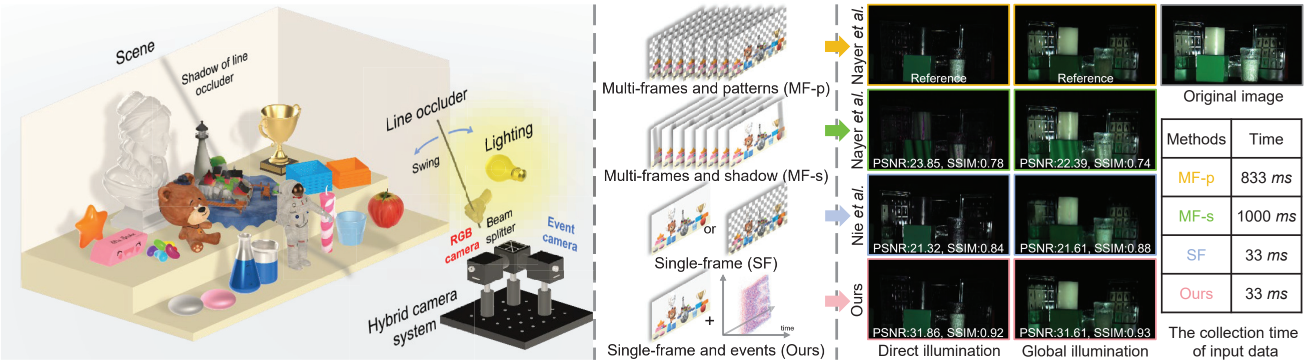
EvDiG: Event-guided Direct and Global Components Separation

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



Methods	Time
MF-p	833 *ms*
MF-s	1000 *ms*
SF	33 *ms*
Ours	33 *ms*

(a) Illustration of our method (b) Different input types (c) Comparison of qualitative results and data collection time

Figure 1. (a) We achieve the separation of direct and global illumination components of a scene by introducing a hybrid system comprising an RGB camera and an event camera. The RGB camera captures an original scene image, while the event camera records rapid illumination changes induced by the shadow of a line occluder sweeping over the scene. We propose EvDiG to combine the two signals for achieving effective and efficient separation. (b) Classification of the input data requirements of existing methods. (c) The results of the corresponding representative methods (Nayer *et al*. [25], Nie *et al*. [26]) and their data collection time (We collect 30 fps videos and input 25/30/1/1 frames into MF-p/MF-s/SF/ours methods respectively, events are collected during the interframe period).

Abstract

Separating the direct and global components of a scene aids in shape recovery and basic material understanding. Conventional methods capture multiple frames under high frequency illumination patterns or shadows, requiring the scene to keep stationary during the image acquisition process. Single-frame methods simplify the capture procedure but yield lower-quality separation results. In this paper, we leverage the event camera to facilitate the separation of direct and global components, enabling video-rate separation of high quality. In detail, we adopt an event camera to record rapid illumination changes caused by the shadow of a line occluder sweeping over the scene, and reconstruct the coarse separation results through event accumulation. We then design a network to resolve the noise in the coarse separation results and restore color information. A real-world dataset is collected using a hybrid camera system for network training and evaluation. Experimental results show superior performance over state-of-the-art methods.

1. Introduction

When a scene is illuminated by a light source, the radiance of each point is a cumulative result of both direct and global components. The direct component signifies the light that travels straight from the source, reflecting just once before reaching the observer. Light also interacts intricately with the environment, resulting in phenomena like interreflections, scattering, and diffusion. These phenomena make up what is known as global illumination.

The separation of direct and global components is a crucial process that can provide material properties of the scene and reveal information about complex shape-material-lighting interactions. For example, eliminating the effect of global illumination can improve shape recovery techniques like photometric stereo [39] and structured light scanning [5]. These techniques predominantly assume direct or low-frequency light transport and do not account for global illumination effects.

How to separate direct and global components accurately is a lasting and challenging problem. The direct components travel the shortest distance among all light paths that

* Corresponding author

arrive at a pixel [40]. The different arrival time of direct and global lighting naturally leads to the development of time-resolved methods. Ultrahigh temporal resolution cameras, such as time-of-flight (ToF) cameras [10, 13] and single photon sensors [19, 30], are employed to distinguish the unique transient behaviors of two illumination components in the order of 10 ps. Despite their effectiveness, the application of time-resolved methods is limited by the high cost of special devices and long periods of data capture [40].

In light of the inherent limitations of time-resolved methods, the separation of direct and global components on RGB images has been explored. The simplest setup is separating two components from a single image, as demonstrated by the single-frame (SF) method illustrated in Fig. 1(b), while it is an obviously ill-posed problem. Thus various priors have been used to enhance separation accuracy, such as sparsity [35], constraints on global illumination types [2], and learned priors [26]. However, achieving reliable illumination separation, especially in challenging scenes such as areas with highly specular reflection, remains difficult when only relying on such priors. To alleviate ill-posedness, Nayar *et al*. [25] propose an active illumination method to separate the direct and global components using high frequency illumination (MF-p and MF-s in Fig. 1(b)). This method achieves promising performance and several methods even treat their results as ground truth (*e.g*., Nie *et al*. [26]). However, it requires capturing multiple frames and keeping the scene static throughout data capture. Achar *et al*. [1] extend the methodology to dynamic scenes with motion compensation, but neglect changes of direct and global components caused by motion within a temporal sliding window.

Event camera [18] is a novel type of neuromorphic sensor that can detect radiance changes and trigger an event whenever its log variation exceeds a threshold. Thanks to their high temporal resolution property, event cameras have been applied to physics-based vision tasks such as shape from polarization [23], photometric stereo [33], and structured light scanning [22], offering solutions that make a great trade-off between speed and accuracy. Several properties of event cameras render them suitable for the separation of direct and global components. Firstly, drawing upon the principles of separation utilizing high frequency illumination [25], only the minimum and the maximum brightness values of each pixel are required to compute the direct and global components. The signal-triggering model of event cameras makes them efficient for capturing the whole radiance change process with significantly **low throughput**. Secondly, the **high temporal resolution** of event cameras can significantly reduce data capture duration, enhancing adaptability to dynamic scenes. Thirdly, their **high dynamic range** can alleviate overexposure issues, particularly in areas with highly specular reflection. These relationships motivate us to think about: *Can we use radiance changes recorded via the event stream to assist the separation of direct and global components?*

Introducing event cameras to the direct-global separation task naturally brings low throughput, shortened data capture periods, and high dynamic range advantages. But processing event streams for the separation task is not a straightforward endeavor. Event cameras have lower signal-to-noise levels and lack color information compared to conventional cameras. Thus, it is necessary to address the impact of noise in event cameras and investigate the separation of color information guided by grayscale event streams.

In this paper, we propose **EvDiG**, to dig out unique properties of Event cameras for effective and efficient Direct and Global components separation, with efforts in solving the above challenges. EvDiG only requires an RGB-event hybrid camera system and a stick occluder for separation, where the combination of a quickly sweeping stick and a fast-sensing event camera enhances **data capture efficiency**. Inputs include a single starting frame of the original scene and corresponding event streams capturing shadow-induced brightness changes (Fig. 1(a)). We first reconstruct the grayscale minimum and maximum brightness value in the shadow-changing process through event accumulation, thereby obtaining coarse separation results. To address problems inherent in the coarse results, we design a two-stage network to refine the coarse separation results and restore color information. Overall, this paper makes the following contributions by proposing:

- the first approach for direct and global separation with a hybrid setup of both an RGB and an event camera, thereby reducing the capture time and adapting to dynamic scenes;
- the event-guided separation framework comprised of two networks designed to enhance the accuracy of noisy coarse separation results and bridge the gap in color representation between the RGB image and events;
- the first dataset for the separation of direct and global components that includes both real-world events and images. Our method outperforms single-frame methods and matches the performance of multi-frame+pattern methods. It maintains a data capture time equivalent to that of single-frame methods and is 20 times more efficient than multi-frame methods (Fig. 1(c)).

2. Related Works

Image-based separation methods. Nayar *et al*. [25] show that direct-global separation can be achieved using high frequency illumination. With active illumination, this separation can be performed with two complementary patterns, but around 20 pattern images are required to obtain satisfactory results with a real digital projector. There have been several extensions proposed, such as direct-global separation for multiple light sources with a multiplexed illumination scheme [9], compensating for motions to allow direct-global separation for dynamic scenes [1], and high-speed capture with temporal dithering of DLP projectors [24]. In addition, O'Toole *et al*. [28, 29] propose a camera-projector system with primal-dual coding which modulates both the illumination and the camera response, to probe the light transport matrix. Based on this system, direct and global components can be directly captured in live videos.To reduce capture time and simplify the hardware setup, single-frame methods have been explored, using either a high-frequency pattern image [6, 35] or the original scene image [26] as input. Due to the ill-posed nature of single-frame separation, these methods strongly rely on priors like sparsity [35] and learned priors [6, 26].

Time-resolved separation methods. The time-resolved method is another paradigm for the separation of direct and global components that exploits ultrafast unconventional cameras and the finite speed of light. Wu *et al*. [40] propose to use temporal delay profiles to analyze global light transport with a streak camera [37]. Similar separation methods can be applied to other ToF camera modifications [13, 16]. Besides ToF cameras, multiple time-resolved methods have been proposed, such as optical interferometer [8] and single-photon avalanche diode (SPAD) sensors [19, 30]. Measuring the full transient images of the scene is a straightforward way to separate the two components. The ultrahigh temporal resolution enables them to distinguish different traveling time of optical components. However, to obtain the whole x-y-t data, multiple captures are required, which increases the overall data capture time. For instance, the streak camera [37] requires one hour capture time for a 600 slices ToF image [40]. Separation using optical interferometer [8] needs to translate the reference mirror to capture frames corresponding to different path-lengths. A comparison between time-resolved sensors and event cameras is presented in Tab. 1.

Event camera for physics-based vision. Event cameras have been introduced into many physics-based vision tasks in recent years. Corresponding methods have demonstrated the distinct advantages of event cameras across various fields, such as shape from polarization [23] and structured light scanning [21, 22], *etc*. MC3D [21] is an event-based structured light 3D scanning technique that simultaneously achieves high resolution, high speed and robust performance. Building on MC3D, ESL [22] enhances noise robustness by exploiting regularities in neighborhoods of event data. Takatani *et al*. [36] utilize an event camera to acquire bispectral difference images using temporally modulated illumination. Chen *et al*. [4] leverage the event camera to alleviate the intensity-distance ambiguity for parametric indoor lighting estimation. Similarly, Chen *et al*. [3] integrate the event camera into a visible light positioning system. By introducing "transient event frequency", Han *et al*. [11] derive precise radiance values from high-temporal-resolution event signals during light activation. Besides, Xu *et al*. [41] propose a method for effectively extracting electric network frequency traces based on the event-sensing. These algorithms showcase the wide range of application scenarios and the potential of event cameras in tasks related to scene understanding.

Table 1. Comparison of typical transient imaging sensors used in direct-global separation [15] with event cameras.

Methods	Overall Acquisition Time	Temporal Resolution	Technology
PMD sensor [13]	90 s	1000 ps	Time-of-flight imaging
Optical interferometry [8]	$> 1\ h$	0.033 ps	Interferometry-based imaging
SPAD [30]	64 s	300 ps	Photon accumulation imaging
Event camera	33 ms	1 μs	Radiance change sensing

3. Method

3.1. Image formation model

Event formation. An event signal $e = (\mathbf{p}, t, \sigma)$ is triggered whenever the logarithmic change of brightness at pixel $\mathbf{p} = (x, y)$ and time t exceeds a preset threshold θ:

$$|\log \mathbf{I}_t(\mathbf{p}) - \log \mathbf{I}_{t-\Delta t}(\mathbf{p})| \geq \theta, \tag{1}$$

where $\mathbf{I}_t(\mathbf{p})$ denotes the intensity of pixel $\mathbf{p}$ at time t, and the previous event of pixel $\mathbf{p}$ is triggered at $t - \Delta t$. Polarity $\sigma \in \{1, -1\}$ indicates whether the intensity changes increase or decrease. Since Eq. (1) applies to each pixel $\mathbf{p}$ independently, pixel indices are omitted henceforth.

Given the instantaneous latent image $\mathbf{I}_{t_1}$, let's assume N_e events occurring between t_1 and t_2, denoted as $\{e_k\}_{k=1}^{N_e}$. We can obtain the latent image $\mathbf{I}_{t_2}$ from the physical model of event cameras:

$$\log \mathbf{I}_{t_2} = \log \mathbf{I}_{t_1} + \theta \cdot \sum_{k=1}^{N_e} \sigma_k. \tag{2}$$

Direct and global components separation. The intensity of each pixel comprises two components: the direct component $\mathbf{I}_d$ and the global component $\mathbf{I}_g$. It is assumed that $\mathbf{I}_d$ and $\mathbf{I}_g$ remain constant at each scene point within a small temporal window. When the scene is lit with high frequency illumination at time t, the contribution of the global illumination component to brightness can be approximated as being scaled by a spatially-uniform factor [25], denoted as $\beta \mathbf{I}_g$. Here, β represents the fraction of activated source pixels. If the scene point is not lit by the source, there will be no direct contribution. Consequently, the intensity at time t

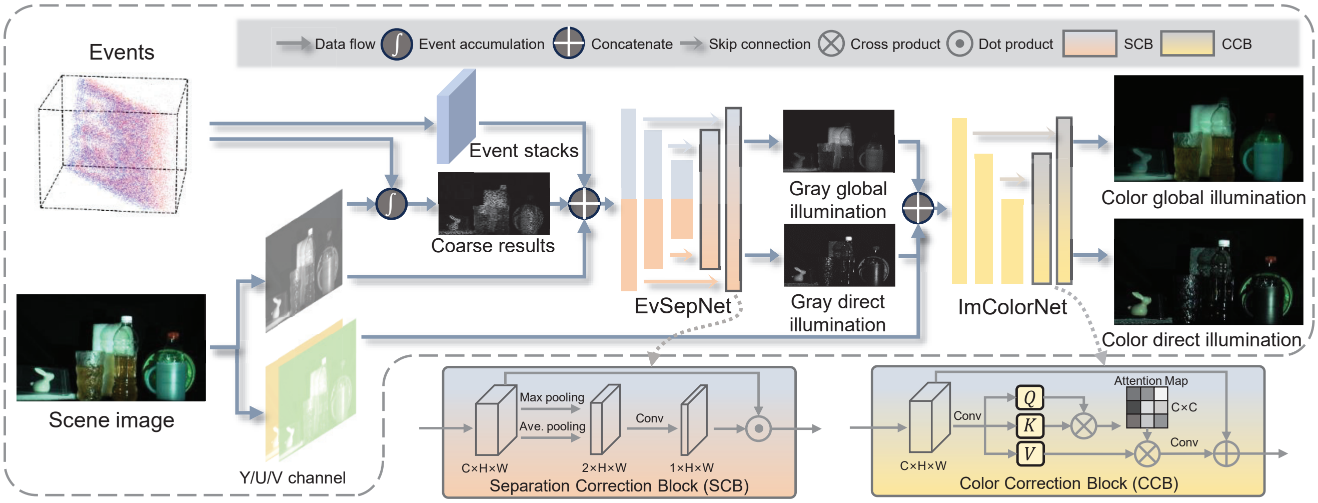


Figure 2. Network architecture of EvDiG. We first obtain coarse separated global components in grayscale channel through event accumulation, then use a U-Net and separation correction block built module EvSepNet to refine grayscale separation, and an ImColorNet module built by color correction blocks is designed to achieve chrominance compensation.

can be expressed as:

$$\mathbf{I}_t = \mathbf{m}_t \odot \mathbf{I}_d + \beta \mathbf{I}_g, \tag{3}$$

where $\mathbf{m}_t$ denotes the illumination pattern at the camera pixels at time t and $\odot$ is element-wise multiplication. Nayar *et al*. [25] propose to utilize a projector to generate the high frequency illumination patterns or use source occluders cast high frequency shadows on the scene.

When changing the projected illumination pattern or sweeping the source occluder across the scene, we will capture a collection of images $\{\mathbf{I}_{t_k}\}_{k=0}^{N}$, where N represents the number of images captured. Defining $\mathbf{I}_{\max}$ and $\mathbf{I}_{\min}$ as the maximum and minimum intensities observed at each scene point within $\{\mathbf{I}_{t_k}\}_{k=0}^{N}$, we can separate the direct and global components from:

$$\mathbf{I}_{\max} = \mathbf{I}_d + \beta \mathbf{I}_g, \quad \mathbf{I}_{\min} = \beta \mathbf{I}_g. \tag{4}$$

The separation of direct and global components using high-frequency illumination ideally requires only two complementary images. In cases using projection illumination patterns, achieving perfect complementary patterns with real digital projectors is challenging, due to light leakages within the projector optics and limited depth of field. In the experimental setting of [25], 25 images are captured to ensure the efficacy of separation results. Similarly, when employing high frequency shadows, to guarantee all scene points have been subjected to the shadow's umbra, a substantial number of images are typically required, often exceeding 20 (*i.e*., $N > 20$).

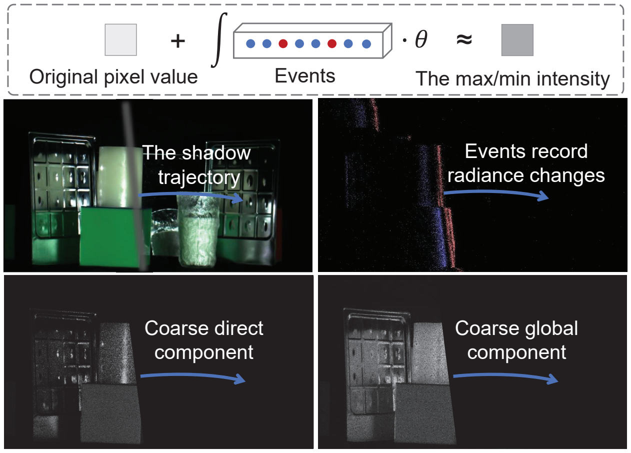


Figure 3. The illustration of the event-based minimum brightness value reconstruction. The θ is the threshold in Eq. (1).

3.2. Event-guided separation

The pipeline of our method is illustrated in Fig. 2. For the sake of versatility, we focus on separation from moving shadows in this work. Initially, the color space of the starting frame is converted from RGB to YUV. Subsequently, we employ event accumulation to reconstruct the maximum and minimum brightness values in grayscale for each pixel, as formulated in Eq. (2). The coarse separated results are derived using Eq. (4). Then **EvSepNet** is employed to refine the grayscale coarse separation results and **ImColorNet** is used to infer the color information, enhancing color accuracy and authenticity.

Coarse separation using event accumulation. We substitute the image collection $\{\mathbf{I}_{t_k}\}_{k=0}^{N}$ with the initial frame $\mathbf{I}_{t_0}$ and events triggered between t_0 and t_N. The event streams, denoted as $\mathcal{E}$, continuously measure the dynamic brightness changes induced by moving shadows. With the grayscale starting frame $\mathbf{I}_{t_0}^{Y}$ and the corresponding events $\mathcal{E}$, we can reconstruct any latent image $\mathbf{I}_{t_k}^{Y}$ within the interval from t_0 to t_N, applying an event accumulation process as shown in Eq. (2) and illustrated in Fig. 3. Following the separation approach presented in Eq. (4), we need to compute the maximum and minimum pixel intensities during the period. Leveraging the asynchronous nature of events, we update the maximum and minimum intensities at each pixel in an event-by-event manner. Then we obtain the coarse estimated direct component $\hat{\mathbf{I}}_d^{Y}$ and global component $\hat{\mathbf{I}}_g^{Y}$.

Separation through event accumulation offers several advantages: 1) With high temporal resolution of event cameras, the capture time can be greatly reduced. Remarkably, even when employing a line occluder, the capture time required is close to single-frame methods. 2) The event-triggering model, as formulated in Eq. (1), optimizes the capture of essential brightness changes, thereby reducing data redundancy in $\{\mathbf{I}_{t_k}\}_{k=0}^{N}$. However, the coarse separation results manifest several issues. Firstly, $\hat{\mathbf{I}}_d^{Y}$ and $\hat{\mathbf{I}}_g^{Y}$ are noisy due to spatial-temporal variations in thresholds [14], and quantization errors inherent in event cameras. Additionally, when employing a simple line occluder to cast high frequency shadows, the smooth global assumption in Eq. (3) does not always hold, leading to inaccuracies in the separation. Moreover, the coarse separation results are in grayscale, devoid of color information. To overcome these challenges, we propose a two-stage network specifically designed to refine the coarse separation results and restore color information.

EvSepNet. To address the noise and inaccuracies present in the coarse separation results $\hat{\mathbf{I}}_d^{Y}$ and $\hat{\mathbf{I}}_g^{Y}$, we employ a U-Net architecture-based network, herein referred to as EvSepNet. This network is specifically designed to denoise and refine the separation results derived from event accumulation, effectively dealing with cases of high-frequency global illumination. We transform the input events to an event stack $\mathbf{E}$, as described in [7]. The refined separation results are obtain from:

$$\mathbf{I}_d^{Y}, \mathbf{I}_g^{Y} = f_r(\mathbf{I}_{t_0}^{Y}, \mathbf{E}, \hat{\mathbf{I}}_d^{Y}, \hat{\mathbf{I}}_g^{Y}), \tag{5}$$

where $\mathbf{I}_d^{Y}$ and $\mathbf{I}_g^{Y}$ represent the refined direct and global components in grayscale respectively, and f_r denotes the implicit function modeled by EvSepNet. The multi-scale architecture has been proven to be effective for event-based video reconstruction [31] and image-event data fusion [42]. The image and event stack features are fused in a multi-scale manner by the EvSepNet. Leveraging the input image as guidance, the network can predict the brightness changes more accurately under spatial-temporal variant thresholds.

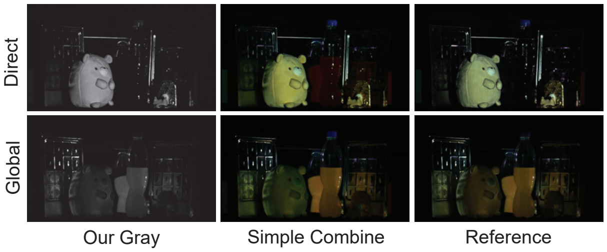


Figure 4. An example (real data) of combining the grayscale separation results (left column) with the U, V channel of the original frame (middle column). Such a straightforward strategy leads to unnatural color. The reference results are obtained by using [25] with shifted checkerboard patterns shown in the right column.

To tackle the issue of high-frequency global illumination, we need to estimate a spatial-variant factor β instead of a spatial-uniform one in Eq. (3). A Separation Correction Block (SCB) is integrated into each decoder stage of EvSepNet to implicitly reweight features at each location, as depicted in Fig. 2. The design of the SCB is inspired by previous works in spatial attention [38]. For a given input feature $\mathbf{F} \in \mathbb{R}^{H \times W \times C}$, we first aggregate its channel information by average and maximum pooling and obtain $\mathbf{F}_{\text{max}}^{s}$ and $\mathbf{F}_{\text{avg}}^{s}$. Then those are concatenated and processed by a convolution layer with 11×11 kernels to obtain the spatial attention map $\mathbf{A}^{s} \in \mathbb{R}^{H \times W \times 1}$, which is formulated as:

$$\mathbf{A}^{s} = \text{Sigmoid}(\text{Conv}([\mathbf{F}_{\text{max}}^{s}; \mathbf{F}_{\text{avg}}^{s}])). \tag{6}$$

Subsequently, spatially corrected features are obtained through element-wise multiplication. The SCB enables EvSepNet to implicitly extract light transport cues encoded in the image and events, facilitating separation correction. The network effectively diminishes artifacts arising from high-frequency global illumination issues in the results.

ImColorNet. This network is designed for chrominance compensation to restore color information in separation results. A preliminary approach to color restoration is to combine the estimated grayscale separation results from EvSepNet with U, V channel of the original frame, denoted as $\mathbf{I}^{U}$, $\mathbf{I}^{V}$ respectively, and convert it back to RGB color space. However, this strategy overlooks the essential fact that the color information inherent in direct and global illuminations differs due to the variance in their light paths. Simply applying the color information in the original frame $\mathbf{I}_{t_0}$ fails to yield satisfactory results. This issue is illustrated in Fig. 4, where directly combining the U, V channel of the original image with the grayscale separation results produces an inaccurate color representation.

Inspired by the chrominance compensation network proposed in [12], we design the color correction network, named ImColorNet, to recover the true color appearance in direct and global components as:

$$\mathbf{I}_d, \mathbf{I}_g = f_c(\mathbf{I}_d^Y, \mathbf{I}_g^Y, \mathbf{I}^U, \mathbf{I}^V), \tag{7}$$

where f_c is the implicit function modeled by ImColorNet, $\mathbf{I}_d$ and $\mathbf{I}_g$ are the final color-corrected results for the direct and global components.

The primary challenge in chrominance compensation involves color correction using the color information in the original image, under the guidance of light transport information encoded in the grayscale separation results. Channel-wise matrix transformation is commonly used in canonical ISP [17] for both color correction and color space conversion. Drawing inspiration from this, we propose a Color Correction Block (CCB) to perform color correction implicitly in feature space. In the decoder of ImColorNet, CCBs are stacked at each stage, as depicted in Fig. 2.

The design of CCB is based on recent advances in transposed self-attention [43]. For a given input feature $\mathbf{F} \in \mathbb{R}^{H\times W\times C}$, the CCB first generates query $\mathbf{Q}$, key $\mathbf{K}$ and value $\mathbf{V}$ projections by applying a 1×1 convolution layer followed by a depth-wise convolution layer. Then $\mathbf{Q}$ and $\mathbf{V}$ are reshaped, then the attention map $\mathbf{M} \in \mathbb{R}^{C\times C}$ is obtained through matrix multiplication, which is defined as:

$$\begin{aligned} \mathbf{Q}, \mathbf{K}, \mathbf{V} &= \mathrm{Reshape}(\mathrm{DConv}(\mathrm{Conv}(\mathbf{F}))), \\ \mathbf{M} &= \mathrm{Softmax}(\mathbf{K} \otimes \mathbf{Q}/\lambda), \\ \mathbf{F}_{\mathrm{out}} &= \mathrm{Conv}(\mathbf{V} \otimes \mathbf{M}) + \mathbf{F}, \end{aligned} \tag{8}$$

where $\mathbf{F}_{\mathrm{out}}$ is the color corrected feature, DConv is a depth-wise convolution layer, and λ is a learnable scaling parameter. The CCB enables ImColorNet to perform global color correction using channel-wise self-attention, supplemented by local color refinement through convolution layers.

3.3. Implementation details

Loss functions. We derive the reference direct and global components obtained from Nayar *et al*. [25] as the pseudo ground truth to train the proposed network. The loss function $\mathcal{L}$ for training is a linear combination of the reconstruction loss $\mathcal{L}_{\mathrm{mse}}$, $\mathcal{L}_{\mathrm{lap}}$ and perceptual loss $\mathcal{L}_{\mathrm{perc}}$:

$$\mathcal{L} = \alpha_1 \mathcal{L}_{\mathrm{mse}} + \alpha_2 \mathcal{L}_{\mathrm{lap}} + \alpha_3 \mathcal{L}_{\mathrm{perc}}, \tag{9}$$

where $\alpha_1 = 10$, $\alpha_2 = 1$ and $\alpha_3 = 0.5$ are hyper-parameters to balance the contributions of different terms. The term $\mathcal{L}_{\mathrm{mse}}$ represents a pixel-wise mean square error (MSE). The $\mathcal{L}_{\mathrm{lap}}$ is the L_1 distance between two Laplacian pyramid representations [27] comprising 5 layers. The perceptual loss $\mathcal{L}_{\mathrm{perc}}$ is defined as the L_2 distance between features extracted from a pre-trained VGG-19 network [34] on the ImageNet dataset [32].

Training details. We implement our approach using the Pytorch framework and run on a single NVIDIA GeForce RTX 3090 GPU. We use AdamW optimizer [20] with the default parameter setting in the training phrase. We configure the batch size to 12. The EvSepNet and ImColorNet are trained together for 100 epochs. The initial learning rate is set to 3×10^{-4}, and a cosine annealing learning rate scheduling strategy is employed. We apply random cropping, horizontal flipping and rotation for spatial-level data augmentation, and incorporate pixel-level augmentations, including color channel shuffle and brightness scaling.

4. Experiments

In this section, we introduce our collected dataset in Sec. 4.1, and qualitatively and quantitatively compare our method with state-of-the-art direct and global components separation methods on our collected dataset (Sec. 4.2), and our real-captured outdoor scenes (Sec. 4.3).

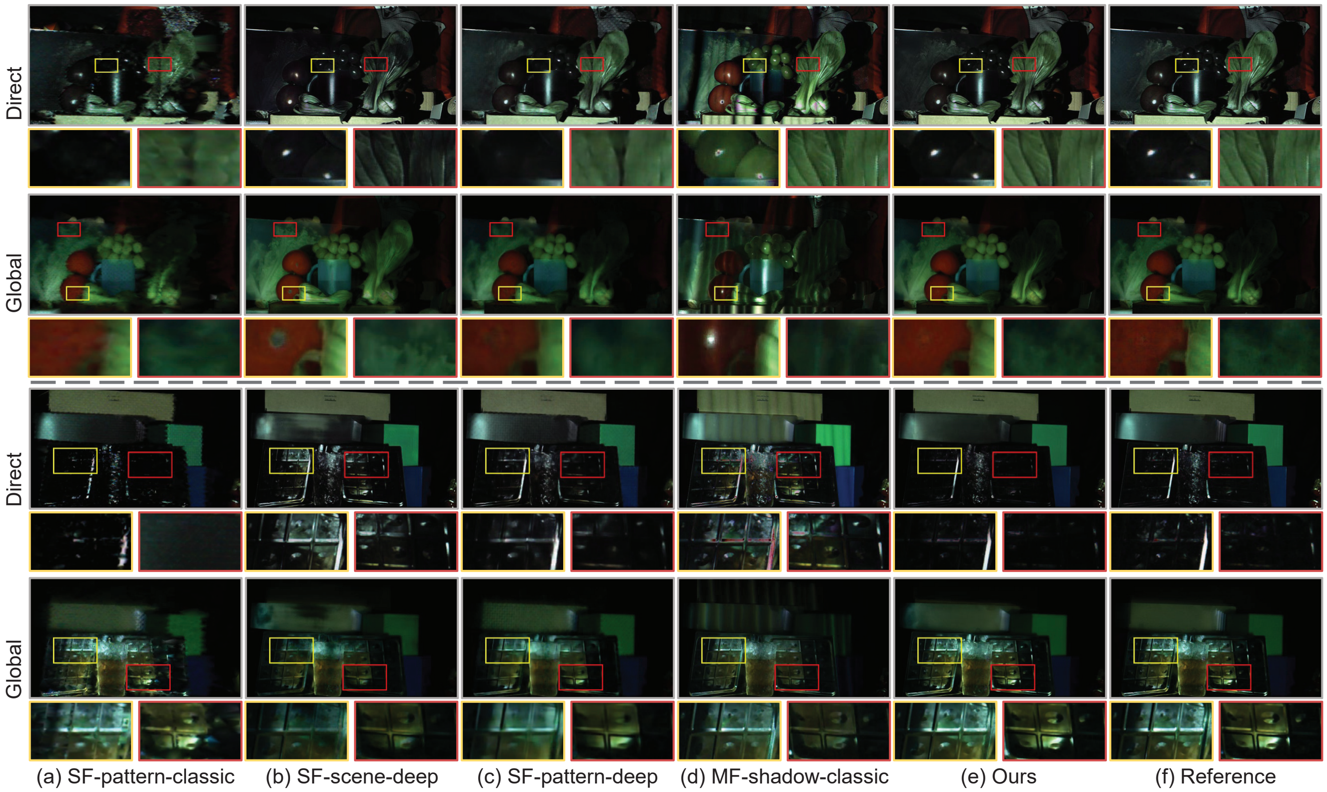
4.1. Dataset collection

Due to the absence of suitable datasets for the separation of direct and global components using event cameras, we collect a dataset consisting of both images and events under controlled indoor scenes. The captured scenes encompass a wide range of daily-life objects, meticulously selected to cover different global illumination effects, such as interreflection and subsurface scattering.

Data collection for each scene is conducted in two steps: 1) Initially, the scene is illuminated with high-frequency checkerboard patterns emitted by a projector. The reference direct and global components are computed using the methodology established by Nayar *et al*. [25]. 2) Subsequently, a white background pattern is projected onto the scene, while source occluders move across the scene at varied speeds. During this phase, both RGB videos and corresponding event signals are recorded. Following the setup outlined in [25], we employ 25 shifted checkerboard patterns for data collection. The images are captured using a 30 fps RGB camera, requiring a minimum total capture duration of 833 ms, with an interframe period of approximately 33 ms, as shown in Fig. 1 (c). In total, we collect 230 distinct scenes and 1800 clips of moving shadow .

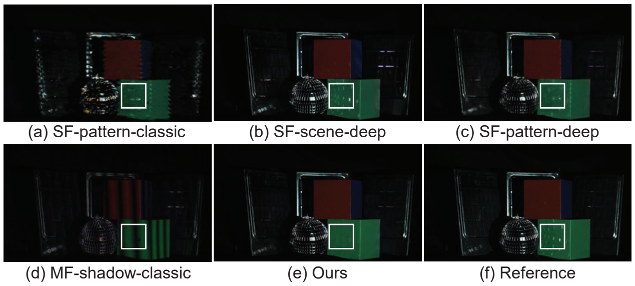
4.2. Comparison on indoor scenes

We compare EvDiG with four image-based methods: learning-based single-frame method that takes the original scene image as input (SF-scene-deep) [26], learning-based single-frame method using one single pattern image (SF-pattern-deep) [6] and two variants of Nayar *et al*. [25] which take one single pattern image (SF-pattern-classic) and multiple shadow-sweeping images (MF-shadow-classic) as input respectively. For the MF-shadow-classic method, we regulate the movement of source occluders to capture more



(a) SF-pattern-classic (b) SF-scene-deep (c) SF-pattern-deep (d) MF-shadow-classic (e) Ours (f) Reference

Figure 5. Direct and global components separation results on our captured real-world indoor scenes. EvDiG is compared to (a) SF-pattern-classic [25], (b) SF-scene-deep [26], (c) SF-pattern-deep [6], and (d) MF-shadow-classic [25], where we name these methods using three encoded elements: single frame (SF) or multiple frames (MF) based methods, original scene image (scene) or pattern-marked image (pattern) or occluder-shadowed image (shadow) as the input, and non-deep learning (classic) or deep learning (deep) based methods.



(a) SF-pattern-classic (b) SF-scene-deep (c) SF-pattern-deep (d) MF-shadow-classic (e) Ours (f) Reference

Figure 6. Direct components separation results in a challenging scene with a disco ball. Note that the interreflections from the disco ball should not appear in the direct components. The compared methods are (a) SF-pattern-classic [25], (b) SF-scene-deep [26], (c) SF-pattern-deep [6], and (d) MF-shadow-classic [25].

Table 2. Quantitative comparison on our 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.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
MF-shadow-classic [25]	20.59/19.98	0.768/0.653	0.203/0.215
SF-pattern-classic [25]	23.73/27.86	0.675/0.762	0.288/0.222
SF-scene-deep [26]	26.05/27.67	0.805/0.807	0.121/0.178
SF-pattern-deep [6]	28.51/**32.96**	0.834/**0.867**	0.208/0.228
Ours	**30.01**/31.66	**0.883**/0.846	**0.077/0.117**

than 30 images, which takes as least 1 s for the entire data collection process, as illustrated in Fig. 1(c). As for EvDiG, the inherent high-speed capabilities of event cameras allow for rapid event data collection. By precisely controlling the source occluder, the event data collection duration can be reduced to fit within the interframe interval, thus achieving a data capture time equivalent to that of single-frame methods. In our dataset, the source occluder is swung at random speeds, and the total capture time of our setup varies from 33 to 2000 ms, demonstrating the robustness of our method across a broad spectrum of dynamic scenarios. For separation results on dynamic scenes, please refer to the supplementary video for a more comprehensive visualization.

The two aforementioned learning-based single-frame methods are retrained on our dataset, leveraging the reference direct and global components as pseudo ground truth for training. We utilize Peak Signal-to-Noise (PSNR),

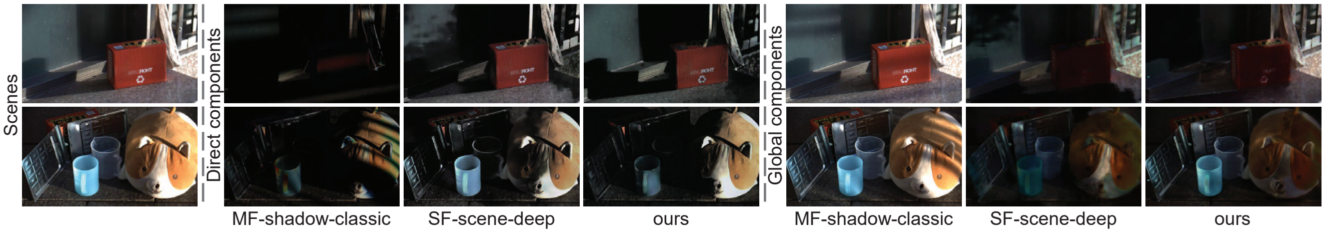


Figure 7. Direct and global components separation results on our captured real-world outdoor scenes. EvDiG is compared to MF-shadow-classic [25] and SF-scene-deep [26] methods.

Structural Similarity (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) between the separation results and reference to evaluate the performance of each method. The quantitative comparison results are listed in Tab. 2, and the qualitative comparisons are shown in Fig. 5. Our method performs well across all metrics, markedly surpassing the two variants of Nayar *et al*. [25] and the SF-scene-deep [26]. The SF-pattern-deep method [6] tends to produce results that, while high in PSNR and SSIM, exhibit blurry textures accompanied by checkerboard-like artifacts. The inherent smoothness of global components contributes to slightly lower scores in terms of PSNR and SSIM for the global components relative to the SF-pattern-deep approach. The artifacts introduced by SF-pattern-deep become apparent when evaluated using the LPIPS metric. In the first case presented in Fig. 5, the specular reflection region on the tomato is accurately separated by our method, attributable to the high dynamic range capability of event cameras. The second example showcased in Fig. 5 demonstrates our method's effectiveness in separating interreflections. Note that the input of compared methods are different, the comparison here is to illustrate the considerable performance gains attainable by incorporating event streams.

In Sec. 3.2, we introduce EvSepNet to address the challenges posed by high-frequency global illuminations. A challenging scene with a disco ball is shown in Fig. 6. Because the smooth global illumination assumption is violated, the reference results obtained from Nayar *et al*. [25] fails to separate the specular interreflections of the disco ball, as shown in Fig. 6 (f). In contrast, our method, trained on the reference results, removes the specular interreflections in the direct components. This improvement shows the efficacy of EvSepNet and highlights the advantages of separating components based on event data.

4.3. Comparison on outdoor scenes

Direct and global separation using source occluders can be utilized for outdoor scenes. To further verify the effectiveness of our method in real-world scenarios, we capture several outdoor scenes with the same hybrid setup. We compare our method with MF-shadow-classic [25] and SF-scene-deep [26], which are applicable in outdoor settings. The visual comparisons are presented in Fig. 7. For pixels never recorded within the umbra of the shadow in the captured image sequence, MF-shadow-classic inaccurately calculates their direct components as zero. However, it is challenging to ensure all pixels have been captured within shadows in a regular 30 fps video using a line occluder. SF-scene-deep, which predicts direct and global components without physical cues, exhibits limited generalization to outdoor scenes, producing severe artifacts in the separation results. In contrast, our method generates satisfactory separation results, showing the robustness to outdoor scenes and rapid-moving occluders.

5. Conclusion

We propose an event-guided direct and global components separation method. Our method takes advantage of the high temporal resolution events to record fast illumination changes, greatly reducing the data capture time close to that of single-frame methods. We propose EvSepNet and ImColorNet to resolve the noise and colorless issues in the coarse separation results. Experimental results show that our method achieves comparable performance with multi-frame methods, whose data capture time is 20 times longer.

Limitations. Since the ground truth of the separation task is difficult to obtain, the results of method [25] are used as references for performance comparison in this paper. Besides, the impact of events caused by motion is not considered in this paper, our method obtains satisfactory performance only when the motion is relatively minor compared to the movement of the shadow in dynamic scenes.

Acknowledgement

This work was supported by National Science and Technology Major Project (Grant No. 2021ZD0109803), Beijing Natural Science Foundation (Grant No. L233024) and National Natural Science Foundation of China (Grand No. 62088102, 62136001, 62276007). Peiqi Duan was also supported by China National Postdoctoral Program for Innovative Talents (Grant No. BX20230010) and China Postdoctoral Science Foundation (Grant No. 2023M740076).

References

[1] Supreeth Achar, Stephen T Nuske, and Srinivasa G Narasimhan. Compensating for motion during direct-global separation. In *Proc. of International Conference on Computer Vision*, 2013. 2, 3

[2] Jiamin Bai, Manmohan Chandraker, Tian-Tsong Ng, and Ravi Ramamoorthi. A dual theory of inverse and forward light transport. In *Proc. of European Conference on Computer Vision*, 2010. 2

[3] Guang Chen, Wenkai Chen, Qianyi Yang, Zhongcong Xu, Longyu Yang, Jörg Conradt, and Alois Knoll. A novel visible light positioning system with event-based neuromorphic vision sensor. *IEEE Sensors Journal*, 20(17):10211–10219, 2020. 3

[4] Zehao Chen, Qian Zheng, Peisong Niu, Huajin Tang, and Gang Pan. Indoor lighting estimation using an event camera. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3

[5] Brian Curless and Marc Levoy. Better optical triangulation through spacetime analysis. In *Proc. of International Conference on Computer Vision*, 1995. 1

[6] 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. 3, 6, 7, 8

[7] Daniel Gehrig, Antonio Loquercio, Konstantinos G Derpanis, and Davide Scaramuzza. End-to-End learning of representations for asynchronous event-based data. In *Proc. of International Conference on Computer Vision*, 2019. 5

[8] Ioannis Gkioulekas, Anat Levin, Frédo Durand, and Todd Zickler. Micron-scale light transport decomposition using interferometry. *ACM Transactions on Graphics*, 34(4):1–14, 2015. 3

[9] Jinwei Gu, Toshihiro Kobayashi, Mohit Gupta, and Shree K Nayar. Multiplexed illumination for scene recovery in the presence of global illumination. In *Proc. of International Conference on Computer Vision*, 2011. 3

[10] Mohit Gupta, Shree K Nayar, Matthias B Hullin, and Jaime Martin. Phasor imaging: A generalization of correlation-based time-of-flight imaging. *ACM Transactions on Graphics*, 34(5):1–18, 2015. 2

[11] Jin Han, Yuta Asano, Boxin Shi, Yinqiang Zheng, and Imari Sato. High-fidelity event-radiance recovery via transient event frequency. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[12] Jin Han, Yixin Yang, Peiqi Duan, Chu Zhou, Lei Ma, Chao Xu, Tiejun Huang, Imari Sato, and Boxin Shi. Hybrid high dynamic range imaging fusing neuromorphic and conventional images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8553–8565, 2023. 6

[13] Felix Heide, Matthias B Hullin, James Gregson, and Wolfgang Heidrich. Low-budget transient imaging using photonic mixer devices. *ACM Transactions on Graphics*, 32(4):1–10, 2013. 2, 3

[14] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbrück. v2e: From video frames to realistic DVS events. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1312–1321, 2021. 5

[15] Adrian Jarabo, Belen Masia, Julio Marco, and Diego Gutierrez. Recent advances in transient imaging: A computer graphics and vision perspective. *Visual Informatics*, 1(1): 65–79, 2017. 3

[16] Achuta Kadambi, Refael Whyte, Ayush Bhandari, Lee Streeter, Christopher Barsi, Adrian Dorrington, and Ramesh Raskar. Coded time of flight cameras: sparse deconvolution to address multipath interference and recover time profiles. *ACM Transactions on Graphics*, 32(6):1–10, 2013. 3

[17] Hakki Can Karaimer and Michael S Brown. A software platform for manipulating the camera imaging pipeline. In *Proc. of European Conference on Computer Vision*, 2016. 6

[18] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128$\times$ 128 120 db 15 μs latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43 (2):566–576, 2008. 2

[19] David B Lindell, Matthew O'Toole, and Gordon Wetzstein. Towards transient imaging at interactive rates with single-photon detectors. In *Proc. of International Conference on Computational Photography*, 2018. 2, 3

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. of International Conference on Learning Representations*, 2017. 6

[21] Nathan Matsuda, Oliver Cossairt, and Mohit Gupta. MC3D: Motion contrast 3d scanning. In *Proc. of International Conference on Computational Photography*, 2015. 3

[22] Manasi Muglikar, Guillermo Gallego, and Davide Scaramuzza. ESL: Event-based structured light. In *Proc. of International Conference on 3D Vision*, 2021. 2, 3

[23] Manasi Muglikar, Leonard Bauersfeld, Diederik Paul Moeys, and Davide Scaramuzza. Event-based shape from polarization. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 3

[24] Srinivasa G Narasimhan, Sanjeev J Koppal, and Shuntaro Yamazaki. Temporal dithering of illumination for fast active vision. In *Proc. of European Conference on Computer Vision*, 2008. 3

[25] 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. 1, 2, 3, 4, 5, 6, 7, 8

[26] 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. 1, 2, 3, 6, 7, 8

[27] Simon Niklaus and Feng Liu. Context-aware synthesis for video frame interpolation. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 6

[28] Matthew O'Toole, Ramesh Raskar, and Kiriakos N Kutulakos. Primal-dual coding to probe light transport. *ACM Transactions on Graphics*, 31(4):39–1, 2012. 3

[29] Matthew O'Toole, Supreeth Achar, Srinivasa G Narasimhan, and Kiriakos N Kutulakos. Homogeneous codes for energy-

efficient illumination and imaging. *ACM Transactions on Graphics*, 34(4):1–13, 2015. 3

[30] Matthew O'Toole, Felix Heide, David B Lindell, Kai Zang, Steven Diamond, and Gordon Wetzstein. Reconstructing transient images from single-photon sensors. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 3

[31] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):1964–1980, 2019. 5

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3): 211–252, 2015. 6

[33] Wonjeong Ryoo, Giljoo Nam, Jae-Sang Hyun, and Sangpil Kim. Event fusion photometric stereo network. *Neural Networks*, 167:141–158, 2023. 2

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. of International Conference on Learning Representations*, 2015. 6

[35] Art Subpa-Asa, Ying Fu, Yinqiang Zheng, Toshiyuki Amano, and Imari Sato. Direct and global component separation from a single image using basis representation. In *Proc. of Asian Conference on Computer Vision*, 2017. 2, 3

[36] Tsuyoshi Takatani, Yuzuha Ito, Ayaka Ebisu, Yinqiang Zheng, and Takahito Aoto. Event-based bispectral photometry using temporally modulated illumination. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3

[37] Andreas Velten, Di Wu, Adrián Jarabo, Belén Masiá, Christopher Barsi, Chinmaya Joshi, Everett Lawson, Moungi Bawendi, Diego Gutierrez, and Ramesh Raskar. Femtophotography: capturing and visualizing the propagation of light. *ACM Transactions on Graphics*, 32(4):44:1–44:8, 2013. 3

[38] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *Proc. of European Conference on Computer Vision*, 2018. 5

[39] Robert J Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1):139–144, 1980. 1

[40] Di Wu, Andreas Velten, Matthew O'Toole, Belen Masia, Amit Agrawal, Qionghai Dai, and Ramesh Raskar. Decomposing global light transport using time of flight imaging. *International Journal of Computer Vision*, 107:123–138, 2014. 2, 3

[41] Lexuan Xu, Guang Hua, Haijian Zhang, Lei Yu, and Ning Qiao. "Seeing" electric network frequency from events. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[42] Yixin Yang, Jin Han, Jinxiu Liang, Imari Sato, and Boxin Shi. Learning event guided high dynamic range video reconstruction. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 5

[43] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 6