DualResPS: Dual-Resolution Photometric Stereo Using a Frame-Event Hybrid Camera

Haotian Zhuang[1*], Bohan Yu[2*], Zhuofeng Wang[2], and Boxin Shi[2†]

[1] Tsinghua University, Beijing, China
zht23@tsinghua.edu.cn
[2] Peking University, Beijing, China
{ybh1998,shiboxin}@pku.edu.cn, wangzf2003@stu.pku.edu.cn

Abstract. As an important 3D sensing method, photometric stereo estimates surface normals from multiple photos taken under varying lighting. The surface normal quality highly relies on large number of input images, which requires long capturing time, high data burden and limits its high-speed applications for photometric stereo. Recently, event-based photometric stereo has emerged as an efficient and rapid approach, but the low resolution of event sensors and the noise of event signals limit its high-quality application. In this paper, we propose DualResPS, a novel photometric stereo pipeline utilizing a frame-event hybrid camera. We design a lighting and capturing strategy tailored to a hybrid-camera setup to utilize observations of different spatial and temporal resolutions. The algorithm explicitly models non-ideal factors including shadow effects and specular reflection to achieve high-resolution, high-quality normal reconstruction. Our proposed DualResPS is validated on both semi-real datasets from the DiLiGenT, DiLiGenT-Pi, and real captured data. The results demonstrate that DualResPS surpasses its frame-based counterpart with 18.1% data bandwidth.

Keywords: Photometric stero · Event camera · Neural inverse rendering

1 Introduction

Photometric stereo [38] is an important method in computer vision for 3D geometry sensing. With photos captured under different lighting directions, surface normals of the objects are reconstructed with fine-grain details aligned to photo resolution. However, the accuracy of reconstructed surface normals highly relies

[1]This work was conducted while Haotian Zhuang was doing internship as a visiting student at Peking university; [2]Bohan Yu, Zhuofeng Wang, and Boxin Shi are with State Key Laboratory of Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University; Boxin Shi is also with PKU-AI[2] Robotics Joint Lab of Embodied AI.
[*] Equal contribution; [†] Corresponding authors.

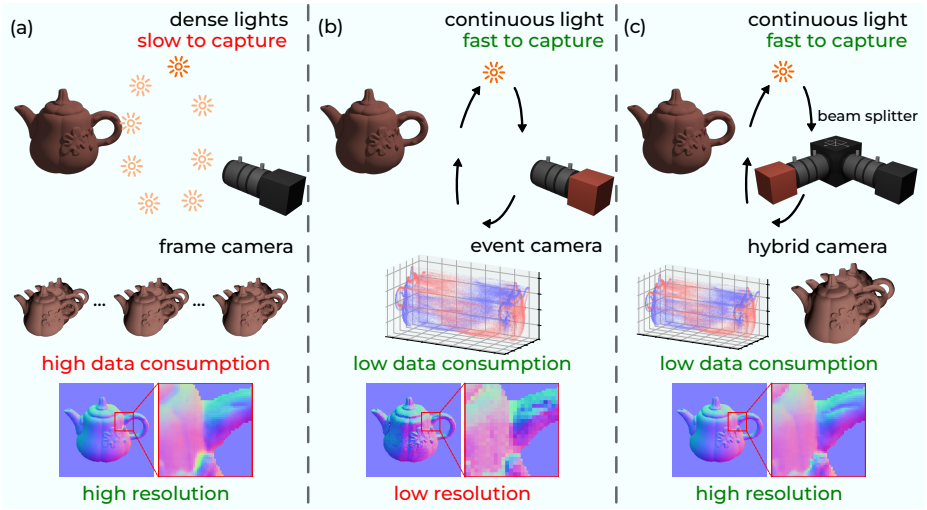

Fig. 1: Comparison between FramePS, EventPS, and the proposed DualResPS. (a) FramePS estimates the surface normal by analyzing images of an object illuminated from multiple directions, which is not only time-consuming but also demands substantial bandwidth for processing. (b) In contrast, EventPS estimates the surface normal by analyzing the events triggered by a continuously rotating light source. It is a rapid and highly bandwidth-friendly approach, but the spatial resolution is limited by event sensors. (c) Our DualResPS utilizes a beam splitter to form a frame-event hybrid camera, which simultaneously takes frames and events from a shared view as input. The rapid speed and low data bandwidth are maintained since we capture sparse and long-exposure frames under continuous light, and the strategy tailored for complementary frames and events enable high-resolution normal estimation.

on the number of input images. Increasing the number of input images also increases the capture time consumption, data burden, and reduces time-sensitive application of photometric stereo. And this limitation is the fundamental property of frame-based cameras. [44]

With the advancement of event cameras, researchers utilize their high dynamic range, high temporal resolution, and efficient event representation in many computer vision tasks [5, 25, 37]. In photometric stereo, EventPS [44] has shown the capability of reconstructing high-quality surface normals in less time and with less data consumption. Unfortunately, the complexity of event trigger circuitry limits the potential reconstruction resolution. By the time of writing, photometric stereo methods using frame-based cameras are proven to work for 4K or higher resolution. However, due to the complexity of the per-pixel event-triggering circuit, the maximum resolution of commercially available event cameras is only 720p. This limitation of sensor resolution undermines the detail reconstruction ability of photometric stereo using event cameras.

Recently, researchers have built hybrid cameras to combine the advantages of both frame-based cameras and event-based cameras. Many event-frame fusion methods are proposed to improve the dynamic range [5, 41], motion robustness [2, 46], frame rate [34], low-light performance [19, 20], etc. The success of hybrid camera systems in these inspires us to explore the possibility of combining the high spatial resolution of frame-based cameras and the high temporal resolution of event cameras. Unfortunately, it is not a trivial task to combine information from hybrid cameras due to the following challenges: i) **Frame degradation** challenge: Under conventional directional light, frame-based pixel intensity is eliminated under shadow or specular high light; ii) **Event resolution** challenge: The spatial resolution of the event camera is lower than surface normal, attached shadow, and specular highlight frequencies; and iii) **Temporal misalignment** challenge: Different from the unified global exposure characteristic of frame-based cameras, event cameras produce asynchronous event streams unaligned with frame capture time.

In this paper, we propose **DualResPS**, a novel photometric stereo pipeline combining two cameras efficiently and predicting high-resolution, high-quality surface normal maps. To solve frame degradation, we propose a capture strategy that simultaneously acquires the events with three long-exposure frames under continuously rotating light. Then, non-ideal shadow and specular highlights in frame images are compensated by events to achieve contamination-free images. To produce high-resolution normal maps aligned to frame images, we explicitly assign low-frequency parameters (cast shadow, material) to the event camera or add smoothness constraints. In this way, the DoF (degree of freedom) of high-frequency parts (albedo, normal) matches the high-resolution frame input. To efficiently utilize asynchronously generated events, we propose a neural inverse rendering pipeline equipped with an importance sampling strategy. More computational power is assigned to instantaneous high-intensity areas (for frames) and high-fluctuation areas (for events). Our proposed pipeline is validated on both semi-real datasets derived from DiLiGenT [29] and DiLiGenT-Pi [35], and real data captured from a hybrid camera prototype. The results indicate that DualResPS surpasses its frame-based counterpart with less than 20% data bandwidth and capturing time.

To summarize, our contributions are as follows:

- We propose DualResPS, which formulates that the surface normals can be efficiently estimated from sparse yet long-exposure frames combined with event stream. It overcomes the limited spatial resolution of event camera without sacrificing much data consumption.
- We propose DualResPS in a self-supervised neural inverse rendering manner, integrating frames with events by efficient sampling to handle shadow effect and specular reflection.
- We build up a validation platform with a frame-event hybrid camera pair, showcasing that the proposed DualResPS estimates high-resolution surface normals with low data burden.

2 Related Work

Photometric stereo estimates surface normal from images captured from a single view under different lighting [38], with both optimization-based and learning-based methods proposed to enhance the performance. Since comprehensive survey can be found in [10,30], this section focuses on photometric stereo by neural inverse rendering and supervised methods, as well as recent advances in event-based photometry and hybrid imaging systems.

Supervised photometric stereo. Supervised photometric stereo relies on deep neural networks. At the early stage, CNN-based architecture is predominant, with learning techniques including observation maps [7], pooling [1], graph-convolution [42], and attention-based weight [11]. Drop-out mechanism [28] and occlusion layer [17] are proposed to handle shadow, while connection table [17] and observation map interpolation [45] are proposed to overcome sparse lighting. Transformer-based architecture [6, 8, 9, 22] becomes the main stream in recent years with the rising of universal photometric stereo [8]. In order to better extract features and decode information, global-attention [14] inspired by VGGT [36] are leveraged, generating high-fidelity results even with sparse input. Moreover, architectures including scale-invariant encoder [9], multi-scale strategies [6], and wavelet branch [14] are explored for details preservation.

Photometric stereo by neural inverse rendering. Neural inverse rendering is introduced to optimization-based photometric stereo to handle non-ideal factors in image formulation. Taniai *et al.* [32] leverages self-supervised convolutional neural networks (CNN) taking images as input, for its capacity to encode the feature of complicated reflectance in photometric stereo. The work is expanded by Kaya *et al.* [12] to deal with inter-reflection implicitly in the context of uncalibrated photometric stereo (UPS). Inspired by Neural Radiance Fields (NeRF) [23] featuring multilayer-perceptron (MLP) and differentiable volume rendering, some work utilizes the coordinately-based and view-dependent 3D representation to solve photometric stereo [40] or even multi-view photometric stereo [39]. By contrast, Li *et al.* [15] proposes view-independent neural surface representation and view-dependent neural specular bases to better fit the characteristic of photometric stereo. Meanwhile, cast shadow is handled by ray marching on depth MLP. Successive work [16, 18] in UPS proposes Spherical Gaussian (SG) specular bases and differentiable shadow handling, achieving comparable results. Our work is also implemented with neural inverse rendering, and it takes events as input in addition to frames.

Event-based photometry and hybrid imaging. The development of neuromorphic sensors fuels the enthusiasm of relavant research including event-based photometry. EventPS [44] pioneers to solve photometric stereo using an event camera, where a fast-rotating light source is set in the far field. It achieves real-time surface normal recovery by the elegant optimization algorithm, as well as learning-based versions for better quality. PS-EIP [13] reconstructs event interval profile with coarse masks for shadow and specularity, and focuses on the ideal part for normal estimation. With the designed scanning pattern of ring light, EventPSR [43] can further estimate surface normal along with metallic

and roughness efficiently. In EventUPS [21], the expansion to UPS and ambient light is realized. All these works demonstrate the superiority of event cameras over frame cameras in speed, data bandwidth, and dynamic range, which is crucial in the application to dynamic objects. Meanwhile, hybrid imaging systems emerge as an approach to exploit the complementary of frames and events in spatial and temporal resolution, noise control, dynamic range and so on. There are multiple works on tasks such as super-resolution [4,24,37], deblurring [25,31,33], motion magnification [2], denoising [37], and high dynamic range (HDR) [5,41]. Regarding photometric stereo, EFPS-Net [27] proposes event fusion observation maps to overcome the issue of dynamic range. Our work pays special attention to the high temporal resolution of events, and compresses the data rate with severely sparse input of frames.

3 Proposed Method

3.1 Problem formulation

Image formulation. The image formation starts from the steady radiometric rendering equation, where the outgoing radiance L_o at world coordinate $\boldsymbol{x}$ in direction $\boldsymbol{\omega}_\mathrm{o}$ is as follows without emissive objects:

$$L_\mathrm{o}(\boldsymbol{x}, \boldsymbol{\omega}_\mathrm{o}) = \int_\Omega f_\mathrm{r}(\boldsymbol{x}, \boldsymbol{\omega}_\mathrm{i}, \boldsymbol{\omega}_\mathrm{o}) L_\mathrm{i}(\boldsymbol{x}, \boldsymbol{\omega}_\mathrm{i}) [\boldsymbol{\omega}_\mathrm{i} \cdot \boldsymbol{n}(\boldsymbol{x})]_+ d\boldsymbol{\omega}_\mathrm{i}. \tag{1}$$

In this equation, $f_\mathrm{r}(\boldsymbol{x}, \boldsymbol{\omega}_\mathrm{i}, \boldsymbol{\omega}_\mathrm{o})$ denotes the bidirectional reflectance distribution function (BRDF); L_i is the incident radiance; $\boldsymbol{\omega}_\mathrm{i}$ is the incident direction over the upper hemisphere Ω; $\boldsymbol{n}(\boldsymbol{x})$ is the unit surface normal; $[\cdot]_+ \doteq \max(0, \cdot)$ clamps negative contributions. The irradiance received by a camera sensor is proportional to the outgoing radiance of a ray from a specific direction. If the radiance field is dynamic, then the irradiance from $\boldsymbol{x}$ is also a function of time:

$$I(\boldsymbol{x}, t) \propto L_\mathrm{o}(\boldsymbol{x}, \boldsymbol{\omega}_\mathrm{o}(\boldsymbol{x}), t). \tag{2}$$

In our work, we mainly focus on the incident radiance carried by the direct illumination of punctual light sources. As the light source sweeps, the incident direction of direct illumination $\boldsymbol{l}$ evolves over time t. By combining Eq. (1) with Eq. (2) and extracting the indirect (inter-reflection) term $I_\mathrm{ind}(\boldsymbol{x}, t)$, the irradiance $I(\boldsymbol{x}, t)$ becomes:

$$I(\boldsymbol{x}, t) = e f_\mathrm{r}(\boldsymbol{x}, \boldsymbol{l}(t)) V(\boldsymbol{x}, \boldsymbol{l}(t)) [\boldsymbol{l}(t) \cdot \boldsymbol{n}(\boldsymbol{x})]_+ + I_\mathrm{ind}(\boldsymbol{x}, t), \tag{3}$$

where:

$$f_\mathrm{r}(\boldsymbol{x}, \boldsymbol{l}(t)) = \rho_\mathrm{d}(\boldsymbol{x}) + \rho_\mathrm{s}(\boldsymbol{x}, \boldsymbol{l}(t)). \tag{4}$$

We express BRDF f_r as diffuse albedo ρ_d plus specular reflectance ρ_s, and substitute $L_\mathrm{i}(\boldsymbol{x}, t)$ with the light intensity e and the visibility of direct illumination $V(\boldsymbol{x}, \boldsymbol{l}(t))$. If the punctual lights can be seen as far-field and spatially-uniform, i.e., directional lights, e is a constant and $\boldsymbol{l} = \boldsymbol{l}(t)$. When a sensor is

exposed to light between t_0 and t_1, the instantaneous irradiance contributions is integrated. We use $A(\boldsymbol{x}, t_0, t_1)$ to represent the temporally accumulated pixel intensity at $\boldsymbol{x}$, which is radiant exposure in precise term:

$$A(\boldsymbol{x}, t_0, t_1) = \int_{t_0}^{t_1} I(\boldsymbol{x}, \tau) d\tau. \tag{5}$$

Event camera model. An event $(\boldsymbol{x}, t, p)$ is triggered when the change in logarithmic irradiance at $\boldsymbol{x}$ reaches a contrast threshold C, where $\boldsymbol{x}$ is the pixel location, t is the timestamp, and $p \in \{+1, -1\}$ is the polarity which represents the increase or decrease of irradiance.

An event camera outputs an asynchronous stream $\mathcal{E}$. We further define the event stream at the pixel $\boldsymbol{x}$ as $\mathcal{E}_{\boldsymbol{x}} = \{(\boldsymbol{x}, t_{\boldsymbol{x}}^{(k)}, p_{\boldsymbol{x}}^{(k)})\}$, where the bracketed superscript (k) indicates the chronological order. Given two temporally adjacent events at $\boldsymbol{x}$, the relationship holds:

$$\log(I(\boldsymbol{x}, t_{\boldsymbol{x}}^{(k)}) + \epsilon) = \log(I(\boldsymbol{x}, t_{\boldsymbol{x}}^{(k-1)} + \eta) + \epsilon) + p_{\boldsymbol{x}}^{(k)} C, \tag{6}$$

$$\text{i.e.,} \ \frac{I(\boldsymbol{x}, t_{\boldsymbol{x}}^{(k)}) + \epsilon}{I(\boldsymbol{x}, t_{\boldsymbol{x}}^{(k-1)} + \eta) + \epsilon} = \exp(p_{\boldsymbol{x}}^{(k)} C), \tag{7}$$

where $\epsilon > 0$ is a small offset to avoid numerical issues, and η approximates the refractory time. For simplicity, we hereafter use $\boldsymbol{x}$ to denote the world coordinate that corresponds to the pixel location, and it is evident that events provide a constraint on the instantaneous irradiance $I(\boldsymbol{x}, t)$.

Dual-resolution photometric stereo. To efficiently recover shape and reflectance, especially surface normal $\boldsymbol{n}(\boldsymbol{x})$ and diffuse albedo $\rho_{\mathrm{d}}(\boldsymbol{x})$ with known intensities e and directions $\boldsymbol{l}(t)$ of directional lights, we design a dual-resolution configuration to obtain information about irradiance $I(\boldsymbol{x}, t)$. This is an efficient variant of photometric stereo that enables low data rate and high capture speed.

3.2 Dual-resolution methodology

Dual-resolution configuration. In our configuration, the illumination varies with time period T per round. We decouple the measurements of irradiance $I(\boldsymbol{x}, t)$ into two streams: high-spatial-resolution frames $\{A_{\mathrm{f}}^{(j)}(\boldsymbol{x})\}$ with a global long-exposure, and low-spatial-resolution event samples $\mathcal{E}$ that capture rapid radiometric changes of $I(\boldsymbol{x}, t_{\boldsymbol{x}}^{(k)})$ asynchronously. The captures of two streams share a single-view and start from t_0:

$$A_{\mathrm{f}}^{(j)}(\boldsymbol{x}) = A\left(\boldsymbol{x},\, t_{j-1},\, t_j\right),\ \boldsymbol{x} \in \mathcal{P}_{\mathrm{f}},\ t_j = t_0 + \frac{T}{3} j, \tag{8}$$

$$\mathcal{E} = \bigcup \mathcal{E}_{\boldsymbol{x}},\ \boldsymbol{x} \in \mathcal{P}_{\mathrm{e}}, \tag{9}$$

where $|\mathcal{P}_{\mathrm{e}}| < |\mathcal{P}_{\mathrm{f}}|$. In practice, $\mathcal{P}_{\mathrm{e}}$ denotes the world coordinates that project onto the grid of event sensors, while $\mathcal{P}_{\mathrm{f}}$ refers to the world coordinates that project

onto the grid of frame sensors. We will refer to their elements as super pixels and frame pixels respectively. Within T, three frames are captured to estimate a normal map, which is a challenging condition for classical photometric stereo. Despite the low temporal resolution of frames, high-temporal-resolution events are concurrently recorded. Our strategy leverages the complementarity of sparse frames and the event stream by navigating the inherent trade-off between spatial fidelity and temporal density under a fixed data budget. It is a spatiotemporally dual-resolution downsampling scheme for efficiently sampling irradiance $I(\boldsymbol{x}, t)$.

Equivalent directional light. Our goal is to decode information from frames $\{A_{\mathrm{f}}^{(j)}(\boldsymbol{x})\}$ and events $\mathcal{E}$. Within T, we gather three frames $(j) = 1, 2, 3$:

$$A_{\mathrm{f}}^{(j)}(\boldsymbol{x}) = \int_{t_{j-1}}^{t_j} e f_{\mathrm{r}}(\boldsymbol{x}, \boldsymbol{l}(t)) V(\boldsymbol{x}, \boldsymbol{l}(t)) [\boldsymbol{l}(t) \cdot \boldsymbol{n}(\boldsymbol{x})]_+ dt + A_{\mathrm{f,ind}}^{(j)}(\boldsymbol{x}). \tag{10}$$

Similar to the representation in [3], we initially consider an ideal image formulation model without some terms:

$$A_{\mathrm{f}}^{(j)}(\boldsymbol{x}) = \int_{t_{j-1}}^{t_j} e\rho_{\mathrm{d}}(\boldsymbol{x}) \boldsymbol{l}(t) \cdot \boldsymbol{n}(\boldsymbol{x}) dt \tag{11}$$

$$= e\rho_{\mathrm{d}}(\boldsymbol{x}) \left(\int_{t_{j-1}}^{t_j} \boldsymbol{l}(t) dt \right) \cdot \boldsymbol{n}(\boldsymbol{x}) \doteq e\rho_{\mathrm{d}}(\boldsymbol{x}) \boldsymbol{l}^{*(j)} \cdot \boldsymbol{n}(\boldsymbol{x}). \tag{12}$$

This equality holds by linearity of both the time integral and the dot product, leading to the definition of the lighting vector $\boldsymbol{l}^*$. Intuitively, $\boldsymbol{l}^*$ acts as an equivalent directional light that aggregates the swept light within the exposure window. These observations connect to classical photometric stereo: if each exposure admits an equivalent directional light, then under a Lambertian reflectance and shadow-free observations, three linearly independent lighting directions suffice to recover the per-pixel surface normal and diffuse albedo.

Non-ideal factors. However, deviations from the assumptions mentioned above break the linear model. There are non-ideal factors including attached shadow, cast shadow, and specular reflection. Taking these into account, we need to propose a modification of the theory and discuss each factor in turn.

Factor 1: Cast shadow. The $V(\boldsymbol{x}, \boldsymbol{l}(t))$ term is an indication of cast shadow, which is determined by lighting and geometry. Another feature of cast shadow is that the pixel intensity will change drastically when getting into or out of it. Owing to the logarithmic sensation, brightness change capture, and high temporal resolution of the event stream, it is possible to infer the instantaneous $V(\boldsymbol{x}, \boldsymbol{l}(t))$ from events $\mathcal{E}$. We therefore define the mask for cast shadow $m_{\mathrm{c}}(\boldsymbol{x}, t)$ as a binary mask under an opaque object assumption. Assuming the mask aligns with the ground truth, substituting $V(\boldsymbol{x}, \boldsymbol{l}(t))$ with $m_{\mathrm{c}}(\boldsymbol{x}, t)$ yields:

$$A_{\mathrm{f}}^{(j)}(\boldsymbol{x}) = \int_{t_{j-1}}^{t_j} e\rho_{\mathrm{d}}(\boldsymbol{x}) V(\boldsymbol{x}, \boldsymbol{l}(t)) \boldsymbol{l}(t) \cdot \boldsymbol{n}(\boldsymbol{x}) dt \tag{13}$$

$$= e\rho_{\mathrm{d}}(\boldsymbol{x}) \left(\int_{t_{j-1}}^{t_j} m_{\mathrm{c}}(\boldsymbol{x}, t) \boldsymbol{l}(t) dt \right) \cdot \boldsymbol{n}(\boldsymbol{x}) = e\rho_{\mathrm{d}}(\boldsymbol{x}) \boldsymbol{l}^{*(j)}(\boldsymbol{x}) \cdot \boldsymbol{n}(\boldsymbol{x}). \tag{14}$$

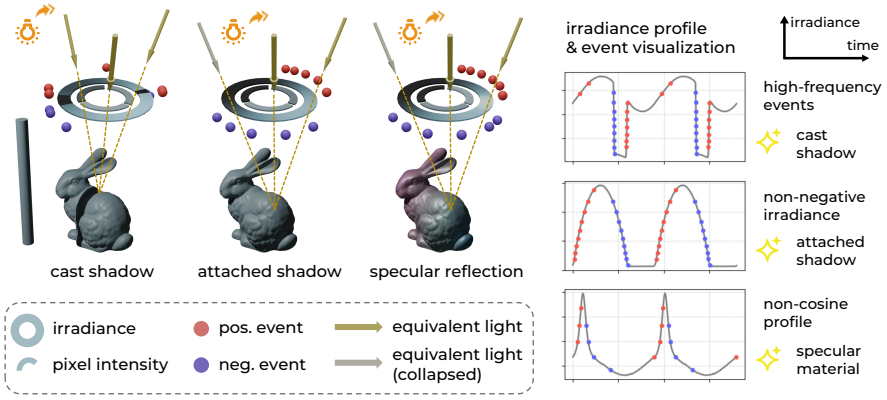

Fig. 2: Visualization of equivalent lights under dual-resolution configuration in consideration of non-ideal factors. On the right, we plot the irradiance profile and triggered events for intuitive understanding of the methodology. The horizontal axis denotes time while the vertical axis denotes irradiance.

In contrast with the ideal condition, the constant vector $\boldsymbol{l}^*$ becomes a vector function $\boldsymbol{l}^*(\boldsymbol{x})$. Surface normal $\boldsymbol{n}(\boldsymbol{x})$ can still be determined by the equivalent lights for each pixel, unless $m_{\mathrm{c}}(\boldsymbol{x}, t) \equiv 0$ for some (j) and the lighting vector $\boldsymbol{l}^* = 0$ (collapses). We will later analyze these cases by the number of effective lighting vectors $\boldsymbol{l}^* \neq 0$.

Factor 2: Attached shadow. $[\cdot]_+$ for $\boldsymbol{l}(t) \cdot \boldsymbol{n}(\boldsymbol{x})$ originates from attached shadow. Its effect can be substituted with a mask for attached shadow:

$$m_{\mathrm{a}}(\boldsymbol{x}, t; \boldsymbol{n}) = \begin{cases} 1, & \text{if } \boldsymbol{l}(t) \cdot \boldsymbol{n}(\boldsymbol{x}) > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

Note that the mask for attached shadow $m_{\mathrm{a}}(\boldsymbol{x}, t; \boldsymbol{n})$ is a parametric function of the surface normal. Therefore, the only free parameter at each pixel is still the normal $\boldsymbol{n}(\boldsymbol{x})$. This dependence motivates the following reformulation (replacing $m_{\mathrm{c}} m_{\mathrm{a}}$ with m for simplicity):

$$A_{\mathrm{f}}^{(j)}(\boldsymbol{x}) = \int_{t_{j-1}}^{t_j} e\rho_{\mathrm{d}}(\boldsymbol{x})[\boldsymbol{l}(t) \cdot \boldsymbol{n}(\boldsymbol{x})]_+ dt = e\rho_{\mathrm{d}}(\boldsymbol{x}) \left(\int_{t_{j-1}}^{t_j} m(\boldsymbol{x}, t; \boldsymbol{n}) \boldsymbol{l}(t) dt \right) \cdot \boldsymbol{n}(\boldsymbol{x}). \tag{16}$$

Similarly, we obtain the vector function $\boldsymbol{l}^*(\boldsymbol{x}; \boldsymbol{n})$ in dependence on surface normal. In our lighting configuration, there is no more than one lighting vector of three that collapses due to attached shadow. The determination of the normal $\boldsymbol{n}(\boldsymbol{x})$ remains possible along with the equivalent lights, despite subtle changes in adaptation. Details are shown in the supplementary material, Sec. 1.1.

Factor 3: Specular reflection. The specular reflectance ρ_s introduces additional specular reflection to the image formation model. It gives:

$$A_\mathrm{f}^{(j)}(\boldsymbol{x}) = \int_{t_{j-1}}^{t_j} e(\rho_\mathrm{d}(\boldsymbol{x}) + \rho_\mathrm{s}(\boldsymbol{x}, \boldsymbol{l}(t))V(\boldsymbol{x}, \boldsymbol{l}(t))[\boldsymbol{l}(t) \cdot \boldsymbol{n}(\boldsymbol{x})]_+ dt \tag{17}$$

$$= e\rho_\mathrm{d}(\boldsymbol{x})\boldsymbol{l}^{*(j)}(\boldsymbol{x}; \boldsymbol{n}) \cdot \boldsymbol{n}(\boldsymbol{x}) + A_\mathrm{f,s}^{(j)}(\boldsymbol{x}; \boldsymbol{n}, \boldsymbol{\beta}), \tag{18}$$

Here we define $\boldsymbol{\beta}(\boldsymbol{x})$ to be the parameter of specular reflectance ρ_s other than normal $\boldsymbol{n}$. It is impossible to distinguish the portion of specular reflection in pixel intensity with three frames alone, but we can fully exploit the information in events $\mathcal{E}$ at the adjacent super pixels to constrain $I(\boldsymbol{x}, t_{\boldsymbol{x}}^{(k)}; \boldsymbol{n}, \boldsymbol{\beta})$. Since the temporal resolution of events is high, the event stream $\mathcal{E}_{\boldsymbol{x}}$ in a period T contains sufficient information to estimate the parameter $\boldsymbol{\beta}(\boldsymbol{x})$ of ρ_s with the smooth nature of material. By removing the effect of the specular term, it is possible to jointly estimate surface normal $\boldsymbol{n}(\boldsymbol{x})$ at each frame pixel.

Analysis by lighting. The recovery of surface normal $\boldsymbol{n}(\boldsymbol{x})$ and diffuse albedo $\rho_\mathrm{d}(\boldsymbol{x})$ at each frame pixel depends on the number of effective lighting vectors. Ideally, the spatial fidelity is aligned with high-spatial-resoltuion frames $\{A_\mathrm{f}^{(j)}(\boldsymbol{x})\}$. When frames are not sufficient, the compromise is to interpolate from super pixels or adjacent frame pixels with more effective lighting vectors by the smooth nature of surface normal and diffuse albedo. Specially, when exactly two effective lighting vectors exist, there is one unconstrained degree of freedom in normal that is estimated by smooth interpolation. Detailed analysis can be found in the supplementary material, Sec. 1.2.

3.3 DualResPS by neural inverse rendering

The dual-resolution methodology theoretically minimizes ambiguity under sparse input. To handle the complex analysis and non-convex optimization, we propose a unified preprocessing and neural inverse rendering pipeline that produces a high-fidelity solution.

Neural representation. We model the surface parameters by a coordinate-based MLP $M(\boldsymbol{x}; \Theta)$:

$$\boldsymbol{n}(\boldsymbol{x}), \rho_\mathrm{d}(\boldsymbol{x}), \boldsymbol{\beta}(\boldsymbol{x}) = M(\boldsymbol{x}; \Theta), \tag{19}$$

where Θ is the trainable parameters of the MLP, and $\boldsymbol{\beta}(\boldsymbol{x})$ is the intrinsic parameters of specular reflectance. The implementation details are provided in the supplementary material, Sec. 2.4. This neural representation bridges dual-resolution input based on the nature of surface smoothness, implicitly reducing a plausible solution without explicit interpolation. Moreover, the compact representation naturally induces priors like material sparsity without hampering high-frequency details. The output of the MLP is passed forward to the physically based model of differentiable rendering, together with preprocessed data including the mask for cast shadow $m_\mathrm{c}(\boldsymbol{x}, t)$.

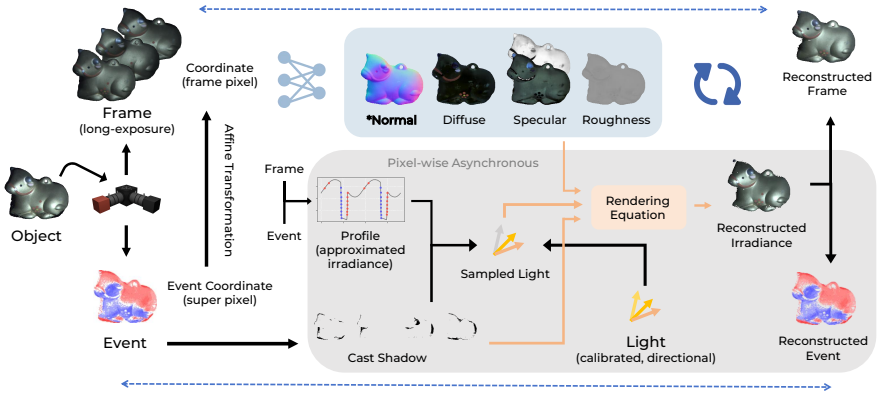

Fig. 3: An overview of our preprocessing and neural inverse rendering pipeline. The cast shadow and sampled light are calculated in the preprocessing stage, while the coordinately based neural network iteratively outputs intrinsic properties including normal estimation, and updates parameters through rendering losses. The neural rendering model is pixel-wise asynchronous.

Shadow handling. By analyzing discrete samples $\mathcal{E}_{\boldsymbol{x}}$ we can infer continuous cast shadow $m_{\mathrm{c}}(\boldsymbol{x}, t)$ at each super pixel, and assign the masks to nearby frame pixels for the low-frequency characteristics of cast shadow. Thanks to the integral process of frame exposure and the high temporal resolution of events, errors of instantaneous cast shadow hardly affect the reconstructed pixel intensity. Therefore, we detect cast shadow merely from events by detecting:

$$f_{\boldsymbol{x}}^{(k)} = \frac{p_{\boldsymbol{x}}^{(k)} C}{t_{\boldsymbol{x}}^{(k)} - t_{\boldsymbol{x}}^{(k-1)}}, \tag{20}$$

where $f_{\boldsymbol{x}}^{(k)}$ is a modification of event interval profile [13] for higher temporal resolution. We also observe a phenomenon in which cast shadow sometimes overlaps with attached shadow, softening one of the two shadow edges. The detection threshold for that edge is therefore iteratively reduced to locate mixed cast shadows. The implementation details are provided in the supplementary material, Sec. 2.1.

Reflectance model. Our dual-resolution configuration requires to sample specular reflectance $\rho_{\mathrm{s}}(\boldsymbol{x}, \boldsymbol{l}, \boldsymbol{v})$ for different light directions for every frame in inverse rendering, so it is important to save computing resource regarding reflectance model. Meanwhile, the sparse inputs make it necessary to strengthen reflectance priors. Thus, our modeling for reflectance is physically based BRDF via neural representation: $\rho_{\mathrm{s}} = \rho_{\mathrm{s}}(\boldsymbol{x}, \boldsymbol{l}(t); \boldsymbol{n}, F_0, r)$, where F_0 is the constant fresnel term with RGB channels, and r is roughness. Material parameters of specular reflectance is described as $\boldsymbol{\beta}(\boldsymbol{x}) = [F_0(\boldsymbol{x}), r(\boldsymbol{x})]$. The implementation details are provided in the supplementary material, Sec. 2.3. The replacement of the conven-

tional parameter metallic is to avoid the misleading of incorrect white balance. F_0 serves as the specular color in the specular-glossiness workflow and boost the generalization of reflectance model for inverse rendering.

Asynchronous importance sampling. The integral of long-exposure frames is calculated by discrete sampling of $I(\boldsymbol{x}, t)$:

$$A_{\mathrm{f}}^{(j)}(\boldsymbol{x}) = \int_{t_{j-1}}^{t_j} I(\boldsymbol{x}, t)dt \approx \sum\nolimits_{t \in \mathcal{T}_{\boldsymbol{x}}^{(j)}} I(\boldsymbol{x}, t)\, \Delta t, \tag{21}$$

$$\tilde{I}(\boldsymbol{x}, t) = e\tilde{f}_{\mathrm{r}}(\boldsymbol{x}, \boldsymbol{l}(t)) m_{\mathrm{c}}(\boldsymbol{x}, t)[\boldsymbol{l}(t) \cdot \tilde{\boldsymbol{n}}(\boldsymbol{x})]_+, \tag{22}$$

where $\tilde{I}$, $\tilde{f}_{\mathrm{r}}$ and $\tilde{\boldsymbol{n}}$ are the reconstructed irradiance, reflectance and surface normal by neural inverse rendering. Meanwhile, it is necessary to sample $I(\boldsymbol{x}, t)$ at event timestamps that is not in cast shadow to reconstruct events. We propose a pixel-wise asynchronous importance sampling scheme, where the sample points are determined and fixed during preprocessing, and sampled iteratively during neural rendering. We perform importance sampling according to the approximated pixel-wise irradiance $\{\hat{I}_{\boldsymbol{x}}^{(k)}\}$ (see Sec. 2.2 of the supplementary material). Specifically, we uniformly sample between adjacent events so that the integral between adjacent sampled timestamps nears a global constant. The number of samples between adjacent events becomes:

$$N_{\boldsymbol{x}}^{(k)} = \left[\mathrm{round}\left(\sigma \sqrt{\Delta \hat{I}_{\boldsymbol{x}}^{(k)}\, \Delta t_{\boldsymbol{x}}^{(k)}} \right) - 1 \right]_+, \tag{23}$$

where σ is the hyperparameter of sampling density. This scheme improves the efficiency of inverse rendering. The intermediate approximated irradiance also provides an added benefit: since events in dark regions are sensitive to noise and unmodeled inter-reflections, they are downweighted by reduced confidence accordingly.

Training. Finally, our loss functions for the dual-resolution data stream become:

$$\mathcal{L}_{\mathrm{f}} = \frac{1}{N_{\mathrm{f}}} \sum |\tilde{A}_{\mathrm{f}}^{(j)}(\boldsymbol{x}) - A_{\mathrm{f}}^{(j)}(\boldsymbol{x})|, \tag{24}$$

$$\mathcal{L}_{\mathrm{e}} = \frac{1}{N_{\mathrm{e}}} \sum w_{\boldsymbol{x}}^{(k)} \left| \log\left(\frac{\tilde{I}(\boldsymbol{x}, t_{\boldsymbol{x}}^{(k)}) + \epsilon}{\tilde{I}(\boldsymbol{x}, t_{\boldsymbol{x}}^{(k-1)}) + \epsilon} \, \frac{1}{\alpha_{\boldsymbol{x}}^{(k)} \exp(p_{\boldsymbol{x}}^{(k)} C)} \right) \right|^2, \tag{25}$$

$$\mathcal{L} = \lambda_{\mathrm{f}} \mathcal{L}_{\mathrm{f}} + \lambda_{\mathrm{e}} \mathcal{L}_{\mathrm{e}} + \lambda_{\mathrm{tv}} \mathcal{L}_{\mathrm{tv}} \tag{26}$$

where $\mathcal{L}_{\mathrm{f}}$ and $\mathcal{L}_{\mathrm{e}}$ denote the frame and event loss terms, respectively; $\tilde{A}_{\mathrm{f}}$ and $\tilde{I}$ are the reconstructed pixel intensity and irradiance; N_{f} and N_{e} are the total numbers of frame pixels and event samples; $\alpha_{\boldsymbol{x}}^{(k)}$ is the scaling factor for refractory compensation; $w_{\boldsymbol{x}}^{(k)}$ is the confidence weight for each event; and λ_{f}, λ_{e}, λ_{tv} are balancing coefficients. The total loss $\mathcal{L}$ is combined with the warm-up (total variation) loss term $\mathcal{L}_{\mathrm{tv}}$ [15] in addition at the early stage of training.

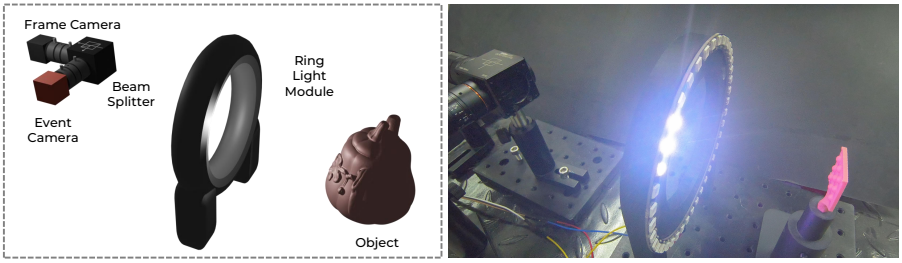

Fig. 4: Components of the proposed validation platform. The rendered image (left) aligns with the viewpoint of the photographed real platform (right). The hybrid camera system includes a beam splitter, a frame camera and an event camera. The illuminance device, a ring light module, is oriented towards the target object.

4 Experiments

4.1 Implementation Details

Algorithms implementation. The input of our framework is sparse (exactly 3 in a lighting cycle) RGB frames and low-spatial-resolution mono event stream, while the target resolution of the output is aligned with input frames. The observation data is assumed to be captured in a dark room. During the neural inverse rendering stage, we iterate in total 1200 epochs when optimizing the network. Our method is implemented in PyTorch and is running on a A6000 GPU. The inference (i.e., training) time for DiLiGenT [29] semi-real dataset at 256×256 target resolution is about 1 min per object. For DiLiGenT-Pi [35] semi-real dataset at 608×608 target resolution, it takes about 6 mins per object.
Validation platform. To verify the performance of the algorithms on real-world objects, we design a high-speed illumination and capturing validation platform as shown in Fig. 4. We use a Prophesee EVK4 HD camera (with an IMX 636 sensor) to capture event, and a global-exposure HIK MV-CA050-12UC camera to capture frames. The spatial resolution is 2048×2448 for frame and 720×1280 for event. To meet the assumption of a single shared view, we utilize a beam splitter to form a frame-event camera pair. The illumination device and hybrid camera pair are controlled by Arduino and camera SDK, with external trigger signals sent to two cameras for synchronization. We set the frame rate as 10 fps to avoid underexposure, which is up to 60 fps given brighter illumination.

4.2 Datasets

Semi-real datasets. We conduct quantitative comparisons on two publicly available real-world datasets: DiLiGenT [29] and DiLiGenT-Pi [35]. DiLiGenT-Pi targets near-planar surfaces with rich details across four material categories: metallic, specular, rough, and translucent. Ground-truth normal maps are obtained via 3D scanning. Following EventPS, we select HDR images from dense

Table 1: Quantitative comparison (MAE in degrees) on the DiLiGenT [29] semi-real dataset. Our self-supervised method outperforms the frame-based method LL22a [15] (best self-supervised method in DiLiGenT [29]) with 18.1% data rate, and matches recent supervised FramePS methods. The last column shows the percentage of data rate compared to one frame. The last row indicates the number of events per round for each data.

Lighting	Self-supervised	Input	Method	Ball	Buddha	Cat	Cow	Goblet	Harvest	Pot1	Pot2	Reader	Average	Data Rate
Cal	×	4 frames	Uni-MS-PS	3.76	7.65	5.53	5.23	6.25	14.71	5.36	4.99	9.05	6.95	17.4%
Uni	×	4 frames	SDM-UniPS	1.83	7.91	6.60	5.73	6.53	13.29	7.28	5.72	9.29	7.13	17.4%
Uni	×	3 frames	LINO	2.41	7.58	6.79	6.45	7.97	11.33	6.72	5.69	10.37	7.26	13.0%
Cal	✓	23 frames	LL22a	2.26	9.88	5.16	5.29	**6.44**	18.54	5.40	5.39	10.95	7.70	100%
Cal	✓	3f + Ev	**DualResPS**	**2.08**	**8.64**	**4.86**	**4.47**	6.88	**18.01**	**5.38**	**5.31**	**10.59**	**7.36**	18.1%
Cal	✓	Events	EventPS	11.00	19.96	12.10	26.14	18.94	38.29	11.54	15.79	26.18	19.99	5.0%
		# Events		148 k	110 k	130 k	146 k	64 k	100 k	85 k	112 k	129 k	114 k	/

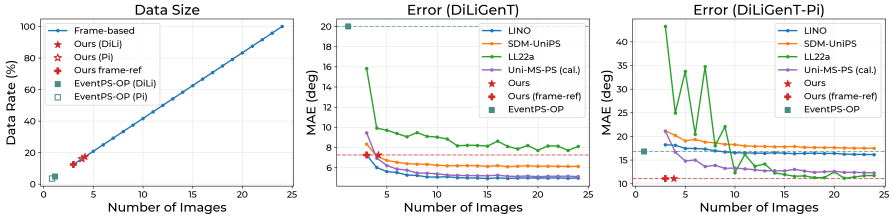

Fig. 5: Comparison of data rate and MAE on DiLiGenT [29] and DiLiGenT-Pi [35] semi-real dataset. On the left, the data rate for FramePS increases linearly as the number of images increases. In contrast, DualResPS has a low and constant data rate paramount to about 4 frame images, similar to EventPS. We also mark the data size of DualResPS's frame input for reference. On the right, the MAE for FramePS decreases with more images, while DualResPS achieves comparable or superior accuracy.

trajectory circles, convert them to event streams using ESIM [26], and aggregate them into long-exposure frames. Images are 2×2 downsampled to match the RGB sensor resolution, and further downsampled to simulate low-resolution events. The resulting semi-real datasets contain LDR-clamped frames, fused long-exposure frames, and event streams.

Prototype data. To validate the performance of the proposed DualResPS method, we fabricate 3 objects and capture a real dataset with ground truth normal maps from 3D model. The real dataset covers simple geometry with strong shadow effect (STAR), shapes with moderate details (LION), and near-planar surfaces with rich details (KING). Each object is captured using our validation platform under few ambient light.

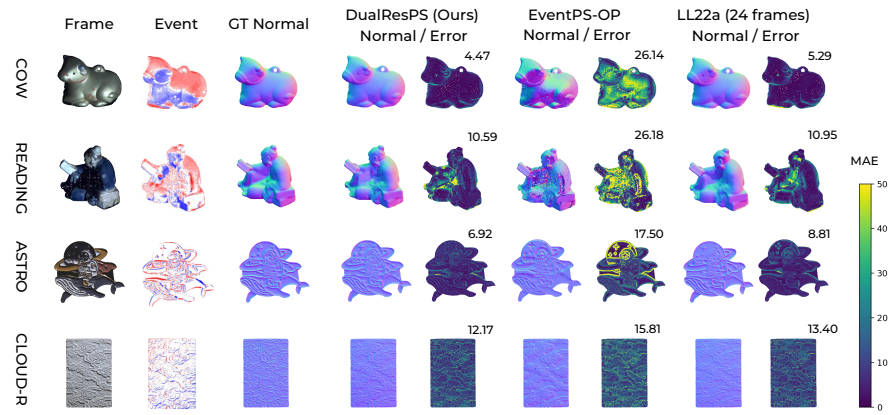

Fig. 6: Results on the DiLiGenT [29] and DiLiGenT-Pi [35] semi-real datasets. Our method produces reconstructions superior to EventPS and the frame-based LL22a [15]. The optimization based on event loss often struggles to converge and causes artifacts as described in EventPSR [43], but our method effectively mitigates the effects.

4.3 Quantitative comparison

We conduct a quantitative comparison of the proposed DualResPS with representative FramePS and EventPS methods on the DiLiGenT [29] and DiLiGenT-Pi [35] semi-real datasets. To compute the data rate required by the event input and frame input, we assume that the event streams employ 16-bit Prophesee EVT 3.06 format, and frame images are captured as 24-bit RGB images. For FramePS algorithms, we uniformly select images from 36 light directions of the trajectory boundary in DiLiGenT dataset, and from 24 light directions of the trajectory circle in DiLiGenT-Pi dataset. For EventPS algorithms, different numbers of low-spatial-resolution events are generated for each scene. The Mean Angular Error (MAE) and data rate comparison are shown in Tab. 1.

As shown in Fig. 5, FramePS shows a linear increase in data rate as the number of input images increases, accompanied by a decrease in normal MAE. In contrast, EventPS and DualResPS both maintain constant data rate and MAE. The self-supervised frame-based counterpart LL22a [15] converges to our accuracy only with significantly more data, while EventPS remains inferior. Supervised FramePS methods match DualResPS on DiLiGenT due to strong pretrained priors, but DiLiGenT-Pi results confirm that DualResPS achieves lower MAE with better data efficiency on diverse shapes and materials. Representatively, four examples in Fig. 6 show that DualResPS produces evenly distributed errors across objects.

4.4 Evaluation on prototype

We evaluate DualResPS on real captured data, with results shown in Fig. 7. EventPS results are warped via the estimated affine transformation between

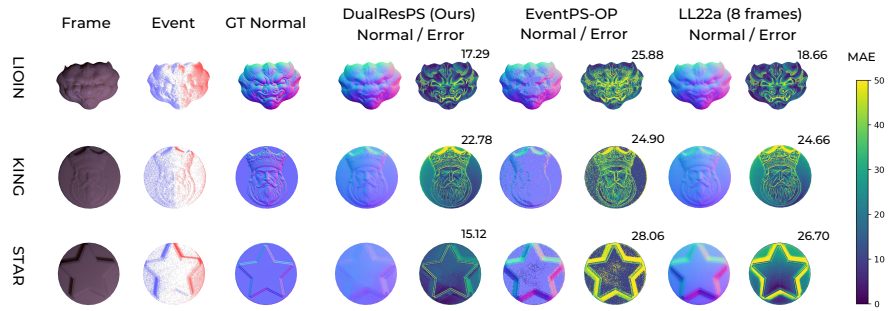

Fig. 7: Results on the prototype data. Our method produces reconstructions superior to EventPS-OP and the frame-based LL22a [15].

frame and event views for fair comparison. Due to different event camera parameters, the event data bandwidth is relatively low (under 0.5 frames). DualResPS achieves superior MAE over both EventPS and LL22a [15] (8 frames), demonstrating the complementarity of frames and events for photometric stereo. Notably, cast shadow effects on STAR are substantially reduced in our results, confirming the effectiveness of shadow handling. Residual errors in high-frequency regions are partly attributed to coarse frame-event alignment and fabrication errors in the 3D-printed ground truth.

5 Conclusion and Discussion

In this paper, we propose DualResPS, a novel photometric stereo model using a hybrid frame-event camera. Our method utilizes the complementary advantages of frame and event cameras in spatial-temporal resolutions, which shows great potential to extend the efficiency and quality for rapid sensing of the surface normal and other intrinsic properties.

Limitations and future work. Our model does not explicitly handle inter-reflections, limiting its applicability to scene-level reconstruction. The self-supervised architecture is sensitive to event camera parameters and requires careful calibration. As hybrid camera systems mature, it is worth exploring an end-to-end model that leverages learned data priors.

Acknowledgements

This work is supported by National Key R&D Program of China (Grant No. 2025YFA1016700), National Natural Science Foundation of China (Grant No. 624B2006), Beijing Natural Science Foundation (Grant No. L233024), Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park (Grant No. Z241100003524012), and Beijing High Innovation Plan (Contract No. 202604841443).

References

1. Chen, G., Han, K., Wong, K.Y.K.: Ps-fcn: A flexible learning framework for photometric stereo. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–18 (2018)
2. Chen, Y., Guo, S., Yu, F., Zhang, F., Gu, J., Xue, T.: Event-based motion magnification. In: European Conference on Computer Vision. pp. 428–444. Springer (2024)
3. Guo, H., Mo, Z., Shi, B., Lu, F., Yeung, S.K., Tan, P., Matsushita, Y.: Patch-based uncalibrated photometric stereo under natural illumination. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(11), 7809–7823 (2021)
4. Han, J., Yang, Y., Zhou, C., Xu, C., Shi, B.: Evintsr-net: Event guided multiple latent frames reconstruction and super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4882–4891 (2021)
5. Han, J., Zhou, C., Duan, P., Tang, Y., Xu, C., Xu, C., Huang, T., Shi, B.: Neuromorphic camera guided high dynamic range imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1730–1739 (2020)
6. Hardy, C., Quéau, Y., Tschumperlé, D.: Uni ms-ps: A multi-scale encoder-decoder transformer for universal photometric stereo. Computer Vision and Image Understanding **248**, 104093 (2024)
7. Ikehata, S.: Cnn-ps: Cnn-based photometric stereo for general non-convex surfaces. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–18 (2018)
8. Ikehata, S.: Universal photometric stereo network using global lighting contexts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12591–12600 (2022)
9. Ikehata, S.: Scalable, detailed and mask-free universal photometric stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13198–13207 (2023)
10. Ju, Y., Lam, K.M., Xie, W., Zhou, H., Dong, J., Shi, B.: Deep learning methods for calibrated photometric stereo and beyond. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(11), 7154–7172 (2024)
11. Ju, Y., Shi, B., Jian, M., Qi, L., Dong, J., Lam, K.M.: Normattention-psn: A high-frequency region enhanced photometric stereo network with normalized attention. International Journal of Computer Vision **130**(12), 3014–3034 (2022)
12. Kaya, B., Kumar, S., Oliveira, C., Ferrari, V., Van Gool, L.: Uncalibrated neural inverse rendering for photometric stereo of general surfaces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3804–3814 (2021)
13. Kitazawa, K., Aoto, T., Ikehata, S., Takatani, T.: Ps-eip: Robust photometric stereo based on event interval profile. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6241–6251 (2025)
14. Li, H., Chen, H., Ye, C., Chen, Z., Li, B., Xu, S., Guo, X., Liu, X., Wang, Y., Zhang, B., et al.: Light of normals: Unified feature representation for universal photometric stereo. arXiv preprint arXiv:2506.18882 (2025)
15. Li, J., Li, H.: Neural reflectance for shape recovery with shadow handling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16221–16230 (2022)

16. Li, J., Li, H.: Self-calibrating photometric stereo by neural inverse rendering. In: European Conference on Computer Vision. pp. 166–183. Springer (2022)
17. Li, J., Robles-Kelly, A., You, S., Matsushita, Y.: Learning to minify photometric stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7568–7576 (2019)
18. Li, Z., Zheng, Q., Shi, B., Pan, G., Jiang, X.: Dani-net: Uncalibrated photometric stereo by differentiable shadow handling, anisotropic reflectance modeling, and neural inverse rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8381–8391 (2023)
19. Liang, G., Chen, K., Li, H., Lu, Y., Wang, L.: Towards robust event-guided low-light image enhancement: a large-scale real-world event-image dataset and novel approach. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23–33 (2024)
20. Liang, J., Yang, Y., Li, B., Duan, P., Xu, Y., Shi, B.: Coherent event guided low-light video enhancement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10615–10625 (2023)
21. Liang, J., Yu, B., Yang, S., Zhuang, H., Ren, J., Duan, P., Shi, B.: Eventups: Uncalibrated photometric stereo using an event camera. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7516–7525 (2025)
22. Liu, H., Yan, Y., Song, K., Yu, H.: Sps-net: Self-attention photometric stereo network. IEEE Transactions on Instrumentation and Measurement **70**, 1–13 (2020)
23. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European Conference on Computer Vision. pp. 405–421. Springer (2020)
24. Mostafavi, M., Nam, Y., Choi, J., Yoon, K.J.: E2sri: Learning to super-resolve intensity images from events. IEEE transactions on pattern analysis and machine intelligence **44**(10), 6890–6909 (2021)
25. Pan, L., Scheerlinck, C., Yu, X., Hartley, R., Liu, M., Dai, Y.: Bringing a blurry frame alive at high frame-rate with an event camera. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6820–6829 (2019)
26. Rebecq, H., Gehrig, D., Scaramuzza, D.: Esim: an open event camera simulator. In: Conference on robot learning. pp. 969–982. PMLR (2018)
27. Ryoo, W., Nam, G., Hyun, J.S., Kim, S.: Event fusion photometric stereo network. Neural Networks **167**, 141–158 (2023)
28. Santo, H., Samejima, M., Sugano, Y., Shi, B., Matsushita, Y.: Deep photometric stereo network. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 501–509 (2017)
29. Shi, B., Mo, Z., Wu, Z., Duan, D., Yeung, S.K., Tan, P.: A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. IEEE TPAMI (2019)
30. Shi, B., Wu, Z., Mo, Z., Duan, D., Yeung, S.K., Tan, P.: A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3707–3716 (2016)
31. Sun, L., Sakaridis, C., Liang, J., Sun, P., Cao, J., Zhang, K., Jiang, Q., Wang, K., Van Gool, L.: Event-based frame interpolation with ad-hoc deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18043–18052 (2023)

32. Taniai, T., Maehara, T.: Neural inverse rendering for general reflectance photometric stereo. In: International Conference on Machine Learning. pp. 4857–4866. PMLR (2018)
33. Teng, M., Zhou, C., Lou, H., Shi, B.: Nest: Neural event stack for event-based image enhancement. In: European Conference on Computer Vision. pp. 660–676. Springer (2022)
34. Tulyakov, S., Gehrig, D., Georgoulis, S., Erbach, J., Gehrig, M., Li, Y., Scaramuzza, D.: Time lens: Event-based video frame interpolation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16155–16164 (2021)
35. Wang, F., Ren, J., Guo, H., Ren, M., Shi, B.: Diligent-pi: Photometric stereo for planar surfaces with rich details-benchmark dataset and beyond. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9477–9487 (2023)
36. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
37. Wang, Z.W., Duan, P., Cossairt, O., Katsaggelos, A., Huang, T., Shi, B.: Joint filtering of intensity images and neuromorphic events for high-resolution noise-robust imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1609–1619 (2020)
38. Woodham, R.J.: Photometric method for determining surface orientation from multiple images. Optical engineering **19**(1), 139–144 (1980)
39. Yang, W., Chen, G., Chen, C., Chen, Z., Wong, K.Y.K.: Ps-nerf: Neural inverse rendering for multi-view photometric stereo. In: European conference on computer vision. pp. 266–284. Springer (2022)
40. Yang, W., Chen, G., Chen, C., Chen, Z., Wong, K.Y.K.: S3-nerf: Neural reflectance field from shading and shadow under a single viewpoint. Advances in Neural Information Processing Systems **35**, 1568–1582 (2022)
41. Yang, Y., Han, J., Liang, J., Sato, I., Shi, B.: Learning event guided high dynamic range video reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13924–13934 (2023)
42. Yao, Z., Li, K., Fu, Y., Hu, H., Shi, B.: Gps-net: Graph-based photometric stereo network. Advances in Neural Information Processing Systems **33**, 10306–10316 (2020)
43. Yu, B., Han, J., Shi, B., Sato, I.: Eventpsr: Surface normal and reflectance estimation from photometric stereo using an event camera. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 11427–11436 (2025)
44. Yu, B., Ren, J., Han, J., Wang, F., Liang, J., Shi, B.: Eventps: Real-time photometric stereo using an event camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9602–9611 (2024)
45. Zheng, Q., Jia, Y., Shi, B., Jiang, X., Duan, L.Y., Kot, A.C.: Spline-net: Sparse photometric stereo through lighting interpolation and normal estimation networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8549–8558 (2019)
46. Zhou, X., Duan, P., Xiaokaiti, Y., Xu, C., Shi, B.: Event-based visual vibrometry. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24666–24676 (2025)

DualResPS: Dual-Resolution Photometric Stereo Using a Frame-Event Hybrid Camera

Supplementary Material

Haotian Zhuang[1*], Bohan Yu[2*], Zhuofeng Wang[2], and Boxin Shi[2†]

[1] Tsinghua University, Beijing, China
zht23@tsinghua.edu.cn
[2] Peking University, Beijing, China
{ybh1998,shiboxin}@pku.edu.cn, wangzf2003@stu.pku.edu.cn

1 Detailed Analysis on Lighting

1.1 Adaptation to attached shadow

We have mentioned that the equivalent light becomes a vector function $\boldsymbol{l}^*(\boldsymbol{x};\boldsymbol{n})$ in dependence on surface normal $\boldsymbol{n}(\boldsymbol{x})$ in consideration of attached shadow, and the normal estimation needs subtle adaptation. Intuitively, the normal determination remains possible since the degree of freedom (DoF) does not increase. Fig. 1 shows that the optimal solution does exist by numerical simulation and visualization. The spheres are normalized in the camera space to represent the direction of lighting and normal vector. (a)-(e) are under directional lights, while (f)-(j) are under continuous lights. (f) shows a discrete sample of the continuous lights: Every sample is a directional light, and their integral becomes a directional light like (a) if the attached shadow for each sample is ignored. The deepened colors of yellow dots indicate three exposure windows of frame cameras $(j) = 1, 2, 3$, and the exposure windows of (a)(f) exactly match so that the equivalent lights without attached shadow are the same. In practice, the lighting configurations of (a) and (f) are proposed by conventional frame-based photometric stereo and DualResPS respectively.

The following subfigures in two rows show the conditions of normal vectors on a sphere under these two categories of lights. (b)(g) show the shading term $i^{(1)} = \boldsymbol{l}^{*(1)} \cdot \boldsymbol{n}$ (normalized) and attached shadow region $i^{(1)} = 0$. (c)(h) show the overlay of attached shadow region from three equivalent lights. (d)(i) show the shading ratio $\frac{i^{(2)}}{i^{(2)}+i^{(3)}}$ and its contours $\{0.5, 0.6, 0.7, 0.8, 0.9\}$. (e)(j) show that contours from two pairs of equivalent lights suffice to determine the surface normal. It can be seen that the adaptation is mainly in the regions close to attached

[1]This work was conducted while Haotian Zhuang was doing internship as a visiting student at Peking university; [2]Bohan Yu, Zhuofeng Wang, and Boxin Shi are with State Key Laboratory of Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University; Boxin Shi is also with PKU-AI[2] Robotics Joint Lab of Embodied AI.
[*] Equal contribution; [†] Corresponding authors.

Table 1: Analysis by the number of effective lighting vectors at each frame pixel.

# $\boldsymbol{l}^*$	unconstr. DoF of $\boldsymbol{n}$	$\boldsymbol{n}$ & $\boldsymbol{\rho}_\mathrm{d}$ constraint	Interpolation of $\boldsymbol{n}$	Interpolation of $\boldsymbol{\rho}_\mathrm{d}$
3	0	Exist	No need	
2	1		From super pixels / frame pixels	From frame pixels
1	2			
0		Inexist		

shadow. Moreover, the attached shadow region is smaller under continous light, which means that continuous lights provide more information to estimate surface normal. As for the regions where less than three effective lighting vectors exist, information from super pixels contributes. (k)(l) focus on a frame pixel A constrained on a contour with only two equivalent lights, and depict the normal estimation by smooth interpolation from adjacent super pixels. The coarse normal estimation at super pixels from events actually provide a constraint on its adjacent frame pixels.

1.2 Analysis by cases

In this section, we continue to analyze the recovery of surface normal $\boldsymbol{n}(\boldsymbol{x})$ and diffuse albedo $\rho_\mathrm{d}(\boldsymbol{x})$ at each frame pixel by the number of effective (omitted below) lighting vectors as shown in Tab. 1. Note that events fail to constrain diffuse albedo $\rho_\mathrm{d}(\boldsymbol{x})$ due to the albedo invariance mentioned in EventPS [7], so the direct interpolation of $\rho_\mathrm{d}(\boldsymbol{x})$ is only from adjacent frame pixels. Meanwhile, there is a constraint between surface normal $\boldsymbol{n}(\boldsymbol{x})$ and diffuse albedo $\rho_\mathrm{d}(\boldsymbol{x})$ at each frame pixel from frames as long as a lighting vector exists.

Case 1: Three lighting vectors. With three equivalent lights, surface normal $\boldsymbol{n}(\boldsymbol{x})$ and diffuse albedo $\rho_\mathrm{d}(\boldsymbol{x})$ at each frame pixel can be determined from three frames $\{A_\mathrm{f}^{(j)}(\boldsymbol{x})\}$ within a period T.

Case 2: Two lighting vectors. With two equivalent lights, surface normal $\boldsymbol{n}(\boldsymbol{x})$ at each frame pixel can be constrained to a trajectory from two of frames $\{A_\mathrm{f}^{(j)}(\boldsymbol{x})\}$ within a period T. As shown in Fig. 1 (k)(l), the remaining free degree of normal $\boldsymbol{n}(\boldsymbol{x})$ and diffuse albedo $\rho_\mathrm{d}(\boldsymbol{x})$ at each frame pixel can be interpolated from events $\mathcal{E}$.

Case 3: One lighting vector. With only one equivalent light and corresponding frame $A_\mathrm{f}^{(j)}(\boldsymbol{x})$, there is no clue for high-fidelity surface normal $\boldsymbol{n}(\boldsymbol{x})$. Estimation of $\boldsymbol{n}(\boldsymbol{x})$ and diffuse albedo $\rho_\mathrm{d}(\boldsymbol{x})$ at each frame pixel is from events $\mathcal{E}$ at the adjacent super pixels.

Case 4: No lighting vector. Without lighting vectors, there is barely effective information in frames $\{A_\mathrm{f}^{(j)}(\boldsymbol{x})\}$. Therefore, estimation of surface normal $\boldsymbol{n}(\boldsymbol{x})$ at each frame pixel depends on events $\mathcal{E}$. Separately, diffuse albedo $\rho_\mathrm{d}(\boldsymbol{x})$ can only be estimated by interpolating estimated $\rho_\mathrm{d}(\boldsymbol{x}')$ where lighting vector exists.

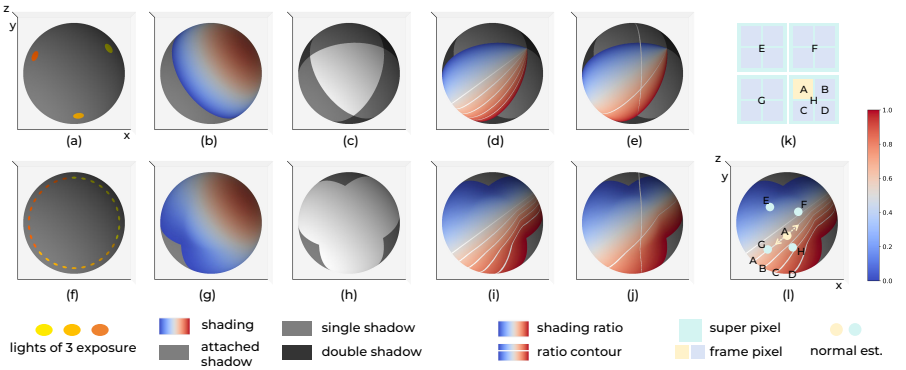

Fig. 1: Visualization of normal estimation and DoF constraint regarding equivalent lights. The spheres are normalized in the camera space to represent the direction of lighting and normal vector. (a)-(e) are under directional lights, while (f)-(j) are under continuous lights. (f) shows a discrete sample of the continuous lights: Every sample is a directional light, and their integral becomes a directional light like (a) if the attached shadow for each sample is ignored. The following subfigures in two rows show the conditions of normal vectors on a sphere under these two categories of lights. (k)(l) focus on a frame pixel A with only two equivalent lights, and depict the normal estimation by smooth interpolation.

2 More Algorithms Details

2.1 Shadow handling

We provide Algorithm 1 to clarify the shadow detection logic. Let $\{(t_{\boldsymbol{x}}^{(k)}, f_{\boldsymbol{x}}^{(k)})\}$ be the modified event interval profile samples at super-pixel $\boldsymbol{x}$. A spike $f_{\boldsymbol{x}}^{(k)}{>}\theta$ indicates a shadow *exit*, while a spike $f_{\boldsymbol{x}}^{(k)}{<}{-}\theta$ indicates a shadow *entry*. The procedure searches matched entry–exit boundaries over one rotation period T. We denote the directed cyclic gap between two timestamps by $\Delta_\rho(t_1, t_2) = \rho(t_1 - t_2) \bmod T$. Accordingly, $\textsc{Opp}_\rho(i^*, \theta, \theta_{\min}, \gamma, D) = \arg\max_j \Delta_\rho(t_{\boldsymbol{x}}^{(i^*)}, t_{\boldsymbol{x}}^{(j)})$ s.t. $0 < \Delta_\rho(t_{\boldsymbol{x}}^{(i^*)}, t_{\boldsymbol{x}}^{(j)}) < D$ and $\rho f_{\boldsymbol{x}}^{(j)} < -\theta'$ for some decayed threshold $\theta' \in [\theta_{\min}, \theta]$, where θ' is reduced by factor γ on failure; similarly, $\textsc{Conf}_\rho(j^*, \theta, D) = \arg\max_k \Delta_\rho(t_{\boldsymbol{x}}^{(j^*)}, t_{\boldsymbol{x}}^{(k)})$ s.t. $0 < \Delta_\rho(t_{\boldsymbol{x}}^{(j^*)}, t_{\boldsymbol{x}}^{(k)}) < D$ and $\rho f_{\boldsymbol{x}}^{(k)} > \theta$. When $\rho{=}+1$, Opp *backward* searches farthest *entry* iteratively for some decayed threshold, while Conf *forward* searches farthest *exit*. The detected $\mathcal{S}_{\boldsymbol{x}}$ is then converted to the cast shadow mask $m_{\mathrm{c}}(\boldsymbol{x}, t)$.

2.2 Irradiance approximation

In the preprocessing stage of our pipeline, the approximated irradiance is calculated by:

$$\hat{I}_{\boldsymbol{x}}^{(k)} = s_{\boldsymbol{x}}^{(k)} \left(\prod_{\kappa=1}^{k} \alpha_{\boldsymbol{x}}^{(\kappa)} \right) \exp\left(\sum_{\kappa=1}^{k} p_{\boldsymbol{x}}^{(\kappa)} C \right), \tag{1}$$

Algorithm 1 FindShadowBoundaries (per super pixel $\boldsymbol{x}$)

Require: Event interval profile $\{(t_{\boldsymbol{x}}^{(k)}, f_{\boldsymbol{x}}^{(k)})\}_k$; threshold θ, minimum threshold $\theta_{\min}$, decay γ, max duration D
Ensure: Shadow interval set $\mathcal{S}_{\boldsymbol{x}}$
Initialize unused samples and $\mathcal{S}_{\boldsymbol{x}} \leftarrow \emptyset$
for $\rho \in \{+1, -1\}$ **do** ▷ +1: exit-first, −1: entry-first
  for each unused anchor i^* with $\rho f_{\boldsymbol{x}}^{(i^*)} > \theta$ **do**
    $j^* \leftarrow \textsc{Opp}_\rho(i^*, \theta, \theta_{\min}, \gamma, D)$; if fail, continue
    $k^* \leftarrow \textsc{Conf}_\rho(j^*, \theta, D)$; if fail, continue
    $(t_l, t_h) \leftarrow (t_{\boldsymbol{x}}^{(j^*)}, t_{\boldsymbol{x}}^{(k^*)})$ if $\rho = +1$, else $(t_{\boldsymbol{x}}^{(k^*)}, t_{\boldsymbol{x}}^{(j^*)})$
    $\mathcal{S}_{\boldsymbol{x}} \leftarrow \mathcal{S}_{\boldsymbol{x}} \cup \{(t_l, t_h)\}$ and mark the interval as used
return $\mathcal{S}_{\boldsymbol{x}}$

where $s_{\boldsymbol{x}}^{(k)}$ is a piecewise-constant scaling factor with respect to (k). It changes value only when $f_{\boldsymbol{x}}^{(k)}$ is sufficiently large, which indicates an abrupt change in irradiance and a potential failure of refractory time compensation. Within one period T, $s_{\boldsymbol{x}}^{(k)}$ is restricted to take no more than three distinct values. This design is consistent with the three exposure intervals of frames, so that the ratios constructed from the corresponding three temporal integrals can still be constrained by the measured frame pixel intensities $\{A_{\mathrm{f}}^{(j)}(\boldsymbol{x})\}$ without introducing an underdetermined system.

2.3 Reflectance model

Our modeling for reflectance is physically based BRDF via neural representation:

$$\rho_{\mathrm{s}}(\boldsymbol{x}, \boldsymbol{l}, \boldsymbol{v}) = \frac{D(\boldsymbol{h})F(\boldsymbol{l}, \boldsymbol{h})G(\boldsymbol{l}, \boldsymbol{v}, \boldsymbol{h})}{4(\boldsymbol{n} \cdot \boldsymbol{l})(\boldsymbol{n} \cdot \boldsymbol{v})}, \tag{2}$$

where $\boldsymbol{v}$ is the viewing (outgoing) direction, $\boldsymbol{h} = \frac{\boldsymbol{l}+\boldsymbol{v}}{\|\boldsymbol{l}+\boldsymbol{v}\|}$ is the half-vector between the incident and outgoing directions. $D(\boldsymbol{h})$ is the normal distribution function (NDF); $F(\boldsymbol{l}, \boldsymbol{h})$ is the Fresnel term; $G(\boldsymbol{l}, \boldsymbol{v}, \boldsymbol{h})$ is the geometry term. We approximate the fresnel term as an angle-independent trainable parameter F_0 with RGB channels. In our implementation, $D(\boldsymbol{h})$ and $G(\boldsymbol{l}, \boldsymbol{v}, \boldsymbol{h})$ follow the definition in [1]. They are parameterized by roughness $r(\boldsymbol{x})$ in addition to $\boldsymbol{n}(\boldsymbol{x})$.

2.4 Network architecture

Following a coordinate-based multi-layer perceptron (MLP) design similar to LL22a [3], we map each pixel coordinate $\boldsymbol{x}$ to the surface normal $\boldsymbol{n}(\boldsymbol{x})$, diffuse albedo $\rho_{\mathrm{d}}(\boldsymbol{x})$, and specular material parameters $\boldsymbol{\beta}(\boldsymbol{x})$. The input coordinate is first transformed by positional encoding, and then processed by an 8-layer MLP with width 256, ReLU activations, and a skip connection at the 4-th layer. The normal head is attached directly to the output of this shared backbone. In

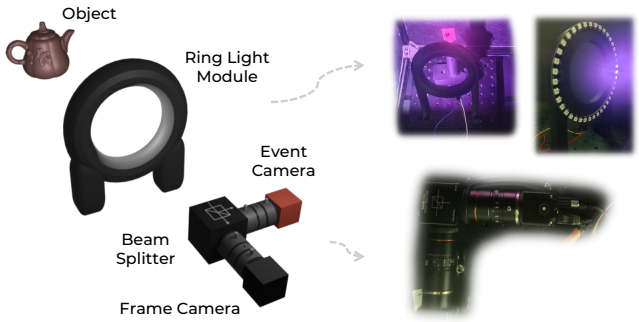

Fig. 2: A diagram of our prototype. The hybrid camera system includes a beam splitter, a frame camera and an event camera. The illuminance device, a ring light module, is oriented towards the target object. The validation platform is ideally in a dark room.

contrast, the material head starts from the same backbone feature but further applies an additional feature projection layer and four fully connected layers of width 128, before predicting $\rho_{\mathrm{d}}(\boldsymbol{x})$ and $\boldsymbol{\beta}(\boldsymbol{x})$. In our setting, $\boldsymbol{\beta}(\boldsymbol{x})$ consists of specular color $F_0(\boldsymbol{x})$ and roughness $r(\boldsymbol{x})$, where roughness $r(\boldsymbol{x})$ is activated by a sigmoid function.

3 Additional Prototype Details

Figure 2 is a diagram of our prototype containing the hybrid camera system and illumination device. For synchronization, an Arduino board controls the illumination device and sends trigger signals to both cameras, so the timing accuracy is limited by the controller precision. Before capturing a target object, we use a non-planar LED array as the calibration reference of affine transformation between event and frame cameras. Afterwards, the hybrid camera pair remains fixed and a ring light module similar to [2] replaces as the illumination device of the target object. The ring light module can simulate continuous light through a smooth transition of intensity between LEDs, so the bottleneck of our capture speed is the frame rate limit and exposure requirement. At the same frame rate, it saves capture time compared with frame-based methods by decreasing the required number of frames. We create a video to visualize the light pattern of the illumination device, and the process of capturing frames and events. The video is available as a separate file named "DualResPS_supp_video.mp4".

In theory, mapping pixel coordinates across the two views with a single affine transformation requires special optical-center conditions; in practice, pixel-level alignment is difficult to achieve exactly. When the non-planar LED array is working, both cameras capture the LED feature points, from which we estimate the affine mapping and reprojection error, then iteratively adjust the camera poses and recompute until the error is sufficiently small. We also support manual annotation of corresponding feature points when automatic detection is unreliable.

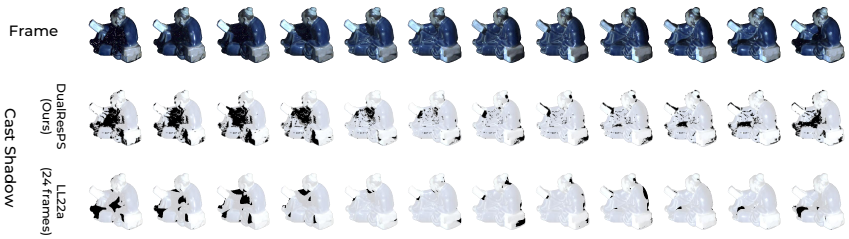

Fig. 3: Qualitative comparison of cast shadow estimation on READING from the DiLiGenT [5] semi-real dataset. DualResPS detects broader shadow regions faithfully, and the isolated wrong pixels are partially corrected by the smoothness regularization from the MLP.

We believe more mature hybrid-camera hardware will further improve alignment robustness and promote practical deployment.

4 Further Analysis and Ablation Studies

4.1 Effectiveness of shadow handling

Previous methods like LL22a [3] estimate cast shadow in inverse rendering by time-consuming ray casting, where depth map or three dimensional representation is required. In contrast, DualResPS proposes a light-weight strategy to detect cast shadow from events. As shown in Fig. 3, we visualize the asynchronous cast shadow mask (including mixed attached shadow) as a synchrounous sequence, and the precise timestamps of shadow edges at each pixel are indicated by the black depth. The results from LL22a [3] in the third row demonstrate that the cast shadow estimated from geometry clues is unstable. DualResPS detects broader shadow regions faithfully, and the isolated wrong pixels are partially corrected by the smoothness regularization from the MLP, which leads to better and stable normal estimation. Ablation study in Tab. 2 is a proof to the effectiveness of the proposed shadow handling strategy.

4.2 Effectiveness of reflectance model

Figure 4 shows the reconstructed maps including surface normal $\boldsymbol{n}(\boldsymbol{x})$, diffuse albedo $\rho_{\mathrm{d}}(\boldsymbol{x})$, specular color $F_0(\boldsymbol{x})$, and roughness $r(\boldsymbol{x})$. Metallic is caculated from diffuse albedo and specular color, and base color can also be calculated for workflow compatibility. The reconstructed diffuse albedo and roughness are locally smooth and contain sharp details. The specular color complementary to diffuse albedo is influenced by biased white balance and light intensities calibration, as well as the clamped dynamic range. Besides, for non-metallic surface and regions where specular reflection is not observed, the estimation of specular color is intrinsically not well constrained, for different color decompositions merge into

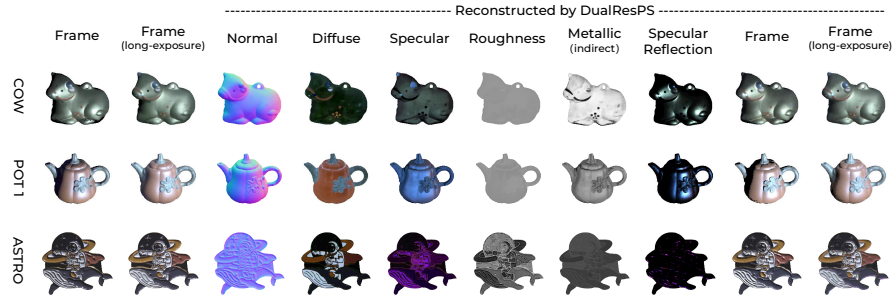

Fig. 4: Qualitative results of neural inverse rendering on the DiLiGenT [5] and DiLiGenT-Pi [6] semi-real datasets. The reconstructed maps include surface normal $\boldsymbol{n}(\boldsymbol{x})$, diffuse albedo $\rho_{\mathrm{d}}(\boldsymbol{x})$, specular color $F_0(\boldsymbol{x})$, and roughness $r(\boldsymbol{x})$. Metallic is cacculated from diffuse albedo and specular color. The specular color complementary to diffuse albedo is influenced by biased white balance and light intensities calibration.

Table 2: Ablation study (MAE in degrees) on the DiLiGenT [5] semi-real dataset. **Bold** and underline denote the best and second-best results. w/o m_{c}: without cast shadow mask; w/ MLP ρ_{s}: replacing physically based BRDF with MLP specular bases in [3]; w/ SG ρ_{s}: replacing with Spherical Gaussian specualr bases in [4]; w/ metallic ρ_{s}: replacing F_0 with metallic parameter.

Method	Ball	Buddha	Cat	Cow	Goblet	Harvest	Pot1	Pot2	Reader	Average
Proposed	2.08	**8.64**	**4.86**	4.47	6.88	**18.01**	5.38	**5.31**	**10.59**	**7.36**
w/o m_{c}	**1.59**	11.64	6.37	5.12	7.69	18.52	7.22	6.25	14.65	8.79
w/ MLP ρ_{s}	2.16	9.59	5.48	**4.41**	**6.44**	30.74	6.23	5.93	12.91	9.32
w/ SG ρ_{s}	1.83	9.19	5.02	4.88	6.93	19.69	**5.08**	5.63	11.74	7.78
w/ metallic ρ_{s}	2.04	9.54	5.18	9.38	8.09	43.68	6.02	7.46	11.95	11.49

similar observations. By comparing to different specular reflectance representations in Tab. 2, we demonstrate the effectiveness of the proposed reflectance model.

4.3 Efficiency of frame samples

In traditional photometric stereo, the more light conditions or frame samples give the better the performance. DualResPS hard-codes the number of input frames to 3 based on theoretical observations, and we conduct experiments on a modified generic version that support flexible frames input to confirm its efficiency. As shown in Tab. 3, accuracy improves and saturates as the number of input frames K in a rotation period increases. This confirms the efficiency of 3 sparse frame samples.

Table 3: Evaluations (MAE in degrees) on the DiLiGenT [5] semi-real dataset regarding the number of input frames. The accuracy improves and saturates as the number of input frames K in a rotation period increases.

K	3	4	5	6	9	18	24
MAE	7.41	7.21	7.08	6.82	6.78	6.79	6.67

Table 4: Evaluations (MAE in degrees) on the DiLiGenT [5] semi-real dataset regarding the robustness to motion blur. The accuracy drops steadily when the total displacement increases from 0 to 8 pixels.

Disp.	0	1	2	3	4	6	8
MAE	7.36	7.59	8.43	9.33	10.16	11.87	13.45

Table 5: Quantitative comparison (MAE in degrees) on the DiLiGenT [5] semi-real dataset. We show the evaluation results on 24 frames for LL22a [3], although its best performance is on 23 frames and is shown in main text. 24 frames are also set as 100% data rate rather than 23 frames.

Lighting	Self-supervised	Input	Method	Ball	Buddha	Cat	Cow	Goblet	Harvest	Pot1	Pot2	Reading	Average	Data Rate
Cal	×	3 frames	Uni-MS-PS	4.15	9.96	7.50	8.32	7.83	19.86	8.50	6.21	12.68	9.45	12.5%
		4 frames		3.76	7.65	5.53	5.23	6.25	14.71	5.36	4.99	9.05	6.95	16.7%
		8 frames		2.46	5.78	3.87	4.39	4.29	12.09	4.41	4.13	7.79	5.47	33.3%
		24 frames		2.36	5.40	3.50	4.05	4.14	11.19	4.12	3.98	7.24	5.11	100%
Uni	×	3 frames	SDM-UniPS	1.88	8.99	7.60	6.11	9.03	15.51	8.43	5.80	11.65	8.33	12.5%
		4 frames		1.83	7.91	6.60	5.73	6.53	13.29	7.28	5.72	9.29	7.13	16.7%
		8 frames		1.75	6.95	6.03	5.03	6.57	11.37	6.51	4.53	8.40	6.35	33.3%
		24 frames		1.74	6.89	5.67	4.71	6.74	10.90	6.26	4.47	7.71	6.12	100%
Uni	×	3 frames	LINO	2.41	7.58	6.79	6.45	7.97	11.33	6.72	5.69	10.37	7.26	12.5%
		4 frames		1.70	6.61	6.25	5.69	5.72	9.56	5.68	4.97	7.85	6.00	16.7%
		8 frames		1.73	6.04	4.91	4.79	4.96	8.43	4.65	4.66	6.86	5.23	33.3%
		24 frames		1.65	5.94	4.32	4.42	4.80	7.89	4.31	4.59	6.77	4.97	100%
Cal	✓	24 frames	LL22a	**1.87**	11.05	5.14	5.75	7.03	19.02	**5.38**	6.92	10.82	8.11	100%
Cal	✓	3f + Ev	**DualResPS**	2.08	**8.64**	**4.86**	**4.47**	**6.88**	**18.01**	5.38	**5.31**	**10.59**	**7.36**	17.3%
Cal	✓	Events	EventPS	11.00	19.96	12.10	26.14	18.94	38.29	11.54	15.79	26.18	19.99	4.8%
		# Events		148 k	110 k	130 k	146 k	64 k	100 k	85 k	112 k	129 k	114 k	/

4.4 Robustness to motion blur

We conduct a experiment regarding the robustness of DualResPS to motion blur by translating raw 612×512 frames in a uniform velocity to simulate moving objects. The fused long-exposure frames and events change accordingly, and we apply the same pipeline to reconstruct transient normal map aligned with the ground truth at the middle timestamp. As shown in Tab. 4, the accuracy drops steadily when the total displacement increases from 0 to 8 pixels.

Table 6: Quantitative comparison (MAE in degrees) on the DiLiGenT-Pi [6] semi-real dataset. The optimization for LL22a [3] on 24 frames still occasionally stucks in poor local minima.

Lighting	Self-supervised	Input	Method	Auto Bugna-R Bugna-T	Bear Bird Cloud-R	Chest-T Crab Fish	Flower Lion-R Lion-T	Lions Lotus-R Lotus-T	Lung Ocean Panda-R	Panda-T Para Queen	Rhino Sail Ship	Sun Tree TV	Taichi Wave Whale	Average	Data Rate
		3 frames		30.52 12.36 18.04	31.22 20.97 13.52	15.92 17.45 16.31	27.64 16.26 19.90	17.43 14.03 17.79	23.90 20.63 14.17	19.38 36.52 14.06	13.39 32.49 34.45	21.65 32.61 14.78	23.58 19.47 22.38	21.10	12.5%
		4 frames		23.58 10.28 14.41	22.29 15.84 12.61	12.41 18.15 32.89	19.93 14.28 14.77	10.02 11.64 12.72	13.57 16.46 12.36	13.19 27.59 13.85	8.66 26.01 15.92	20.49 26.80 11.64	19.45 11.89 16.15	16.66	16.7%
Cal	×	8 frames	Uni-MS-PS	13.82 9.33 12.99	17.31 17.99 11.73	11.29 16.89 21.81	16.62 12.03 10.42	6.92 10.72 9.24	9.49 14.35 11.54	9.79 30.11 9.53	7.67 23.35 19.69	20.57 17.98 9.39	13.34 9.24 11.49	13.89	33.3%
		24 frames		10.46 9.09 12.57	14.86 14.65 11.58	11.17 13.33 15.79	13.51 11.73 9.58	7.07 10.60 9.00	8.64 12.27 11.42	9.58 26.97 6.44	6.42 18.85 16.03	19.70 14.48 8.83	11.97 9.95 12.20	12.29	100%
		3 frames		16.09 14.27 17.62	15.63 36.81 19.51	22.42 23.81 33.92	28.09 16.99 17.22	32.17 10.94 17.80	14.77 14.42 16.11	18.05 14.17 34.63	32.91 26.46 33.18	16.34 22.60 19.47	16.03 15.26 17.39	21.17	12.5%
		4 frames		15.42 13.30 15.98	15.53 35.77 19.36	21.21 22.57 37.04	23.82 14.94 15.45	27.62 9.76 15.29	14.15 14.21 15.81	16.30 12.94 36.75	30.86 26.14 33.38	17.57 22.42 17.33	15.21 15.50 14.70	20.21	16.7%
Uni	×	8 frames	SDM-UniPS	14.97 13.68 15.65	14.27 32.78 19.54	20.08 20.20 36.98	20.63 14.35 14.40	26.13 9.85 13.01	13.01 13.15 15.48	14.57 15.45 32.62	23.02 19.59 30.70	14.13 19.89 17.73	13.91 14.82 14.48	18.64	33.3%
		24 frames		14.89 13.71 15.49	13.52 28.79 19.48	19.92 19.08 35.20	20.23 14.32 14.26	21.30 10.02 13.14	12.37 12.54 15.48	14.50 15.89 26.23	20.79 16.54 26.04	12.86 19.00 17.34	13.27 14.27 14.14	17.49	100%
		3 frames		15.28 13.04 17.26	16.20 23.17 17.21	17.35 25.92 35.59	26.34 15.91 18.18	11.99 13.32 19.07	13.18 12.17 15.18	18.02 10.15 33.62	13.21 16.86 33.89	15.67 21.92 17.96	12.63 13.33 13.26	18.23	12.5%
		4 frames		15.16 12.20 17.28	15.33 29.46 14.21	18.29 29.95 36.41	20.06 13.55 16.38	14.02 12.33 15.24	12.10 10.81 14.90	18.83 9.44 32.44	13.38 28.54 36.56	13.52 22.29 15.86	11.34 12.29 11.61	18.13	16.7%
Uni	×	8 frames	LINO	13.82 12.92 15.32	13.90 30.58 13.96	18.99 29.16 36.80	14.33 13.01 15.24	14.25 12.80 13.17	11.30 11.12 14.32	17.07 8.78 33.18	9.63 22.87 34.40	13.23 17.97 15.01	10.41 11.25 10.27	16.97	33.3%
		24 frames		13.28 12.93 15.16	13.35 19.08 13.96	20.46 26.74 33.65	13.24 13.23 15.12	14.12 13.08 13.61	11.06 10.89 14.16	17.05 7.94 33.47	8.49 21.92 32.78	13.17 17.56 14.21	10.33 10.83 9.82	16.16	100%
Cal	✓	24 frames	LL22a	8.81 12.40 **15.78**	**6.15** **9.74** 13.40	16.05 **5.07** **6.96**	**6.48** 18.85 21.44	14.98 10.69 **12.71**	9.49 **5.41** 11.59	14.13 **6.27** 20.80	**7.53** 11.00 **5.47**	32.73 12.85 10.38	**9.24** 7.55 **8.00**	11.73	100%
Cal	✓	3f + Ev	**DualResPS**	**6.92** **11.54** 16.05	8.59 10.68 **12.17**	**15.98** 7.39 8.96	8.41 **16.30** **20.41**	**14.68** **10.09** 13.22	**8.21** 5.76 **11.44**	**14.06** 7.69 **19.92**	9.36 **11.02** 10.14	**7.44** **12.11** **9.36**	10.05 **6.95** 8.27	**11.11**	16.1%
Cal	✓	Events	EventPS	17.50 15.38 17.24	10.13 21.43 15.81	17.66 12.53 16.54	16.51 21.72 24.23	23.17 14.20 14.79	14.74 10.22 13.64	15.20 11.40 23.67	16.79 23.70 15.99	13.92 20.22 19.69	19.77 12.90 13.46	16.81	3.6%
		# Events		466 k 383 k 241 k	258 k 772 k 322 k	144 k 345 k 478 k	454 k 863 k 630 k	811 k 441 k 271 k	593 k 550 k 316 k	189 k 335 k 1034 k	575 k 724 k 466 k	404 k 516 k 519 k	280 k 530 k 352 k	475 k	/

5 Complete Evaluation Results

Complete quantitative results on DiLiGenT [5] and DiLiGenT-Pi [6] semi-real datasets are shown in Tab. 5 and Tab. 6. For supervised methods, we show their performance on 3 (similar capturing time as DualResPS), 4 (similar data rate as DualResPS), 8, and 24 frames. Complete visual results on DiLiGenT [5] semi-real dataset are shown in Fig. 5.

Fig. 5: Complete visual results on the DiLiGenT [5] semi-real dataset. Our method produces reconstructions superior to EventPS and the frame-based LL22a [3]. The results of LL22a [3] are on 23 frames in fact for better performance.

References

1. Karis, B., Games, E.: Real shading in unreal engine 4. Proc. Physically Based Shading Theory Practice **4**(3), 1 (2013)
2. Kitazawa, K., Aoto, T., Ikehata, S., Takatani, T.: Ps-eip: Robust photometric stereo based on event interval profile. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6241–6251 (2025)
3. Li, J., Li, H.: Neural reflectance for shape recovery with shadow handling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16221–16230 (2022)
4. Li, J., Li, H.: Self-calibrating photometric stereo by neural inverse rendering. In: European Conference on Computer Vision. pp. 166–183. Springer (2022)

5. Shi, B., Mo, Z., Wu, Z., Duan, D., Yeung, S.K., Tan, P.: A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. IEEE TPAMI (2019)
6. Wang, F., Ren, J., Guo, H., Ren, M., Shi, B.: Diligent-pi: Photometric stereo for planar surfaces with rich details-benchmark dataset and beyond. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9477–9487 (2023)
7. Yu, B., Ren, J., Han, J., Wang, F., Liang, J., Shi, B.: Eventps: Real-time photometric stereo using an event camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9602–9611 (2024)