48, 52, 54], with which the camera-side optical elements and the computational algorithm are jointly optimized in a deep-learning-based framework. Our work is pioneering in applying deep-optics to event cameras.

3. Proposed Method

3.1. Background and Basics

A schematic diagram of our camera is shown in Fig. 1 (left). All the light rays coming into the camera can be parameterized by (x, y, u, v), where (u, v) and (x, y) denote the positions on the aperture and imaging planes, respectively. Therefore, a light field L is defined over (x, y, u, v). We assume that each light ray has only a monochrome intensity; but the extension to RGB color is straight-forward in theory. We also assume that x, y, u, v take discretized integer values; thus, L is equivalent to a set of multi-view images, where (x, y) and (u, v) respectively denote the pixel position and viewpoint. The arrangement of the viewpoints is assumed to be 8×8 ($u, v \in \{1, \ldots, 8\}$). Each element of L is described using subscripts as $L_{x,y,u,v}$. The goal of our method is to reconstruct L of the target scene from the data measured on the camera in a single exposure.

Before describing our imaging method in Section 3.2, we mention two previous imaging methods that use *a coded aperture with a frame-based camera* (no events available).

Coded-aperture imaging [10, 14, 22, 30, 38]: To optically encode a light field L along the viewpoint dimension, a sequence of light-attenuating patterns, $a^{(1)}, \ldots, a^{(N)}$, where $a^{(n)}_{u,v} \in [0, 1]$ and $u, v \in \{1, \ldots, 8\}$, is placed at the aperture plane. Each pixel under the n-th coding pattern is described as the weighted sum of the light rays over (u, v):

$$I^{(n)}_{x,y} = \sum_{u,v} a^{(n)}_{u,v} L_{x,y,u,v}. \tag{1}$$

The light field L is computationally reconstructed from N images taken with different coding patterns, $I^{(1)}, \ldots, I^{(N)}$. The images under different coding patterns would have parallax with each other, which is essential for 3-D/light-field reconstruction. As demonstrated in previous studies [10, 14, 42], a relatively small N (e.g. $N = 4$) is sufficient for accurate reconstruction. However, since N (more than one) images should be acquired in sequence, the lengthy measurement time remains an issue.

Joint aperture-exposure coding [28, 41, 42, 46]: This is an advanced form of coded-aperture imaging; N coding patterns are applied to both the aperture and imaging planes synchronously *during* a single exposure. The coding patterns for the imaging plane are described as $p^{(1)}, \ldots, p^{(N)}$, where $p^{(n)}_{x,y} \in \{0, 1\}$. The imaging process is described as

$$I_{x,y} = \sum_{n} \sum_{u,v} a^{(n)}_{u,v}\, p^{(n)}_{x,y}\, L_{x,y,u,v}. \tag{2}$$

While the light field L can be reconstructed from a single observed image alone, the increased complexity of the coding scheme (Eq. (2)) makes its hardware implementation very difficult.

3.2. Combining Coded Aperture and Events

As shown in Fig. 1, our method uses *a coded aperture with an event camera* that can obtain both the image frames and events simultaneously (e.g., DAVIS 346 camera). Similar to the baseline coded-aperture imaging method (Eq. (1)), we apply N aperture-coding patterns $(a^{(1)}, \ldots, a^{(N)})$ in sequence. However, we do so *in a single exposure* for an image frame. Therefore, we do not directly observe individual coded-aperture images, $I^{(1)}, \ldots, I^{(N)}$, but we have their sum as the image frame:

$$\bar{I}_{x,y} = \sum_{n} I^{(n)}_{x,y}. \tag{3}$$

The camera also measures the events at each pixel (x, y) asynchronously during the exposure. We denote an event occurring at time t as $e_{x,y,t} \in \{+1, -1\}$, where the positive/negative signs correspond to the increase/decrease in the intensity at pixel (x, y). Since the target scene is assumed static, the events are caused exclusively by the change in the coding patterns. Although the pattern changes instantly, the camera responds gradually in a very short but no-zero transient time. We denote the transient time between $a^{(n)}$ and $a^{(n+1)}$ as $T^{(n,n+1)}$. We sum the events during the transient time to obtain an event stack as

$$E^{(n,n+1)}_{x,y} = \sum_{t \in T^{(n,n+1)}} e_{x,y,t}. \tag{4}$$

As shown in Fig. 1, $E^{(n,n+1)}$ includes the parallax information between $I^{(n)}$ and $I^{(n+1)}$, which is essential for 3-D reconstruction. Our goal is to reconstruct the original light field L from a single image frame $\bar{I}$ and $(N - 1)$ event stacks, $E^{(1,2)}, \ldots, E^{(N-1,N)}$, all of which are measured during a single exposure.

Quasi-equivalence. We theoretically relate our imaging method to the baseline coded-aperture imaging method. The camera records an event at time t when the intensity change (in log scale) exceeds a contrast threshold τ as

$$|\log(I_{x,y,t}) - \log(I_{x,y,t_0})| > \tau \tag{5}$$

where t_0 is the time when the previous event was recorded. On the basis of Eq. (5), we derive an approximate relation