Deep Polarization Reconstruction with PDAVIS Events

Haiyang Mei[1,2] Zuowen Wang[2] Xin Yang[1] Xiaopeng Wei[1] Tobi Delbruck[2]
[1]Dalian University of Technology, Dalian, China
[2]Institute of Neuroinformatics, University of Zürich and ETH Zürich, Zürich, Switzerland
haiyang.mei@outlook.com, {xinyang, xpwei}@dlut.edu.cn, {zuowen, tobi}@ini.uzh.ch

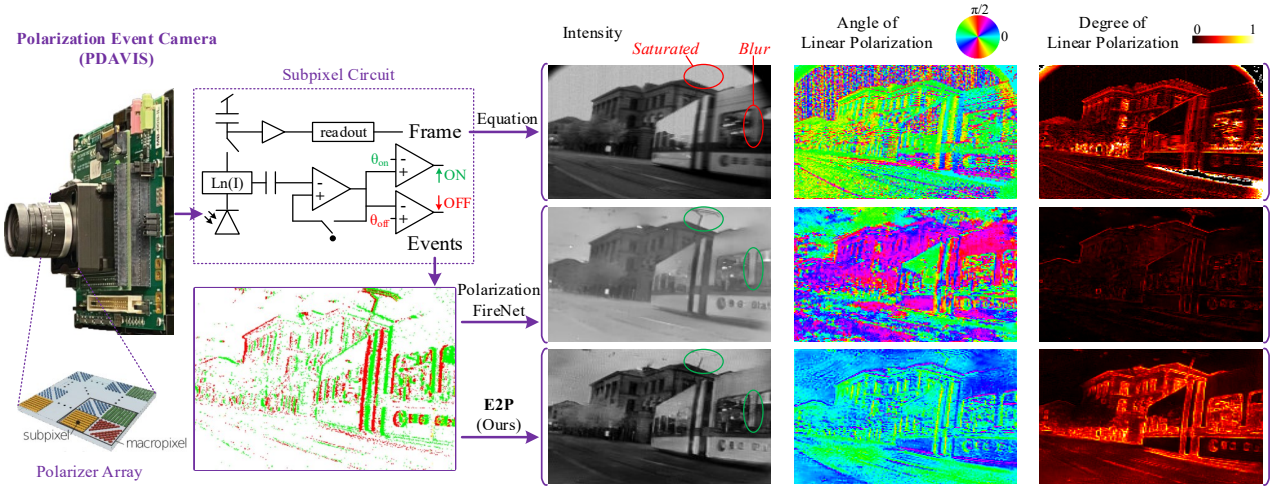

Figure 1. The polarization dynamic and active pixel vision sensor (PDAVIS) [15] is integrated with a nanowire polarizer array. It concurrently outputs conventional polarization intensity frames and asynchronous polarization brightness change events with submillisecond latency over a million-fold illumination range. Taking polarization events as input, the existing state of the art (**SOA**) method *Polarization FireNet* [15,41] can reconstruct the intensity video with high dynamic range (**HDR**) and less motion blur, but fails to correctly reconstruct the angle and degree of linear polarization. Our Events to Polarization (**E2P**) achieves sharp output and HDR with a more accurate polarization reconstruction.

Abstract

*The polarization event camera PDAVIS is a novel bio-inspired neuromorphic vision sensor that reports both conventional polarization frames and asynchronous, continuously per-pixel polarization brightness changes (polarization events) with **fast temporal resolution** and **large dynamic range**. A deep neural network method (Polarization FireNet) was previously developed to reconstruct the polarization angle and degree from polarization events for bridging the gap between the polarization event camera and mainstream computer vision. However, Polarization FireNet applies a network pre-trained for normal event-based frame reconstruction independently on each of four channels of polarization events from four linear polarization angles, which ignores the correlations between channels and inevitably introduces content inconsistency between the four reconstructed frames, resulting in unsatisfactory polarization reconstruction performance. In this work, we strive to train an effective, yet efficient, DNN model that directly outputs polarization from the input raw polarization events. To this end, we constructed the first large-scale event-to-polarization dataset, which we subsequently employed to train our events-to-polarization network E2P. E2P extracts rich polarization patterns from input polarization events and enhances features through cross-modality context integration. We demonstrate that E2P outperforms Polarization FireNet by a significant margin with no additional computing cost. Experimental results also show that E2P produces more accurate measurement of polarization than the PDAVIS frames in challenging fast and high dynamic range scenes. Code and data are publicly available at: https://github.com/SensorsINI/e2p.*

1. Introduction

Visual information is encoded in light by intensity, color, and polarization [12]. Polarization is a property of transverse light waves that specifies the geometric orientation of

the oscillations (which can be described by the Angle of Linear Polarization (**AoLP**) and the Degree of Linear Polarization (**DoLP**)), providing strong vision cues and enabling solutions to challenging problems in medical [27], underwater [34], and remote sensing [53] applications. Existing polarization digital cameras capture synchronous polarization frames with a linear photo response [14], while biological eyes tend to perceive asynchronous and sparse data with a compressed non-linear response [12].

Inspired by the mantis shrimp visual system [29], the novel neuromorphic vision sensor called Polarization Dynamic and Active pixel VIsion Sensor (**PDAVIS**) illustrated in Figure 1 was developed to concurrently record a high-frequency stream of asynchronous polarization brightness change events under four polarization angles (*i.e.*, 0°, 45°, 90°, and 135°) over a wide range of illumination. PDAVIS also outputs low-frequency synchronous frames like conventional polarization cameras [15].

Even though the stream of polarization events has advantages of low latency and HDR, it is not friendly to human observation and traditional computer vision due to the sparse, irregular, and unstructured properties. To better exploit the advantages of PDAVIS, an intuitive solution is to reconstruct polarization from polarization events, which can bridge off-the-shelf frame-based algorithms and PDAVIS. Gruev *et al*. [15] proposed the Polarization FireNet, which first runs the FireNet [41] pre-trained for normal event-based intensity frame reconstruction on each of four types of polarization events under four different polarization angles, and then computes the polarization from four reconstructed intensity frames via mathematical formulas. Since this method treats four polarization angle channels independently, the correlation between channels is ignored and inconsistency between the four reconstructed frames hinders accurate measurement of polarization.

In this work, we make the first attempt to train an accurate yet efficient DNN model tailored for event-to-polarization reconstruction. We approach this twofold. First, we construct the first large-scale event-to-polarization synthetic-real mixed dataset, dubbed Events to Polarization Dataset (**E2PD**), which contains 5 billion polarization events and corresponding 133 thousand polarization video frames. The diversity and practicality of E2PD are ensured by including diverse real-world road scenes under different weather conditions (rainy and sunny) in different cities. Second, we design an E2P network that consists of three branches to reconstruct intensity, AoLP, and DoLP, respectively, from the raw polarization events directly. E2P is built on two key modules: (i) a Rich Polarization Pattern Perception (**RPPP**) module that effectively harvests features from raw polarization events and (ii) a Cross-Modality Attention Enhancement (**CMAE**) module that explores cross-modality contextual cues for feature enhancement.

We perform extensive validation experiments to demonstrate the efficacy of our method and show that the network trained on our E2PD is more accurate than all previously reported PDAVIS methods, and produces more accurate polarization compared with polarization computed from the PDAVIS frames in challenging scenes (*e.g.*, Figure 1). In summary, our contributions are:

1. the first attempt to solve the event-to-polarization problem using an end-to-end trained deep neural network with polarization events as input, intensity, AoLP and DoLP as outputs;
2. a new and unique large-scale event-to-polarization dataset containing both synthetic and real data; and
3. a novel network that perceives rich polarization patterns from raw polarization events and enhances features via a cross-modality attention mechanism.

2. Background and Related Work

Polarization describes the orientation of the transverse electric field in light. Within a non-zero finite time of observation, this orientation can be randomly distributed (unpolarized), biased toward a single direction (linearly polarized), or in between the two extremes (partially linearly polarized). Objects in the real world can produce polarized signals that are related to the nature of materials throughout the process of light reflection, scattering, and transmission [53]. Polarization can reveal intrinsic physical properties of the object [6] and thus can benefit a wide range of applications in computer vision tasks such as estimating shape and/or surface normals [1–3, 10, 18, 42], reflection separation [22, 23, 50], detection [6, 7], and segmentation [19, 24, 30, 52]. It is important to accurately estimate DoLP because AoLP is only meaningful when DoLP is large.

Existing polarization-array CMOS sensors simultaneously record four linear polarization states of light: I_{0°, I_{45°, I_{90°, and I_{135°, where I_ϕ describes the intensity image filtered by a linear polarizer at the angle ϕ, which are then used to compute AoLP and DoLP, defined as:

$$\begin{aligned} S_0 &= I_{0^\circ} + I_{90^\circ} = I_{45^\circ} + I_{135^\circ}, \\ S_1 &= I_{0^\circ} - I_{90^\circ}, \quad S_2 = I_{45^\circ} - I_{135^\circ}, \\ \text{AoLP} &= \frac{1}{2}\arctan\left(\frac{S_2}{S_1}\right), \quad \text{DoLP} = \frac{\sqrt{S_1^2 + S_2^2}}{S_0}, \end{aligned} \tag{1}$$

where S_0, S_1, and S_2 are Stokes elements: S_0 stands for total light intensity and S_1/S_2 describes the ratio between the 0°/45° linear polarization and its perpendicular counterpart. In addition to I_ϕ, the neuromorphic PDAVIS [15] records polarization channel brightness change events E_ϕ with fast temporal resolution and large dynamic range. Our goal is to generate high-quality (*i.e.*, less motion blur and higher dynamic range) polarization video from E_ϕ.

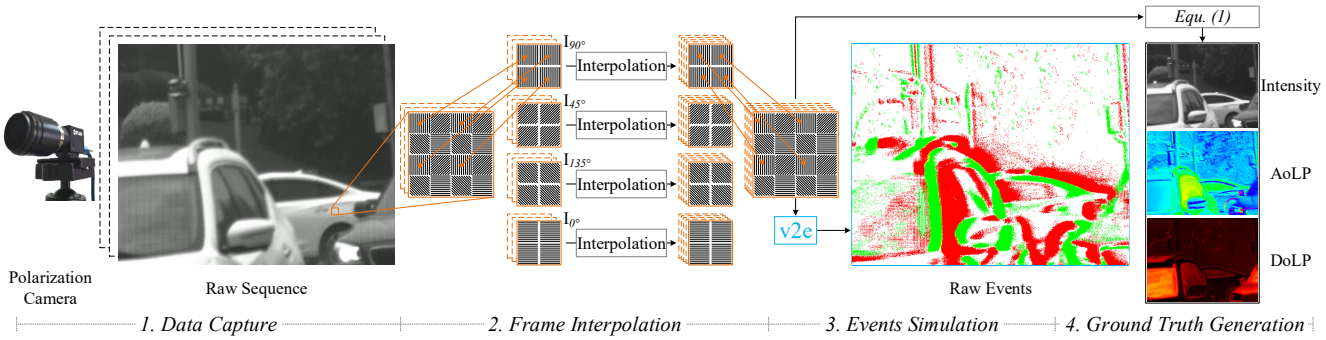

Figure 2. The polarization events data synthesis process.

Event-Based Video Reconstruction converts input events streams to video composed of sequential intensity frames, which bridges the existing frame-based algorithms and event cameras. Early attempts to approach this problem are based on hand-crafted features (*e.g.*, optimization [4], regularization [31], and temporal filtering [39, 40]) or SLAM [11,20,35]. Recently, Barua *et al*. [5] introduced the first learning approach to this task. Then, sparse dictionary learning [48], generative adversarial network (GAN) [44], and fusion between events and frames [47] are also used for event-based video reconstruction. Rebecq *et al*. [36, 37] presented a new events-to-video reconstruction framework called E2VID which is a fully convolutional recurrent UNet architecture inspired by [38, 55] and is trained on a large amount of synthetic event data. Scheerlinck *et al*. developed a fast and lightweight version of E2VID, named FireNet [41]. Weng *et al*. [49] took advantage of the powerful Transformer [43] to improve image reconstruction accuracy, and Zhu *et al*. [57] designed a spiking neural network (SNN) [8] for event-based video reconstruction. [15] showed that the complementary filter method of [39] could fuse PDAVIS frames and events to produce polarization information when many pixels were averaged, but the accuracy was limited by the non-idealities of the real camera output.

Polarization FireNet [15] is the SOA network used as the baseline for our comparisons. It first reconstructs the individual I_ϕ channels from E_ϕ and then calculates polarization using (1). However, we experimentally demonstrate that this method cannot handle complex scenes and using a more powerful event-based video reconstruction method brings significant polarization reconstruction accuracy improvement. We show that training an end-to-end network that directly outputs polarization from polarization events yields more satisfactory results.

3. Event-to-Polarization Dataset

To train the polarization reconstruction network, we construct the first large-scale dataset, named E2PD, which contains 200 polarization video clips and the corresponding 5 billion polarization events.

Synthetic / Real	Videos	Frames (K)	Events (M)
Train	92 / 56	91 / 9	3019 / 680
Test	29 / 23	29 / 4	1087 / 308
Total	121 / 79	120 / 13	4107 / 988

Table 1. Number statistics of our event-to-polarization dataset.

A straightforward way to get polarization events and corresponding ground-truth polarization frames is to use the polarization event camera PDAVIS [15]. However, the polarization frames acquired by PDAVIS [15] cannot provide good ground truth in HDR and high-speed scenes (*e.g.*, Figure 1). Therefore, we further include synthetic data generated from the v2e Dynamic Vision Sensor (**DVS**) simulator [16] and train the network on the synthetic-real mixed data, and subsequently show that the network trained on our dataset can even perform better than real PDAVIS frames in challenging scenes (Section 5.2).

Data Synthesis: the existing largest normal event-to-video dataset [37] is generated by mapping MS-COCO images [26] to a 3D plane and simulating the events triggered by random camera motion within this simple 3D scene. However, we cannot follow this paradigm to obtain reliable polarization events from the raw polarization frames, because the polarization states would change with the movement of the camera. To solve this, we develop a method dedicated to the synthesis of polarization events. As shown in Figure 2, we use a high-quality polarizer-array camera (LUCID PHX050S) that records four linear-polarization directions (0°, 45°, 90°, and 135°) to capture raw polarization videos with a frame rate of 25 Hz and each video lasts 5 seconds. We then generate synthetic polarization events with the v2e tool [16], which accurately models DVS non-idealities. v2e [16] is designed to simulate events from intensity (but not polarization) videos. We first split the raw polarization frame sequence into four channels according to the four linear polarization directions. Each pixel in the resulted single channel polarization frame corresponds to one subpixel of 2×2 macropixels (bottom left of Figure 1). We perform frame interpolation [17] for each channel independently. Then the interpolated frames are rearranged to the raw polarization pattern and the resulting frames are fed into

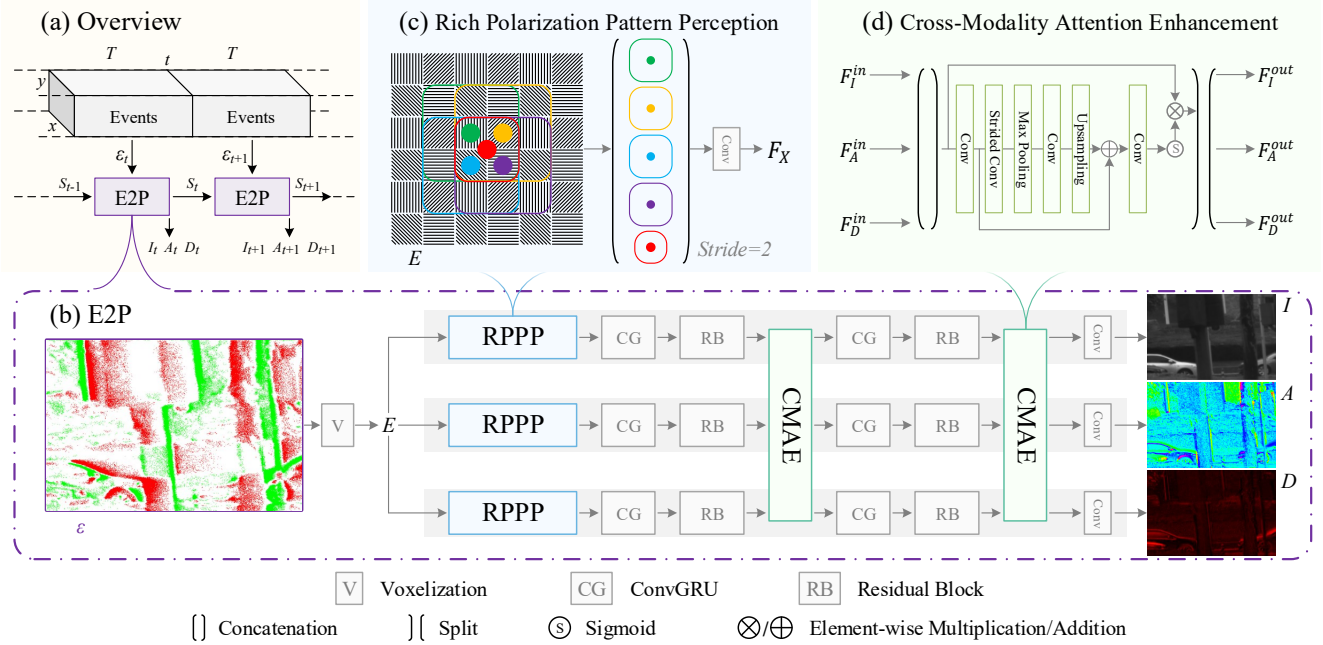

Figure 3. The overview of our polarization reconstruction method (a), the architecture of E2P network (b), and its two main building components: (c) the RPPP module and (d) the CMAE module.

v2e [16] to generate polarization events. Applying (1) to the interpolated polarization frames, we get the corresponding ground-truth polarization (*i.e.*, intensity, AoLP, and DoLP).

Dataset Property: (i) polarization contributes informative cues for various computer vision tasks, especially for road scenes (*e.g.*, car detection [7] and segmentation [52]). Therefore, we include diverse road scenes under different weather conditions (rainy and sunny) in different cities in our dataset, by mounting the camera on the top of a driving car; (ii) for the frame interpolation, we set the scale factor to 8, resulting in a high-speed (200 Hz) polarization video; (iii) in v2e [16] polarization events synthesis, the ON/OFF brightness change event thresholds for individual pixels are sampled from a realistic normal distribution of 0.11 ± 0.02 natural log units, which is more realistic compared with an ideal constant threshold and could help prevent the network from learning to naively integrate events; and (iv) our E2PD is a large-scale dataset which contains 133k polarization frames and 5 billion polarization events. The detailed train/test set split for synthetic/real data is listed in Table 1.

4. Event-to-Polarization Reconstruction

In reality, the brightness change threshold triggering events is neither constant nor uniform [16, 25, 37], which varies depending on factors such as the event rate [9], the sign of brightness change [13], and the temperature [32]. Thus, the polarization events cannot be directly integrated to recover accurate polarization. Instead, we implement the reconstruction function using our recurrent convolutional neural E2P network.

4.1. Overview

As illustrated in Figure 3 (a), given a stream of incoming polarization events, we use a fixed time duration $T = 5$ms to partition it into non-overlapping spatio-temporal windows $\varepsilon_t = \{\mathbf{e}_i\}$, where $\mathbf{e}_i$ is the i-th event in the current window with a format of (t, x, y, p), reporting the timestamp, spatial coordinates, and polarity (-1 or +1), respectively. E2P takes the polarization events ε_t as input and the previous hidden state S_{t-1}, and outputs the intensity frame I_t, AoLP A_t and DoLP D_t, as well as an updated state S_t.

Figure 3 (b) shows the architecture of E2P, which first performs events voxelization (Section 4.2) to convert input polarization events ε into voxel tensor E, enabling the following feature extraction of convolutional recurrent network, and then uses three branches to predict intensity I, AoLP A, and DoLP D, respectively. Each branch consists of an RPPP module (Section 4.3) to extract abundant features from the polarization event tensor E, a Convolutional Gated Recurrent Unit Network (**CG**) to integrate temporal information, a Residual Block (**RB**) that contains two 3×3 convolution layers with skip connection to harvest spatial features, a CMAE module (Section 4.4) to enhance features by exploring cross-modality contextual cues, and a final 3×3 convolution layer for prediction. And the consecutive operations of CG, RB and CMAE are performed twice in each branch for better feature extraction.

4.2. Event Voxelization

We follow prior works [37,41,46,56] to convert input polarization events ε into voxel tensor $E \in \mathbb{R}^{B \times H \times W}$, where

B is the number of temporal bins and H and W are the height and width of ε. Each event (t_i, x_i, y_i, p_i) contributes its polarity to its two closest temporal bins according to:

$$E(b,h,w) = \sum_i p_i \max(0, 1-|b-b_i|), \tag{2}$$

$$b_i = \frac{(t_i - t_{start})}{(t_{end} - t_{start})}(B-1), \tag{3}$$

where $b \in \{0, 1, ..., B-1\}$ is the index of the temporal bin; $p_i \in \{-1, 1\}$ is the polarity and $b_i \in [0, B-1]$ is the normalized timestamp of the i-*th* event; and t_{start}/t_{end} denotes the start/end time of ε. The number of bins B is set to 5 and each has a duration of 1 ms, which integrates from the two neighboring bins with interpolation. Therefore, at each time step the network is fed 5 ms of input events.

4.3. Rich Polarization Pattern Perception

Polarization events differ from normal events in that they report the brightness changes of polarization light filtered by polarizer array as shown in the bottom left of Figure 1. Thus, the polarization pattern should be taken into account when extracting features from raw polarization events, which inspires our RPPP module.

As depicted in Figure 3 (c), RPPP takes as input the raw polarization event tensor E and outputs the extracted features $F_X \in \mathbb{R}^{C\times\frac{H}{2}\times\frac{W}{2}}$, where $X \in \{I, A, D\}$ and C is the channel number. Five convolution layers are applied to E to match five different polarization patterns in the feature extraction process. The stride of these convolutions is set to 2 to match the stride of the polarizer array, which enables learning a set of translation-invariant features. The extracted features are then concatenated and fused by a 1×1 convolution. Formally, the RPPP module can be denoted as:

$$F_X = \psi^r_{31}(\langle \psi^r_{32}(\overset{\nwarrow}{E}), \psi^r_{32}(\overset{\nearrow}{E}), \psi^r_{32}(\overset{\swarrow}{E}), \psi^r_{32}(\overset{\searrow}{E}), \psi^r_{22}(E)\rangle), \tag{4}$$

where ψ^r_{ks} is a $k\times k$ convolution layer with a stride of s followed by a ReLU activation function; $\langle\ \rangle$ denotes the concatenation operation over the channel dimension; and $\nwarrow$, $\nearrow$, $\swarrow$, and $\searrow$ indicate padding the tensor with one zero in top-left, top-right, bottom-left, and bottom-right boundaries, respectively, which enables the four ψ^r_{32} to extract features under four different polarization patterns.

4.4. Cross-Modality Attention Enhancement

Channel-/spatial-wise attention mechanisms [21, 28, 51] are widely used in image reconstruction works to force the deep features to be more focused on the important channels/regions, because they can effectively select useful features. Our CMAE module enhances features via concurrent channel and spatial attention for better polarization reconstruction. The key point of the CMAE module is the observation that the intensity frame, AoLP, and DoLP depicting the same scene can have similar features (*e.g.*, shape and edges) and complementary information (*e.g.*, in Figure 3 (b), the black car on the bottom right of the scene shares similar intensity with its surroundings and presents distinctive intensity to the bottom-left white car, while the opposite is true in the AoLP and DoLP images), which reveals the potential that the features in the three modalities could help each other for better feature enhancement.

Given the input features from three modalities $F^{in}_I, F^{in}_A, F^{in}_D \in \mathbb{R}^{C\times\frac{H}{2}\times\frac{W}{2}}$, the CMAE module, illustrated in Figure 3 (d), outputs both channel-wise and spatial-wise attention enhanced features F^{out}_I, F^{out}_A, and F^{out}_D, by multiplying the input features with a same-shape attention matrix which is obtained via cross-modality context exploration. We adopt the enhanced spatial attention (ESA) block [28] in our CMAE to explore cross-modality contextual cues in a lightweight and efficient manner. Mathematically, CMAE is defined as:

$$\begin{aligned} &F_c = \langle F^{in}_I, F^{in}_A, F^{in}_D\rangle,\ \ F'_c = \psi^r_{11}(F_c),\\ &W = \sigma(\psi_{11}(F'_c + \mathcal{U}(\psi^r_{31}(\mathcal{P}_{44}(\psi^r_{42}(F'_c)))))),\\ &F^{out}_I, F^{out}_A, F^{out}_D = \rangle W * F_c\langle, \end{aligned} \tag{5}$$

where $W \in \mathbb{R}^{3C\times\frac{H}{2}\times\frac{W}{2}}$ is an element-wise attention matrix; σ is the Sigmoid function; $\mathcal{U}$ is a bilinear upsampling; $\mathcal{P}_{mn}$ is a max pooling with a kernel size of m and stride of n; and $\rangle\ \langle$ denotes the split over the channel dimension.

4.5. Loss Function

We follow previous event-to-video methods (*e.g.*, [37, 41]) to use the LPIPS distance function [54] as the reconstruction loss for the intensity, *i.e.*, $\mathcal{L}_i = \ell_{lpips}$. For the reconstructed AoLP and DoLP images, we combine the LPIPS loss [54] with MSE and SSIM loss [45], *i.e.*, $\mathcal{L}_a = \mathcal{L}_d = \alpha\ell_{lpips} + \beta\ell_{mse} + \gamma\ell_{ssim}$, to force the E2P to output the absolute polarization values. We empirically set the balancing parameters α, β, and γ to 1, 50, and 1, respectively. Finally, the overall loss function is:

$$\mathcal{L}_{overall} = w_i\mathcal{L}_i + \mathcal{L}_a + \mathcal{L}_d, \tag{6}$$

where a weighting parameter $w_i = 3$ is used to adjust the magnitude of the intensity loss $\mathcal{L}_i$.

5. Experiments

5.1. Training Setup

We implement E2P in PyTorch [33] and train it for 80 epochs with a batch size of 10 using the Adam optimizer. We set the initial learning rate to 0.001 and decay it by $\eta = 0.3$ at the 50*th* and 70*th* epochs. We augment the training data using random cropping with a crop size of 112×112 while keeping the upper left pixel of the patch

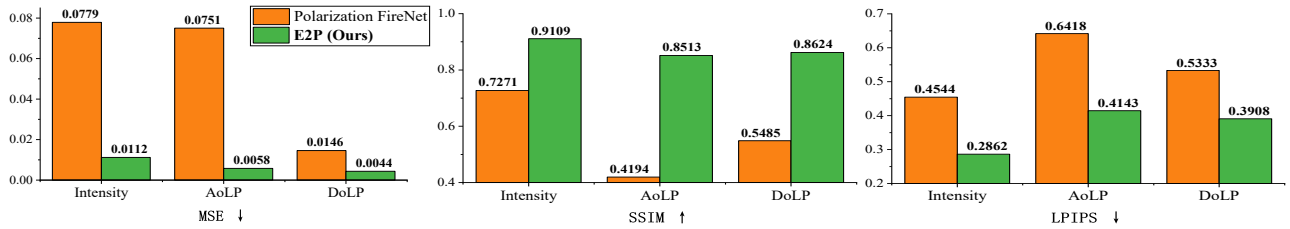

Figure 4. Quantitative comparison of E2P against the SOA event-to-polarization method [15] (Table 2 a) on E2PD synthetic testing set.

with the polarization angle of 90°. The initial training takes 15 hours on a Tesla V100 GPU. We use three metrics for the validation and ablation of our method: mean squared error (MSE), structural similarity index measure (SSIM) [45], and perceptual similarity (LPIPS) [54]. For SSIM, higher is better, while for MSE and LPIPS, lower is better. We compare the effectiveness of E2P to the SOA event-to-polarization method Polarization FireNet [15] on both synthetic and real polarization events in the comparisons shown in Figures 4, 5, and 6 and ablation study in Table 2. We compared both original and retrained FireNets.

5.2. Comparison to Prior Work

Qualitative Comparison. Figure 5 qualitatively demonstrates the advantages of our method: (i) our E2P is able to reconstruct intensity frames with a HDR (*e.g.*, the bottom left example) and less motion blur (*e.g.*, the top right example); (ii) the AoLP images from our method have less noise for the background regions; and (iii) our E2P succeeds in correct DoLP measurement of the scenes, especially for highly polarized objects/regions such as the car windshield and wet roads (pointed by a yellow circle) which pose great challenges for Polarization FireNet [15].

Quantitative Comparison. From Figure 4 we can see that our E2P (Table 2 j) produces more accurate polarization than the state-of-the-art event-to-polarization method Polarization FireNet [15] (Table 2 a) by a significant margin for all three modalities in terms of all three evaluation metrics on the E2PD synthetic testing set.

Validation on Real PDAVIS Data. We also tested E2P on real polarization events recorded by PDAVIS [15] to validate its generalizability. Two scenes are shown in Figure 6, which are more challenging than the ones shown in Figure 5 due to factors including high dynamic range. Because of the high dynamic range of the events, both Polarization FireNet [15] and our E2P can recover the intensity details in overexposed regions (*e.g.*, the car and buildings in the left example) and underexposed parts (*e.g.*, the person in the right scene). Our E2P produces more accurate measurement of polarization than Polarization FireNet [15], which tends to reconstruct noisy AoLP and inaccurate DoLP. And E2P does much more accurate measurement of polarization than the PDAVIS frames in difficult lighting conditions (*e.g.*, in the right example, the DoLP of the person in darkness from the PDAVIS frame is zero due to the limited dynamic range of the polarization frame, while our method can reconstruct the reasonable DoLP from the polarization events). This shows the great potential of our E2P to benefit downstream practical applications such as car/person detection in challenging scenes.

5.3. Comparisons with Stock and Retrained FireNets

Table 2 a, b, c, and d detail the results of a series of experiments to explore whether E2P provides more accurate polarization compared with simpler architectures. First, we replaced the original FireNet [41] used in the Polarization FireNet [15] (a) with a more powerful event-to-video network E2VID [37] (b, Polarization E2VID). The comparison results in Table 2 a and b show that under the paradigm of handling four types of polarization events independently, the use of a more powerful event-to-video network brings only a minor accuracy improvement. Second, we retrained the FireNet architecture to use three FireNet branches to independently predict Stokes elements (*i.e.*, $S0$, $S1$ and $S2$) (c, FireNet-S) or polarization (*i.e.*, intensity, AoLP and DoLP) (d, FireNet-P). These both produce better results than both a and b, indicating that directly estimating Stokes elements or polarization is a better way to reconstruct polarization. These results led us to our development of E2P (j) which estimates polarization from the polarization events using the RPPP features and CMAE integration blocks; it clearly is more accurate than any FireNet variant by a significant margin.

5.4. Ablation Study

How important are the elements of E2P?

Effectiveness of Rich Polarization Pattern Perception. We first define and train a base model (Base, Table 2 e) that is based on E2P (j) but with CMAE removed and RPPP replaced with a 2×2 convolution layer with a stride of 2. Base has the same architecture as FireNet-P but with twice the number of feature channels (*i.e.*, 32 versus 16). By comparing e and d we can conclude that more feature channels can help improve accuracy. Adding our RPPP module to Base improves reconstruction accuracy (*i.e.*, g is better than e). How important is the structure of RPPP? RPPP has a receptive field of 4×4 and so we replaced the RPPP in 'Base + RPPP' by a vanilla 4×4 convolution layer with a stride of 2 (**K4**) and show the results in f. The perfor-

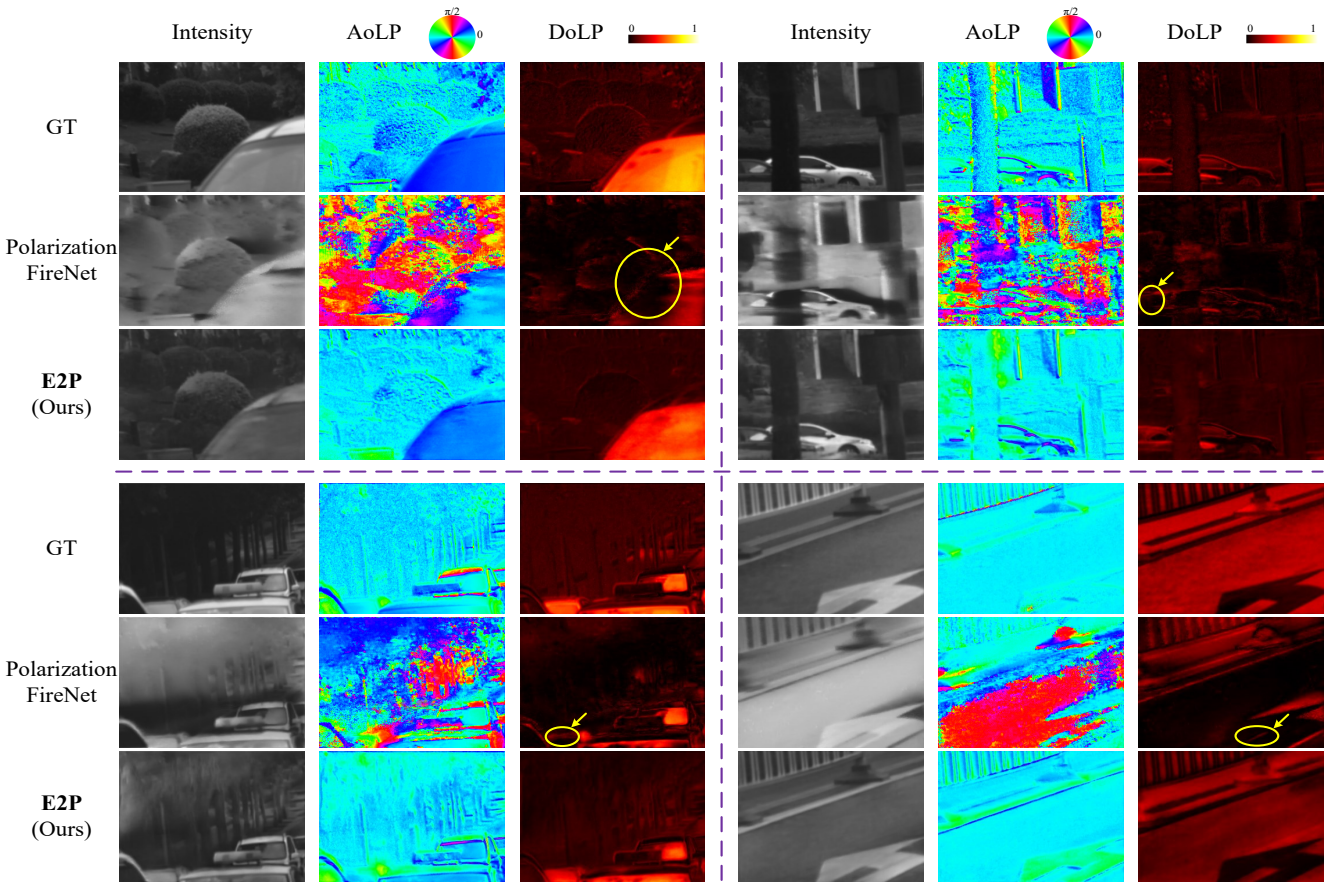

Figure 5. Qualitative comparison of our E2P against SOA event-to-polarization method [15] on the E2PD synthetic testing data.

mance of 'Base + K4' is worse than that of 'Base + RPPP', showing that RPPP can better perceive polarization patterns from polarization events for polarization reconstruction.

Effectiveness of Cross-Modality Attention Enhancement. How important is CMAE for integrating across modalities of polarization? Including CMAE to Base can benefit polarization reconstruction (h is better than e). This conclusion holds when adding CMAE on 'Base + RPPP', *i.e.*, E2P j is better than g. Furthermore, CMAE is effective; replacing CMAE by three independent attention enhancement (IAE) blocks where each has the same architecture as CMAE but takes single modality features as input (*i.e.*, 'Base + RPPP + IAE', i) performs better than 'Base + RPPP' (g) but worse than E2P (j) which has cross-modality attention enhancement, showing that feature enhancement is helpful for polarization reconstruction, and features from three modalities can help each other for better feature enhancement.

5.5. Computational Cost

Our E2P is an end-to-end process that directly outputs intensity, AoLP and DoLP, unlike the Polarization FireNet [15] which first runs FireNet [41] for each of the four types of polarization events and then calculates polarization using (1). All three methods were measured on the same PC with an NVIDIA GeForce RTX 3080 GPU. Table 3 shows that E2P is cheaper in FLOPs and quicker than both Polarization FireNet [15, 41] and Polarization E2VID [37].

5.6. Discussion

E2P has the same limitation as other pure DVS reconstruction methods. When regions of the input lack incoming events for a long time (*e.g.*, the static background from our stationary car) E2P's memory gradually forgets its previous input. Figure 7 shows such a case. Adaptive integration of frames and events, as in the hand-crafted complementary filter method of [15] may address this problem and would be a promising future work. As the first attempt to train a network for more robust event-to-polarization reconstruction, we focus on demonstrating the effectiveness of this idea. Exploring more architectures (*e.g.*, the powerful Transformer [43, 46, 49] or the biologically inspired spiking neural networks [8, 57]) for more accurate/efficient polarization reconstruction is also an interesting research topic.

6. Conclusion

We present E2P, a fully convolutional recurrent neural network tailored for event-to-polarization reconstruc-

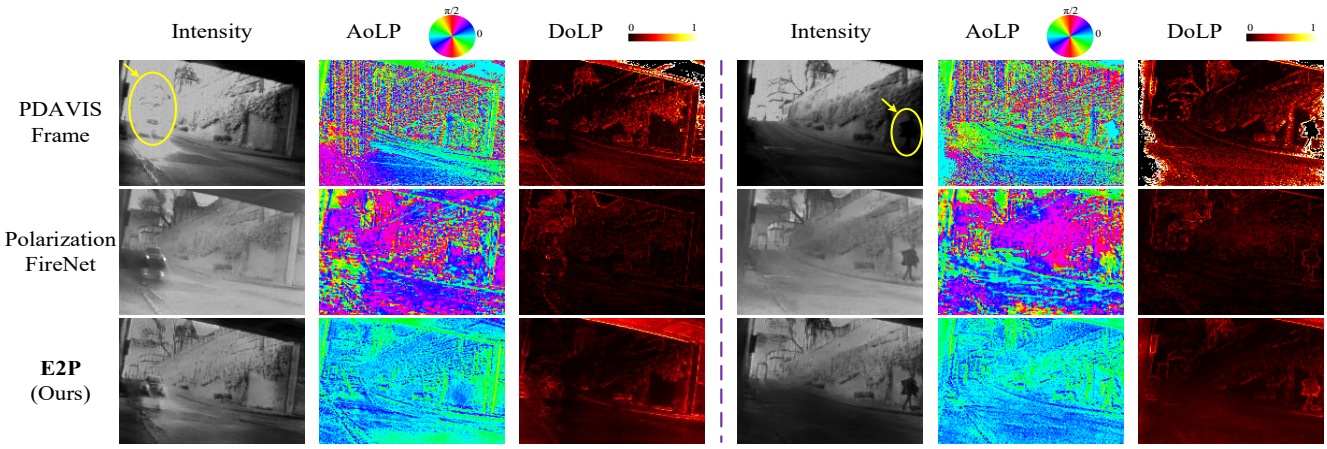

Figure 6. Qualitative comparison of our E2P against SOA event-to-polarization method [15] on the E2PD real testing data.

	Networks	Intensity			AoLP			DoLP		
		MSE ↓	SSIM ↑	LPIPS ↓	MSE ↓	SSIM ↑	LPIPS ↓	MSE ↓	SSIM ↑	LPIPS ↓
a	Polarization FireNet [15, 41]	0.0779	0.7271	0.4544	0.0751	0.4194	0.6418	0.0146	0.5485	0.5333
b	Polarization E2VID [37]	0.0971	0.7086	0.4640	0.0546	0.5728	0.5932	0.0160	0.5883	0.5303
c	Retrained FireNet-S	0.0277	0.7839	0.3939	0.0254	0.6005	0.5582	0.0157	0.6791	0.5180
d	Retrained FireNet-P	0.0254	0.8303	0.3848	0.0074	0.8318	0.5011	0.0128	0.7699	0.4998
e	Base	0.0171	0.8558	0.3565	0.0065	0.8446	0.4908	0.0088	0.8145	0.4674
f	Base + K4	0.0170	0.8418	0.3717	0.0074	0.8253	0.5003	0.0077	0.7905	0.4749
g	Base + RPPP	0.0134	0.8901	0.3099	0.0064	0.8464	0.4708	0.0062	0.8389	0.4413
h	Base + CMAE	0.0128	0.8788	0.3579	0.0066	0.8435	0.4533	0.0044	0.8500	0.4229
i	Base + RPPP + IAE	0.0127	0.8932	0.3301	0.0057	0.8507	0.4466	0.0049	0.8464	0.4215
j	Base + RPPP + CMAE (**E2P**)	0.0112	0.9109	0.2862	0.0058	0.8513	0.4143	0.0044	0.8624	0.3908

Table 2. Quantitative ablation results on the E2PD synthetic testing set. Each column codes accuracy from worse (red) to better (green).

Methods	FLOPs (G)	Params (K)	Time (ms)
Polarization FireNet [15, 41]	46.4	**38**	5.3
Polarization E2VID [37]	558.4	10712	10.5
E2P (Ours)	**36.4**	517	**5.1**

Table 3. Computational efficiency comparison of different event-to-polarization methods for the input polarization events with a spatial size of 640×480. For Polarization FireNet [15] and Polarization E2VID [37], the FLOPs and the inference time are four times as much as FireNet [41] and E2VID [37], respectively, since inference is needed for each of the four types of polarization events each with a spatial size of 320×240.

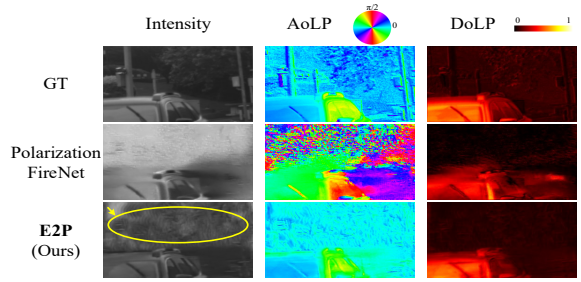

Figure 7. E2P's effectiveness is reduced for the static regions that lack continuous input events.

tion. Our solution is the first deep learning method to directly reconstruct the polarization angle and degree from the PDAVIS [15] polarization events. E2P builds on two key components: a rich polarization pattern perception module that effectively extracts features from raw polarization events and a cross-modality attention enhancement block that explores cross-modality contextual cues for feature enhancement. We also introduce the first large-scale event-to-polarization dataset to train E2P and stimulate further research in this area. We show that our E2P trained on our synthetic-real mixed dataset significantly outperforms the existing event-to-polarization method on the synthetic data and even performs better than the PDAVIS [15] frames using real polarization events in challenging scenes.

Acknowledgments. This work was supported in part by National Key Research and Development Program of China (2022ZD0210500), the National Natural Science Foundation of China under Grant 61972067/U21A20491/U1908214, and the Distinguished Young Scholars Funding of Dalian (No. 2022RJ01). Haiyang Mei was supported by a Chinese Scholarship Council (CSC) grant. This project was also supported by Swiss SNF grant SCIDVS (200021_185069/1). We thank Prof. Viktor Gruev, Univ. of Illinois at Urbana-Champaigne, for initial discussions. We thank Cedric Scheerlinck and Henri Rebecq for making the useful code repository from [41] public.

References

[1] Gary A Atkinson and Edwin R Hancock. Multi-view surface reconstruction using polarization. In *ICCV*, 2005. 2

[2] Gary A Atkinson and Edwin R Hancock. Recovery of surface orientation from diffuse polarization. *IEEE TIP*, 2006. 2

[3] Gary A Atkinson and Edwin R Hancock. Two-dimensional BRDF estimation from polarisation. *CVIU*, 2008. 2

[4] Patrick Bardow, Andrew J Davison, and Stefan Leutenegger. Simultaneous optical flow and intensity estimation from an event camera. In *CVPR*, 2016. 3

[5] Souptik Barua, Yoshitaka Miyatani, and Ashok Veeraraghavan. Direct face detection and video reconstruction from event cameras. In *WACV*, 2016. 3

[6] Rachel Blin, Samia Ainouz, Stéphane Canu, and Fabrice Meriaudeau. Road scenes analysis in adverse weather conditions by polarization-encoded images and adapted deep learning. In *IEEE ITSC*, 2019. 2

[7] Rachel Blin, Samia Ainouz, Stéphane Canu, and Fabrice Meriaudeau. A new multimodal RGB and polarimetric image dataset for road scenes analysis. In *CVPRW*, 2020. 2, 4

[8] Sander M Bohte, Joost N Kok, and Han La Poutre. Error-backpropagation in temporally encoded networks of spiking neurons. *Neurocomputing*, 2002. 3, 7

[9] Christian Brandli, Lorenz Muller, and Tobi Delbruck. Real-time, high-speed video decompression using a frame-and event-based DAVIS sensor. In *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2014. 4

[10] Tongbo Chen, Hendrik P. A. Lensch, Christian Fuchs, and Hans-Peter Seidel. Polarization and phase-shifting for 3D scanning of translucent objects. In *CVPR*, 2007. 2

[11] Matthew Cook, Luca Gugelmann, Florian Jug, Christoph Krautz, and Angelika Steger. Interacting maps for fast visual interpretation. In *IJCNN*, 2011. 3

[12] Thomas W Cronin, Sönke Johnsen, N Justin Marshall, and Eric J Warrant. Visual ecology. In *Visual Ecology*. Princeton University Press, 2014. 1, 2

[13] Guillermo Gallego, Jon EA Lund, Elias Mueggler, Henri Rebecq, Tobi Delbruck, and Davide Scaramuzza. Event-based, 6-DOF camera tracking from photometric depth maps. *IEEE TPAMI*, 2017. 4

[14] Missael Garcia, Christopher Edmiston, Radoslav Marinov, Alexander Vail, and Viktor Gruev. Bio-inspired color-polarization imager for real-time in situ imaging. *Optica*, 2017. 2

[15] Viktor Gruev, Germain Haessig, Damien Joubert, Justin Haque, Yingkai Chen, Moritz Milde, and Tobi Delbruck. Division of focal plane asynchronous polarization imager. In *Polarization: Measurement, Analysis, and Remote Sensing XV*. SPIE, 2022. 1, 2, 3, 6, 7, 8

[16] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *CVPRW*, 2021. 3, 4

[17] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *CVPR*, 2018. 3

[18] Achuta Kadambi, Vage Taamazyan, Boxin Shi, and Ramesh Raskar. Polarized 3D: High-quality depth sensing with polarization cues. In *ICCV*, 2015. 2

[19] Agastya Kalra, Vage Taamazyan, Supreeth Krishna Rao, Kartik Venkataraman, Ramesh Raskar, and Achuta Kadambi. Deep polarization cues for transparent object segmentation. In *CVPR*, 2020. 2

[20] Hanme Kim, Stefan Leutenegger, and Andrew J Davison. Real-time 3D reconstruction and 6-DoF tracking with an event camera. In *ECCV*, 2016. 3

[21] Fangyuan Kong, Mingxi Li, Songwei Liu, Ding Liu, Jingwen He, Yang Bai, Fangmin Chen, and Lean Fu. Residual local feature network for efficient super-resolution. In *CVPRW*, 2022. 5

[22] Chenyang Lei, Xuhua Huang, Mengdi Zhang, Qiong Yan, Wenxiu Sun, and Qifeng Chen. Polarized reflection removal with perfect alignment in the wild. In *CVPR*, 2020. 2

[23] Rui Li, Simeng Qiu, Guangming Zang, and Wolfgang Heidrich. Reflection separation via multi-bounce polarization state tracing. In *ECCV*, 2020. 2

[24] Yupeng Liang, Ryosuke Wakaki, Shohei Nobuhara, and Ko Nishino. Multimodal material segmentation. In *CVPR*, 2022. 2

[25] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 dB 15 μs latency asynchronous temporal contrast vision sensor. *IEEE J. Solid-State Circuits*, 43(2):566–576, Feb. 2008. 4

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 3

[27] Chenyang Liu, Chengyong Shi, Taisheng Wang, Hongxin Zhang, Lei Jing, Xiya Jin, Jia Xu, and Hongying Wang. Bio-inspired multimodal 3D endoscope for image-guided and robotic surgery. *Optics Express*, 2021. 2

[28] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *CVPR*, 2020. 5

[29] N Justin Marshall. A unique colour and polarization vision system in mantis shrimps. *Nature*, 1988. 2

[30] Haiyang Mei, Bo Dong, Wen Dong, Jiaxi Yang, Seung-Hwan Baek, Felix Heide, Pieter Peers, Xiaopeng Wei, and Xin Yang. Glass segmentation using intensity and spectral polarization cues. In *CVPR*, 2022. 2

[31] Gottfried Munda, Christian Reinbacher, and Thomas Pock. Real-time intensity-image reconstruction for event cameras using manifold regularisation. *IJCV*, 2018. 3

[32] Y Nozaki and T Delbruck. Temperature and parasitic photocurrent effects in dynamic vision sensors. *IEEE Trans. Electron Devices*, 64(8):3239–3245, Aug. 2017. 4

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 5

[34] Samuel B Powell, Roman Garnett, Justin Marshall, Charbel Rizk, and Viktor Gruev. Bioinspired polarization vision en-

ables underwater geolocalization. *Science advances*, 2018. 2

[35] Henri Rebecq, Timo Horstschäfer, Guillermo Gallego, and Davide Scaramuzza. EVO: A geometric approach to event-based 6-DOF parallel tracking and mapping in real time. *IEEE Robotics and Automation Letters*, 2016. 3

[36] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *CVPR*, 2019. 3

[37] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE TPAMI*, 2019. 3, 4, 5, 6, 7, 8

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 3

[39] Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Continuous-time intensity estimation using event cameras. In *ACCV*, 2018. 3

[40] Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Asynchronous spatial image convolutions for event cameras. *IEEE Robotics and Automation Letters*, 2019. 3

[41] Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Fast image reconstruction with an event camera. In *WACV*, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[42] Vimal Thilak, David G. Voelz, and Charles D. Creusere. Polarization-based index of refraction and reflection angle estimation for remote sensing applications. *Applied Optics*, 2007. 2

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3, 7

[44] Lin Wang, Yo-Sung Ho, Kuk-Jin Yoon, et al. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *CVPR*, 2019. 3

[45] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004. 5, 6

[46] Zuowen Wang, Yuhuang Hu, and Shih-Chii Liu. Exploiting spatial sparsity for event cameras with visual transformers. In *ICIP*, 2022. 4, 7

[47] Zihao W Wang, Weixin Jiang, Kuan He, Boxin Shi, Aggelos Katsaggelos, and Oliver Cossairt. Event-driven video frame synthesis. In *ICCVW*, 2019. 3

[48] Yijing Watkins, Austin Thresher, David Mascarenas, and Garrett T Kenyon. Sparse coding enables the reconstruction of high-fidelity images and video from retinal spike trains. In *Proceedings of the International Conference on Neuromorphic Systems*, 2018. 3

[49] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *ICCV*, 2021. 3, 7

[50] Patrick Wieschollek, Orazio Gallo, Jinwei Gu, and Jan Kautz. Separating reflection and transmission images in the wild. In *ECCV*, 2018. 2

[51] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *ECCV*, 2018. 5

[52] Kaite Xiang, Kailun Yang, and Kaiwei Wang. Polarization-driven semantic segmentation via efficient attention-bridged fusion. *Optics Express*, 2021. 2, 4

[53] Lei Yan, Taixia Wu, and Xueqi Wang. Polarization remote sensing for land observation. In *Understanding of Atmospheric Systems with Efficient Numerical Methods for Observation and Prediction*. IntechOpen London, UK, 2018. 2

[54] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5, 6

[55] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. EV-FlowNet: Self-supervised optical flow estimation for event-based cameras. In *Robotics: Science and Systems*, 2018. 3

[56] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *CVPR*, 2019. 4

[57] Lin Zhu, Xiao Wang, Yi Chang, Jianing Li, Tiejun Huang, and Yonghong Tian. Event-based video reconstruction via potential-assisted spiking neural network. In *CVPR*, 2022. 3, 7