RGB-Event Fusion for Robust Lane Detection

Jingtao Dong
jingtaodong0616@bit.edu.cn
Hao Zhuang
haozhuang@bit.edu.cn
Hao Yang
hao.yang@bit.edu.cn
Liyuan Pan†
liyuan.pan@bit.edu.cn

School of Computer Science & Technology
Beijing Institute of Technology
Beijing, China

Abstract

Lane detection is crucial for autonomous driving, yet standard RGB cameras struggle under extreme lighting conditions. Event cameras, with their high temporal resolution and low latency, offer a promising solution in such environments, motivating the fusion of RGB and event data to enhance lane detection robustness. To address this challenge, we propose a pipeline incorporating the Multiple Feature Fusion Blocks (MFFB), which leverages the temporal correlation of lane markings by maintaining the current state and providing crucial information for subsequent timesteps. Additionally, we integrate edge features to mitigate the insufficiency of lane edge information caused by insufficient motion in event cameras. Despite the availability of event camera datasets for autonomous driving, no existing dataset simultaneously includes both event and RGB data with annotation for lane detection. To address this gap, we extend the DSEC dataset by introducing lane annotations and performing homographic alignment, resulting in the DSEC-Lane dataset. This dataset provides RGB frames and event data, captured in parallel, from autonomous driving scenes under challenging lighting conditions across urban, suburban, and rural environments. Extensive experiments demonstrate that our approach surpasses SOTA methods, highlighting the effectiveness of event-image fusion in challenging lane detection scenarios.

1 Introduction

Autonomous driving has gained significant attention in academia and industry, focusing on achieving detailed perception through sensor fusion and advanced control systems. With the rapid advancement of deep learning[53, 62, 63, 73], remarkable achievements have also been made in this field. Key tasks, such as lane detection [9, 18, 35, 46, 56, 66], object detection [12, 47, 68] and traffic sign recognition [33, 51], are integral to this process, with lane detection being particularly foundational. However, standard RGB cameras face limitations in detecting lane markings under extreme lighting conditions.

† Corresponding author.

Figure 1: Visualization of the proposed DSEC-Lane dataset under diverse and challenging driving scenarios. From top to bottom: RGB images, event frames, ground-truth (GT) lane annotations, and model predictions. The dataset covers a wide range of lighting conditions, including daytime, nighttime, and backlit scenes.

Unlike frame-based methods, which degrade under challenging lighting conditions [24, 59, 65], event-based bio-inspired cameras excel in low-light environments due to their high temporal resolution and dynamic range [7, 23, 31]. By capturing illumination changes, event cameras [52] effectively mitigate motion blur while reducing energy consumption, bandwidth usage, and computational demands [13], making them particularly well-suited for lane detection tasks. However, their effectiveness is limited when detecting lane markings with low contrast or subtle variations, reducing their reliability in certain scenarios. To address this challenge, we fuse event cameras with standard RGB cameras, leveraging the complementary strengths of both modalities to enhance robustness and accuracy in diverse conditions.

In this paper, we propose a Multiple Feature Fusion Block (MFFB) that effectively leverages the complementary strengths of asynchronous event streams and conventional RGB frames for robust lane detection, as shown in Fig. 2. This module utilizes the temporal continuity of lane markings over time by preserving fused features from the previous timestep, aiding the extraction of structural lane features at the current timestep and enhancing the robustness of lane detection. Additionally, considering the limitations of event-based models for detecting distant lane markings due to insufficient relative motion and the subsequent loss of event edge information [13, 19, 64], we incorporate edge features into the fusion process. This allows the model to better capture lane boundaries, which are often weakened in traditional event and image-based representations, as edges are typically transformed into high-level features by feature extractors, neglecting low-level edge information. By guiding the fusion of these two modalities with edge information, our approach enhances the structural features of lane markings.

Despite the availability of event camera datasets for autonomous driving, there remains a lack of lane detection datasets that integrate both event and standard camera modalities, significantly hindering research on the application of event cameras in this domain. To bridge this gap, we introduce DSEC-Lane, a dataset comprising 11,862 frames extracted from the

DSEC dataset [15]. This newly introduced dataset captures a diverse range of challenging driving scenarios, including various lighting conditions and complex traffic environments, as illustrated in Fig. 1. To ensure temporal consistency, we provide timestamp-synchronized data from both sensors and apply a homography transformation to align standard camera images with the event camera's perspective [47], accounting for resolution differences.

Our contributions are summarized as follows,

- We introduce DSEC-Lane, the first dataset for lane detection. DSEC-Lane comprises 11,862 annotated samples captured across dynamic urban, suburban, and rural traffic scenes under both day and night conditions. Each sample includes temporally aligned event streams, RGB frames, and precise lane-line annotations.
- We propose a Multiple Feature Fusion Block (MFFB), which fuses RGB frames with asynchronous event streams through time-aware self-attention and edge cues, delivering robust lane estimates under extreme lighting.
- We establish the first benchmark for RGB-Event fusion in lane detection, evaluate diverse frame-based and fusion baselines, and achieve state-of-the-art performance on the real-world DSEC-Lane dataset using our event-image fusion strategy.

2 Related Work

Lane Detection. Lane detection has long been a critical component of autonomous driving systems[14], initially addressed by traditional methods that relied on handcrafted features and classical machine learning algorithms [5, 11, 27]. With the rise of deep learning, a new wave of lane detection methods has emerged [3, 17, 18, 46, 66], methods such as LaneATT [46], which uses attention mechanisms to highlight relevant lane features, and CLRNet [66], which incorporates multi-scale feature fusion for improved lane boundary prediction. To support the development of these deep learning-based approaches, a series of large-scale datasets have been introduced [2, 4, 9, 22, 39, 49, 61]. Datasets like Tusimple [49], which focuses on highway driving scenarios, and CULane [39], one of the largest datasets featuring diverse traffic, weather and lighting conditions. Despite the success of these methods in conventional driving environments, significant challenges persist in extreme scenarios, where traditional RGB-based methods exhibit limited robustness.

Event-based Vision. Event cameras, with their advantages of low latency and high dynamic range (HDR), have been widely applied in tasks such as object detection [8, 25, 34, 68, 69], segmentation [1, 34, 43, 72], optical flow estimation [10, 30, 38, 71] and deblurring [20, 37, 44, 67]. Meanwhile, a series of event-based camera datasets for autonomous driving have emerged [9, 15, 70], such as DSEC [15], which provides synchronized event streams, RGB frames, and LiDAR data from real-world driving scenarios. MVSEC [70] offers synchronized event data along with stereo camera images, captured in outdoor environments, and supports tasks like visual odometry and depth estimation. In this work, we integrate event-based vision into lane detection, enabling complementary information fusion between the RGB and event modalities.

Multimodal Fusion. With the development of event cameras, the fusion of event-based and RGB data has gained increasing attention [32, 44, 45, 47, 48, 68], Zhou *et al.*[68] explicitly models both modality-specific features and shared representations, Sun *et al.*[44] further leverages the transformer's attention to better guide the feature fusion. We analyze

Table 1: Comparison of various lane detection datasets, including the year of release, modality, and dataset characteristics. The columns indicate whether datasets feature certain attributes, such as E-I Pair (Event-Image Pairing), MI (Multi-Illumination), MS (Multi-Scenes), ML (Multi-Line Type), and MM (Multiple Moments). The annotation types, scale and resolution of each dataset are also listed.

Dataset	Year	Modality	E-I Pair	MI	MS	ML	MM	Real	Annotations	Scale	Resolution
Caltech Lanes [2]	2008	RGB	✗	✗	✗	✗	✗	✓	Key point coordinates	1,225	640×480
Tusimple [49]	2017	RGB	✗	✗	✗	✗	✗	✓	Key point coordinates	6,498	1280×720
CULane [39]	2017	RGB	✗	✗	✗	✗	✓	✓	Key point coordinates	133,235	1640×590
VPGNet [27]	2017	RGB	✗	✗	✗	✓	✓	✓	Key point coordinates	21,097	346×260
LLAMAS [1]	2019	RGB	✗	✗	✗	✓	✓	✓	Pixel level	100,442	128×128
CurveLanes [51]	2020	RGB	✗	✓	✓	✓	✓	✓	Key point coordinates	150,000	1280×800
DET [9]	2020	Event	✗	✓	✓	✓	✗	✓	Pixel level	5,424	1280×800
DSEC-Lane	2025	RGB+Event	✓	✓	✓	✓	✓	✓	Key point coordinates	11,862	640×480

the strong temporal correlation of lane detection and the important structural information of lane markings, and propose a method that incorporates temporal information and utilizes edge features extracted from images to guide the fusion of event and image data. This approach effectively leverages the strengths of both sensors, while improving lane detection performance under extreme conditions.

3 Methodology

3.1 Dataset

Data Preparation. Our DSEC-Lane, based on the DSEC dataset [15], uses stereo monochrome event cameras with a 640×480 resolution and a 60 cm baseline, paired with color cameras at 1440×1080 resolution and a 51 cm baseline. This setup leverages the high spatial resolution of the RGB cameras while retaining the temporal precision of the event cameras. The dataset was recorded in urban, suburban, and rural areas of Switzerland, covering both daytime and nighttime conditions, with diverse and dynamic traffic scenarios that present significant challenges for lane marking detection.

We compress the raw event stream along the time dimension with $\Delta t = 50ms$, which corresponds to the frame rate of an RGB camera, allowing for better alignment. Δt represents the time span of the event stream corresponding to a single image. A total of 11,862 images were carefully selected from all sequences in the DSEC dataset, encompassing a variety of traffic scenarios for annotation. Tab. 1 presents a comparison between our dataset and other lane detection datasets.

Data Annotation. The dataset contains manually annotated lane line instance point sets, with points sampled at consistent height intervals. The lane line annotation protocol adheres to the guidelines established by Tusimple [49]. In addition, we classify the types of each lane line and annotate the road boundaries, as both are essential for constraining the vehicle's motion in autonomous driving systems.

Table 2: Distribution of lane annotation counts per image in our dataset.

Lanes	One	Two	Three	Four	Total
Count	566	4033	4232	3031	11862
Percent (%)	4.77	34.00	35.68	25.55	100

We apply a homography transformation [47] to map the scene from the RGB frame to the

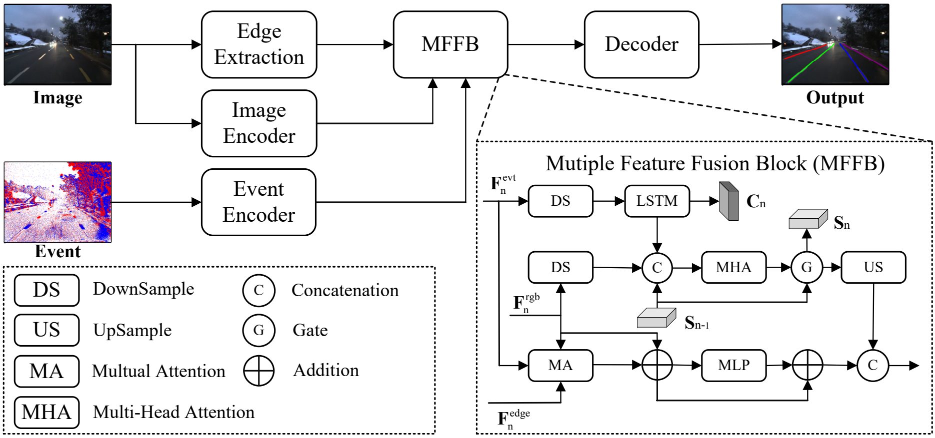

Figure 2: Overview of the proposed pipeline for lane detection using RGB and event camera data. The system consists of an Image Encoder, Event Encoder, and a Multiple Feature Fusion Block (MFFB). The event features $\mathbf{F}_n^{\text{evt}}$ at time step n are processed through DownSampling (DS) and an LSTM to maintain the contextual state $\mathbf{C}_n$. These features are then fused with the previous state $\mathbf{S}_{n-1}$ and the RGB features $\mathbf{F}_n^{\text{rgb}}$ through Multi-head Attention (MHA) and a gating mechanism to update the current state $\mathbf{S}_n$. The edge features $\mathbf{F}_n^{\text{edge}}$, extracted from grayscale images, undergo Mutual Attention (MA) and are fused with both the $\mathbf{F}_n^{\text{rgb}}$ and $\mathbf{F}_n^{\text{evt}}$ features. The fused features are processed through a Multi-Layer Perceptron (MLP) layer with residual connections, followed by concatenation with the upsampled $\mathbf{S}_n$ and sent to the Decoder to produce the final lane detection output.

event camera frame, followed by annotating the lane lines on the transformed RGB image. Each annotation includes up to four lane lines, and Tab. 2 provides the statistical details of lane line quantities within the dataset.

3.2 Framework Overview

Our model is built upon the CLRNet [66] architecture, extending it into a dual-stream feature extraction network and multi-scale fusion network to effectively process both event and image data. Given the asynchronous nature of event data, integrating it with traditional image frames poses unique challenges. To address this, we treat event volumes directly as event frames, which are then combined with image frames and fed into our feature extraction network to extract shallow features from both modalities.

The extracted features are passed through a Multiple Feature Fusion Block (MFFB), which is designed to effectively integrate complementary spatiotemporal information from both RGB images and event-based data. By leveraging the high temporal resolution of events and the rich spatial details of RGB frames, MFFB enhances the quality of fused representations. In addition, edge information is incorporated to guide the fusion process, improving robustness in regions with low motion where event may be sparse or degraded. The fused features are then forwarded to the decoder for final prediction.

As illustrated in Fig. 2, the Multiple Feature Fusion Block (MFFB) operates directly on the global feature map extracted from the backbone, rather than region-specific (ROI) features. This design choice is intentional: lane detection is inherently a dense, scene-level task that requires modeling long-range spatial continuity of lane markings, rather than localized object-centric semantics. By applying MFFB at the global scale, our framework can exploit complementary spatial–temporal cues from both event and image modalities, enabling the fused representation to capture holistic structural information across the entire driving scene. In contrast, restricting fusion to ROI features would compromise contextual integrity and fail to leverage the global consistency that is critical for robust lane localization under challenging conditions.

3.3 Multiple Feature Fusion Block

To integrate complementary cues from both event and RGB modalities, we propose a Multiple Feature Fusion Block (MFFB) , as illustrated in Fig. 2. Given sequences of RGB frames $\{\mathbf{X}_n^{\text{rgb}}\}_{n=1}^T$ and event frames $\{\mathbf{X}_n^{\text{evt}}\}_{n=1}^T$ (we follow [21, 36, 40, 71] and adopt a 4-channel event representation as the network input), modality-specific features $\mathbf{F}_n^{\text{rgb}}$ and $\mathbf{F}_n^{\text{evt}}$ are first extracted using independent encoders, i.e., $\mathbf{F}_n^{\text{rgb}} = \mathcal{E}_{\text{rgb}}(\mathbf{X}_n^{\text{rgb}})$ and $\mathbf{F}_n^{\text{evt}} = \mathcal{E}_{\text{evt}}(\mathbf{X}_n^{\text{evt}})$, where $\mathcal{E}_{\text{rgb}}(\cdot)$ and $\mathcal{E}_{\text{evt}}(\cdot)$ denote the RGB and event encoders, respectively. Here, the subscript n denotes the n-th time step.

To account for the strong temporal continuity inherent in lane trajectories, we first downsample both the event and RGB features, obtaining $\mathbf{D}_n^{\text{evt}}$ and $\mathbf{D}_n^{\text{rgb}}$, respectively. To model temporal dependencies, We apply a ConvLSTM block to $\mathbf{D}_n^{\text{evt}}$, maintaining a temporal memory cell $\mathbf{C}_n$. Its output, together with the previous features and $\mathbf{D}_n^{\text{rgb}}$, is fused and fed into a multi-head attention module:

$$\mathbf{H}_n = \text{MHA}\big(\text{Concat}(\ \mathbf{D}_n^{\text{rgb}},\ \mathbf{S}_{n-1},\ \mathbf{C}_n)\big), \tag{1}$$

where MHA$(\cdot)$ represents the Multi-Head Attention mechanism [50], and $\mathbf{S}_{n-1}$ denotes the fused information from the previous time step. the updated representation $\mathbf{H}_n$ is fused with the previous output $\mathbf{S}_{n-1}$ through a gating unit, which adaptively controls the information flow and updates the current state $\mathbf{S}_n$:

$$\mathbf{S}_n = \mathcal{G}\left(\mathbf{H}_n, \mathbf{S}_{n-1}\right). \tag{2}$$

Here $\mathcal{G}(\cdot)$ denotes a gating unit. Event data provides high temporal resolution and low latency, making it particularly well-suited for capturing fine-grained motion dynamics in lane detection. Its asynchronous nature enables the network to effectively model temporal continuity, which is crucial for detecting lane trajectories that evolve smoothly over time. However, event streams may experience edge degradation in areas with insufficient relative motion, resulting in incomplete structural cues. To mitigate this issue, we introduce edge information extracted from grayscale images as a complementary signal. This additional structural guidance helps maintain the integrity of lane boundaries, particularly in low-motion or static scenes where event activity is sparse, thus enhancing the robustness of the fusion process. To further improve this, we design a Mutual Attention (MA) component, as illustrated in Fig. 2.

Specifically, given grayscale image features $\mathbf{X}_n^{\text{gray}}$, we extract edge features via $\mathbf{F}_n^{\text{edge}} = \text{Sobel}(\mathbf{X}_n^{\text{gray}})$, where Sobel$(\cdot)$ refers to the sobel operator introduced earlier. We then normalize and modulate the edge features as follows:

$$\gamma, \beta = f(\mathbf{F}_n^{\text{edge}}), \quad \mathbf{A}_n = \text{GN}(\mathbf{F}_n^{\text{evt}}) \odot (1+\gamma) + \beta. \tag{3}$$

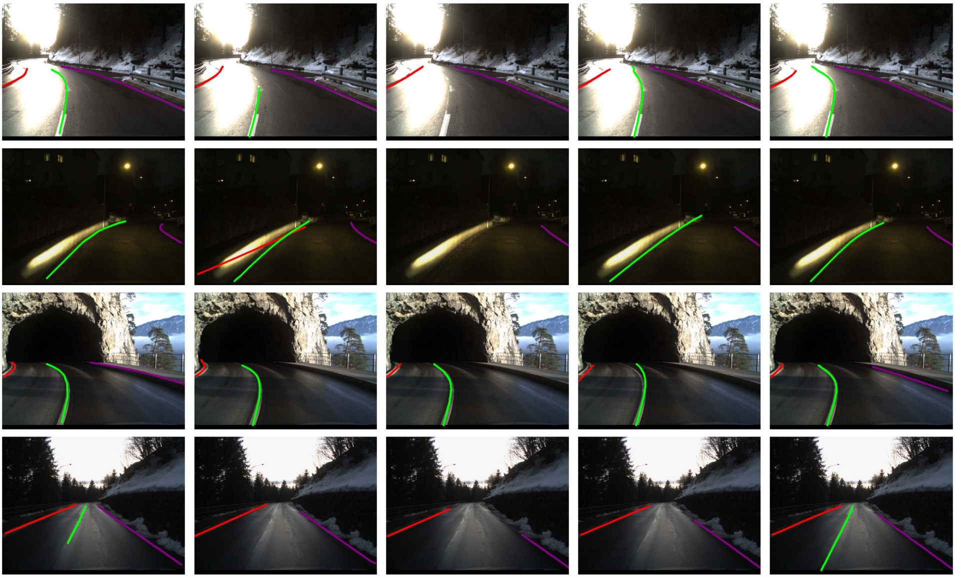

Figure 3: Qualitative comparison of lane detection results. From left to right: Ground Truth (GT), LaneATT [46], CLRNet [66], CLRKDNet [41], and our method. Our method demonstrates superior lane detection accuracy, particularly in challenging lighting and environmental conditions.

Here $\mathbf{F}_n^{\mathrm{evt}}$ denotes the event feature, and $\mathbf{F}_n^{\mathrm{edge}}$ is the corresponding edge feature extracted from the grayscale image. $\mathrm{GN}(\cdot)$ applies Group Normalization [58], and $f(\cdot)$ is a modulation network composed of multi-scale dilated convolutions that outputs spatially-adaptive scaling γ and bias β, $\odot$ denotes element-wise multiplication, $\mathbf{A}_n$ represents the final result.

We further leverage the adaptive representation $\mathbf{A}_n$ as the basis for attention computation. Specifically, we project $\mathbf{A}_n$ to generate the keys and values, while the grayscale feature $\mathbf{F}_n^{\mathrm{gray}}$ is projected to obtain the query. The final cross-modal attention is computed as:

$$\begin{aligned} \mathbf{Q}_n &= \mathbf{W}_Q \mathbf{F}_n^{\mathrm{rgb}}, \mathbf{K}_n = \mathbf{W}_K \mathbf{A}_n, \mathbf{V}_n = \mathbf{W}_V \mathbf{A}_n, \\ \mathbf{O}_n &= \mathbf{V}_n \cdot \mathrm{Softmax}\left(\frac{\mathbf{Q}_n^{\top}\mathbf{K}_n}{\tau}\right). \end{aligned} \tag{4}$$

Here $\mathbf{F}_n^{\mathrm{rgb}}$ denotes the RGB feature map used to compute the query $\mathbf{Q}_n$, and $\mathbf{A}_n$ is the adaptive feature fused from event and edge information. $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are learnable projection matrices applied to obtain the query, key, and value embeddings, respectively. τ is a temperature parameter that controls the sharpness of the attention distribution. The output $\mathbf{O}_n$ is the attention-enhanced representation.

To enhance feature expressiveness while preserving identity information, we adopt a residual feed-forward structure following the attention module. The attention-enhanced feature $\mathbf{O}_n$ is first added to the original RGB feature $\mathbf{F}_n^{\mathrm{rgb}}$, yielding an intermediate representation $\mathbf{H}_n = \mathbf{F}_n^{\mathrm{rgb}} + \mathbf{O}_n$, which is further refined by a multilayer perceptron (MLP) with an additional residual connection:

$$\mathbf{F}_n^{\mathrm{out}} = \mathbf{H}_n + \mathrm{MLP}(\mathbf{H}_n). \tag{5}$$

Here MLP$(\cdot)$ consists of two linear layers with a GELU activation and dropout in between [16]. Finally, we concatenate the upsampled state feature $\mathbf{S}_n$ with the fused output $\mathbf{F}_n^{\text{out}}$, and feed the combined representation $\mathbf{Z}_n$ into the decoder for final lane prediction.

Training Loss. We follow [66] and use classification, regression, LineIoU, and segmentation losses for our lane detection task.

$$\mathcal{L}_{total} = w_{cls}\mathcal{L}_{cls} + w_{reg}\mathcal{L}_{reg} + w_{seg}\mathcal{L}_{seg} + w_{LIoU}\mathcal{L}_{LIoU}. \tag{6}$$

4 Experiments

Dataset. We conduct experiments on the DSEC-Lane dataset, derived from the DSEC dataset, which contains 11,862 image frames captured in real-world autonomous driving scenarios across different times of day and night. Each image frame is paired with its corresponding event frame, ensuring accurate alignment. The dataset includes a wide range of environments, such as urban, suburban, and rural settings. To thoroughly assess the robustness of our fusion approach under challenging lighting conditions, we categorize the dataset into three groups based on image exposure: well-exposed, overexposed, and underexposed frames. Of the total frames, 9,856 are used for training, while the remaining 2,006 frames are designated for testing. All image and event frames have a resolution of 640 × 480.

Implementation Details. Our method is fully implemented using Pytorch, and all experiments are conducted on a single NVIDIA RTX 3090 GPU. We adopt the ResNet as our pre-trained backbone. All input images are resized to 320 × 640. For data augmentation, following established methods [26, 42], we apply random affine transformations (including translation, rotation, and scaling) and random horizontal flips. Model optimization is performed using the AdamW [29] optimizer with an initial learning rate of 1×10^{-3} and a cosine decay learning rate strategy [28], with the power set to 0.9. We set the lane-prior point count to $N = 72$, and the sampled number $N_p = 36$, training the model for 100 epochs on the DSEC-Lane dataset.

Evaluation Metrics. We evaluate lane detection performance following the methods of [49] and [39], using the Accuracy and F1-score metrics. Accuracy quantifies the proportion of correctly identified lane points within 20 pixels of the ground truth. F1-score balances precision and recall, considering Intersection over Union (IoU) above a set threshold for true positives.

Evaluation Results. In this section, we evaluate the performance of our proposed method on the DSEC-Lane dataset under various exposure conditions. As shown in Tab. 3, Ours-B achieves state-of-the-art performance, surpassing CLRKDNet[41] by 0.7% in accuracy (87.32% vs. 86.62%) and 2.25% in F1-score (81.57% vs. 79.32%) across all conditions. In the overexposure scenario, Ours-B outperforms CLRNet [66] by 0.08% in accuracy and CLRKDNet by 0.19%. In the underexposure case, Ours-B shows a more significant improvement, surpassing CLRNet by 0.91% and CLRKDNet by 0.63% in accuracy. Additionally, in normal conditions, Ours-B consistently outperforms both CLRNet and CLRKDNet, with an accuracy increase of 1.03% and 0.87%, respectively (Ours-S represents the model using ResNet-18 architecture, while Ours-B corresponds to the model utilizing ResNet-34 architecture). These results demonstrate the effectiveness and robustness of our approach across a wide range of lighting conditions, particularly in extreme exposure scenarios.

We further visualize the experimental results in Figure 3. As depicted in the first row, CLRNet fails to detect lane lines under overexposure, while CLRKDNet loses key structural features of the lanes. Furthermore, in third and fourth rows, LaneATT [46], CLRNet

Table 3: Lane-detection results on the DSEC-Lane dataset for each method under overall, normal, overexposure, and underexposure conditions. Acc.(%) represents Accuracy(%), and F1(%) represents F1-score(%). Ours-S and Ours-B represent the small and base versions of our proposed method, respectively.

Method	Overall		Normal		Overexposure		Underexposure	
	Acc.(%)↑	F1(%)↑	Acc.(%)↑	F1(%)↑	Acc.(%)↑	F1(%)↑	Acc.(%)↑	F1(%)↑
LaneATT [46]	84.80	73.42	85.48	70.04	83.16	73.11	85.05	74.21
CLRNet [66]	86.48	79.97	86.32	79.95	87.97	78.51	85.82	81.28
ADNet [60]	86.38	79.41	86.02	79.42	87.73	77.36	86.75	81.16
SRLane [6]	86.39	79.74	86.52	80.51	87.94	77.20	84.42	78.61
CLRKDNet [41]	86.62	79.32	86.48	79.70	87.86	77.07	86.10	79.69
Ours-S	86.92	81.41	86.67	81.48	87.99	**79.07**	86.10	**82.80**
Ours-B	**87.32**	**81.57**	**87.35**	**82.14**	**88.05**	78.03	**86.73**	82.21

Table 4: Ablation study on different Event and RGB fusion methods.

Method	Acc. (%)↑	F1 (%)↑
Add	86.30	80.44
Concat	86.81	80.34
CBAM [57]	86.58	80.38
ECANet [54]	86.24	80.79
CMT [55]	86.54	80.70
Ours	**87.32**	**81.57**

Table 5: Ablation study on edge guidance in mutual attention. w/o denotes the use of only event and image features in Mutual Attention (MA), while w/ incorporates edge guidance to enhance the attention mechanism.

Guidance	Operator	Acc. (%)↑	F1 (%)↑
w/o edge	-	86.00	80.49
w/ edge	Canny	86.45	80.91
	Sobel	**87.32**	**81.57**

and CLRKDNet exhibit missing lane detections under blackkit conditions. In contrast, our method consistently demonstrates superior robustness, maintaining accurate lane detection even under these challenging extreme lighting conditions.

Ablation of fusion method. In Tab. 4., we compare our proposed fusion method with several baseline approaches, including addition, concatenation, CBAM [57], ECANet [54] and CMT [55]. Our method outperforms all baseline methods in both accuracy and F1-score, demonstrating the superior effectiveness of our fusion module.

Ablation of edge gudiance. The ablation study in Tab. 5. examines the impact of edge guidance in the fusion of event and image data for lane detection. Without the edge component, the model performs poorly, emphasizing the importance of edge guidance. Using the Canny operator improves lane boundary detection, and the Sobel operator yields the best performance. Compared to the implicit and high-level features extracted by an image encoder, Sobel provides explicit gradient-based edge cues, which highlight boundary structures more effectively and preserve fine details critical for lane markings. These results show that different edge methods provide valuable structural cues, enhancing lane detection. Incorporating edge guidance alongside event and image data is crucial for more accurate and robust lane detection.

Table 6: Ablation study on model components: Baseline is concat fusion, Time refers to the temporal information, and Edge represents the edge features.

Baseline	Time	Edge	Acc.(%)↑	F1(%)↑
✓	✗	✗	86.81	80.34
✓	✓	✗	86.70	81.27
✓	✗	✓	87.22	81.00
✓	✓	✓	**87.32**	**81.57**

Ablation of component. The ablation study in Tab. 6. demonstrates the importance of temporal and edge components in RGB and Event fusion for lane detection. The temporal module captures the dynamic nature of lane markings, while the edge component addresses the loss of boundary details in low-motion areas, ensuring accurate lane detection. Both components are essential for improving robustness and accuracy in challenging scenarios.

Ablation of representation. In Tab. 7, the timestamp representation consistently outperforms the voxel approach for lane detection. While voxel representation aggregates events into discrete time bins, potentially losing crucial temporal details, timestamp encoding retains the exact time of each event, preserving fine-grained information crucial for accurately capturing the structural continuity of lane markings. Voxelization, by focusing on the overall scene transformation and motion, does not emphasize the edge details as effectively. In contrast, timestamp representation excels at maintaining edge information, which is key for detecting lane boundaries under dynamic lighting and motion conditions. This makes timestamp encoding a more effective choice for lane detection, where precise edge and boundary details are essential.

Table 7: Ablation study on different event representations.

Representation	Acc. (%)↑	F1 (%)↑
Voxel(bin=3)	86.01	80.38
Voxel(bin=6)	86.31	80.60
Ours(Timestamp)	**87.32**	**81.57**

5 Conclusion

In this work, we design a pipeline for lane detection that combines RGB and event camera data, addressing the limitations of standard RGB cameras in challenging lighting conditions. The pipeline incorporates Multiple Feature Fusion Blocks (MFFB) to capture temporal correlations in lane markings and integrates edge features to compensate for the lack of lane boundary information in event cameras under low motion. To support this approach, we introduce DSEC-Lane, the first dataset with synchronized RGB and event data for lane detection. Extensive experiments demonstrate that our method outperforms state-of-the-art techniques, highlighting the effectiveness of event-image fusion in enhancing lane detection robustness under extreme lighting conditions.

Acknowledgement. This work was supported by the National Natural Science Foundation of China under grant No. 62302045, and the Beijing Institute of Technology Special-Zone.

References

[1] Inigo Alonso and Ana C Murillo. Ev-segnet: Semantic segmentation for event-based cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[2] Mohamed Aly. Real time detection of lane markers in urban streets. In *2008 IEEE intelligent vehicles symposium*, pages 7–12. IEEE, 2008.

[3] Meghana Basavaraj, Upendra Suddamalla, and Shenxin Xu. Lanenet++: Uncertainty-aware lane detection for autonomous vehicle. In *International Symposium on Visual Computing*, pages 245–258. Springer, 2023.

[4] Karsten Behrendt and Ryan Soussan. Unsupervised labeled lane markers using maps. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

[5] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.

[6] Chao Chen, Jie Liu, Chang Zhou, Jie Tang, and Gangshan Wu. Sketch and refine: Towards fast and accurate lane detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1001–1009, 2024.

[7] Guang Chen, Hu Cao, Jorg Conradt, Huajin Tang, Florian Rohrbein, and Alois Knoll. Event-based neuromorphic vision for autonomous driving: A paradigm shift for bio-inspired visual sensing and perception. *IEEE Signal Processing Magazine*, 37(4):34–49, 2020.

[8] Yingjie Chen, Jiarui Zhang, Tao Wang, and Yun Liang. Eventformer: Au event transformer for facial action unit event detection. In *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*. BMVA, 2023.

[9] Wensheng Cheng, Hao Luo, Wen Yang, Lei Yu, Shoushun Chen, and Wei Li. Det: A high-resolution dvs dataset for lane extraction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[10] Ziluo Ding, Rui Zhao, Jiyuan Zhang, Tianxiao Gao, Ruiqin Xiong, Zhaofei Yu, and Tiejun Huang. Spatio-temporal recurrent networks for event-based optical flow estimation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 525–533, 2022.

[11] Richard O Duda and Peter E Hart. Use of the hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15(1):11–15, 1972.

[12] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaeser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1341–1360, 2020.

[13] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020.

[14] Aditya Nalgunda Ganesh and Gowri Srinivasa. D^3nav: Data-driven driving agents for autonomous vehicles in unstructured traffic. In *35th British Machine Vision Conference 2024, BMVC 2024, Glasgow, UK, November 25-28, 2024*. BMVA, 2024.

[15] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021.

[16] Ian J Goodfellow. Piecewise linear multilayer perceptrons and dropout. *arXiv preprint arXiv:1301.5088*, 2013.

[17] Jianhua Han, Xiajun Deng, Xinyue Cai, Zhen Yang, Hang Xu, Chunjing Xu, and Xiaodan Liang. Laneformer: Object-aware row-column transformers for lane detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 799–807, 2022.

[18] Hiroto Honda and Yusuke Uchida. Clrernet: improving confidence of lane detection with laneiou. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1176–1185, 2024.

[19] Kunping Huang, Sen Zhang, Jing Zhang, and Dacheng Tao. Event-based simultaneous localization and mapping: A comprehensive survey. *arXiv preprint arXiv:2304.09793*, 2023.

[20] Zhe Jiang, Yu Zhang, Dongqing Zou, Jimmy Ren, Jiancheng Lv, and Yebin Liu. Learning event-based motion deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3320–3329, 2020.

[21] Xavier Lagorce, Garrick Orchard, Francesco Galluppi, Bertram E Shi, and Ryad B Benosman. Hots: a hierarchy of event-based time-surfaces for pattern recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1346–1359, 2016.

[22] Seokju Lee, Junsik Kim, Jae Shin Yoon, Seunghak Shin, Oleksandr Bailo, Namil Kim, Tae-Hee Lee, Hyun Seok Hong, Seung-Hoon Han, and In So Kweon. Vpgnet: Vanishing point guided network for lane and road marking detection and recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1947–1955, 2017.

[23] Jianing Li, Siwei Dong, Zhaofei Yu, Yonghong Tian, and Tiejun Huang. Event-based vision enhanced: A joint detection framework in autonomous driving. In *2019 ieee international conference on multimedia and expo (icme)*, pages 1396–1401. IEEE, 2019.

[24] Jiarong Lin and Fu Zhang. R 3 live: A robust, real-time, rgb-colored, lidar-inertial-visual tightly-coupled state estimation and mapping package. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 10672–10678. IEEE, 2022.

[25] Xinxu Lin, Mingxuan Liu, Kezhuo Liu, and Hong Chen. Spike-slr: An energy-efficient parallel spiking transformer for event-based sign language recognition. 2024.

[26] Lizhe Liu, Xiaohao Chen, Siyu Zhu, and Ping Tan. Condlanenet: a top-to-down lane detection framework based on conditional convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3773–3782, 2021.

[27] A López, J Serrat, C Canero, F Lumbreras, and T Graf. Robust lane markings detection and road geometry computation. *International Journal of Automotive Technology*, 11: 395–407, 2010.

[28] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[30] Xinglong Luo, Ao Luo, Zhengning Wang, Chunyu Lin, Bing Zeng, and Shuaicheng Liu. Efficient meshflow and optical flow estimation from event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19198–19207, 2024.

[31] Ana I Maqueda, Antonio Loquercio, Guillermo Gallego, Narciso García, and Davide Scaramuzza. Event-based vision meets deep learning on steering prediction for self-driving cars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5419–5427, 2018.

[32] Nico Messikommer, Stamatios Georgoulis, Daniel Gehrig, Stepan Tulyakov, Julius Erbach, Alfredo Bochicchio, Yuanyou Li, and Davide Scaramuzza. Multi-bracket high dynamic range imaging with event cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 547–557, 2022.

[33] Weidong Min, Ruikang Liu, Daojing He, Qing Han, Qingting Wei, and Qi Wang. Traffic sign recognition based on semantic scene understanding and structural traffic sign location. *IEEE Transactions on Intelligent Transportation Systems*, 23(9):15794–15807, 2022.

[34] Anton Mitrokhin, Cornelia Fermüller, Chethan Parameshwara, and Yiannis Aloimonos. Event-based moving object detection and tracking. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–9. IEEE, 2018.

[35] Davy Neven, Bert De Brabandere, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Towards end-to-end lane detection: an instance segmentation approach. In *2018 IEEE intelligent vehicles symposium (IV)*, pages 286–291. IEEE, 2018.

[36] Anh Nguyen, Thanh-Toan Do, Darwin G Caldwell, and Nikos G Tsagarakis. Real-time 6dof pose relocalization for event cameras with stacked spatial lstm networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[37] Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, and Yuchao Dai. 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*, pages 6820–6829, 2019.

[38] Liyuan Pan, Miaomiao Liu, and Richard Hartley. Single image optical flow estimation with an event camera. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1669–1678. IEEE, 2020.

[39] Xingang Pan, Jianping Shi, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Spatial as deep: Spatial cnn for traffic scene understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[40] Paul KJ Park, Baek Hwan Cho, Jin Man Park, Kyoobin Lee, Ha Young Kim, Hyo Ah Kang, Hyun Goo Lee, Jooyeon Woo, Yohan Roh, Won Jo Lee, et al. Performance

improvement of deep learning based gesture recognition using spatiotemporal demosaicing technique. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1624–1628. IEEE, 2016.

[41] Weiqing Qi, Guoyang Zhao, Fulong Ma, Linwei Zheng, Jun Ma, and Ming Liu. Clrkdnet: Speeding up lane detection with knowledge distillation. In *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, pages 679–686. IEEE, 2024.

[42] Zhan Qu, Huan Jin, Yang Zhou, Zhen Yang, and Wei Zhang. Focus on local: Detecting lane marker from bottom up via key point. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14122–14130, 2021.

[43] Timo Stoffregen, Guillermo Gallego, Tom Drummond, Lindsay Kleeman, and Davide Scaramuzza. Event-based motion segmentation by motion compensation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7244–7253, 2019.

[44] Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, and Luc Van Gool. Event-based fusion for motion deblurring with cross-modal attention. In *European conference on computer vision*, pages 412–428. Springer, 2022.

[45] Zhaoning Sun, Nico Messikommer, Daniel Gehrig, and Davide Scaramuzza. Ess: Learning event-based semantic segmentation from still images. In *European Conference on Computer Vision*, pages 341–357. Springer, 2022.

[46] Lucas Tabelini, Rodrigo Berriel, Thiago M Paixao, Claudine Badue, Alberto F De Souza, and Thiago Oliveira-Santos. Keep your eyes on the lane: Real-time attention-guided lane detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 294–302, 2021.

[47] Abhishek Tomy, Anshul Paigwar, Khushdeep S Mann, Alessandro Renzaglia, and Christian Laugier. Fusing event-based and rgb camera for robust object detection in adverse conditions. In *2022 International conference on robotics and automation (ICRA)*, pages 933–939. IEEE, 2022.

[48] Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17755–17764, 2022.

[49] TuSimple. Tusimple dataset. https://cvpr2017.tusimple.ai/, 2017.

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[51] Aayush Atul Verma, Bharatesh Chakravarthi, Arpitsinh Vaghela, Hua Wei, and Yezhou Yang. etram: Event-based traffic monitoring dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22637–22646, 2024.

[52] Celyn Walters and Simon Hadfield. Ceril: Continuous event-based reinforcement learning. In *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*. BMVA, 2023.

[53] Hanqing Wang, Wei Liang, Luc Van Gool, and Wenguan Wang. Dreamwalker: Mental planning for continuous vision-language navigation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10873–10883, 2023.

[54] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11534–11542, 2020.

[55] Xiao Wang, Jianing Li, Lin Zhu, Zhipeng Zhang, Zhe Chen, Xin Li, Yaowei Wang, Yonghong Tian, and Feng Wu. Visevent: Reliable object tracking via collaboration of frame and event flows. *IEEE Transactions on Cybernetics*, 54(3):1997–2010, 2023.

[56] Ze Wang, Weiqiang Ren, and Qiang Qiu. Lanenet: Real-time lane detection networks for autonomous driving. *arXiv preprint arXiv:1807.01726*, 2018.

[57] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[58] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[59] Zongwei Wu, Guillaume Allibert, Christophe Stolz, Chao Ma, and Cédric Demonceaux. Modality-guided subnetwork for salient object detection. In *2021 International Conference on 3D Vision (3DV)*, pages 515–524. IEEE, 2021.

[60] Lingyu Xiao, Xiang Li, Sen Yang, and Wankou Yang. Adnet: Lane shape prediction via anchor decomposition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6404–6413, 2023.

[61] Hang Xu, Shaoju Wang, Xinyue Cai, Wei Zhang, Xiaodan Liang, and Zhenguo Li. Curvelane-nas: Unifying lane-sensitive architecture search and adaptive point blending. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 689–704. Springer, 2020.

[62] Hao Yang, Liyuan Pan, Yan Yang, and Wei Liang. Language-driven all-in-one adverse weather removal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24902–24912, 2024.

[63] Ruikun Zhang, Hao Yang, Yan Yang, Ying Fu, and Liyuan Pan. Lmhaze: intensity-aware image dehazing with a large-scale multi-intensity real haze dataset. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia*, pages 1–1, 2024.

[64] Fengan Zhao, Qianang Zhou, and Junlin Xiong. Edge-guided fusion and motion augmentation for event-image stereo. In *European Conference on Computer Vision*, pages 190–205. Springer, 2024.

[65] Weikun Zhen and Sebastian Scherer. Estimating the localizability in tunnel-like environments using lidar and uwb. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4903–4908. IEEE, 2019.

[66] Tu Zheng, Yifei Huang, Yang Liu, Wenjian Tang, Zheng Yang, Deng Cai, and Xiaofei He. Clrnet: Cross layer refinement network for lane detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 898–907, 2022.

[67] Chu Zhou, Minggui Teng, Jin Han, Jinxiu Liang, Chao Xu, Gang Cao, and Boxin Shi. Deblurring low-light images with events. *International Journal of Computer Vision*, 131(5):1284–1298, 2023.

[68] Zhuyun Zhou, Zongwei Wu, Rémi Boutteau, Fan Yang, Cédric Demonceaux, and Dominique Ginhac. Rgb-event fusion for moving object detection in autonomous driving. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7808–7815. IEEE, 2023.

[69] Zhuyun Zhou, Zongwei Wu, Danda Pani Paudel, Rémi Boutteau, Fan Yang, Luc Van Gool, Radu Timofte, and Dominique Ginhac. Event-free moving object segmentation from moving ego vehicle. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8960–8965. IEEE, 2024.

[70] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3(3):2032–2039, 2018.

[71] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Ev-flownet: Self-supervised optical flow estimation for event-based cameras. *arXiv preprint arXiv:1802.06898*, 2018.

[72] Lin Zhu, Xianzhang Chen, Lizhi Wang, Xiao Wang, Yonghong Tian, and Hua Huang. Continuous-time object segmentation using high temporal resolution event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[73] Hao Zhuang, Zheng Fang, Xinjie Huang, Kuanxu Hou, Delei Kong, and Chenming Hu. Ev-mgrflownet: Motion-guided recurrent network for unsupervised event-based optical flow with hybrid motion-compensation loss. *IEEE Transactions on Instrumentation and Measurement*, 73:1–15, 2024.