PEOD: A Pixel-Aligned Event-RGB Benchmark for Object Detection Under Challenging Conditions

Luoping Cui*[1], Hanqing Liu*[1], Mingjie Liu[1], Endian Lin[1], Donghong Jiang[1], Yuhao Wang[1], Chuang Zhu[1†]

[1]School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China
{ lpcui, hanqingliu, LMJ, ledgogo, donghongjiang, wyhao, czhu }@bupt.edu.cn

Abstract

Robust object detection for challenging scenarios increasingly relies on event cameras, yet existing Event-RGB datasets remain constrained by sparse coverage of extreme conditions and low spatial resolution ($\leq 640 \times 480$), which prevents comprehensive evaluation of detectors under challenging scenarios. To address these limitations, we propose PEOD, the first large-scale, pixel-aligned and high-resolution (1280×720) Event-RGB dataset for object detection under challenging conditions. PEOD contains 130+ spatiotemporally aligned sequences and 340k manually annotated bounding boxes, with 57% of the data captured under low-light, overexposure, and high-speed motion. Furthermore, we benchmark 14 methods across three input configurations (Event-based, RGB-based, and Event-RGB fusion) on PEOD. On the full test set and normal subset, fusion-based models achieve excellent performance. However, in illumination challenge subset, the top event-based model outperforms all fusion models, while fusion models still outperform their RGB-based counterparts, indicating limits of existing fusion methods when the frame modality is severely degraded. PEOD establishes a realistic, high-quality benchmark for multimodal perception and facilitates future research.

Datasets — https://github.com/bupt-ai-cz/PEOD

Introduction

Object detection is a critical perception task for intelligent systems, including robotics (Falanga, Kleber, and Scaramuzza 2020), autonomous vehicles (Teng et al. 2023), and surveillance (Wei et al. 2021; Du et al. 2023). However, conventional frame-based cameras suffer from inherent limitations in exposure time and dynamic range, leading to low-quality images and significant information loss in challenging scenarios such as high-speed motion and varying illumination (Dai et al. 2023). Event cameras, which asynchronously report per-pixel brightness changes, offer a new paradigm with their microsecond-level temporal resolution and high dynamic range, demonstrating superior performance in extreme conditions (Prophesee 2024). However, event data lack the rich texture and static scene information that are core strengths of RGB cameras. Consequently, fusing RGB frames and event streams is a highly promising approach for building robust, all-day, all-scenario detection systems (Gehrig and Scaramuzza 2024).

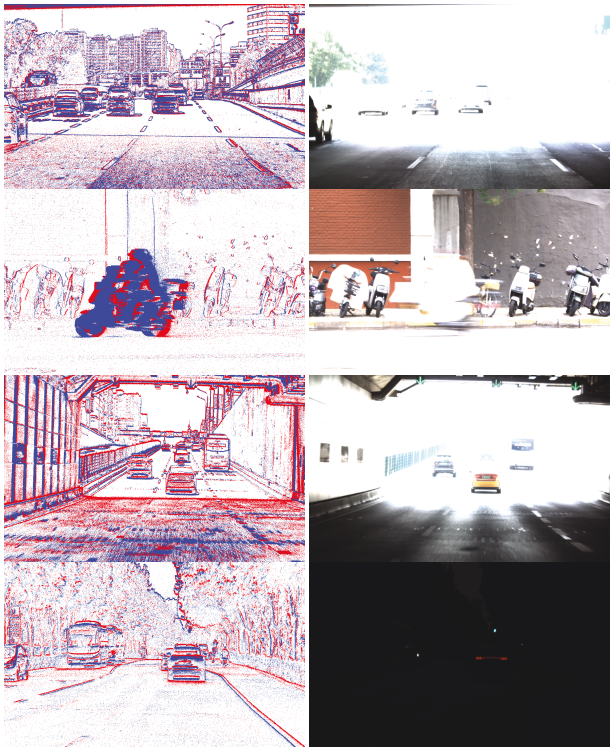


Figure 1: PEOD examples under diverse challenging conditions. Overexposure (rows 1-3), motion blur (row 2), and low-light (row 4). Each row presents the event stream (left) with its pixel-aligned RGB frame (right).

While existing dual-modality datasets such as DSEC (Gehrig et al. 2021), PKU-SOD (Li, Tian, and Li 2023), have enabled initial research, they suffer from critical limitations: **1)** ***Scarcity of Extreme Scenarios***: Challenging data (e.g., night, overexposure, motion blur) constitutes less than 20% of existing datasets, hindering the proper evaluation of model robustness. **2)** ***Low Resolution***: Most datasets feature

*These authors contributed equally.
†Corresponding author.

resolutions like 346×260 or 640×480, which are insufficient for modern detectors that require fine-grained detail.

To overcome these bottlenecks, we introduce a new, large-scale pixel-aligned Event-RGB dataset for object detection (PEOD), as Figure 1. It is the first dataset of its kind, captured with a high-resolution (1280×720) EVK4 event camera (Prophesee 2024) and an RGB camera, using a beam-splitter optical system and a hardware signal generator to achieve spatiotemporal synchronization. The dataset covers diverse driving environments, with over 57% of the data captured in extreme conditions, including low-light, overexposure, and high-speed motion. We establish a comprehensive benchmark to unify the evaluation of existing event and Event-RGB fusion algorithms on the PEOD dataset. Our main contributions can be summarized as follows:

- We introduce a first-of-its-kind dataset containing more than 57% of the data collected under extreme conditions (low-light, overexposure, and high-speed motion).
- We provide the first high-resolution (1280×720), pixel-aligned Event-RGB dataset, with 340k manually annotated bounding boxes in six traffic-related object classes.
- We establish a comprehensive benchmark by evaluating 14 classical and state-of-the-art object detectors under RGB-based, event-based, and Event-RGB fusion settings on both the full dataset and challenging subsets.

Related Work

Event and Event-RGB Vision Datasets

Numerous event-based datasets have been introduced to address a diverse set of vision tasks, including object detection, object tracking, and anomaly detection. For object detection task, Gen1 dataset (De Tournemire et al. 2020) offers limited resolution, but only annotates two object categories. The 1 Mpx dataset offers event streams accompanied by annotations mapped from frames. eTram dataset (Verma et al. 2024) focuses on monitoring of traffic scenes, providing an all-day dataset. PEDRo dataset (Boretti et al. 2023) is dedicated to pedestrian detection in service robotics. DSEC dataset (Gehrig et al. 2021) employs synchronized Gen3.1 stereo event cameras along with RGB, LiDAR, and RTK-GPS, collecting data across diverse environments. PKU-SOD dataset (Li, Tian, and Li 2023), collected with a DAVIS346 sensor, is a large-scale Event-RGB benchmark with three-class annotations. For object tracking task, EventVOT dataset (Wang et al. 2024) provides a large-scale, high-resolution benchmark, comprising 1,141 videos that cover 19 object categories. For 3D perception tasks, MVSEC dataset (Zhu et al. 2018) extends to various platforms, and provides synchronized IMU and LiDAR ground truth. For gesture recognition task, EvRealHands dataset provides a real-world event-based resource for 3D hand pose estimation. (Jiang et al. 2024). For anomaly detection task, UCF-Crime-DVS (Qian et al. 2025) dataset captures event streams for video anomaly detection based on the UCF-Crime dataset. For image reconstruction task, RLED dataset (Liu et al. 2024a) uses a coaxial imaging setup to capture 64k synchronized RGB frames and event streams, offering a benchmark for nighttime reconstruction.

Existing event and Event-RGB datasets, although valuable, are still constrained by low spatial resolution, sparse annotations, or a narrow sampling of scenes captured under challenging illumination, which together limit their utility for comprehensive benchmarking. Therefore, our goal is to construct a dataset that simultaneously provides high spatial resolution, densely annotated ground truth, and large-scale coverage of adverse lighting scenarios.

Object Detection with Event Cameras

Research in event-based object detection now follows two main directions: event-based detectors and Event-RGB fusion detectors. Event-based detectors range from CNN slice-based approaches (Li et al. 2022; Fan et al. 2024a) to Transformer variants such as RVT (Gehrig and Scaramuzza 2023) and SAST (Peng et al. 2024), as well as SNN-based models (Luo et al. 2024; Fan et al. 2024b), yet all remain constrained by an inherent insensitivity to texture details and a limited ability to detect slow-moving or stationary objects in event data. These inherent limitations motivate the second paradigm: Event-RGB fusion. Fusion detectors overcome these limitations by combining dense RGB texture information with event-derived motion cues (Gehrig and Scaramuzza 2024). Initial fusion methods rely on a cascade policy of converting event streams into pseudo-frames for simple concatenation with RGB features, which offers marginal improvements in robustness (Tomy et al. 2022). To facilitate more effective feature interaction, attention mechanisms have been introduced into fusion frameworks, employing temporal Transformers and asynchronous cross-modal attention for bimodal feature integration (Li, Tian, and Li 2023). Moreover, the incorporation of advanced strategies like multi-scale feature aggregation, bi-directional calibration, and illumination-adaptive compensation has further boosted model accuracy and robustness (Liu et al. 2024b).

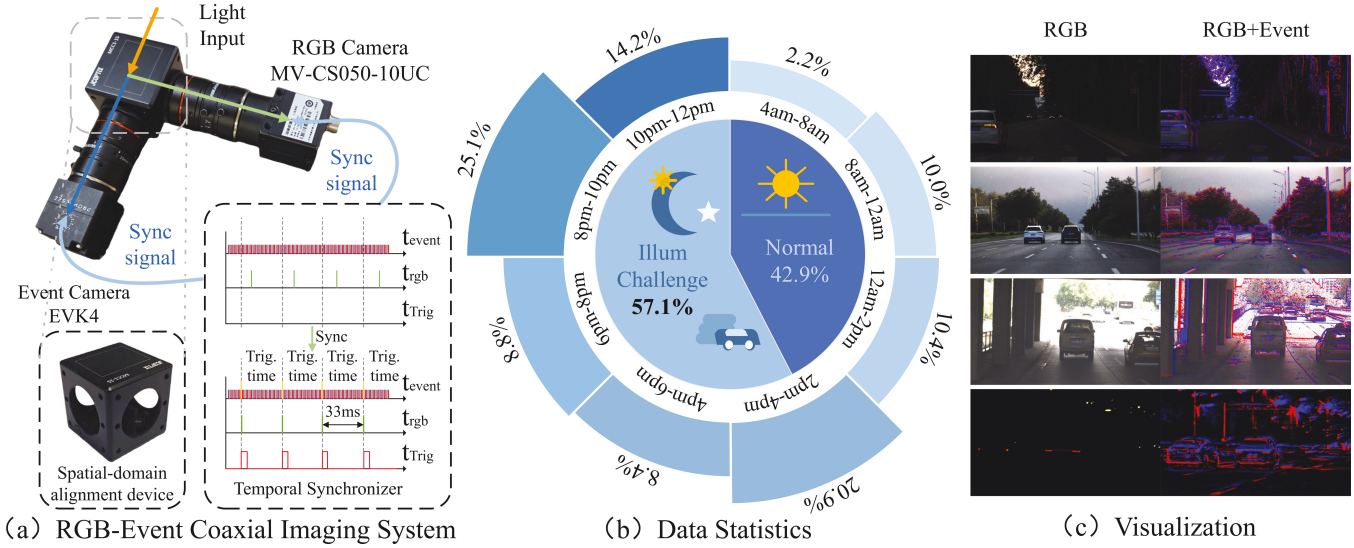
Overall, while event-based detectors excel in handling challenging scenarios such as extreme lighting and motion blur, they still face limitations in detecting texture details and static objects. By exploiting complementary information from RGB frames and event streams, Event-RGB fusion detectors achieve robust detection performance across diverse illumination conditions.

PEOD Dataset

In this section, we systematically introduce a high-resolution, **P**ixel-aligned **E**vent-RGB **O**bject **D**etection Dataset under challenging scenarios. We provide a comprehensive analysis of our data acquisition system, collection methodology, and the composition of the dataset.

Dual-Camera Sync System

To construct a high-resolution Event-RGB dataset with strict spatiotemporal alignment and a unified imaging scale, we developed a coaxial optical system. This system comprises a JCOPTIX OSB25R55-T5 non-polarizing plate beam splitter (50:50 split ratio) and an MCC1-1S $10mm$ coaxial cube, as illustrated in Figure 2(a). This configuration allows an



(a) RGB-Event Coaxial Imaging System (b) Data Statistics (c) Visualization

Figure 2: Overview of PEOD dataset and acquisition system. (a) The coaxial imaging system used to capture spatiotemporally aligned Event and RGB data. (b) Temporal distribution of the dataset, with 57.1% captured under challenging illumination conditions. (c) Sample aligned Event-RGB pairs from diverse driving scenarios.

event camera and an RGB camera to share the same optical path, enabling pixel-level spatial alignment through a calibration-guided rectification procedure. For precise temporal synchronization, a single square-wave signal generator provides hardware trigger pulses to both cameras, achieving microsecond-level accuracy. The event stream is captured by a Prophesee EVK4 HD camera (1280×720), while RGB frames are synchronously acquired by a Hikvision MV-CS050-10UC industrial camera ($2448 \times 2048, 60FPS$). To eliminate discrepancies in focal length and lens distortion, we equipped both cameras with identical Hikvision $25mm$ C-mount fixed-focal-length lenses and maintained a fixed aperture setting for all recordings.

Data Collection and Annotation

Using the acquisition system mounted behind a car's front windshield, we recorded all sequences at 30Hz. Data were collected continuously from 04:00 h to 24:00 h, covering lighting conditions that range from dawn to nighttime, and across diverse environments such as urban roads, suburban roads, complex intersections, tunnels, and highways.

Given the prevalence of challenging conditions such as high speeds, low-light, and overexposure, we adopted a hybrid annotation strategy to ensure the accuracy of the labels. For normal conditions, annotations were performed directly on the high-quality RGB frames. For challenging conditions, we leveraged an advanced reconstruction algorithm, NER-Net, to generate grayscale images from the asynchronous event streams, matching the RGB camera's frequency. Annotations were then carried out on these clear, reconstructed frames. Under normal lighting, bounding boxes were annotated directly on the raw RGB frames. Under challenging conditions, we directly annotated the high-clarity reconstructions generated by NER-Net (Liu et al. 2024a). Our team manually annotated six common classes (car, bus, truck, two-wheeler, three-wheeler, and person), each of which appears in more than 30% of the recorded instances. All annotations underwent a rigorous cross-checking review process to ensure high quality and consistency.

Data Statistics

The PEOD dataset comprises over 130 sequences, with the longest lasting nearly 90 s, and includes more than 72k annotated frames and 340k bounding boxes across six categories, along with fixed training and test splits for evaluation.

Motion-blurred and sharp frames frequently coexist within the same driving sequence, and our statistics indicate that roughly 40-50% of frames in nominally high-speed segments are blur-free. Assigning blur labels at the frame level would fragment temporal context and compromise sequence-wise evaluation. Therefore, we center our dataset split on illumination, the factor where RGB sensors degrade the most yet event cameras excel. To quantitatively identify and categorize these conditions, we define an underexposure score (S_{LL}) and an overexposure score (S_{OE}) based on the pixel saturation in each grayscale frame F. The formulations are as follows:

$$S_{LL} = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} \mathbf{I}(F_{ij} < T_{\text{dark}}) \tag{1}$$

$$S_{OE} = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} \mathbf{I}(F_{ij} > T_{\text{bright}}) \tag{2}$$

where I is the indicator function, and T_{dark} and T_{bright} are predefined thresholds for dark and bright saturation, respectively. Using thresholds of 30 and 250, a frame is classified into a specific subset (e.g., low-light, overexposed) if

Dataset	Year	Resolution	Modality	Boxes	Classes	Manual	Real	HS	LL ($\mu Y_{709} < 1$)	Ext.(%)
Gen1	2020	304×240	Event	255K	2	✓	✓	✗	✓	-
1 Mpx	2020	1280×720	Event	25M	7	✗	✓	✗	✓	-
PEDRo	2023	346×260	Event	43K	1	✓	✓	✗	✗	-
eTraM	2024	1280×720	Event	2M	3	✓	✓	✗	✓	-
SEVD	2024	800×600	Event	9M	6	✗	✗	-	-	-
Event-KITTI	2024	1333×401	Event	80K	8	✓	✗	-	-	-
DSEC	2021	640×480	Frame, Event	390K	8	✗	✓	✗	✗	20
PKU-SOD	2022	346×260	Frame, Event	1080K	3	✗	✓	✓	✗	13
PEOD(Ours)	2025	$\mathbf{1280 \times 720}$	**Frame, Event**	**340K**	6	✓	✓	✓	✓	**57**

Table 1: Comparison with existing object detection datasets. PEOD dataset is the first to provide a high-resolution, 6-class benchmark annotated at 30Hz, with over 57% of data focusing on extreme scenarios. LL: Mean BT.709 luminance of frames (μY_{709}) in the 0-255 domain. HS: High-speed scenarios. Ext.(%): Proportion of sequences collected under extreme conditions.

its corresponding score exceeds a certain percentage threshold. Based on illumination conditions, we divide the dataset into two subsets: **1)** ***Illumination Challenge Subset***: sequences recorded under challenging illumination conditions (e.g., low-light, overexposure, abrupt lighting changes). **2)** ***Normal Subset***: sequences recorded under standard lighting conditions. The approximate distribution across these subsets is 57.1% extreme lighting scenarios and 42.9% normal lighting scenarios.

Comparison with Other Datasets

In Table 1, we compare our PEOD dataset with other object detection datasets. In contrast, other large-scale public event-based datasets, such as Gen1 (De Tournemire et al. 2020) and 1 Mpx dataset (Perot et al. 2020), offer only long-duration event streams. This limitation hinders the development of high-precision, all-day object detection systems, particularly in static or extremely slow-moving scenarios. Datasets like PEDRo (Boretti et al. 2023) and eTram (Verma et al. 2024) are tailored for niche applications, focusing on event-based pedestrian detection and traffic flow monitoring from a surveillance perspective, respectively. Furthermore, while SEVD (Aliminati et al. 2024) and Event-KITTI (Zhou, Chang, and Shi 2024) provide large-scale event data, they rely on event simulators, creating a significant domain gap compared to data captured by real-world event cameras. More importantly, all the aforementioned datasets exclusively provide event streams. Among datasets that offer aligned RGB and event data, DSEC (Gehrig et al. 2021) and PKU-SOD (Li, Tian, and Li 2023) are constrained by their low spatial resolution, which limits the performance of detection algorithms that require fine-grained detail. Moreover, challenging scenarios constitute less than 20% of their data, rendering them inadequate for a comprehensive evaluation of model robustness under adverse conditions.

Overall, our PEOD dataset offers four key advantages: **1) Challenging Scenarios:** Over 57% of the dataset comprises sequences captured in difficult conditions. **2) High Spatiotemporal Resolution:** The dataset features high spatial resolution of 1280×720, complemented by the microsecond-level temporal resolution. **3) High Dynamic Range:** The event camera ensures an HDR of over 120dB, preserving signal integrity in extreme Illumination conditions. **4) Dense and Diverse Annotations:** The data stream is continuously annotated with 6 object classes at 30Hz.

Evaluation and Benchmark

Experimental Setup

All experiments are conducted on our proposed PEOD dataset, adhering strictly to the official train-test splits detailed in Section 3. In addition to evaluating model performance on the complete test set, we specifically assess robustness under distinct illumination conditions using the Illumination Challenge and Normal subsets. Three distinct categories of detectors are systematically evaluated: Event-based, RGB-based, and Event-RGB fusion approaches. For Event-based models such as RVT (Gehrig and Scaramuzza 2023), SAST (Peng et al. 2024), and SMamba (Yang et al. 2025), asynchronous event streams are encoded into tensor representations using a stacked-histogram method (Gehrig and Scaramuzza 2023), accumulating events at the original spatial resolution within a fixed 33 ms temporal window partitioned into 10 bins. These models are uniformly trained for 120k iterations. For the SNN-based detector (SpikingYOLO), we strictly follow its original methodology by collecting events within the 250ms preceding each annotation, evenly splitting this interval into two temporal segments for event integration. For RGB-based and Event-RGB fusion detectors, events are first transformed into event images with an integration window of 33ms. Detectors in these two categories are trained for 20 epochs. All three categories (Event-based, RGB-based, and Event-RGB fusion) are trained on 4× NVIDIA RTX 4090 GPUs. Hyperparameter and training settings mostly follow the original configurations.

Evaluation Metrics. To enable fair comparison across the three detector categories, we report 1) COCO-style mean Average Precision (mAP) on the PEOD benchmark, 2) inference time per image (ms), and 3) model size measured by the number of parameters. For mAP, we provide scores at multiple IoU thresholds: $\text{mAP}_{50:95}$, mAP_{50}, and mAP_{75}.

Input	Method	Pub. & Year	Backbone	mAP	mAP_{50}	mAP_{75}	Param(M)	FLOPs(G)	T(ms)
Event	YOLOX(Event)	arXiv'21	CSPDN	16.1	28.3	16.0	8.9	32.1	6.5
	RVT	CVPR'23	Transformer	20.1	38.4	18.8	4.4	31.2	9.3
	SAST	CVPR'24	Transformer	18.1	37.8	16.7	4.5	37.1	19.1
	SpikingYOLO	ECCV'24	SNN	10.2	21.8	7.9	23.1	136.7	53.8
	SMamba	AAAI'25	SSM	22.9	43.8	19.9	23.7	72.8	38.7
RGB	RetinaNet	arXiv'20	ResNet-50	12.6	22.1	12.5	36.4	198.2	14.5
	YOLOX	arXiv'21	CSPDN	17.4	34.0	14.2	8.9	32.1	6.5
	YOLOv7	CVPR'23	CSPDN	19.8	37.1	19.1	6.1	31.5	6.17
	YOLOv8	2023	CSPDN	18.9	36.3	18.8	11.1	34.4	5.4
	RT-DETR	CVPR'24	ResNet-18	21.8	38.5	21.3	23.1	61.6	10.9
Event+RGB	FPN-Fusion	ICRA'22	ResNet-50	14.6	29.9	12.6	65.6	283.7	32.2
	RENet	ICRA'23	CSPDN	24.5	41.6	22.8	37.7	102.7	28.5
	SODFormer	TPAMI'23	ResNet-50	17.8	36.3	15.0	86.5	279.7	48.7
	EOLO	ICRA'24	SNN + CSPDN	26.1	45.8	26.2	46.2	100.9	22.6
	SFNet	TITS'24	CSPDN	19.5	37.9	17.7	38.7	103.8	23.8
	CAFR	ECCV'24	ResNet-50	21.4	39.6	19.3	82.1	319.9	52.4

Table 2: Results on the PEOD benchmark comparing RGB, Event, and Event-RGB detectors. (CSPDN denotes the CSPDarknet backbone). The fusion detectors show a clear performance advantage, yielding the highest mAP over single-modality methods.

Benchmark Evaluation

We conduct a comprehensive evaluation on our PEOD dataset, comparing three categories of detectors. The results, summarized in Table 2, quantitatively characterize the performance of each category.

Evaluation on Event-based Detectors. To evaluate performance on the event stream, we benchmark several detectors that represent distinct model classes, including Transformer-based (RVT and SAST), CNN-based (YOLOX) (Ge et al. 2021), SNN-based (SpikingYOLO) (Luo et al. 2024), and Mamba-based (SMamba). As reported in Table 2, SMamba attains the highest detection accuracy of 22.9%, whereas RVT and SAST exhibit marginally lower performance. Although SMamba clearly excels at capturing long-term temporal dependencies, its 38.7ms inference latency and 72.8 GFLOPs indicate that the enhanced representational capacity is achieved at the cost of considerably greater computational overhead. By contrast, SpikingYOLO achieves limited detection accuracy, a limitation attributable to the still-maturing training frameworks for SNN and their specialized hardware requirements.

Evaluation on RGB-based Detectors. We benchmark several classical frame-based object detectors, including RetinaNet (Lin et al. 2017), RT-DETR (Zhao et al. 2024) and multiple YOLO variants (Wang, Bochkovskiy, and Liao 2023; Jocher 2023; Ge et al. 2021), which frequently serve as baselines for subsequent fusion strategies. RT-DETR achieves the best performance among these models, albeit with higher latency (10.9ms) and more parameters (23.1M). On the other hand, the YOLO variants present a more balanced performance profile. The results obtained by these frame-based detectors underscore the effectiveness of conventional cameras under favorable conditions where rich texture and color information are available.

Evaluation on Event-RGB Fusion Detectors. We evaluate 6 fusion detectors, FPN-Fusion (Tomy et al. 2022), RENet (Zhou et al. 2022), SODFormer (Li, Tian, and Li 2023), EOLO (Cao et al. 2024b), SFNet (Liu et al. 2024b) and CAFR (Cao et al. 2024a). Integrating event data with RGB frames produces substantial performance gains relative to single-modality detectors. EOLO achieves the highest detection accuracy of 26.1%, representing an absolute improvement of 4.3% over RGB-based detectors and 3.2% over event-based model. Comparative analysis shows that architectures such as EOLO and RENet, which employ modality-aware fusion mechanisms, markedly surpass simple feature concatenation detectors exemplified by FPN-style fusion. These results underscore that sophisticated schemes capable of jointly leveraging frame texture features and motion cues captured by event streams are indispensable for realizing the full potential of multimodal input.

Across the three detector categories, fusion models exhibit higher latency than RGB-based detectors. However, methods such as EOLO (22.6ms) and SFNet (23.8ms) still provide practical inference speeds. Consequently, Event-RGB fusion constitutes an effective means of attaining significant improvements in detection accuracy.

Condition-Specific Evaluation

A comparative performance analysis of three detector categories was conducted on the Illumination Challenge and Normal subsets. The comprehensive results are reported in Table 3 and Table 4, respectively.

Experimental results demonstrate the superior robustness of event-based detectors. Under both normal and extreme illumination, the event-based detectors maintain essentially constant performance, attributable to the high dynamic range of event cameras and their asynchronous, sparsity-driven data representation. By contrast, RGB-based detec-

Input	Method	mAP	mAP_{50}	mAP_{75}
Event	YOLOX(Event)	17.5	33.1	16.3
	RVT	20.4	39.1	19.0
	SAST	19.1	38.5	16.5
	SpikingYOLO	10.2	22.6	7.2
	SMamba	23.2	44.5	20.3
RGB	RetinaNet	8.0	16.4	7.0
	YOLOX	10.5	24.6	8.0
	YOLOv7	12.6	31.1	13.0
	YOLOv8	11.9	27.7	11.1
	RT-DETR	13.2	30.1	13.9
Event+RGB	FPN-Fusion	11.2	23.6	9.0
	RENet	10.8	24.4	8.3
	SODFormer	10.4	23.6	7.7
	EOLO	11.5	25.2	9.2
	SFNet	11.3	28.1	8.6
	CAFR	11.2	25.2	8.6

Table 3: Results on Illumination Challenge Subset (e.g., low-light, overexposure, abrupt illumination changes) comparing RGB-based, Event-based, and Event-RGB detectors.

Input	Method	mAP	mAP_{50}	mAP_{75}
Event	YOLOX(Event)	14.0	23.2	12.0
	RVT	19.0	36.0	18.1
	SAST	16.7	35.6	17.4
	SpikingYOLO	10.1	20.1	9.7
	SMamba	21.9	41.2	19.1
RGB	RetinaNet	20.0	32.5	21.7
	YOLOX	31.7	49.8	28.7
	YOLOv7	32.9	50.5	32.5
	YOLOv8	29.3	49.2	30.6
	RT-DETR	36.5	51.2	37.6
Event+RGB	FPN-Fusion	21.3	37.8	21.8
	RENet	43.7	65.8	43.7
	SODFormer	30.8	54.0	28.8
	EOLO	45.2	66.7	48.4
	SFNet	38.9	58.3	41.3
	CAFR	28.5	47.9	29.9

Table 4: Results on the Normal Subset comparing RGB-based, Event-based, and Event-RGB detectors.

tors suffer a pronounced drop when exposed to low-light, overexposed, or high-speed conditions, where conventional frame sensors yield degraded images. In the challenging illumination scenarios, this image degradation explains the sharp decline in RGB-based detectors.

However, Event-RGB fusion detectors outperform their corresponding RGB baselines by approximately 2% in mAP, effectively compensating for missing appearance cues with event-derived edge information. Conversely, in normal scenes, rich color and texture allow RGB-based detectors to reach high accuracy, while the event-based detectors underperform because of their limited fine-grained appearance cues. Event-RGB fusion detectors attain the top performance by integrating rich texture information in RGB frames with the motion cues inherent in event data.

In normal illumination, fusion detectors effectively integrate the complementary strengths of RGB and event streams, yielding a substantial improvement in detection accuracy. In extreme illumination, fusion detectors still exceed RGB-based detectors but remain inferior to the strongest event-based detectors. The divergent performance across the two subsets exposes intrinsic weaknesses in current fusion strategies: **1)** Most architectures rely on shallow feature concatenation, lacking mechanisms to suppress the noise introduced by degraded RGB. **2)** During training, the weighting scheme disproportionately favors the texture-rich RGB branch, so event features cannot take precedence when lighting conditions deteriorate. **3)** Existing designs fail to exploit the high temporal resolution and motion cues that are unique to event data. Closing the performance gap in adverse scenarios will require reliability-aware, deeply coupled frameworks that adaptively reweight the two modalities and fully leverage event representations, thereby delivering consistent and robust perception under all illumination conditions.

Qualitative Analysis

Figure 3 contrasts representative results from an RGB-based detector (YOLOv8), an event-based detector (RVT), and two Event-RGB fusion detectors (CAFR and EOLO), alongside the ground truth. The analysis spans four characteristic scenes: (a) a normal traffic intersection; (b) a tunnel exit exhibiting severe overexposure; (c) a low-light intersection; and (d) high-speed two-wheelers affected by motion blur. These qualitative results substantiate the scenario-specific conclusions discussed earlier. **(a) Normal scene.** In well-illuminated conditions, RGB frames provide rich texture and color cues, enabling the RGB-based detector to localize most medium-sized objects reliably. In contrast, the event stream contains sparse edge-like responses and limited texture, which leads the event-based model to miss small or heavily occluded instances and to produce fragmented boxes. Fusion detectors leverage complementary cues and recover several small or partially occluded targets that are absent from the event-based predictions, while maintaining precise localization. **(b) Overexposure.** At the tunnel exit, saturation in RGB frames suppresses object contrast and the RGB-based detector consequently fails to detect multiple vehicles. Event data, however, remains informative due to its high dynamic range, allowing the event-based detector to retain object contours and maintain detections. Fusion detectors inherit this robustness: both CAFR and EOLO correctly detect vehicles that the RGB-based detector misses, illustrating how event cues compensate for severe illumination degradation. **(c) Low-light.** Under nighttime conditions, RGB frames lose contrast and texture, causing the RGB-based detector to miss targets. The event-based detector benefits from strong responses to intensity changes (e.g., headlights, motion edges) and therefore detects the main traffic participants. Fusion produces competitive results, but when the RGB modality is extremely degraded, rigid fu-

Figure 3: Representative visualization results on our PEOD dataset. (a) Traffic intersection in normal scenario. (b) Rushing cars in overexposure scenario. (c) Traffic intersection in low-light scenario. (d) High-speed moving two-wheelers with motion blur. While the RGB-based detector (YOLOv8) effectively utilizes rich textures in the normal scene (a), fusion detectors (CAFR, EOLO) and the event-based detector show decisive advantages by leveraging event data in the challenging overexposure (b), low-light (c), and motion-blur (d) conditions where the RGB-based detector fails.

sion can inject noise from the RGB branch and slightly undercut the discriminative signal of the event stream. This modality-conflict effect is consistent with our quantitative results on the Illumination Challenge subset, where current fusion strategies may underperform strong event-based baselines. **(d) Motion blur.** Pronounced blur in RGB frames smears spatial details and leads to missed or poorly localized two-wheeler instances for the RGB-based detector. Event streams preserve sharp motion-induced edges at microsecond resolution, enabling the event-based detector to retain detections. Fusion further improves the stability and tightness of bounding boxes by synergizing motion cues from the event stream with the residual textural or structural information still discernible in the blurred RGB frame.

These qualitative examples highlight the complementary nature of events and frames: RGB excels in texture-rich, well-lit scenes but degrades under saturation, darkness, or blur, while events remain reliable in those adverse conditions yet struggle with small, distant, or slow targets due to weak texture and sparse contours. By furnishing pixel-aligned Event-RGB pairs captured across a spectrum of extreme lighting and motion scenarios, PEOD supplies a large-scale, high-fidelity dataset and benchmark for systematically closing the Event-RGB fusion gap under such conditions.

Discussion and Outlook

Rich in challenging scenarios, our high-resolution PEOD dataset also promises to catalyze progress in several related fields:**(1) Image Reconstruction.** Learning-based event-to-image reconstruction remains hindered by synthetic training data that falter in real-world settings. PEOD can bridge this sim-to-real gap enabling high-fidelity reconstruction in truly adverse scenarios. **(2) Object Tracking.** Existing datasets, with their simple trajectories, lead to an overestimation of multimodal tracker performance. PEOD closes this evaluation gap by offering long, continuous sequences rich with real-world challenges like frequent occlusions and extreme lighting changes, providing a platform for long-term identity preservation, model robustness, and real-time efficiency.

Conclusion

In this paper, we introduce PEOD, a high-resolution (1280×720), pixel-aligned Event-RGB dataset and benchmark for object detection, with extensive coverage of extreme scenarios. By addressing the inadequate resolution and scarcity of adverse conditions found in existing resources, PEOD provides a solid foundation for next-generation robust perception. Our comprehensive evaluation reveals that while fusion methods achieve the best overall performance, they still fail to fully harness the event stream in challenging conditions, highlighting the need for more sophisticated fusion mechanisms. Furthermore, PEOD enables image reconstruction to bridge the sim-to-real gap, while its challenging sequences advance robust multi-object tracking especially under extreme illumination shifts.

Acknowledgments

This work was supported by the National Key R&D Program of China (2021ZD0109802); BUPT Innovation and Entrepreneurship Support Program (2025-YC-T026); High-performance Computing Platform of BUPT.

References

Aliminati, M. R.; Chakravarthi, B.; Verma, A. A.; Vaghela, A.; Wei, H.; Zhou, X.; and Yang, Y. 2024. Sevd: Synthetic event-based vision dataset for ego and fixed traffic perception. *arXiv preprint arXiv:2404.10540*.

Boretti, C.; Bich, P.; Pareschi, F.; Prono, L.; Rovatti, R.; and Setti, G. 2023. Pedro: an event-based dataset for person detection in robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4065–4070.

Cao, H.; Zhang, Z.; Xia, Y.; Li, X.; Xia, J.; Chen, G.; and Knoll, A. 2024a. Embracing events and frames with hierarchical feature refinement network for object detection. In *European Conference on Computer Vision*, 161–177. Springer.

Cao, J.; Zheng, X.; Lyu, Y.; Wang, J.; Xu, R.; and Wang, L. 2024b. Chasing day and night: Towards robust and efficient all-day object detection guided by an event camera. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 9026–9032. IEEE.

Dai, P.; Zhang, Y.; Yu, X.; Lyu, X.; and Qi, X. 2023. Hybrid neural rendering for large-scale scenes with motion blur. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 154–164.

De Tournemire, P.; Nitti, D.; Perot, E.; Migliore, D.; and Sironi, A. 2020. A large scale event-based detection dataset for automotive. *arXiv preprint arXiv:2001.08499*.

Du, W.; Ye, J.; Gu, J.; Li, J.; Wei, H.; and Wang, G. 2023. Safelight: A reinforcement learning method toward collision-free traffic signal control. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 14801–14810.

Falanga, D.; Kleber, K.; and Scaramuzza, D. 2020. Dynamic obstacle avoidance for quadrotors with event cameras. *Science Robotics*, 5(40): eaaz9712.

Fan, L.; Li, Y.; Shen, H.; Li, J.; and Hu, D. 2024a. From dense to sparse: low-latency and speed-robust event-based object detection. *IEEE Transactions on Intelligent Vehicles*.

Fan, Y.; Zhang, W.; Liu, C.; Li, M.; and Lu, W. 2024b. Sfod: Spiking fusion object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17191–17200.

Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.

Gehrig, D.; and Scaramuzza, D. 2024. Low-latency automotive vision with event cameras. *Nature*, 629(8014): 1034–1040.

Gehrig, M.; Aarents, W.; Gehrig, D.; and Scaramuzza, D. 2021. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3): 4947–4954.

Gehrig, M.; and Scaramuzza, D. 2023. Recurrent vision transformers for object detection with event cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13884–13893.

Jiang, J.; Li, J.; Zhang, B.; Deng, X.; and Shi, B. 2024. Evhandpose: Event-based 3d hand pose estimation with sparse supervision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9): 6416–6430.

Jocher, G. 2023. Yolov8. https://github.com/ultralytics/ultralytics/tree/main. Accessed: 2025-08-02.

Li, D.; Tian, Y.; and Li, J. 2023. Sodformer: Streaming object detection with transformer using events and frames. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11): 14020–14037.

Li, J.; Li, J.; Zhu, L.; Xiang, X.; Huang, T.; and Tian, Y. 2022. Asynchronous spatio-temporal memory network for continuous event-based object detection. *IEEE Transactions on Image Processing*, 31: 2975–2987.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Liu, H.; Peng, S.; Zhu, L.; Chang, Y.; Zhou, H.; and Yan, L. 2024a. Seeing motion at nighttime with an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25648–25658.

Liu, Z.; Yang, N.; Wang, Y.; Li, Y.; Zhao, X.; and Wang, F.-Y. 2024b. Enhancing traffic object detection in variable illumination with rgb-event fusion. *IEEE Transactions on Intelligent Transportation Systems*.

Luo, X.; Yao, M.; Chou, Y.; Xu, B.; and Li, G. 2024. Integer-valued training and spike-driven inference spiking neural network for high-performance and energy-efficient object detection. In *European Conference on Computer Vision*, 253–272. Springer.

Peng, Y.; Li, H.; Zhang, Y.; Sun, X.; and Wu, F. 2024. Scene adaptive sparse transformer for event-based object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16794–16804.

Perot, E.; De Tournemire, P.; Nitti, D.; Masci, J.; and Sironi, A. 2020. Learning to detect objects with a 1 megapixel event camera. *Advances in Neural Information Processing Systems*, 33: 16639–16652.

Prophesee. 2024. Metavision EVK4 – HD. https://www.prophesee.ai/eventcamera-evk4/. Accessed: 2024-07-18.

Qian, Y.; Ye, S.; Wang, C.; Cai, X.; Qian, J.; and Wu, J. 2025. UCF-Crime-DVS: A Novel Event-Based Dataset for Video Anomaly Detection with Spiking Neural Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6577–6585.

Teng, S.; Hu, X.; Deng, P.; Li, B.; Li, Y.; Ai, Y.; Yang, D.; Li, L.; Xuanyuan, Z.; Zhu, F.; et al. 2023. Motion planning

for autonomous driving: The state of the art and future perspectives. *IEEE Transactions on Intelligent Vehicles*, 8(6): 3692–3711.

Tomy, A.; Paigwar, A.; Mann, K. S.; Renzaglia, A.; and Laugier, C. 2022. Fusing event-based and rgb camera for robust object detection in adverse conditions. In *2022 International Conference on Robotics and Automation (ICRA)*, 933–939. IEEE.

Verma, A. A.; Chakravarthi, B.; Vaghela, A.; Wei, H.; and Yang, Y. 2024. etram: Event-based traffic monitoring dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22637–22646.

Wang, C.-Y.; Bochkovskiy, A.; and Liao, H.-Y. M. 2023. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7464–7475.

Wang, X.; Wang, S.; Tang, C.; Zhu, L.; Jiang, B.; Tian, Y.; and Tang, J. 2024. Event stream-based visual object tracking: A high-resolution benchmark dataset and a novel baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19248–19257.

Wei, H.; Xu, D.; Liang, J.; and Li, Z. J. 2021. How do we move: Modeling human movement with system dynamics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 4445–4452.

Yang, N.; Wang, Y.; Liu, Z.; Li, M.; An, Y.; and Zhao, X. 2025. SMamba: Sparse Mamba for Event-based Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9229–9237.

Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; and Chen, J. 2024. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16965–16974.

Zhou, H.; Chang, Y.; and Shi, Z. 2024. Bring event into rgb and lidar: Hierarchical visual-motion fusion for scene flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26477–26486.

Zhou, Z.; Wu, Z.; Boutteau, R.; Yang, F.; Demonceaux, C.; and Ginhac, D. 2022. Rgb-event fusion for moving object detection in autonomous driving. *arXiv preprint arXiv:2209.08323*.

Zhu, A. Z.; Thakur, D.; Özaslan, T.; Pfrommer, B.; Kumar, V.; and Daniilidis, K. 2018. The multivehicle stereo event camera dataset: An event camera dataset for 3D perception. *IEEE Robotics and Automation Letters*, 3(3): 2032–2039.