EFFICIENT RGB-EVENT FUSION FOR MOVING OBJECT DETECTION IN AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

In autonomous driving, low-light conditions and motion blur degrade RGB appearance representations, thereby impairing detection accuracy. RGB–Event fusion introduces event-based motion information to complement RGB appearance information when RGB imaging is degraded. However, camera ego-motion generates substantial background event responses, which interfere with target representation during cross-modal fusion. Notably, although existing multi-window methods improve detection accuracy, they struggle to balance accuracy and efficiency, ultimately constraining the reliability and real-time performance of autonomous driving perception systems. To address these challenges, we propose a single-window RGB–Event fusion network for object detection in autonomous driving. The network comprises hierarchical cross-modal fusion and multi-window-to-single-window knowledge distillation (M2S-KD). Hierarchical cross-modal fusion enhances complementary information through bidirectional feature modulation and improves deep feature fusion through Bidirectional Attention. M2S-KD transfers detection knowledge from a multi-window teacher to a single-window student through cross-head distillation, while only the student network is retained during inference to improve efficiency. Experimental results on DSEC-MOD show that the proposed method outperforms the second-best method by 2.55 percentage points in Frame mAP@0.5, while efficiency experiments demonstrate a 115.41% increase in inference speed compared with the multi-window model. These results provide a practical approach to balancing detection accuracy and deployment efficiency for object detection in autonomous driving.

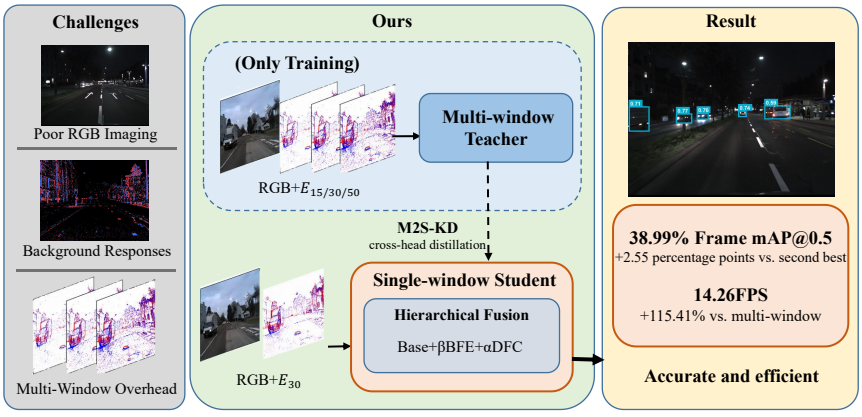


Figure 1: Overview of the proposed framework.

1 INTRODUCTION

In autonomous driving, timely and accurate moving object detection is fundamental to environmental perception and safety-critical decision-making (Feng et al., 2021). Existing road object detection methods predominantly rely on RGB frames, which provide rich texture, color, and semantic infor-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

mation. However, their performance is highly dependent on image quality, and challenging illumination and high-speed motion may cause imaging degradation, weakening object representation and detection performance (Dodge & Karam, 2016; Zhou et al., 2023).

Event cameras have recently been introduced into visual perception systems. Their high dynamic range and temporal resolution enable them to capture clear motion boundaries and instantaneous intensity changes (Gallego et al., 2022). However, event data have limitations in appearance and texture representation. Overall, event data and RGB frames are complementary in motion perception and appearance representation, providing a viable approach to road object detection under challenging illumination and high-speed motion.

RGB–Event fusion methods have evolved from simple feature aggregation to more sophisticated cross-modal interaction. For example, the early FPN-Fusion method directly concatenates RGB and event features at corresponding scales (Tomy et al., 2022), whereas the more recent SFNet adaptively incorporates event features according to the illumination distribution of the input RGB frame (Liu et al., 2024b). Although existing methods achieve competitive performance in challenging scenarios, further improvements are limited by computational efficiency and effective information selection. On the one hand, some methods employ multiple event accumulation windows to capture motion information at different temporal scales (Zhou et al., 2023). Although multi-window event inputs improve detection performance, they introduce computational cost. On the other hand, existing methods emphasize enhancing cross-modal interaction between RGB and event, but pay limited attention to the distinction between target-relevant information and background interference, particularly background event responses induced by camera ego-motion (Vasco et al., 2017). Therefore, enhancing target-relevant complementary information while mitigating background interference, and preserving the detection benefits of multi-window inputs while reducing inference overhead, is key to improving the accuracy and efficiency of RGB–Event detection.

To address these challenges, we propose an RGB–Event moving object detection framework that combines hierarchical bidirectional cross-modal fusion with Multi-Window-to-Single-Window Knowledge Distillation (M2S-KD) to balance detection accuracy and inference efficiency, as summarized in Figure 1.

Our main contributions are summarized as follows:

1. We develop a hierarchical cross-modal fusion mechanism comprising Bidirectional Feature Enhancement (BFE) and Deep Feature Calibration (DFC). BFE enhances complementary information through bidirectional feature modulation across all six feature levels, whereas DFC calibrates deep cross-modal representations through Bidirectional Attention and feature statistics. Their outputs are combined with the base fusion features through weighted residual aggregation to mitigate background event responses and noise.
2. We propose M2S-KD, which distills motion and detection knowledge across multiple temporal windows from a multi-window teacher network into a single-window student network during training. During inference, only an RGB frame and a single-window event frame are required, improving single-window detection accuracy while avoiding the computational overhead introduced by multi-window inputs.

2 Related Work

2.1 RGB–Event Fusion

Early RGB–Event fusion methods mainly relied on direct feature aggregation. For example, FPN-Fusion concatenates corresponding RGB and event features and performs multi-scale fusion through a feature pyramid (Tomy et al., 2022). Subsequent studies have primarily explored adaptive modality weighting and stronger cross-modal interaction. For modality weighting, EOLO dynamically adjusts the contributions of the two modalities under different illumination conditions (Cao et al., 2024b). For cross-modal interaction, RENet employs bidirectional spatial and channel calibration (Zhou et al., 2023), while FRN combines bidirectional cross-modal interaction with feature-statistics-based channel alignment (Cao et al., 2024a). Beyond synchronous feature fusion, SODFormer uses a Transformer (Vaswani et al., 2017) to model asynchronous interactions between RGB frames and event streams (Li et al., 2023). Meanwhile, FlexEvent investigates event-frame detection at vary-

ing operational frequencies (Lu et al., 2025). However, ego-motion on vehicle-mounted platforms produces background event responses that interfere with cross-modal fusion (Vasco et al., 2017). Accordingly, our method employs bidirectional feature enhancement and deep feature calibration to mitigate this interference.

2.2 Knowledge Distillation

Knowledge distillation provides students with supervision beyond ground-truth annotations (Hinton et al., 2015; Gou et al., 2021) and has been extended from output predictions to intermediate features and feature relations (Romero et al., 2015; Park et al., 2019). For object detection, FGD transfers foreground, background, and global relational information (Yang et al., 2022); DFD imposes region-specific constraints based on teacher–student feature discrepancies (Liu et al., 2024a); and CrossKD reuses the frozen teacher detection head to generate cross-head predictions (Wang et al., 2024). Event-based studies primarily investigate cross-modal knowledge transfer. EvDistill uses an RGB-frame teacher to guide event-based segmentation and recognition (Wang et al., 2021), while COFD, EA-DETR, and M^{2}C-EvDet transfer object-centric, event-aware, or frequency-decoupled knowledge from frame-based teachers to event-based students (Li et al., 2024; Rossi et al., 2025; Bao et al., 2026). These methods mainly transfer knowledge across modalities and retain only event inputs during inference. In contrast, knowledge transfer across different event accumulation windows remains underexplored. We therefore use a multi-window teacher network to supervise a single-window student network during training and retain only the student network during inference to improve inference efficiency.

3 Method

3.1 Event Accumulation Window

Event streams are accumulated within a temporal window Δt to generate event frames aligned with RGB frames. Specifically, events within a fixed time interval before each RGB frame are aggregated to construct the corresponding event representation. In this work, the teacher network uses multiple event accumulation windows, whereas the student network uses a single event accumulation window.

3.2 Overall Architecture

The proposed framework employs a single-window RGB–Event detector with hierarchical cross-modal fusion as the student network. During training, a frozen multi-window teacher network supervises the student network through M2S-KD. During inference, the teacher network and distillation pathway are removed, leaving only the single-window student network.

The student network takes an RGB frame and its temporally aligned 30 ms event frame as inputs. The two modalities are processed by ResNet-101 and ResNet-18 feature extractors (He et al., 2016), respectively, producing RGB and event features at six feature levels. These features are integrated through Base Fusion, BFE, and DFC, and the branch outputs are combined through weighted residual aggregation. The fused features are then fed into the detection head to predict the center heatmap, temporal center offsets, and bounding-box width and height (Li et al., 2020b). During training, an additional teacher network takes an RGB frame together with event frames accumulated over 15, 30, and 50 ms. The student network is supervised by both the ground-truth annotations and cross-head distillation based on the teacher predictions, while the teacher network remains frozen.

3.3 Hierarchical Cross-Modal Fusion

The RGB and event frames are processed by separate feature extractors to obtain six-level feature R and E, respectively. Before fusion, the channel dimension of each event feature is aligned with that of the corresponding RGB feature:

$$\bar{E} = \psi(E). \tag{1}$$

where ψ denotes a 1×1 convolutional mapping for channel alignment. The aligned feature pairs R and $\bar{E}$ are processed by the Base Fusion, BFE, and DFC branches, whose outputs are combined through weighted residual aggregation. BFE operates on all six feature levels, whereas DFC is applied only to the deep features at Levels 4–6. The overall structure is illustrated in Figure 2.

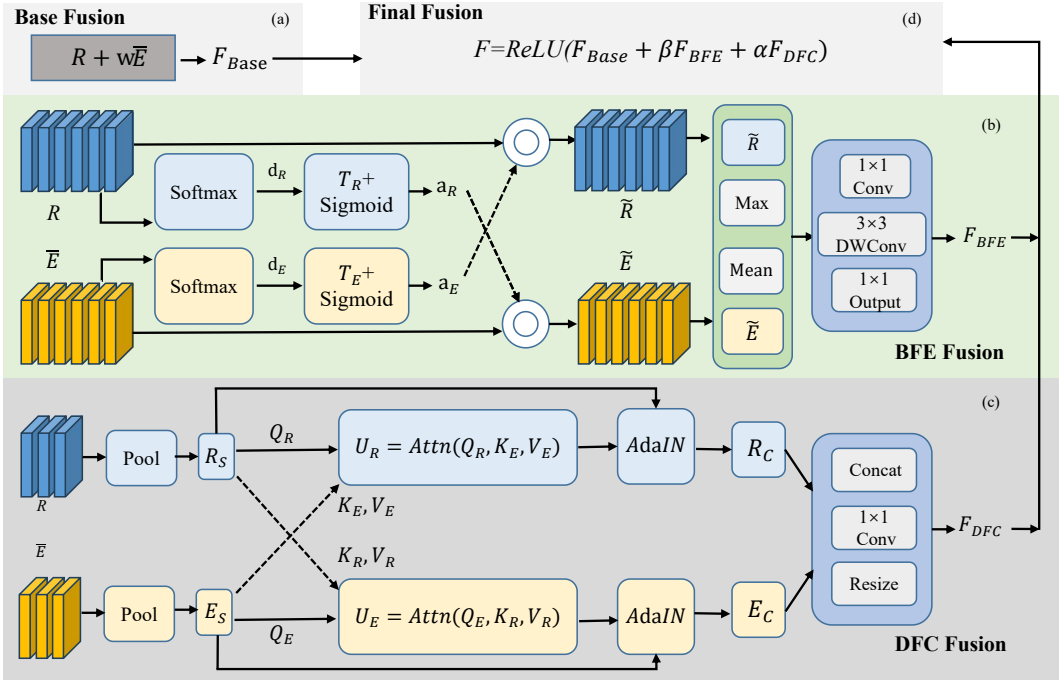


Figure 2: Hierarchical cross-modal fusion.

Base Fusion Module. The Base Fusion Module is illustrated in Figure 2(a). It aggregates the RGB features and weighted event features to obtain the base fusion features:

$$F_{\text{Base}} = R + w\bar{E}. \tag{2}$$

Bidirectional Feature Enhancement Module. As shown in Figure 2(b), BFE is applied to all six feature levels. Each modality uses the global response of the other modality to modulate its own features, enhancing complementary information while reducing unreliable responses caused by ego-motion or degraded RGB inputs.

The features of the two modalities are first projected to a shared channel dimension:

$$R_p = \phi_r(R), \qquad E_p = \phi_e(\bar{E}). \tag{3}$$

For either projected feature $X \in \{R_p, E_p\}$, learnable spatial scores aggregate informative responses into a global response d_X, which is then transformed into channel-wise modulation coefficients a_X:

$$d_X = \sum_{j=1}^{HW} \text{softmax}_j(m_X(x_j))\, x_j, \qquad a_X = 2\,\text{sigmoid}(T_X(d_X))\,. \tag{4}$$

Here, H and W denote the spatial dimensions, x_j is the feature vector at location j, and m_X and T_X denote the spatial scoring and channel modulation transformations, respectively. The two modalities use independent parameters (Hu et al., 2018).

The modulation coefficients generated from the event features are applied to the RGB features, while those generated from the RGB features are applied to the event features:

$$\widetilde{R} = R_p \odot a_{E_p}, \qquad \widetilde{E} = E_p \odot a_{R_p}. \tag{5}$$

Here, $\odot$ denotes element-wise multiplication. Since $a_X \in (0, 2)$, coefficients above and below 1 enhance and suppress the corresponding feature channels, respectively, allowing modality to be modulated according to the response of the other.

The modulated features, together with their element-wise maximum and mean, are concatenated along the channel dimension to form the fused feature Z, which is further processed through channel mixing and local spatial refinement to produce the BFE-enhanced feature F_{BFE}:

$$Z = \mathrm{Concat}\Big(\widetilde{R}, \widetilde{E}, \mathrm{Max}(\widetilde{R}, \widetilde{E}), \mathrm{Mean}(\widetilde{R}, \widetilde{E})\Big). \tag{6}$$

$$F_{\mathrm{BFE}} = \phi^o(\mathrm{DWConv}_{3\times3}(\phi^m(Z))). \tag{7}$$

Here, ϕ^m and ϕ^o denote channel mixing and output projection, respectively, while $\mathrm{DWConv}_{3\times3}$ denotes depthwise convolution for local spatial refinement.

Deep Feature Calibration Module. As shown in Figure 2(c), RGB features emphasizing object semantics and event features describing motion structures and edge variations may exhibit discrepancies in spatial correspondence and response distribution at deep feature levels. DFC is therefore applied to Levels 4–6 to calibrate the features of both modalities through Bidirectional Attention, thereby adjusting their feature response distributions.

DFC first reuses the channel-compressed features from BFE and applies adaptive average pooling to reduce the computational cost of attention:

$$R_s = \mathrm{Pool}(R_p), \qquad E_s = \mathrm{Pool}(E_p). \tag{8}$$

Independent 1×1 convolutions are then used to generate the query Q, key K, and value V:

$$\begin{aligned} Q_R &= W_q^R R_s, \quad K_R = W_k^R R_s, \quad V_R = W_v^R R_s, \\ Q_E &= W_q^E E_s, \quad K_E = W_k^E E_s, \quad V_E = W_v^E E_s. \end{aligned} \tag{9}$$

Here, W_q, W_k, and W_v denote learnable linear projections.

The query from each modality is used to compute correlations with the key from the other modality, and the corresponding values from the other modality are then weighted and aggregated:

$$U_R = \mathrm{softmax}\left(\frac{Q_R K_E^{\mathsf{T}}}{\sqrt{d}}\right) V_E, \qquad U_E = \mathrm{softmax}\left(\frac{Q_E K_R^{\mathsf{T}}}{\sqrt{d}}\right) V_R. \tag{10}$$

Here, d denotes the channel dimension of each attention head.

The mean and standard deviation of U_R and U_E are subsequently used to calibrate the feature statistics of the corresponding modalities. (Cao et al., 2024a; Huang & Belongie, 2017):

$$R_c = \mathrm{AdaIN}(R_s, U_R), \qquad E_c = \mathrm{AdaIN}(E_s, U_E). \tag{11}$$

$$\mathrm{AdaIN}(X, Y) = \frac{\sqrt{\sigma^2(Y) + \varepsilon}}{\sqrt{\sigma^2(X) + \varepsilon}} (X - \mu(X)) + \mu(Y). \tag{12}$$

Here, μ and σ denote the mean and standard deviation, and ε ensures numerical stability.

Finally, the calibrated RGB and event features are concatenated and fused through convolution, after which bilinear interpolation is used to restore their spatial resolution:

$$F_{\mathrm{DFC}} = \mathrm{Resize}(\Phi(\mathrm{Concat}(R_c, E_c))) \,. \tag{13}$$

Here, Φ denotes feature fusion and output projection, while Resize denotes bilinear spatial resizing.

Weighted Residual Aggregation Module. As shown in Figure 2(d), the base fusion, BFE-enhanced, and DFC-calibrated features are combined as follows:

$$F = \mathrm{ReLU}(F_{\mathrm{Base}} + \beta F_{\mathrm{BFE}} + \alpha F_{\mathrm{DFC}}) \,. \tag{14}$$

Here, β and α denote the learnable weights of the BFE and DFC branches, respectively. This design preserves the base fusion information while controlling the contributions of the two complementary feature branches to the final output.

3.4 M2S-KD: Multi-Window-to-Single-Window Knowledge Distillation

As illustrated in Figure 3, M2S-KD uses a frozen multi-window teacher network to supervise a single-window student network through cross-head distillation during training, while retaining only the student network during inference to improve efficiency.

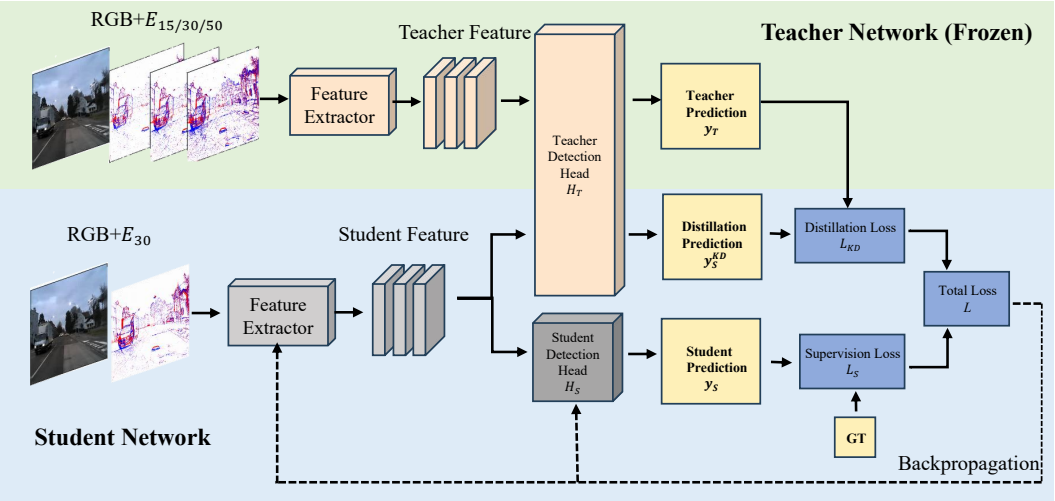


Figure 3: Multi-window-to-single-window knowledge distillation.

We adopt a pretrained multi-window RENet (Zhou et al., 2023) as the teacher network and the proposed single-window RGB–Event detection network as the student network. Their respective inputs are defined as follows:

$$X_T = [\mathrm{RGB}, E_{15}, E_{30}, E_{50}], \qquad X_S = [\mathrm{RGB}, E_{30}]. \tag{15}$$

Here, X_T and X_S denote the teacher and student inputs, respectively, while E_{15}, E_{30}, and E_{50} denote the event frames constructed using the corresponding accumulation windows.

The teacher and student inputs are processed by their respective feature extraction and fusion networks, after which the corresponding detection heads generate the teacher prediction y_T and student prediction y_S:

$$F_T = G_T(X_T), \quad y_T = H_T(F_T), \quad F_S = G_S(X_S), \quad y_S = H_S(F_S). \tag{16}$$

Directly enforcing consistency between y_S and y_T requires the student prediction to fit both the ground-truth annotations GT and the teacher prediction y_T, thereby subjecting the student detection

head to two supervisory targets. We therefore adopt cross-head distillation (Wang et al., 2024), in which the student features are additionally passed through the frozen teacher detection head.

Because the teacher and student networks receive different inputs, the student features are aligned using the channel-wise statistics of the teacher features. The aligned features $\widehat{F}_S$ are then passed through the frozen teacher detection head to obtain the distillation prediction y_S^{KD}:

$$\widehat{F}_S = \frac{F_S - \mu(F_S)}{\sigma(F_S) + \varepsilon} \odot \sigma(F_T) + \mu(F_T), \qquad y_S^{\mathrm{KD}} = H_T(\widehat{F}_S). \tag{17}$$

The teacher, student, and distillation predictions, denoted by y_T, y_S, and y_S^{KD}, respectively, each contain three branches $\mathcal{B} = \{\mathrm{hm}, \mathrm{mov}, \mathrm{wh}\}$. The student loss compares y_S with the ground-truth annotations (GT), while the distillation loss compares y_S^{KD} with y_T:

$$L_S = \sum_{b \in \mathcal{B}} w_b L_b, \qquad L_{\mathrm{KD}} = \sum_{b \in \mathcal{B}} w_b L_b^{\mathrm{KD}}, \qquad L = L_S + \lambda L_{\mathrm{KD}}. \tag{18}$$

Here, w_b is the weight of branch b, and λ is the distillation loss weight. During backpropagation, only the student network is updated, while the teacher network and its detection head remain frozen.

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation Metrics. This work focuses on RGB–Event fusion-based moving object detection for autonomous driving, with DSEC-MOD (Zhou et al., 2023; Gehrig et al., 2021) serving as the default dataset throughout the paper. PKU-DDD17-Car (Binas et al., 2017; Li et al., 2019) is additionally used to evaluate the proposed method under different road data conditions.

For DSEC-MOD, event streams are accumulated into event frames using windows of 15 , 30 , and 50 ms. Following its evaluation protocol (Zhou et al., 2023), we adopt frame mean average precision (Frame mAP, abbreviated as F.mAP) and video mean average precision (Video mAP, abbreviated as V.mAP) to evaluate single-frame and continuous-sequence detection performance. For PKU-DDD17-Car, event streams are accumulated into event frames using windows of 10 , 15 , and 20 ms. Since its annotations are provided for discrete frames, F.mAP is used as the evaluation metric.

Teacher Network. We employ the publicly available pretrained multi-window RENet (Zhou et al., 2023) as the teacher network. For DSEC-MOD, the teacher network takes an RGB frame together with event frames constructed using 15 ms, 30 ms, and 50 ms windows, whereas the student network takes an RGB frame and the 30 ms event frame. For PKU-DDD17-Car, the teacher network uses event windows of 10 , 15 , and 20 ms, whereas the student network uses the 15 ms window. During distillation training, both the parameters and batch-normalization statistics of the teacher network remain frozen, and its predictions provide additional supervision for the student network.

4.2 Comparison with Existing Methods

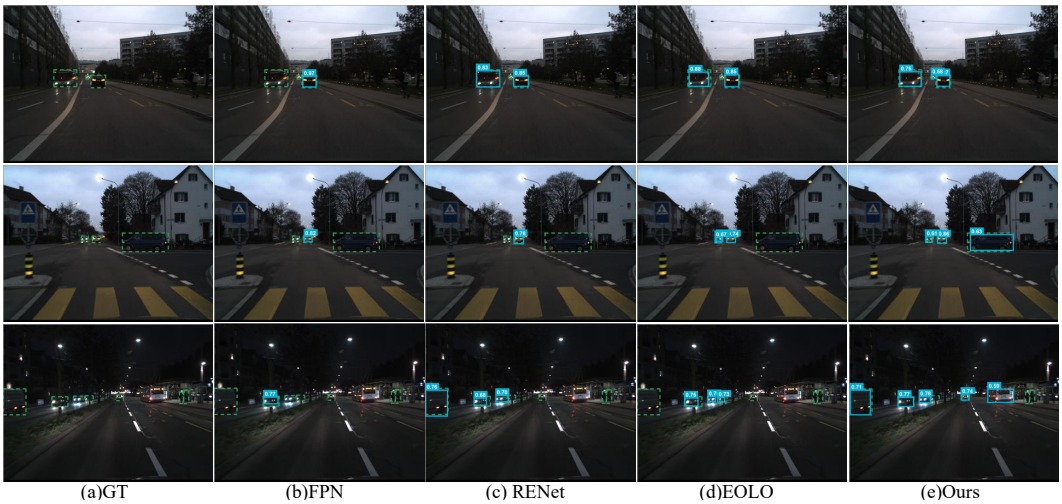
DSEC-MOD. On DSEC-MOD, we compare the proposed method with representative RGB–Event fusion detectors, including FAOD (Zhang et al., 2024), FPN-Fusion (Tomy et al., 2022), FRN (Cao et al., 2024a), RENet (Zhou et al., 2023), EOLO (Cao et al., 2024b), and SFNet (Liu et al., 2024b), under the same dataset split and evaluation protocol.

As shown in Table 1, the proposed method achieves 38.99%, 25.08%, and 7.52% in F.mAP@0.5, V.mAP@0.2, and V.mAP@0.5 on DSEC-MOD, respectively, achieving the best results across all three metrics. Compared with the best competing result for each metric, our method improves performance by 2.55, 2.56, and 0.91 percentage points, respectively. Compared with the multi-window teacher network RENet, the corresponding improvements are 3.73, 6.07, and 2.08 percentage points. These results demonstrate that the proposed method achieves the best performance in both single-frame and continuous-sequence detection, with the latter being particularly important for autonomous driving.

Table 1: Comparison on DSEC-MOD and PKU-DDD17-Car (%).

	DSEC-MOD			PKU-DDD17-Car
Method	F.mAP@0.5	V.mAP@0.2	V.mAP@0.5	F.mAP@0.5
FAOD	29.68	18.90	6.61	77.49
FPN-Fusion	29.53	20.82	4.93	58.46
FRN	32.02	19.37	5.07	84.77
RENet	35.26	19.01	5.44	85.65
EOLO	35.71	19.89	2.43	84.22
SFNet	36.44	22.52	5.30	86.87
Ours	**38.99**	**25.08**	**7.52**	**90.50**

Figure 4 presents qualitative comparisons in three representative road scenes involving distant small objects, low-contrast targets, and nighttime low-light conditions. Compared with competing methods, the proposed method exhibits fewer missed detections and localization errors, demonstrating better adaptability to these challenging conditions.



(a)GT (b)FPN (c) RENet (d)EOLO (e)Ours

Figure 4: Qualitative comparison on DSEC-MOD.

PKU-DDD17-Car. To further evaluate the proposed method, we compare it with the aforementioned methods on PKU-DDD17-Car. These results further validate the applicability of the proposed method across different road scenarios.

4.3 Ablation Studies

Individual Components. To evaluate the effectiveness of each component, we use the single-window student network with only Base Fusion as the baseline and evaluate different combinations of M2S-KD, BFE, and DFC. All other training and testing settings remain unchanged.

Table 2 reports ablations using the single-window student with Base Fusion as the baseline. M2S-KD improves F.mAP@0.5 from 26.74% to 28.94%, while introducing BFE and DFC without distillation yields 34.05%. Combining all components achieves the best F.mAP@0.5, V.mAP@0.2, and V.mAP@0.5 of 38.99%, 25.08%, and 7.52%, respectively. These results demonstrate the complementary contributions of multi-window supervision and hierarchical cross-modal fusion.

Table 2: Component ablation on DSEC-MOD (%).

#	Base	M2S-KD	BFE	DFC	F.mAP@0.5	V.mAP@0.2	V.mAP@0.5
1	✓				26.74	14.34	4.11
2	✓	✓			28.94	16.55	5.31
3	✓		✓	✓	34.05	19.45	5.07
4	✓	✓		✓	34.40	19.38	5.76
5	✓	✓	✓		35.58	24.49	6.47
6	✓	✓	✓	✓	**38.99**	**25.08**	**7.52**

4.4 Robustness Experiments

We evaluate robustness by applying low illumination, overexposure, reduced contrast, motion blur, and fog to RGB frames at three severity levels while keeping the event input unchanged (Hendrycks & Dietterich, 2019). As shown in Table 3, the proposed method achieves average retention rates ranging from 83.10% to 97.90% across the five degradation types, with an overall average of 89.40%, demonstrating stable detection performance under RGB degradation.

Table 3: Robustness to RGB degradation on DSEC-MOD.

Degradation	Mild	Moderate	Severe	Retention (%)
Low illumination	38.411	37.316	30.953	91.20
Overexposure	38.946	38.846	36.722	97.90
Reduced contrast	38.613	37.085	27.219	87.99
Motion blur	37.720	34.093	25.385	83.10
Fog	38.620	37.414	25.491	86.80

4.5 Inference Efficiency Evaluation

We construct a multi-window control model by replacing the fusion module of the RENet teacher network with the proposed fusion module. Meanwhile, a single-window control model is constructed by replacing the proposed fusion module with RENet fusion to control for differences in fusion design. Inference latency is measured from raw data input to the generation of detection results. As shown in Table 4, the proposed model achieves 14.26 FPS, improving upon the multi-window and single-window control models by 115.41% and 7.95%, respectively. Therefore, the efficiency gain primarily stems from reducing the event input from multiple accumulation windows to a single accumulation window.

Table 4: Inference efficiency on DSEC-MOD.

Model	Event windows	FPS
Multi-window + BFE/DFC	15/30/50 ms	6.62
Single-window + RENet fusion	30 ms	13.21
Single-window + BFE/DFC **(Ours)**	30 ms	**14.26**

5 Conclusion

We present a single-window RGB–Event detector for autonomous driving that combines hierarchical cross-modal fusion with M2S-KD. BFE and DFC enhance complementary information and calibrate deep cross-modal representations, while M2S-KD transfers multi-window detection knowledge to the single-window student during training. The proposed method achieves 38.99% and 90.50% F.mAP@0.5 on DSEC-MOD and PKU-DDD17-Car, respectively, and improves inference speed by 115.41% over the multi-window model with the same fusion module. Future work will reduce end-to-end event-processing latency and extend knowledge transfer across event accumulation windows to other event-based vision tasks.

AI USE STATEMENT

Generative AI tools were used for literature retrieval and reference verification, research ideation and experimental-design discussions, language translation and refinement, and drafting parts of the manuscript. All AI-assisted content was reviewed and revised by the authors, who made the final methodological decisions, conducted the experiments, validated the results, and take full responsibility for this work.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure the reproducibility of the results reported in this paper. The network architecture, fusion modules, training and distillation procedures, implementation details, and evaluation protocols are described in the main paper and appendix. The source code of the proposed model is provided in the supplementary material. All experiments are conducted on the publicly available DSEC-MOD and PKU-DDD17-Car datasets, following their established evaluation protocols.

REFERENCES

Wei Bao, Siqi Li, Shouan Pan, Yi Xie, and Yue Gao. M^2C-EvDet: Multi-domain multi-order cross-modal knowledge distillation for event-based object detection, 2026.

Jonathan Binas, Daniel Neil, Shih-Chii Liu, and Tobi Delbrück. DDD17: End-to-end DAVIS driving dataset, 2017.

Hu Cao, Zehua Zhang, Yan Xia, Xinyi Li, Jiahao Xia, Guang Chen, and Alois Knoll. Embracing events and frames with hierarchical feature refinement network for object detection. In *European Conference on Computer Vision*, pp. 161–177, 2024a.

Jiahang Cao, Xu Zheng, Yuanhuiyi Lyu, Jiaxu Wang, Renjing Xu, and Lin Wang. Chasing day and night: Towards robust and efficient all-day object detection guided by an event camera. In *IEEE International Conference on Robotics and Automation*, pp. 9026–9032, 2024b.

Samuel Dodge and Lina Karam. Understanding how image quality affects deep neural networks. In *International Conference on Quality of Multimedia Experience*, 2016.

Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Gläser, 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, 2021.

Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2022.

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.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision*, 2017.

Dianze Li, Yonghong Tian, and Jianing Li. SODFormer: Streaming object detection with transformer using events and frames. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):14020–14037, 2023.

Jianing Li, Siwei Dong, Zhaofei Yu, Yonghong Tian, and Tiejun Huang. Event-based vision enhanced: A joint detection framework in autonomous driving. In *IEEE International Conference on Multimedia and Expo*, pp. 1396–1401, 2019.

Lei Li, Alexander Liniger, Mario Millhäusler, Vagia Tsiminaki, Yuanyou Li, and Dengxin Dai. Object-centric cross-modal feature distillation for event-based object detection. In *IEEE International Conference on Robotics and Automation*, pp. 15440–15447, 2024.

Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21002–21012, 2020a.

Yixuan Li, Zixu Wang, Limin Wang, and Gangshan Wu. Actions as moving points. In *European Conference on Computer Vision*, 2020b.

Kang Liu, Yingyi Zhang, Jingyun Zhang, Jinmin Li, Jun Wang, Shaoming Wang, Chun Yuan, and Rizen Guo. DFD: Distilling the feature disparity differently for detectors. In *International Conference on Machine Learning*, pp. 32421–32430, 2024a.

Zhanwen Liu, Nan Yang, Yang Wang, Yuke Li, Xiangmo Zhao, and Fei-Yue Wang. Enhancing traffic object detection in variable illumination with RGB-event fusion. *IEEE Transactions on Intelligent Transportation Systems*, 2024b.

Dongyue Lu, Lingdong Kong, Gim Hee Lee, Camille Simon Chane, and Wei Tsang Ooi. FlexEvent: Towards flexible event-frame object detection at varying operational frequencies. In *Advances in Neural Information Processing Systems*, 2025.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

Djessy Rossi, Pascal Vasseur, Fabio Morbidi, Cédric Demonceaux, and François Rameau. Event-aware distilled DETR for object detection in an automotive context. In *IEEE Intelligent Vehicles Symposium*, pp. 1849–1855, 2025.

Abhishek Tomy, Anshul Paigwar, Khushdeep Singh Mann, Alessandro Renzaglia, and Christian Laugier. Fusing event-based and RGB camera for robust object detection in adverse conditions. In *IEEE International Conference on Robotics and Automation*, 2022.

Valentina Vasco, Arren Glover, Elias Mueggler, Davide Scaramuzza, Lorenzo Natale, and Chiara Bartolozzi. Independent motion detection with event-driven cameras. In *International Conference on Advanced Robotics*, pp. 530–536, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Jiabao Wang, Yuming Chen, Zhaohui Zheng, Xiang Li, Ming-Ming Cheng, and Qibin Hou. CrossKD: Cross-head knowledge distillation for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16520–16530, 2024.

Lin Wang, Yujeong Chae, Sung-Hoon Yoon, Tae-Kyun Kim, and Kuk-Jin Yoon. EvDistill: Asynchronous events to end-task learning via bidirectional reconstruction-guided cross-modal knowledge distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 608–619, 2021.

Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4643–4652, 2022.

Haitian Zhang, Xiangyuan Wang, Chang Xu, Xinya Wang, Fang Xu, Huai Yu, Lei Yu, and Wen Yang. Frequency-adaptive low-latency object detection using events and frames, 2024.

Zhuyun Zhou, Zongwei Wu, Rémi Boutteau, Fan Yang, Cédric Demonceaux, and Dominique Ginhac. RGB-event fusion for moving object detection in autonomous driving. In *IEEE International Conference on Robotics and Automation*, pp. 7808–7815, 2023.

A Appendix

A.1 Event Accumulation Window

For an RGB frame timestamped at t_i, an event frame with an accumulation window of Δt represents the events within $[t_i - \Delta t, t_i)$. Event frames aligned with the same RGB frame therefore differ in their temporal coverage. In the DSEC-MOD experiments, the teacher network takes event frames with accumulation windows of 15, 30, and 50 ms, while the student network takes a single 30 ms event frame.

A.2 Loss Computation

The detector produces three prediction branches, $y = \{y^{\mathrm{hm}}, y^{\mathrm{mov}}, y^{\mathrm{wh}}\}$, corresponding to the center heatmap, temporal center offsets, and bounding-box width and height. The supervised student loss is computed branch-wise between the student prediction y_S and the corresponding ground-truth targets:

$$L_S = \sum_{b \in \mathcal{B}} w_b \, \ell_b\big(y_S^b, GT^b\big), \qquad \mathcal{B} = \{\mathrm{hm}, \mathrm{mov}, \mathrm{wh}\}. \tag{19}$$

The center heatmap is supervised by a classification loss, while the temporal center offset and bounding-box regression losses are evaluated at the annotated object-center locations. The distillation loss is computed in the same branch-wise manner between the distillation prediction y_S^{KD} and the teacher prediction y_T:

$$L_{\mathrm{KD}} = \sum_{b \in \mathcal{B}} w_b \, \ell_b^{\mathrm{KD}}\Big(y_S^{\mathrm{KD},b}, y_T^b\Big). \tag{20}$$

For distillation, Quality Focal Loss (Li et al., 2020a) is applied to the center heatmap using the teacher prediction as the soft target. The temporal center offset and bounding-box branches use teacher-confidence-weighted regression losses. The total training loss is

$$L = L_S + \lambda L_{\mathrm{KD}}, \tag{21}$$

where w_b is the weight of prediction branch b, and λ is the distillation loss weight.

A.3 Datasets and Implementation Details

DSEC-MOD contains 16 road sequences, with 11 used for training and five for testing. The dataset provides 13,314 paired RGB and event samples at an original resolution of 640×480. PKU-DDD17-Car contains 14 sequences and 3,154 annotated samples, with seven sequences used for training and seven for testing. Its original resolution is 346×260.

All RGB and event frames are resized to 288×288. The RGB and event feature extractors use ResNet-101 and ResNet-18, respectively, initialized from COCO-pretrained weights. On DSEC-MOD, the student network is trained for 15 epochs using Adam, with an initial learning rate of 5×10^{-4} and a batch size of 7.

The implementation uses PyTorch 2.0.1 and CUDA 11.7. All experiments are conducted on a single NVIDIA GeForce RTX 3090 GPU. During distillation, the parameters and batch-normalization statistics of the teacher network remain frozen.

A.4 Evaluation Metrics

For Frame mAP, frame-level detections are first processed by non-maximum suppression and ranked by confidence. Each detection is then greedily matched with an unmatched ground-truth box according to spatial IoU. AP is computed from the resulting precision–recall curve and averaged across object classes. F.mAP@0.5 uses an IoU threshold of 0.5.

For Video mAP, frame-level detections are linked into temporal tubes and evaluated using spatiotemporal IoU. This measure combines the mean spatial IoU over the shared temporal interval with the temporal overlap between two tubes. We report V.mAP@0.2 and V.mAP@0.5. Since PKU-DDD17-Car provides discrete frame annotations, only F.mAP@0.5 is reported on this dataset.

A.5 Event Accumulation Window Ablation

Table 5: Effect of the student event accumulation window on DSEC-MOD.

Event accumulation window	F.mAP@0.5 (%)
15 ms	34.18
30 ms	**38.99**
50 ms	35.93

Table 5 shows that the 30 ms window achieves the highest F.mAP@0.5. The 15 ms window accumulates too few events to form complete motion structures. In contrast, the 50 ms window provides more event information but also introduces additional background event responses and motion trails. The 30 ms window provides a better balance between event sparsity and the completeness of motion information and is therefore used as the student input window in the DSEC-MOD experiments.

A.6 Robustness Evaluation Settings

We evaluate robustness to five types of RGB imaging degradation, each with three severity levels. Only the RGB frames are degraded, while the event input remains unchanged. Table 6 lists the degradation parameters.

Table 6: Severity settings for RGB imaging degradation.

Degradation type	Mild	Moderate	Severe
Low illumination (brightness factor)	0.75	0.50	0.25
Overexposure (gain)	1.25	1.50	2.00
Contrast reduction (contrast factor)	0.75	0.50	0.25
Motion blur (kernel size)	5×5	9×9	15×15
Fog (strength)	0.15	0.30	0.50

For degradation type d, the average retention rate across the three severity levels is

$$R_d = \frac{1}{3} \sum_{k=1}^{3} \frac{M_{d,k}}{M_{\text{clean}}} \times 100\%, \tag{22}$$

where $M_{d,k}$ is the F.mAP@0.5 under severity level k of degradation d, and M_{clean} is the F.mAP@0.5 obtained from clean RGB input. The overall retention rate is the mean of R_d across the five degradation types.