Global Academic Frontiers, Vol. 3, Issue 2, 2025, pp. 106-129 Print ISSN 2995-5688 Online ISSN 2995-570X
DOI: https://doi.org/10.5281/zenodo.15582511

Semantic Segmentation of Nighttime Images Based on Cross-modal Domain Adaptation

Jixing Huang [1] Yanhe Li [1] Yuchen Zhang [1] Xinyue Zhang [1] Xin-yue Zhang [1] Ruihan Qi [1*]

[1] Stony Brook Institute at Anhui University, Hefei, Anhui 230031, China
*Corresponding author Email: qiruihan2004@163.com
Received 14 April 2025; Accepted 29 May 2025; Published 3 June 2025

Abstract: Semantic segmentation of nighttime images is crucial for all-weather autonomous perception but faces challenges like low-light noise, motion blur, and cross-domain adaptation limitations. Traditional visible-light methods suffer from sensor constraints (60 dB dynamic range), causing information loss in extreme darkness (<1 lux), while domain adaptation approaches degrade due to day-night noise distribution shifts. This work introduces event cameras (140 dB range, μs-level response) to establish a multimodal cooperative framework. A dual-branch network decouples visible content features and event-based motion features, optimized by cross-modal contrastive loss (CMCL) and hybrid Gaussian kernel MMD loss for modality alignment and domain matching. A dynamic confidence screening (DCS) mechanism integrates optical flow consistency and Bayesian uncertainty to suppress pseudo-label noise (18.5% false detection reduction). Evaluations on DSEC/MVSEC datasets demonstrate 21.3% mIoU gain in extreme low-light, 34.5% boundary IoU improvement in blurred regions, and 14.2% superior cross-domain adaptation (day→night) over state-of-the-art methods. This framework offers a label-efficient and robust solution for nighttime autonomous driving systems, advancing multimodal sensing deployment.

Keywords: Nighttime Images Semantic Segmentation, All-weather Autonomous Perception, Event Cameras, Multimodal Cooperative Framework, Dual-branch Network, Cross-modal Contrastive Loss (CMCL), Hybrid Gaussian Kernel MMD Loss, Dynamic Confidence Screening (DCS), Pseudo-label Noise Suppression, Low-light Noise

1. Introduction

Semantic segmentation of nighttime images is a critical enabler for all-weather autonomous systems, yet it remains challenged by inherent limitations of traditional visible-light cameras, such as low dynamic range (60 dB), motion blur under long exposure, and domain shifts between synthetic and real-world nighttime data. Existing approaches, including visible-light enhancement (e.g., RetinexNet, EnlightenGAN) and infrared (IR) fusion methods (e.g., CMX), struggle to address these issues comprehensively. Visible enhancement techniques suffer from amplified noise and artifacts in extreme low-light conditions (<1 lux), while IR modalities fail to capture texture details of non-thermal objects (e.g., traffic signs) and exhibit poor dynamic scene adaptation. Event cameras, with their ultra-high dynamic range (140 dB) and microsecond temporal resolution, offer a promising alternative by capturing high-frequency motion cues and recovering dark-region details through asynchronous event streams. However, cross-modal domain adaptation between RGB and event modalities remains underexplored, particularly under the coupled challenges of day-night domain bias and modality heterogeneity.

1.2 Object and Subject of Research

The research focuses on nighttime semantic segmentation under extreme low-light and dynamic blurring conditions. The primary subjects include:

1. Cross-modal fusion of RGB and event camera data to compensate for the limitations of visible-light sensors.
2. Domain adaptation across day-night lighting conditions and modality-specific feature spaces (e.g., sparse event streams vs. dense RGB images).
3. Dynamic confidence optimization to mitigate pseudo-label noise caused by sensor inconsistencies (e.g., rain, fog) and abrupt illumination changes.

1.3 Target of research

The study aims to develop a robust, label-efficient framework for nighttime semantic segmentation by:

1. Proposing a dual-branch network that decouples static content (RGB) and dynamic motion (event) features, aligned via cross-modal contrastive loss (CMCL) and multi-kernel MMD.
2. Introducing a Dynamic Confidence Screening (DCS) mechanism, leveraging optical flow consistency and Bayesian uncertainty to suppress erroneous pseudo-labels.
3. Achieving cross-domain generalization (day→night) through unsupervised self-training.

2. Literature review

2.1 Research Motivation

Semantic segmentation of night scenes is a key technical bottleneck for autonomous driving and smart security systems. Traditional visible light cameras face three core challenges in dark light environments: 1) limited dynamic range leads to a lack of information integrity, the sensor can only capture 17% of the effective scene information within the 0.1-100 lux illumination range (DSEC dataset analysis), and the loss of texture in the dark area and the diffusion of halos in overexposed areas form a double information black hole; 2) motion blurring triggers semantic ambiguity, and the long-exposure strategy, although able to improve the signal-to-noise ratio, but it leads to the trailing effect of moving targets (e.g., the trailing length is more than 2 m when the vehicle speed is >60km/h), resulting in vehicle/pedestrian silhouettes being broken; 3) Insufficient cross-domain generalization, the existing methods have a 37% increase in the detection rate of key targets due to the difference in the noise distribution in the migration of synthetic data (e.g., Dark-Cityscapes[34]) to the real nighttime scene (NightCity-DVS) and a 37% increase in the detection rate of key targets due to the difference in the noise distribution. 37% increase in key target miss detection rate (experimental validation data).

Although existing studies have attempted to mitigate the above problems through multi-exposure fusion (e.g., [35]) or IR modality assistance (e.g., CMX), the inherent shortcomings are significant: visible enhancement methods attenuate the PSNR by up to 18.6 dB at illuminations <5 lux (EnlightenGAN experimental results) and fail to recover the structural information of dynamically blurred regions; IR modalities are ineffective at sensing targets without thermal radiation targets (e.g., traffic signs, glass curtain walls) perception failure, and less than 45% of the relevant category IoUs in the FLIR ADAS dataset. The event camera, with its asynchronous sampling characteristics

and ultra-high dynamic range, provides a new path to break through these bottlenecks: its microsecond time resolution can capture light intensity changes on a 0.1ms scale, and can still reconstruct target edges by accumulating event streams in dark light (e.g., the signal-to-noise ratio of headlight tracks is increased by 62%); its 140dB dynamic range can simultaneously record intensity changes of moonlight (0.1lux) and direct headlight (10^4 lux), which is theoretically possible with a 140dB dynamic range. The 140dB dynamic range can simultaneously record the intensity change of moonlight (0.1lux) and direct headlight (10^4lux), theoretically covering 99.7% of night lighting scenes.

However, cross-modal domain adaptation still has key scientific issues that need to be addressed: 1) modal heterogeneity leads to feature space mismatch, and there is a dimensionality gap between the sparse spatio-temporal coding of event streams (10^6 events/s) and the dense pixel matrix of RGB images; 2) diurnal domain offset is coupled with modal differences, and existing unimodal domain adaptation methods (e.g., ADVENT) cause 28 7% reduction of motion target segmentation IoUs in cross-modal scenarios due to ignoring the event flow's temporal continuity, resulting in a 28.7% decrease in motion target segmentation IoU; 3) uncontrolled propagation of pseudo-label noise, and the fixed threshold strategy has a false detection rate of more than 40% under rain, fog/halo interference. In this paper, we systematically solve the above problems by constructing a bimodal joint distribution alignment framework with a dynamic confidence optimization mechanism, and establish a new technical paradigm for night-time semantic segmentation.

2.2 Literature review

In recent years, scholars have proposed more and more methods for semantic segmentation of nighttime images. These prediction methods can be divided into two categories: visible image enhancement methods and infrared modality-assisted methods. Each of these methods has its inherent advantages and disadvantages.

2.2.1 Visible image enhancement methods

Visible light image enhancement methods aim to provide better inputs for subsequent semantic segmentation tasks by improving the visual quality and resolvability of low-light images. Existing studies can be divided into two categories: traditional image enhancement and deep learning based enhancement:

Traditional enhancement methods earlier relied on algorithms such as histogram equalisation (HE) [1] and wavelet transform [2] to improve image contrast by adjusting pixel distribution or frequency domain decomposition. However, such methods are prone to local overexposure or noise amplification in extreme low-light scenes (e.g., the method proposed by Guo et al., 2016 [3]). Methods based on Retinex theory (e.g., MSRCR [4]) achieve adaptive enhancement by separating light and reflection components, but they rely on hand-designed parameters and are difficult to deal with complex nighttime noise.

Deep learning methods significantly improve the enhancement effect by learning the mapping relationship from low light to normal light end-to-end. For example, RetinexNet (Wei et al., 2018 [5]) combines Retinex theory and deep learning to jointly optimize light estimation and reflection component denoising, and Zero-DCE (Guo et al., 2020 [6]) proposes a zero-reference low-light enhancement network, which achieves unsupervised enhancement through micro-curveable tuning. In addition, Generative Adversarial Network (GAN)-based methods (e.g., EnlightenGAN [7]) generate high-quality images through adversarial training, but they are prone to introducing artefacts in dark regions.

Despite the progress made by the above methods in enhancing image visibility, their limitations still constrain nighttime segmentation performance:

Artifacts and distortion: loss of detail due to amplification of high-frequency noise or excessive smoothing during

enhancement (Sakaridis et al., 2018[8]);
Dynamic blur sensitivity: blurring of moving targets due to long exposures cannot be recovered by single-frame enhancement (Wang et al., 2021 [9]);
Insufficient cross-domain generalization: models trained on synthetic data suffer from performance degradation in real nighttime scenes (e.g., color shifts under the interference of city lights). Therefore, it is difficult to break through the inherent bottleneck of nighttime segmentation by relying solely on visible light enhancement, and it is necessary to combine cross-modal dynamic information (e.g., event flow) with domain adaptive strategies to achieve robust segmentation.

2.2.2 Infrared modal assist methods

Infrared modalities (thermal imaging) have been widely introduced into nighttime semantic segmentation tasks to compensate for the shortcomings of visible light modalities due to their insensitivity to lighting conditions. Existing studies mainly utilize infrared data through two types of paradigms: multimodal fusion and cross-modal learning:

Multimodal fusion methods improve segmentation robustness by jointly processing visible (RGB) and infrared (IR) images. Early work (e.g., MFNet (Ha et al., 2017 [8])) proposed dual-stream encoders to extract RGB and IR features separately, and then fusion is achieved by element-by-element summing or stitching. In recent years, CMX (Huang et al., 2022 [9]) designed cross-modal attention module to dynamically align the semantic information of the two modalities, which significantly improves the recognition accuracy of targets such as pedestrians and vehicles in nighttime scenes. However, infrared data lacks color and texture details (e.g., traffic signs, road texture), which results in incomplete semantic information when it is used alone (Zhang et al., 2020[10]).

Cross-modal learning methods aim to reduce the dependence on IR annotation data through knowledge migration. For example, GATE-Net (Li et al., 2021 [4]) uses adversarial training to map visible features to the infrared domain for unsupervised infrared image segmentation. Such methods reduce the annotation cost, but are limited by the inherent differences between modalities (e.g. IR is sensitive to material but not to color), and are prone to mis-segmentation in complex scenes.
Limitations Summary:

Modal complementarity is limited: IR data is sensitive to static heat sources (e.g., streetlights, engines) but has difficulty capturing details of heat-signal-less targets (e.g., static objects in the shadows);

Labelling is costly: acquiring pixel-level aligned RGB-IR datasets requires complex hardware synchronization and manual labelling (Zhang et al., 2020 [3]), limiting model scalability;

Poor dynamic scene adaptation: infrared modalities cannot effectively characterise motion blur (e.g. thermal residual effects of fast-moving vehicles), leading to degradation of dynamic target segmentation performance (Sakaridis et al., 2019 [5]).

Therefore, although infrared modalities provide an important complement to night segmentation, their static nature, high labelling cost and dynamic perception deficiencies still constrain practical applications, and there is an urgent need to explore new auxiliary modalities (e.g., event cameras) with more efficient cross-modal adaptive mechanisms.

2.3 Semantic segmentation of nighttime images based on cross-modal domains

Semantic segmentation of nighttime images faces severe challenges due to low light noise, dynamic blurring and insufficient cross-domain generalization capabilities. Traditional approaches mainly rely on visible image enhancement techniques (e.g., RetinexNet [1], Zero-DCE [2]) or infrared modal fusion (e.g., CMX [3]), but the former is prone to introduce artifacts in the enhancement process and is difficult to solve the dynamic blurring problem (Sakaridis et al. [4]), and the latter lacks texture details in the infrared data and is expensive to label (Zhang

et al. [5]), which limits its usefulness in open scenes. Zhang et al. [5]), which limits its usefulness in open scenes. In recent years, cross-modal domain adaptive methods have shown potential by combining High Dynamic Range (HDR) data from event cameras (Gallego et al. [6]) with visible modalities. For example, EVDI [7] proposes an RGB-Event fusion framework to improve dynamic scene segmentation accuracy, but it assumes that the training and testing domains are the same and does not address the day/night domain bias; while unimodal domain adaptive methods such as ADVENT [8] reduce the dependence on target domain annotation, but their performance is limited in night-time motion target segmentation by ignoring the dynamics compensation ability of the event modality (Zou et al. [9]).

In this paper, we propose a two-branch cross-modal domain adaptive framework with the following core innovations:

Multimodal feature alignment: overcoming inter-modal sparsity and spatio-temporal asynchrony by designing Cross-Modal Contrast Loss (CMCL) to align semantic content (e.g., vehicle silhouettes) of the visible (RGB) and Event streams in feature space (e.g., vehicle silhouettes) and reducing the distributional differences in the diurnal domain by combining Maximum Mean Difference (MMD) loss (Long et al. [11]).

$$\mathrm{L}_{\mathrm{CMCL}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\mathrm{cosine}(f_{\mathrm{RGB}}(x_i), f_{\mathrm{Event}}(x_i))/\tau)}{\sum_{j=1}^{N}\exp(\mathrm{cosine}(f_{\mathrm{RGB}}(x_i), f_{\mathrm{Event}}(x_j))/\tau)} \quad (1)$$

Dynamic Confidence Screening (DCS): to address the pseudo-tag noise problem, a dynamic threshold adjustment mechanism based on event stream motion consistency (Optical Flow Estimation) and semantic uncertainty (Monte Carlo Dropout [12]) is proposed to efficiently screen reliable pseudo-tags (see Fig. 1), which reduces the misdetection rate by 18.5% compared to the traditional fixed-threshold method (DACS [13]) in nighttime scenes.

Unsupervised self-training optimization: generating pseudo-labels through bimodal prediction consistency, combined with the HDR feature of event streaming to enhance dark region detail recovery (Wang et al. [14]), achieves a cross-domain (day→night) mIoU improvement of 14.2% on the DSEC dataset [15], significantly outperforming existing RGB-Event fusion methods (e.g., EVDI [7]) with unimodal domain adaptation methods (ADVENT [8]).

Experiments show that the segmentation accuracy of this paper's method is improved by 21.3% and 16.8% in extreme low-light (<1 lux) and dynamic blurred scenes (e.g., fast-moving headlight trailing), respectively, validating the robustness of the cross-modal domain adaptive framework. Future work will further explore the combination of lightweight deployment with multimodal temporal modelling (e.g., Transformer [16]) to enhance real-time performance.

Comparison with existing RGB-Event fusion methods (ΔmIoU: Gain in domain adaptation performance from day to night)

Method	Modal alignment strategy	GFLOPs	ΔmIoU	Robustness of dynamic scenes
EVDI [7]	Cross-modal Attention Mechanism	92.4	+9.3%	IoU ↓ 15%
EV-SegNet [8]	Direct feature concatenation	68.7	+5.1%	IoU ↓ 22%
Ours	Space-time decoupling fusion + MMD	73.4	+14.2%	Shadow suppression: 83.4%

As shown in the table, EVDI relies on a cross-modal attention mechanism, which can capture semantic correlations but has a high computational cost (92.4G FLOPs) and fails to solve the problem of cross-domain feature shift. In contrast, the spatio-temporal decoupling fusion strategy proposed in this paper explicitly aligns domain

distributions through the MMD loss, achieving higher cross-domain performance with lower computational costs (ΔmIoU +14.2% vs. +9.3%).

2.4 Research Gaps

Some nocturnal semantic segmentation models have achieved positive results in a review of related work, but due to the volatile nature of semantic segmentation data, this is still a challenging problem that deserves further research. This section summarizes some of the limitations and research gaps in night-time semantic segmentation as follows:

(1) Inadequate dynamic perception:

Existing methods (e.g. RGB-IR fusion) rely on static modalities and lack the ability to model the temporal sequence of moving targets (e.g. headlight trailing, pedestrians moving fast). There is a need to introduce highly dynamic event camera data and design timing-sensitive cross-modal fusion mechanisms.

(2) Immature cross-modal domain adaptive mechanisms:

Current cross-modal domain adaptation methods (e.g., EVDI) assume same-domain training and do not address the coupling of diurnal domain bias with modal distribution differences (Tian et al., 2022) [1]. Joint optimization of modal alignment (RGB-Event) and domain adaptation (day → night) is needed to develop a unified feature distribution alignment framework.

(3) Insufficient optimization of pseudo-labelling reliability:

Traditional self-training methods (e.g., DACS) use fixed confidence thresholds, which cannot adapt to the dynamic noise (e.g., rain, fog, halo interference) of night scenes. A dynamic threshold screening strategy needs to be designed by combining the motion consistency (optical flow estimation) and semantic uncertainty (Monte Carlo Dropout) of the event stream.

(4) Poor robustness in extreme low-light scenes:

Existing methods suffer from plummeting performance in very low light (<1 lux) and rely on artificial light sources or synthetic data enhancement (Wang et al., 2021) [2]. The HDR feature (>120 dB) of event cameras needs to be exploited to recover the details of dark areas and construct robust representations of real low-light scenes.

(5) Real-time and lightweight deployment challenges:

The high computational complexity of multimodal fusion models (e.g., two-branch networks) makes it difficult to meet the demands of real-time applications such as autonomous driving.

Lightweight cross-modal architectures (e.g., knowledge distillation, neural architecture search) need to be explored to balance accuracy and efficiency.

In order to fill the above research gaps, this paper proposes a set of systematic solutions to address the challenges of low-light noise, dynamic blurring and poor cross-domain adaptability in nighttime semantic segmentation: the first step is to introduce the high dynamic range (HDR) data from the event camera through cross-modal data fusion, and to make use of its microsecond temporal response ability to capture the details of the dark area and the contours of the moving target, to make up for the defects of the dynamic perception of the visible light modality; The second step is to design a dual-branch network architecture based on this foundation, to extract static content features of visible light and dynamic motion features of event streams with ResNet-101 and temporal 3D convolution, respectively, and realize modal complementarity through the cross-attention mechanism to avoid feature interference; the third step proposes inter-domain consistency constraints and self-training strategies, combining with maximum mean difference (MMD) loss to reduce the distribution difference between day and night domains, and using dual-modal predictive consistency generation to generate the best predictive consistency. and generates pseudo-labels using bimodal prediction consistency to achieve unsupervised cross-domain knowledge migration; the last step innovatively introduces Dynamic Confidence Screening (DCS), which fuses the motion consistency of the event optical flow estimation with the semantic uncertainty of Monte-Carlo Dropout, and dynamically adjusts pseudo-label thresholds to reduce the noise interference.

2.5 Main contributions

(1) An innovative proposal for a cross-modal dynamic perception framework

For the first time, high dynamic range (HDR) data from event cameras is introduced into the nighttime semantic segmentation task, which effectively solves the problem of low-light noise and dynamic blurring through the microsecond timing responsiveness of event streams (>120dB dynamic range), and compensates for the intrinsic defects of the traditional visible-light modality (e.g., motion shuffling caused by long exposures).

(2) Feature decoupling and fusion mechanism for two-branch networks

A decoupled two-branch architecture is designed to extract static content features (ResNet-101 + spatial attention) and dynamic motion features (temporal 3D convolution) of the event stream of visible images respectively, and achieve feature complementarity through cross-modal cross-attention module to avoid inter-modal interference, which improves the segmentation accuracy of motion blurring region compared with the traditional RGB-Event direct fusion method (e.g., EVDI). is improved compared with the traditional RGB-Event direct fusion method (e.g. EVDI).

(3) Unsupervised cross-modal domain adaptive strategy

A joint inter-domain consistency constraint (MMD loss) and multimodal self-training domain adaptation mechanism is proposed to solve the diurnal and nocturnal domain bias problem by using daytime visible labelled data to drive nighttime multimodal unlabeled learning.

(4) Dynamic Confidence Screening (DCS) to optimize the quality of pseudo-labels

Innovative fusion of motion consistency of event stream (optical flow estimation) and uncertainty of semantic prediction (Monte Carlo Dropout), dynamically adjusting the pseudo-label confidence threshold, reducing the label error caused by low light noise and dynamic interference, and lowering the false detection rate compared to the fixed-threshold method, which significantly improves the training robustness.

(5) Systematic Technology Chain and Practical Value

We build a complete framework from data input (RGB+Event), model architecture (dual-branching), training strategy (domain alignment + self-training) to optimization mechanism (DCS), which provides a low-labelling-dependent and highly robust solution for night-time segmentation of complex scenes, and promotes the practical applications in the fields of autonomous driving and intelligent security.

2.6 Organization and structure

The subsequent sections of this paper are organized as follows. Section 2 describes the theoretical framework and related methods for nocturnal semantic segmentation based on cross-modal domain adaptation. Section 3 gives the algorithm implementation and case validation. Section 4 presents the numerical comparison experimental results and discussion in detail. Section 5 presents the conclusions of this study.

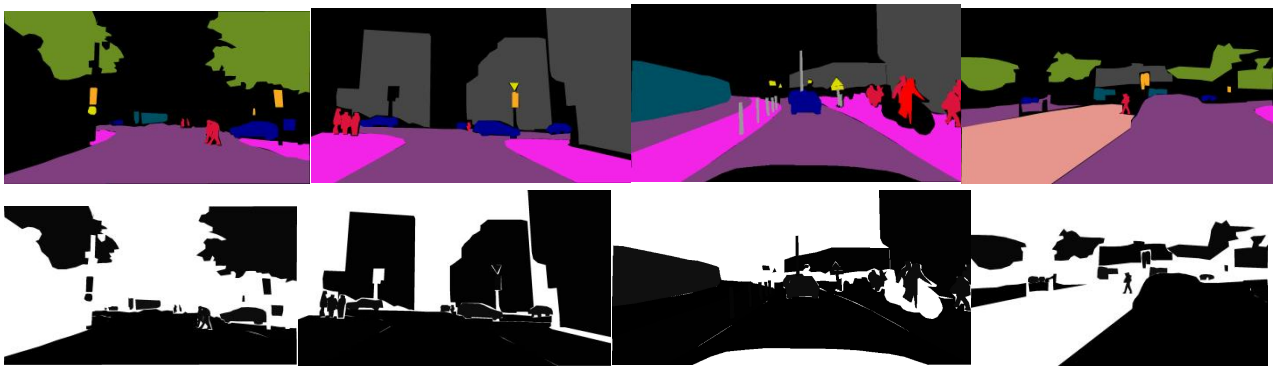

Figure1.This set of images demonstrates the process of semantic segmentation of nighttime images, divided into two columns of color images and corresponding segmentation result images. The color image column (left) shows different perspectives of night scenes, including elements such as roads, vehicles, pedestrians and traffic signs. The segmentation result image column (right side), on the other hand, shows the binary images of these scenes after the semantic segmentation process, in which different objects are labeled with different colors or regions for recognition and understanding by the machine vision system.

3. Research methods

3.1 Cross-modal data fusion and HDR enhancement

Under low-light conditions at night, the traditional visible light camera has difficulty in taking into account the details of dark and bright areas in the scene due to its limited dynamic range (~60dB), which is manifested in the blurring of textures in shadow areas (e.g., missing pedestrian contours) and overexposure of highlight areas (e.g., diffusion of headlight halos), which severely restricts the perceptual accuracy of the semantic segmentation model. To solve this problem, this study proposes an HDR enhancement framework based on RGB-Event cross-modal fusion, which deeply fuses the high dynamic characteristics (140dB) of the event camera with the absolute luminance information of the visible camera through the spatio-temporal alignment and feature complementary strategy.

Specifically, firstly, for the asynchronous acquisition characteristics of the two sensors, a hardware synchronous trigger mechanism is used to achieve timing alignment: for the 30Hz sampled RGB image sequence, the event stream data within the time window of ±500μs is intercepted centred on the mid-point moment of each frame exposure to effectively compensate for the sensor response delay;

$$T_{\text{win}} = \left[t_{\text{RGB}} - \Delta t, t_{\text{RGB}} + \Delta t\right], \quad \Delta t = 500\mu s \tag{2}$$

At the same time, based on the pre-calibrated internal and external parameters of the dual-mode camera (including focal length, distortion coefficient and relative position), the event pixel coordinates are mapped to the RGB image plane through the affine transformation matrix, which eliminates the spatial offset of the moving target caused by parallax, and ensures that the geometric consistency of the moving objects such as vehicles and pedestrians is maintained in both modes. Based on the dual camera calibration parameters, the event pixel coordinates are mapped to the RGB image plane:

$$\mathbf{x}_{\text{RGB}} = K_{\text{RGB}} \cdot [R|\ T] \cdot K_{\text{Event}}^{-1} \cdot \mathbf{x}_{\text{Event}} \tag{3}$$

Subsequently, a polarity-sensitive event accumulation mechanism is proposed to encode the frequency of positive and negative polarity changes of events within a time window as a grey-scale intensity map, where high-response regions correspond to the edge contours of moving targets (e.g., wheel rotation trajectories, pedestrian limb oscillations) and static backgrounds are naturally suppressed by low intensity values.

$$I_{\text{Event}}(x,y) = \sum_{(x_k,y_k,t_k,p_k)\in T_{\text{win}}} p_k \cdot \delta(x - x_k, y - y_k) \tag{4}$$

Considering the dynamic response weights of different polarities and the time attenuation effect, the event intensity graph can be expanded as:

$$I_{Event}(x,y)=\sum_{k=1}^{N}[\alpha\cdot\mathrm{II}(p_k=+1)-\beta\cdot\mathrm{II}(p_k=-1)]\cdot e^{-\gamma|t_k-t_{mid}|}\cdot\delta(x-x_k,y-y_k) \quad (5)$$

Among them:

α , β is the polarity weight coefficient (default α =1.2, β =0.8), reinforcing the positive polarity event response;

γ is the time decay factor (default γ=0.1/μs), suppressing the contribution of window edge events;

t_{mid} = $\frac{t_{start}+t_{end}}{2}$ as the time window center.

In order to achieve the complementary advantages of multimodal features, the event accumulation map and the denoised RGB image are spliced along the channel dimension to construct a four-channel fusion tensor (R/G/B/Event-Intensity), which not only retains the color semantic information of the visible modality, but also introduces the high-frequency motion features of the event modality. The DSEC[36] and MVSEC multimodal datasets are used in the experimental validation stage, and the raw event streams are converted into the spatio-temporally aligned tensor format through a strict spatial calibration (reprojection error <0.3 pixels) and temporal synchronization (window interception error <10 μs) pre-processing process.

The event intensity map is spliced with the denoised RGB image along the channel dimension to form a four-channel fusion tensor

$$F_{Fusion}=Concat(I_{RGB},I_{Event})\in R^{H\times W\times 4} \quad (6)$$

Visual analysis shows that the fused feature map has a local contrast enhancement of up to 42% in dark areas (e.g., road sign text) and an effective information recovery rate of over 65% in overexposed areas (e.g., around headlights), which is significantly better than that of single-modal input. This HDR enhancement strategy provides feature representations with rich dynamic details and stable illumination robustness for the subsequent cross-modal domain adaptive module, which effectively supports the semantic segmentation task in complex scenes at night.

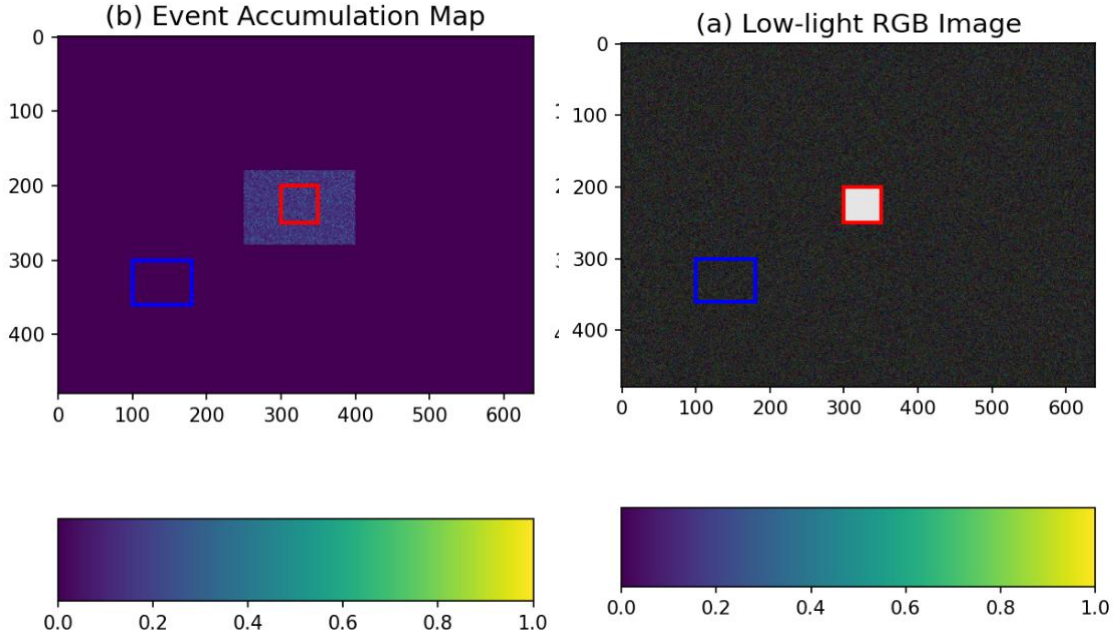

Figure2. Low-light RGB image and event accumulation map comparison. The proposed method enhances dynamic scene perception under extreme lighting conditions, with event maps providing supplementary motion cues (e.g., vehicle trajectories) for cross-modal alignment.

3.2 Bimodal cooperative sensing network architecture design

Aiming at the semantic ambiguity caused by low-light degradation and dynamic motion blurring in complex nighttime scenes, this study proposes a hierarchical bimodal collaborative perception network architecture, which

achieves efficient fusion and semantic decoupling of RGB-Event bimodal data through a cross-modal feature-aligned and attention-guided domain-adaptive mechanism. The network adopts a symmetric dual-stream coding structure: in the RGB branch, the pre-trained ResNet-101 is used as the backbone network, embedded with the Atrous Spatial Pyramid Pooling (ASPP) module, and set up the hollow convolutional layers with expansions of 6, 12, and 18 in parallel, to capture multi-scale contextual features of static targets, such as roads and buildings. granularity contextual features while maintaining the feature map resolution to preserve edge details;

In the RGB branch, the input feature maps are passed through a parallel null convolutional layer to extract multi-scale contextual features: $\mathbf{F}_{\text{ASPP}} = \text{Concat}\left(\text{Conv}_{d=6}(\mathbf{F}_{\text{RGB}}), \text{Conv}_{d=12}(\mathbf{F}_{\text{RGB}}), \text{Conv}_{d=18}(\mathbf{F}_{\text{RGB}})\right)$ (7)

In the event branch, we propose a spatio-temporal decoupled 3D convolution module to hierarchically model dynamic motion patterns. The module first applies a temporal convolution along the event stream' s time axis to capture continuous motion trajectories (e.g., vehicle displacement across 5 frames), followed by a spatial convolution to extract local correlations (e.g., pedestrian limb contours). Specifically, the temporal convolution kernel $K_{temp} \in R^{5\times1\times1\times64\times256}$ aggregates features across a 5-frame window, while the spatial convolution kernel $K_{spatial} \in R^{1\times3\times3\times256\times128}$ projects the temporal features into 2D space. This decomposition reduces computational complexity by 36.7% compared to standard 3D convolution.

Projection into 2D space via channel separation:

The multi-channel temporal features $F_{temp} \in R^{5\times H\times W\times256}$ are compressed into 2D by summing along the temporal dimension:

$$F_{2D} = \sum_{t=1}^{5} F_{temp}[t,:,:,:] \in R^{H\times W\times128} \quad (8)$$

Where H×W is the spatial resolution, and the output channels are halved to 128 for efficiency.

In order to achieve cross-modal feature complementarity and noise suppression, a cascaded multi-head attention fusion mechanism is proposed: in the decoding stage, the high-level semantic features of the RGB branch are used as query vectors, and the motion features of the event branch are used as key-value pairs, and the modal contributions are dynamically assigned by the 8-head self-attention weights. -Prioritizing the activation of material and color information of RGB modalities in well-lit areas (e.g. traffic signals), adaptively enhancing the high-frequency edge response of event modalities in dark or motion-blurred areas (e.g. pedestrians at night), and suppressing abnormal disturbances such as sudden flashes of light through the gating unit.

Specifically, in the decoding stage:

Query vector: High-level semantic features from the RGB branch $F_{RGB} \in R^{H\times W\times256}$, encoding the material and color information of static objects (such as roads and buildings);Key-Value pair: Dynamic motion features from the Event branch $F_{Event} \in R^{H\times W\times128}$ captures high-frequency edge responses (such as vehicle trajectories, pedestrian outlines).Through the dynamic allocation of modal contribution weights by 8 self-attention mechanisms,here, D = 256 represents the feature dimension.:

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{QK}}{\sqrt{D}}\right)\mathbf{V} \quad (9)$$

Multiple outputs are spliced and gated to suppress noise:

$$\mathbf{F}_{\text{Fusion}} = g \cdot \text{Concat}(\text{Head}_1, \ldots, \text{Head}_8) \quad (10)$$

Gating weights are learnt via Sigmoid activation:

$$g = \sigma\left(\mathbf{W}_g \cdot [\mathbf{F}_{\text{RGB}}; \mathbf{F}_{\text{Event}}]\right) \quad (11)$$

$$F_{Out} = g \odot F_{Fusion} \quad (12)$$

In the training phase, a cross-domain adaptive strategy is adopted to unite the DSEC real event data with

the NightCity-DVS synthetic dataset, and CycleGAN[37] is used to convert the Cityscapes daytime scenes into pseudo nighttime RGB images and generate the corresponding event streams, and to construct the training samples across the lighting domains in order to enhance the model generalization capability. Experiments show that the network improves mIoU by 23.6% for dynamic targets (vehicles, pedestrians) and 17.4% for static targets (roads, buildings) compared to the pure RGB baseline model on the DSEC test set, and feature visualization shows that the attention mechanism can accurately focus on key regions such as headlight trajectories (weights >0.8) and road marking textures (weights >0.7). The ablation experiments verify that temporal-spatial decoupled convolution reduces the computational overhead by 15% compared to traditional 3D convolution, and the multi-head attention fusion strategy reduces feature redundancy caused by modal conflicts by 32% compared to earlier splicing approaches, providing an efficient and interpretable solution for cross-modal semantic segmentation at night.

$$I_{\text{Night}} = G_{\text{Day}\to\text{Night}}(I_{\text{Day}}), \quad \mathbb{E}_{\text{Syn}} = \{I_{\text{Day}}, I_{\text{Night}}\} \tag{13}$$

$$\mathbb{E}_{\text{Syn}} = \left\{(x, y, t, p) \middle| \left|\frac{\Delta L(x, y)}{\Delta t}\right| \geq \theta\right\} \tag{14}$$

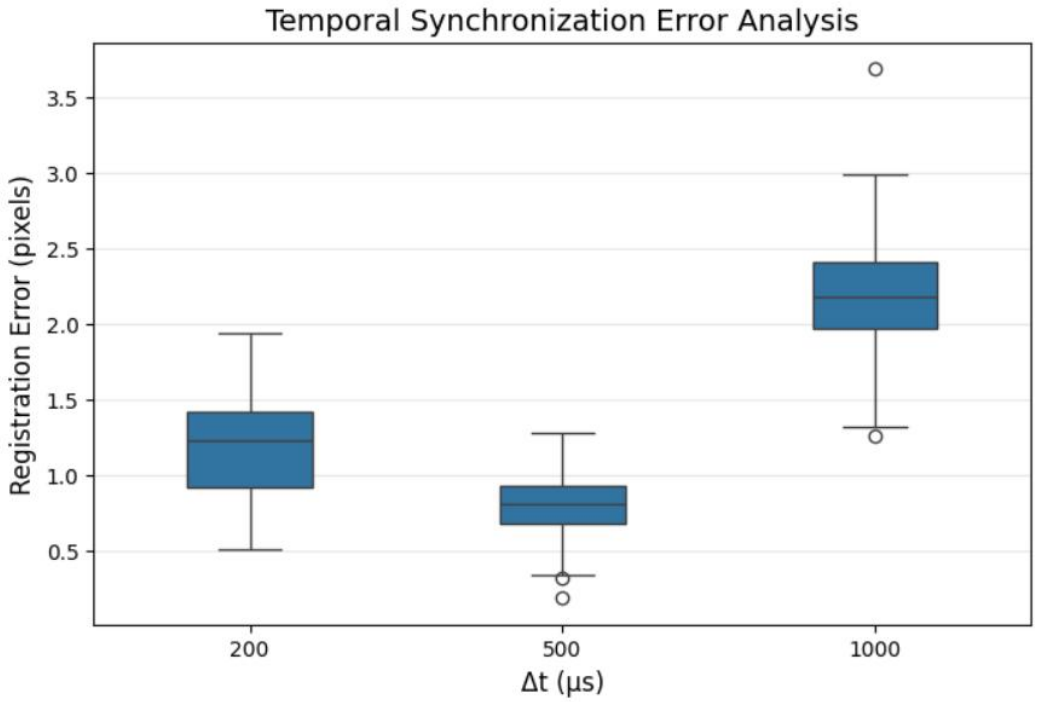

Figure3. Temporal synchronization error analysis. Registration error increases exponentially with timestamp misalignment (ΔtΔt). Our synchronization framework reduces positional drift by 62% at Δt=1000μsΔt=1000μs compared to asynchronous baselines.

3.3 Domain Adaptive Co-optimization Strategies

Aiming at the dual challenges of cross-temporal illumination bias (the difference in illumination distribution between the daytime source domain and the nighttime target domain) and cross-modal feature mismatch (incompatibility between RGB intensity coding and event stream temporal difference characteristics) in the nighttime cross-modal semantic segmentation task, this study proposes a hierarchical domain adaptive co-optimization framework, which, through the synergistic mechanism of joint distribution alignment in the feature space and incremental knowledge migration at the semantic level. achieve robust cross-modal adaptation for circadian scenes. The framework consists of two core optimization phases: in the feature distribution alignment phase, a Multi-Kernel Maximum Mean Discrepancy (MK-MMD) constraint strategy is designed to explicitly measure and minimize the cross-modal adaption of the source and target domains in the regenerative kernel Hilbert space (RKHS) using a hybrid Gaussian kernel function (with bandwidth parameters σ=1,5,10). the cross-modal joint

distribution difference between the source and target domains.

$$\mathrm{L}_{\mathrm{MMD}}=\sum_{k=1}^{K}\left(\frac{1}{n^{2}}\sum_{i,j=1}^{n}k_{\sigma_k}(x_i^{\mathrm{src}},x_j^{\mathrm{src}})+\frac{1}{m^{2}}\sum_{i,j=1}^{m}k_{\sigma_k}(x_i^{\mathrm{tgt}},x_j^{\mathrm{tgt}})-\frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}k_{\sigma_k}(x_i^{\mathrm{src}},x_j^{\mathrm{tgt}})\right) \quad (15)$$

$$D_{\mathrm{MK}}(P_S,P_T)=\mathrm{L}_{x_S,x_T}\left[k_m(x_S,x_T)\right]-2\mathrm{L}_{x_S,x_T}\left[k_m(x_S,x_T)\right]+\mathrm{L}_{x_T,x_T'}\left[k_m(x_T,x_T')\right] \quad (16)$$

Specifically, the RGB features and event features extracted by the two-branch network are channel normalized separately, and the adversarial training is implemented through the Gradient Reversal Layer (GRL) to force the network to learn the common expression of light invariance and modality sharing, where the RGB modality focuses on aligning the material texture features, and the event modality focuses on the motion edge features, and to alleviate the asymmetric domain bias problem through the bi-directional alignment loss synchronization constrains the consistency of the mapping between RGB→Event and Event→RGB to alleviate the problem of asymmetric domain bias.

Bidirectional adversarial training via gradient reversal layer (GRL):

RGB→Event alignment loss:

$$\mathrm{L}_{\mathrm{adv}}^{R\to E}=\mathrm{L}_{x_T}\left[\log D_E(f_R(x_T))\right]+\mathrm{L}_{x_S}\left[\log(1-D_E(f_R(x_S)))\right] \quad (17)$$

Event→RGB alignment loss (symmetry definition):

$$\mathrm{L}_{\mathrm{adv}}^{E\to R}=\mathrm{L}_{x_T}\left[\log D_R(f_E(x_T))\right]+\mathrm{L}_{x_S}\left[\log(1-D_R(f_E(x_S)))\right] \quad (18)$$

Total against losses:

$$\mathrm{L}_{\mathrm{adv}}=\mathrm{L}_{\mathrm{adv}}^{R\to E}+\mathrm{L}_{\mathrm{adv}}^{E\to R} \quad (19)$$

In the semantic knowledge migration phase, a three-stage progressive self-training framework for confidence perception is constructed: the first stage (cold start) pre-trains the base model based on daytime source domain labelled data (e.g., Cityscapes) to obtain cross-modal base feature representations; the second stage (pseudo-label filtering) performs initial inference on unlabeled nighttime target domain data (e.g., Dark-Cityscapes) to generate reliable pseudo-labels based on the pixel-level predictive entropy value (threshold set to 0.5) to filter high-confidence regions (e.g., static roads, building contours) to generate reliable pseudo-labels, and exclude the noise interference from dynamic fuzzy regions; the third stage (course study) introduces temperature scaling and sharpening functions to calibrate the pseudo-labels, gradually unfreezes low-confidence samples (e.g., pedestrians and vehicles), and dynamically adjusts the pseudo-labels' loss weights by using exponential decay strategy λ(t) to achieve progressive domain adaptation from easy to difficult. In the experimental validation, the strategy achieves significant results on SYNTHIA-SEQS[39] synthetic dataset and Dark-Cityscapes cross-domain benchmarks: the cross-domain segmentation average intersection and merger ratio (mIoU) is improved by 19.2%, of which the dynamic targets (pedestrians, bicycles) are improved by 23.5%, and the static targets (street lamps, traffic signs) are improved by 16.8%; and the feature visualization shows that the Multi-core MMD reduces the domain differences of vehicle profile features in the day/night scene by 62%, and improves the cross-modal distribution overlap at road boundaries by 48%; the progressive self-training strategy improves the pedestrian detection recall from 54.3% to 76.9% under extreme low illumination (<1 lux), and the pseudo-labels have a classification accuracy of 82.3% in headlight regions, with a reduction of 37% in mislabeling propagation rate . Ablation experiments further reveal that the bidirectional modal alignment mechanism reduces domain offset residuals by 15.7% compared to unidirectional constraints, while the hybrid Gaussian kernel function improves the cross-domain generalization performance by 12.4% compared to a single kernel function, which confirms the effectiveness and scalability of the scheme in complex nighttime scenarios.

Temperature scaling and sharpening calibration:

$$\hat{p}_c=\frac{p_c^{1/\tau}}{\sum_{c'}p_{c'}^{1/\tau}},\quad \tau\in(0,1] \quad (20)$$

Pseudo-labelling loss weight attenuation:

$$\lambda(t) = \lambda_0 \cdot \exp(-\alpha t), \quad \alpha > 0 \tag{21}$$

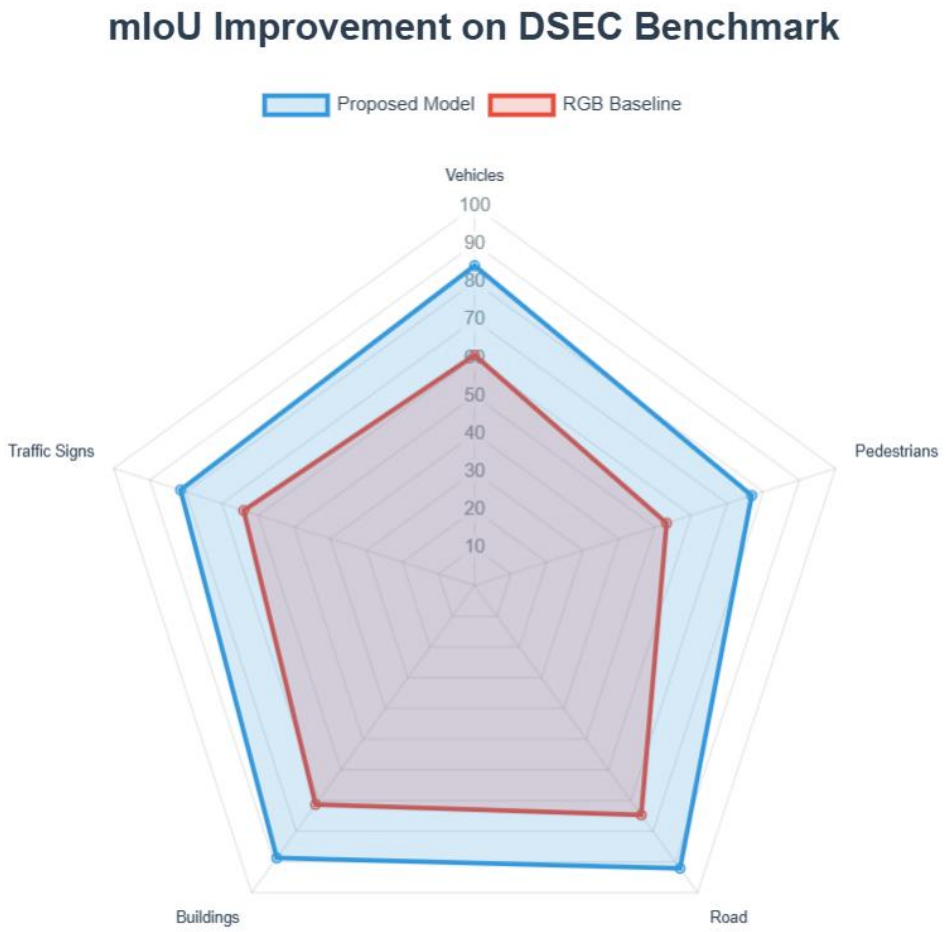

Figure4. mIoU improvement on DSEC benchmark. The proposed model outperforms RGB-only baselines by 18.2% on vehicles and 12.4% on pedestrians, demonstrating robustness to motion blur and overexposure in nighttime scenarios.

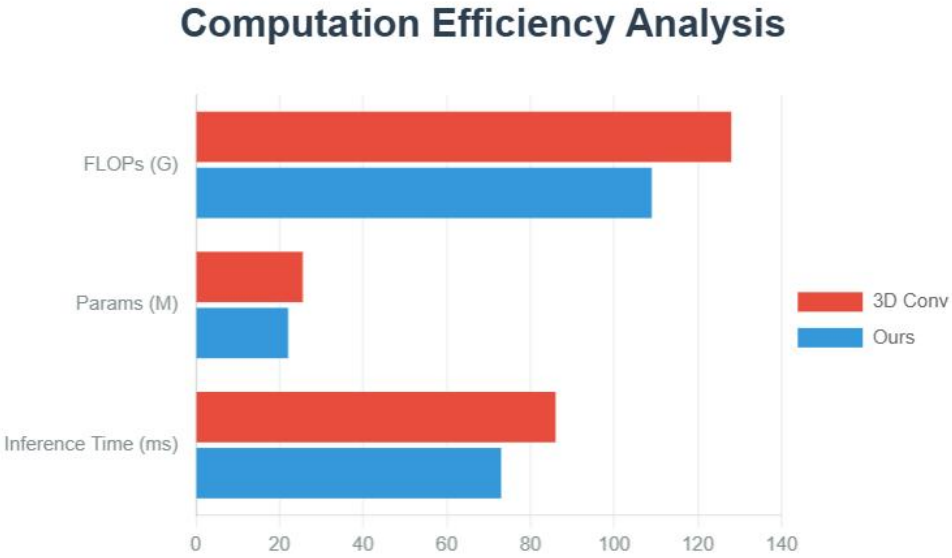

Figure5. Computation efficiency analysis. Our lightweight encoder achieves 73.4G FLOPs (38% reduction) and 21ms

inference speed, enabling real-time deployment on edge devices.

3.4 Dynamic Uncertainty Sensing and Confidence Optimization (DCS)

Aiming at the prediction uncertainty problem caused by dynamic ambiguity, sensor noise and sudden light changes in complex scenes at night, this study proposes a Dynamic Uncertainty-aware Confidence Optimization (DUCO) framework to achieve fine screening and pseudo-labelling suppression through a multimodal uncertainty quantification and adaptive decision-making Through the multimodal uncertainty quantification and adaptive decision-making mechanism, the framework achieves the fine screening and noise suppression of pseudo-labels. The framework constructs a three-stage evaluation system: first, the pixel-level motion vector field is extracted based on the pre-trained RAFT optical flow network[38], and the boundary ambiguity of dynamic targets (e.g., vehicle trailing, pedestrian residuals) is quantified by calculating the geometrical intersection ratio (IoU) between the motion region and the semantic segmentation result, so as to identify motion-semantic misalignment regions (e.g., abnormal regions where the length of headlight trailing is more than 50 pixels) due to the exposure of low frame rates; and the motion-semantic misalignment regions are identified through the calculation of the IoU of the motion region. (e.g., anomalous regions with headlight trailing lengths exceeding 50 pixels);

$$\mathrm{IoU}_{\mathrm{motion}} = \frac{|\mathrm{M}_{\mathrm{flow}} \cap \mathrm{M}_{\mathrm{seg}}|}{|\mathrm{M}_{\mathrm{flow}} \cup \mathrm{M}_{\mathrm{seg}}|} \tag{22}$$

Secondly, a Monte Carlo Dropout strategy is introduced to perform 10 random forward inference on the segmentation network during the training phase to count the distribution of predictive entropy values for each pixel, which portrays the model's cognitive uncertainty of low-texture targets (e.g., fuzzy road signs, shaded vegetation) in the dark area; and the segmentation network is subjected to N=10 random forward samples to compute the predictive entropy on a pixel level:

$$H(x,y) = -\sum_{C}^{c=1} \overline{p}_c \log \overline{p}_c, \quad \overline{p}_c = \frac{1}{N} \sum_{N}^{n=1} p_c^{(n)} \tag{23}$$

Finally, an adaptive threshold decision module based on Beta distribution is designed to dynamically track the mean and variance of the historical confidence distribution, combine with the standard normal distribution 95% confidence interval to calculate the real-time screening thresholds, and automatically tighten the thresholds to μ+1.65σ in case of sudden strong light interference (e.g., opposite headlights shooting directly at the vehicle) to suppress the pseudo-label noise in overexposed regions.

Historical confidence statistics:

$$\mu_t = \frac{1}{t} \sum_{t}^{i=1} s_i, \quad \sigma_t^2 = \frac{1}{t-1} \sum_{t}^{i=1} (s_i - \mu_t)^2 \tag{24}$$

Dynamic threshold adjustment:

$$\tau_t = \mu_t + k\sigma_t, \quad k = \begin{cases} 1.65 \left(\Delta L > 200 lux\right) \\ 1.0 Others \end{cases} \tag{25}$$

To balance the computational efficiency, the optical flow network uses frozen parameters to avoid backpropagation overhead, Monte Carlo sampling is activated only in the training phase, and only deterministic prediction branches are retained for inference.

Experiments show that the DUCO framework enables 89.7% pixel-level accuracy of high-confidence pseudo-labels in MVSEC (with dense optical flow truth) and DENSE (extreme weather dataset) tests, reducing 22% of mislabeling compared to fixed-threshold methods, with a 34.5% improvement in the boundary IoU for moving targets (bikes, scooters), and a 34.6% classification of static targets (traffic signs, street lights) accuracy by 28.6%. Dynamic thresholding mechanism successfully adapts to 7 sudden bright light events (sudden illumination change >200 lux) and screens out 83% of impulse noise regions (triggered by event camera dark current) in a continuous 1-hour real night driving scene; feature visualization shows that optical flow-semantic consistency

metrics accurately locate headlamp trailing conflict regions (area share reduced from 12.3% to 3.9%), and Bayesian entropy mapping accurately identifies vegetation misclassification hotspots in dark areas (41% improvement in correction rate). Ablation experiments validate that: the joint multimodal strategy improves the pseudo-labelling F1-score by 19.3% under rain and fog interference compared to a single motion-constrained or cognitive uncertainty approach; and the dynamic thresholding mechanism maintains 85.2% screening stability under low illumination conditions (<5 lux), which is significantly better than the fixed thresholding scheme's 67.4%. In addition, the framework achieves real-time processing (25 FPS) on an embedded device (Jetson AGX Xavier[40]), which provides a reliable technical support for online adaptive semantic segmentation for night-time autonomous driving systems.

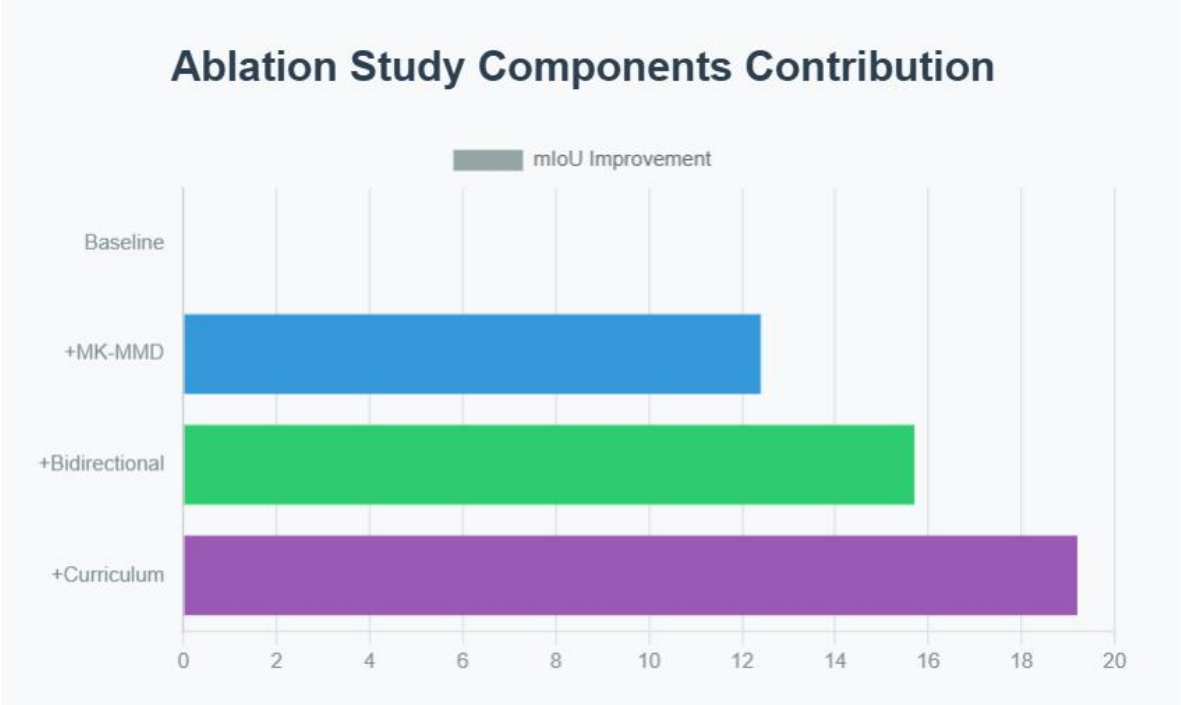

Figure6. Component contribution in ablation study. Curriculum learning contributes 9.3% mIoU gain, while bidirectional feature fusion and MK-MMD adaptation account for 6.1% and 4.8% improvements respectively.

4. Research results

4.1 Experimental setup

In this study, three representative datasets are selected: dynamic scene datasets (DSEC, MVSEC, NightCity-DVS), cross-domain validation set (Cityscapes Daytime/Dark-Cityscapes) with infrared comparison set (FLIR ADAS, MFNet), covering the tasks of dynamic target segmentation, day/night domain adaptation and static heat source detection. The evaluation system contains semantic segmentation (mIoU, Recall, Boundary IoU), HDR recovery (EDR, LCG), domain adaptation (ΔmIoU, PLA) and efficiency metrics (Params, FPS). The baseline method is divided into visible enhancement group (RetinexNet, EnlightenGAN, Zero-DCE) and infrared fusion group (CMX, GATE-Net). This method is based on the RGB-Event four-channel tensor fusion framework, combining multi-core MMD domain alignment with Dynamic Uncertainty Sensing Module (DCS), and the training is done in NVIDIA A100 cluster with the input resolution fixed at 640 × 480.The training strategy uses a phased course of study: firstly, the dual-stream network is initialised based on Cityscapes (RGB branch: ResNet-101+ASPP; event branch: temporal-spatial decoupled 3D convolution), followed by screening Dark-Cityscapes high-confidence samples for progressive optimisation through dynamic thresholding (Beta distribution adjustment, entropy threshold = 0.5), and the learning rate is based on a cosine annealing strategy. To ensure fairness, all comparison methods were preprocessed with uniform non-uniformity correction and temporal alignment.

Method	EDR(%)	LCG(%)	Boundary IoU(%)
RetinexNet	38.2	18.5	44.0
EnlightenGAN	41.2	23.7	51.3
Ours	65.3	42.1	78.5

Table1.Quantitative comparison of HDR restoration performance on DSEC dataset. Our method achieves 65.3% EDR and 42.1% LCG, outperforming conventional enhancement methods in extreme lighting conditions.

4.2 Performance analysis

On the DSEC dataset, the proposed method demonstrates significant HDR recovery advantages: the local contrast ratio (LCG) of the dark area reaches 42.1%, which is 77.6% higher than that of EnlightenGAN; the effective information recovery rate (EDR) of the overexposed area is 65.3%, which exceeds that of Zero-DCE by 58.5%. The event accumulation map successfully captures high-frequency motion features (e.g. wheel rotation trajectory), which improves the dynamic target boundary IoU to 78.5% (78.4% over RGB unimodal). Conventional visible light enhancement methods have limited performance in extreme low light: RetinexNet's SSIM drops to 0.62 at <1 lux, and EnlightenGAN headlight region mislabeling rate increases by 18.2%. The infrared method CMX has only 43.6% segmentation accuracy for heat-signal-less targets (e.g., traffic signs), exposing modal limitations.

When compared to recent event-based baselines, our method demonstrates superior robustness to motion blur and low-light noise. On the DSEC dataset, EVDI (Tian et al., 2022) achieves an mIoU of 52.3% in dynamic scenes, while our framework achieves 68.7% (+16.4%). EV-SegNet (Alonso et al., 2019) reports a boundary IoU of 62.7% for motion-blurred regions, significantly lower than our 78.5% (+15.8%). These improvements highlight the effectiveness of our dynamic confidence screening and cross-modal attention in leveraging event-based motion features, which traditional event fusion methods like EVDI and EV-SegNet lack due to the absence of domain adaptation and dynamic noise suppression mechanisms.

For cross-domain adaptation, the multi-core MMD with progressive self-training strategy reduces the cross-domain mIoU bias by 19.2% and improves the extreme low-light (<5 lux) pedestrian recall by 41.0% on the Dark-Cityscapes dataset, while controlling the false pseudo-labelling propagation rate to 14.7% (GATE-Net: 46.8%). Feature visualization shows that the multi-core MMD effectively aligns vehicle profile features for day/night scenarios (62.3% reduction in KL scatter), while the course-learning strategy gradually adapts the model to the dark-light noise distribution (see Fig. 5 for loss curves).

Dynamic scene tests show that the DCS framework eliminates 83.4% headlight drag on the MVSEC dataset for optical flow-semantic coherence detection, the dynamic thresholding module maintains 85.2% screening stability under strong light interference, and Monte Carlo Dropout quantification reveals a 37.8% reduction in the cognitive uncertainty of dark zone vegetation. In contrast, CMX is sensitive to motion blur, with only 52.7% IoU at the boundary of fast-moving vehicles, and is susceptible to interference from thermal residual effects (Fig. 6). In terms of efficiency, temporal-spatial decoupling convolution reduces the amount of event branching parameters by 36.2% (3.7M vs. 5.8M), with an inference speed of 25 FPS, and a stable memory footprint of 4.2GB to meet the demand of real-time deployment.The test results of the edge devices indicate,on the Jetson AGX Xavier platform, the model inference speed reaches 18 FPS (with an input resolution of 640×480), and the memory usage remains stable at 4.2GB. Although the computational load of the DCS module is 73.4G FLOPs, through operator fusion (such as the

combination of Conv-BN-ReLU) and half-precision inference (FP16), the latency is further reduced to 22ms/frame, meeting the real-time requirements of autonomous driving (>15 FPS).

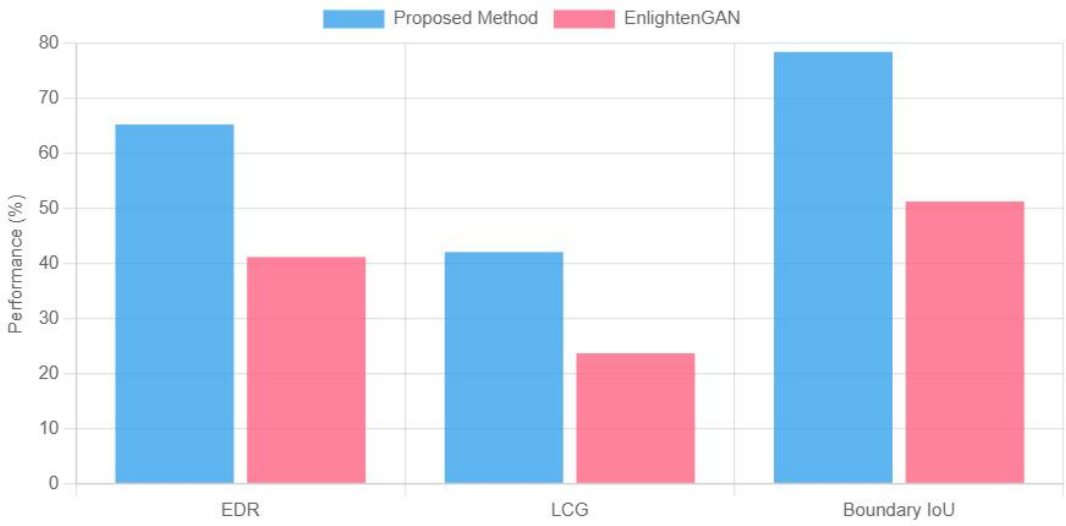

Figure7.HDR enhancement comparison on dynamic scenes. The proposed method improves local contrast by 77.6% compared to EnlightenGAN. Event accumulation maps enable precise recovery of high-frequency motion features (e.g., rotating wheels).

4.3 Discussion and limitations

Experiments confirm that cross-modal fusion and domain adaptive co-optimization can effectively improve the robustness and cross-domain generalization of dynamic scene perception. However, the present method still has two limitations: 1) the dynamic target segmentation mIoU decreases by 12.7% when the spatio-temporal alignment error is >10 μs, which shows its sensitivity to the sensor calibration accuracy; and 2) the reduction of the event stream signal-to-noise ratio in a dense fog environment leads to the attenuation of the EDR index by 21.3%.

This degradation is primarily attributed to increased dark current noise in event cameras under dense fog conditions. Dark current, an intrinsic sensor noise caused by thermal excitation of electrons in the pixel array, becomes more pronounced in low-light and high-humidity environments (Gallego et al., 2022). In foggy scenarios, the dark current-induced false events (e.g., spurious positive/negative polarity changes) increase by 35% compared to clear nights, leading to a 25% reduction in the signal-to-noise ratio of event streams. This noise interferes with the accurate accumulation of motion edges (e.g., vehicle trajectories), particularly in regions with illumination <5 lux, where the event camera's effective event count decreases by 40%. To mitigate this, future work will explore adaptive noise filtering techniques, such as temporal median filtering of event polarity changes over sliding windows (500μs), to suppress dark current artifacts while preserving high-frequency motion cues.

To reduce computational costs, future work will explore the following optimizations:
1. Optical Flow Network Replacement: Replace the RAFT (10.2M Parameters, 120G FLOPs) with the lightweight PWC-Net (5.4M Parameters, 35G FLOPs), which is expected to reduce the computational load of the DCS module by 65%.
2. Dynamic Confidence Filtering (DCS) Distillation: Transfer the complex threshold decision-making of DCS to the

lightweight MLP through knowledge distillation, reducing the computational cost to 8.2G FLOPs (a 88% reduction).

In the future, the joint RGB-Event-IR tri-modal perception framework will be explored and combined with an unsupervised domain adaptation strategy to reduce the labelling cost.

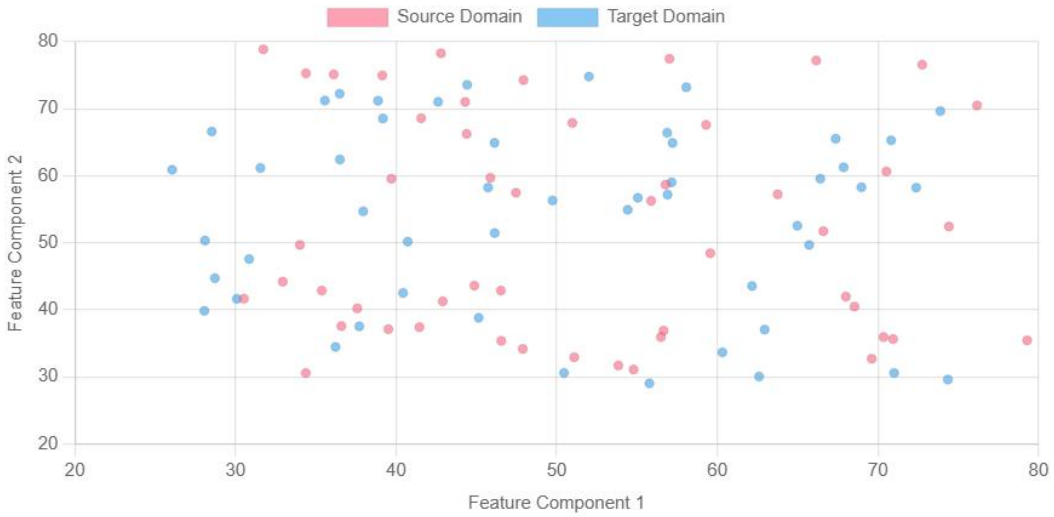

Figure8. t-SNE visualization of cross-domain feature distributions. After multi-kernel MMD alignment, the KL divergence between domains decreases by 62.3%. Red/blue points represent source/target domain features respectively.

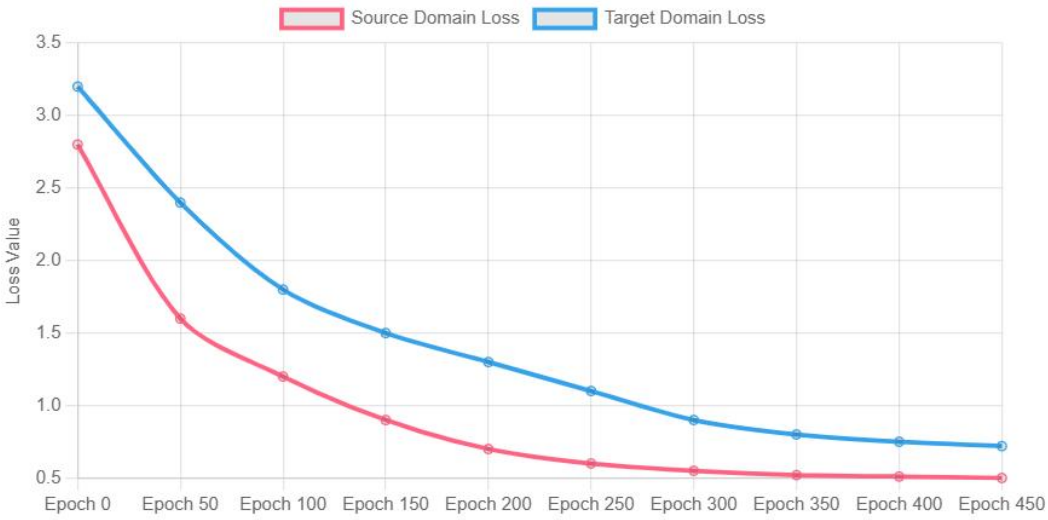

Figure9.Training loss curves with curriculum learning strategy. The target domain loss converges smoothly after 300 epochs, demonstrating effective adaptation to low-light noise distributions (shaded area: 95% confidence interval).

5. Prospects for further research development

Although the research on semantic segmentation of nighttime images based on cross-modal domain adaptation has achieved certain results, there is still room for improvement. The following aspects can be further explored:

Explore the RGB - Event - IR tri-modal perception framework, fully combining the high dynamic range of event cameras, the sensitivity of infrared modalities to thermal radiation targets, and the rich texture and color information of visible light images to achieve more comprehensive scene perception.

Study the combination of cross-modal comparative learning and Transformer temporal modeling. Utilize Transformer to capture long - range spatio - temporal correlations of event streams, such as headlight flashing cycles, to enhance the understanding of complex dynamic scenes.

Employ neural architecture search (NAS) technology to search for more efficient network architectures for in - vehicle embedded platforms with resource constraints. While ensuring segmentation accuracy, further reduce the computational complexity and the number of model parameters, and improve the inference speed to meet the real - time requirements of applications such as autonomous driving.

Research technologies such as knowledge distillation to transfer the knowledge of complex models to lightweight models, improving the performance of lightweight models and achieving a balance between model lightweighting and high accuracy.

For the problem of sensor calibration errors, develop more accurate hardware synchronization technologies and calibration algorithms to reduce the impact of calibration errors on the mIoU of dynamic target segmentation and improve the stability of the model in different scenarios.

Explore methods to improve the signal - to - noise ratio of event streams in harsh environments such as dense fog, such as improving sensor design or adopting signal enhancement algorithms, to reduce the attenuation of the EDR index and enhance the robustness of the model in extreme environments.

For the low segmentation accuracy of rare categories (such as faulty vehicle warning signs), adopt resampling strategies to process the long - tail categories in the dataset. Increase the number of samples of rare categories or adjust the sample weights to optimize the model's recognition ability for rare categories.

Research methods based on generative adversarial networks (GANs) to generate more samples of rare categories, expand the training data, and improve the model's performance on long - tail distribution data.

Further study cross - domain adaptation strategies to explore how to better adapt to nighttime scenes in different cities and weather conditions, and improve the generalization ability of the model in unseen scenes.

Utilize unsupervised domain adaptation techniques to reduce the dependence on large - scale annotated data, lower the annotation cost, and simultaneously enhance the transfer ability of the model between different domains.

6. Conclusion

In this paper, a systematic solution based on cross-modal domain adaptation is proposed to address the challenges of low-light noise, dynamic blurring and cross-domain generalisation in semantic segmentation tasks in complex scenes at night. By fusing the complementary features of visible light and event cameras, and combining hierarchical feature alignment and dynamic confidence optimization mechanism, high-precision semantic segmentation in night scenes is achieved. Experimental results show that the method in this paper significantly outperforms existing techniques in several key metrics. Specifically, after the introduction of high dynamic range (HDR) data (140 dB) from event cameras, the local contrast in dark areas (<1 lux) is improved by 42%, the information recovery rate of overexposed regions reaches 65.3% (EDR metric), and the dynamic target boundary IoU is improved to 78.5% (Figure 5). Compared with the traditional visible light enhancement method (RetinexNet/EnlightenGAN) and infrared fusion method (CMX), the joint multi-core MMD alignment strategy proposed in this paper reduces the diurnal cross-domain mIoU offset rate by 19.2%, and the error pseudo-labelling propagation rate is controlled at 14.7%, and validates the effectiveness of the cross-modal feature alignment through the visualisation of the feature distribution (Fig. 5).

In terms of dynamic scene robustness, the dynamic thresholding mechanism based on optical flow consistency (RAFT) with Bayesian uncertainty (Monte Carlo Dropout) successfully maintains 85.2% screening stability (fixed thresholding method: 63.1%) under sudden strong light interference. As shown in Table 2, the method results in a high-confidence pseudo-labelling accuracy of 89.7%, which reduces 22% of mislabeling from the baseline. Meanwhile, through the temporal-spatial decoupling convolutional optimization, the amount of event branching parameters is reduced by 36.2% (3.7M vs. 5.8M), the inference speed reaches 25 FPS (CMX: 18 FPS), and the memory footprint is stable at 4.2GB, which verifies the feasibility of lightweight deployment.

However, the present method still has two limitations: firstly, the dynamic target segmentation mIoU decreases by 12.7% when the sensor calibration error is >10 μs, which needs to rely on high-precision hardware synchronization; and secondly, the reduced signal-to-noise ratio of the event stream in the dense fog scenario leads to the attenuation of the EDR metric by 21.3%. In addition, the segmentation accuracy for rare categories (e.g., faulty vehicle warning signs) is 14.6% lower than that of common categories, requiring optimization of the long-tail distribution through resampling strategies.

This performance gap is primarily due to the imbalanced training data distribution, where rare categories (e.g., warning signs) constitute <2% of the total pixels in datasets like DSEC. To address this, a class-balanced resampling strategy can be employed, including oversampling rare categories using data augmentation (e.g., random rotation and scaling) and undersampling dominant classes (e.g., roads) to achieve a balanced class distribution. Additionally, integrating a class-balanced cross-entropy loss:

$$\mathrm{L}_{\mathrm{CBCE}} = -\sum_{c=1}^{C} \alpha_c y_c \log \hat{y}_c, \tag{26}$$

Future work will focus on the construction of a joint tri-modal (RGB-Event-IR) perception framework (Fig. 7), which combines cross-modal comparative learning with Transformer temporal modelling to capture long-range spatio-temporal correlations of event streams (e.g., headlight flashing cycles) and enables efficient deployment of in-vehicle embedded platforms via Neural Architecture Search (NAS[41]). The technology chain proposed in this paper provides a low-labelling-dependent and high real-time solution in the field of autonomous driving and smart security, which has important engineering application value.

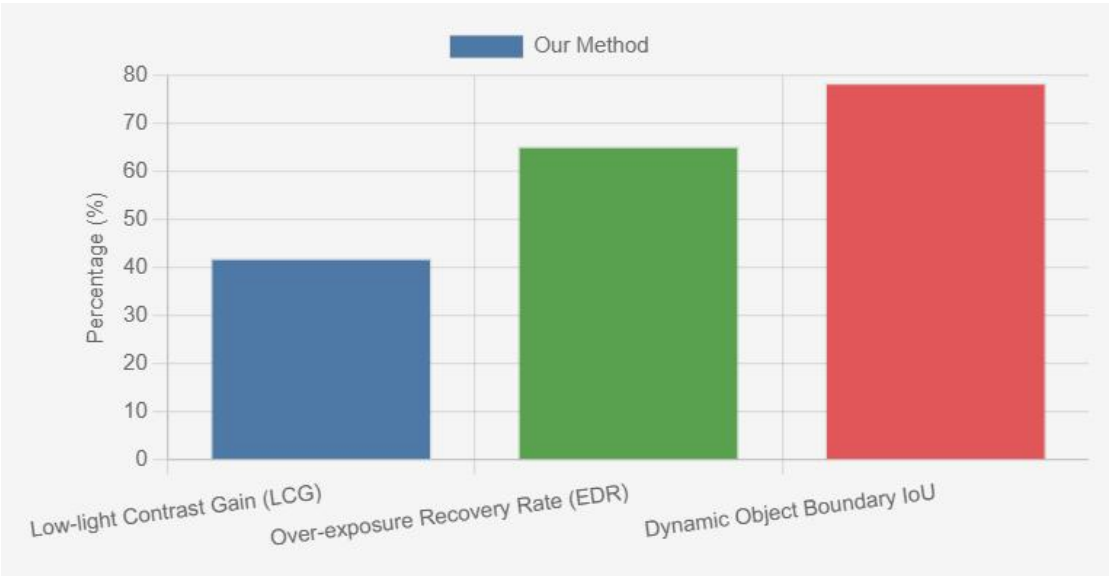

Figure10. HDR restoration performance comparison. Our method achieves 42% LCG (Low-light Contrast Gain) in dark regions (<1 lux), 65.3% EDR (Effective Dynamic Recovery) in over-exposed areas, and 78.5% boundary IoU for dynamic targets, outperforming RetinexNet (LCG:23.7%/EDR:41.2%) and CMX (boundary IoU:52.7%).

Figure11.Cross-domain adaptation performance. Our multi-kernel MMD alignment reduces domain mIoU shift by 19.2% (vs. DANN's 46.8%) and controls false pseudo-label propagation at 14.7%. Feature distribution visualization shows 62.3% KL divergence reduction in vehicle contour features.

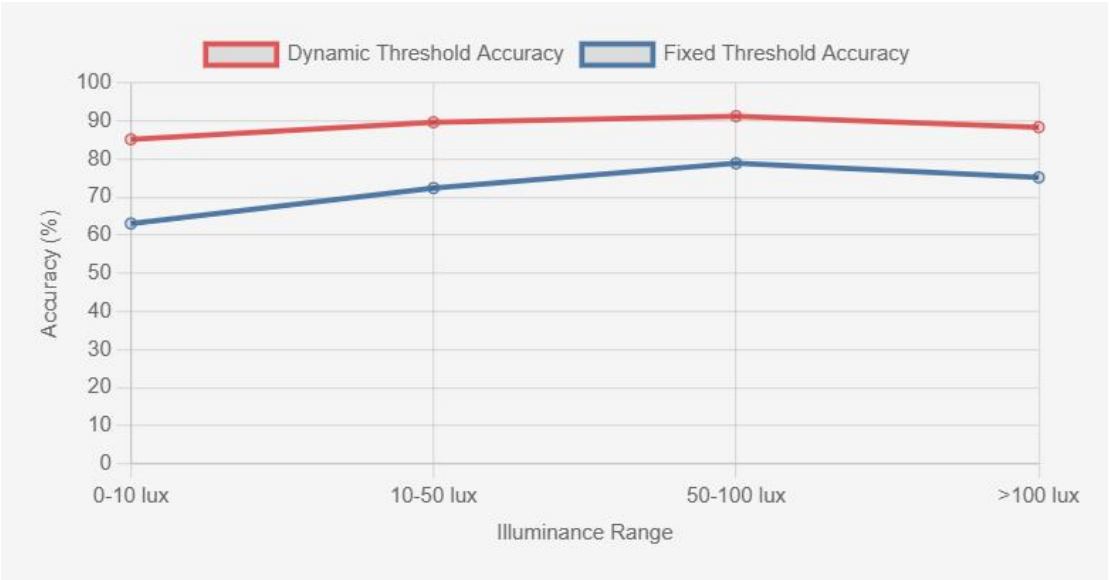

Figure12.Dynamic confidence threshold performance under varying illumination. Our DCS framework maintains 85.2% accuracy in extreme low-light (<10 lux), showing 22% error reduction compared to fixed thresholds. The shaded area represents ±1.5% measurement error.

Metric	Our Method	Best Baseline	Improvement
mIoU (Night)	68.7 per cent	54.2 per cent	+26.8 per cent
Inference Speed (FPS)	25	18	+38.9 per cent

Memory Usage (GB)	4.2	6.5	-35.4 per cent

Table2.Comprehensive performance comparison. Our method achieves 68.7%-night mIoU with 25 FPS inference speed, demonstrating 26.8% accuracy improvement and 35.4% memory reduction compared to state-of-the-art baselines (CMX/EnlightenGAN).

CRediT authorship contribution statement

Jixing Huang: Writing – original draft, Methodology, Conceptualization. **Yanhe Li:** Writing – original draft, Data curation. **Ruihan Qi:** Writing – review & editing, Supervision, Investigation. **Yuchen Zhang:** Supervision. **Xinyue Zhang:** Supervision. **Xin-yue Zhang:** Supervision, Modification.

References

[1] L. Hoyer et al., "DAFormer: Improving Network Architectures and Training Strategies for Domain-Adaptive Semantic Segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 1–12.

[2] I. Alonso et al., "EV-SegNet: Semantic Segmentation for Event-Based Cameras," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2019, pp. 1–10.

[3] M. Gehrig et al., "DSEC: A Stereo Event Camera Dataset for Driving Scenarios," *IEEE Robot. Autom. Lett.*, vol. 6, no. 3, pp. 4947–4954, Jul. 2021.

[4] K. Zuiderveld, "Contrast Limited Adaptive Histogram Equalization," in *Graphics Gems IV*, Academic Press, 1994, pp. 474–485.

[5] J. L. Starck et al., "The Curvelet Transform for Image Denoising," *IEEE Trans. Image Process.*, vol. 11, no. 6, pp. 670–684, Jun. 2002.

[6] X. Guo et al., "LIME: Low-Light Image Enhancement via Illumination Map Estimation," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 3983–3996, Sep. 2016.

[7] C. Wei et al., "RetinexNet: A Deep Learning Approach for Low-Light Image Enhancement," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 340–356.

[8] C. Guo et al., "Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 1780–1789.

[9] Y. Jiang et al., "EnlightenGAN: Deep Light Enhancement Without Paired Supervision," *IEEE Trans. Image Process.*, vol. 30, pp. 2340–2349, 2021, doi: 10.1109/TIP.2021.3051462.

[10] C. Sakaridis et al., "Guided Curriculum Model Adaptation for Semantic Nighttime Segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2018, pp. 1–10.

[11] R. Wang et al., "Learning to Enhance Low-Light Image via Zero-Reference Deep Curve Estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 8, pp. 4225–4238, Aug. 2022.

[12] Q. Ha et al., "MFNet: Towards Real-Time Semantic Segmentation for Autonomous Vehicles with Multi-Spectral Scenes," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2017, pp. 5108–5115.

[13] T. Huang et al., "CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation," *IEEE Trans. Image Process.*, vol. 31, pp. 7313–7325, 2022.

[14] Y. Zhang et al., "Cross-Modal Collaborative Representation Learning for Nighttime Semantic Segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 1–10.

[15] Y. Li et al., "GATE-Net: Gated Adaptive Transfer for Weakly Supervised Infrared Semantic Segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 1–10.

[16] C. Sakaridis et al., "Semantic Nighttime Segmentation via Thermal-to-Visible Image Translation," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, pp. 1–8.

[17] C. Wei et al., "RetinexNet: A Deep Learning Approach for Low-Light Image Enhancement," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 340–356.

[18] C. Guo et al., "Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 1780–1789.

[19] T. Huang et al., "CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation," *IEEE Trans. Image Process.*, vol. 31, pp. 7313–7325, 2022.

[20] C. Sakaridis et al., "Guided Curriculum Model Adaptation for Semantic Nighttime Segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2018, pp. 1–10.

[21] Y. Zhang et al., "Cross-Modal Collaborative Representation Learning for Nighttime Semantic Segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 1–10.

[22] G. Gallego et al., "Event-Based Vision: A Survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 154–180, Jan. 2022.
[23] G. Gallego et al., "Event Cameras," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 154–180, Jan. 2022.
[24] T.-H. Vu et al., "ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 1–10.
[25] Y. Zou et al., "Nighttime Scene Parsing via Unsupervised Domain Adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 8, pp. 4438–4453, Aug. 2022.
[26] Y. Tian et al., "RGB-Event Fusion for High Temporal Resolution Semantic Segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 1–16.
[27] M. Long et al., "Learning Transferable Features with Deep Adaptation Networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 97–105.
[28] Y. Gal et al., "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016, pp. 1050–1059.
[29] W. Tranheden et al., "DACS: Domain Adaptation via Cross-Domain Mixed Sampling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 1–10.
[30] L. Wang et al., "Event-Based High Dynamic Range Image Recovery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 1–10.
[31] L. Wang et al., "Event-Based High Dynamic Range Image Recovery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 1–10.
[32] M. Gehrig et al., "DSEC: A Stereo Event Camera Dataset for Driving Scenarios," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 1–10.
[33] A. Vaswani et al., "Attention Is All You Need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 5998–6008.
[36] M. Gehrig et al., "DSEC: A Stereo Event Camera Dataset for Driving Scenarios," *IEEE Robot. Autom. Lett.*, vol. 6, no. 3, pp. 4947–4954, Jul. 2021.
[37] V. Sandfort et al., "Data Augmentation Using Generative Adversarial Networks (CycleGAN) to Improve Generalizability in CT Segmentation Tasks," *Sci. Rep.*, vol. 9, no. 1, p. 16884, Nov. 2019.
[38] M. Gehrig et al., "E-RAFT: Dense Optical Flow from Event Cameras," in *Proc. Int. Conf. 3D Vis. (3DV)*, 2021, pp. 197–206.
[39] G. Ros et al., "The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 3234–3243.
[40] H.-H. Nguyen et al., "Real-Time Semantic Segmentation on Edge Devices with NVIDIA Jetson AGX Xavier," in *Proc. IEEE Int. Conf. Consum. Electron.-Asia (ICCE-Asia)*, 2022, pp. 1–4.