

EDM:Event-guided Diffusion Model for Video Shadow Detection in Complex Dynamic Scenes

Boyang Gao¹, Jiayi Guo³, Ziyu Wang¹, Yinbin Cui¹, and Zhan Tu²(✉) 

¹ Tianmushan Laboratory, Beihang University, Hangzhou, China

² Institute of Unmanned System, Beihang University, Beijing, China

³ Tsinghua University, Beijing, China

{bygao, zhantu}@buaa.edu.cn

Abstract. Video shadow detection (VSD) aims to robustly identify shadow regions in videos and serves as a key task for improving scene understanding. Although substantial progress has been made in recent years, existing methods still suffer notable performance degradation in challenging scenarios involving fast motion and drastic illumination changes. To address these limitations, we introduce event cameras into the VSD task and propose an Event-guided Diffusion Model (EDM) tailored for complex dynamic environments. Based on the high temporal resolution and asynchronous sensing characteristics of event cameras, we design a Motion Trajectory Attention (MTA) module and a Bidirectional Mask Generation (BMG) module to capture event-driven temporal cues. These cues are further integrated through a Dual-Modal Temporal Guidance (DMTG) module and injected into the diffusion model as conditional guidance to steer the denoising process. By incorporating event streams as a complementary modality, EDM effectively captures fast scene dynamics and abrupt lighting variations, leading to more robust and accurate video shadow detection. Furthermore, to address the lack of event modalities in existing VSD benchmarks, we construct a multimodal extension of the ViSha dataset, termed Event-ViSha, and collect a new real-world multimodal dataset, the Event-based Video Shadow Dataset (EVSD). Extensive experiments demonstrate that our method consistently outperforms state-of-the-art approaches across multiple benchmarks.

Keywords: Event Camera · Diffusion Model · Video Shadow Detection

1 Introduction

Shadows are ubiquitous in real-world environments and exhibit a dual nature in visual perception. On the one hand, their intrinsic properties can be effectively exploited to enhance various computer vision tasks, including but not limited to depth estimation [1], object shape recovery [28], and light source estimation [10]. On the other hand, shadows often introduce ambiguities and misinterpretations that significantly degrade the reliability of vision systems. In particular, in tasks

such as object detection [13, 14], object tracking [3], and semantic segmentation [19, 25], shadows are frequently misclassified as physical objects, leading to false positives and inaccurate localization. For instance, in autonomous driving scenarios, vehicle shadows cast on the road may be mistakenly perceived as obstacles [27]. Therefore, achieving robust and accurate shadow detection in complex dynamic scenes remains a critical and highly challenging problem in computer vision.

Although significant progress has been made in video shadow detection, existing methods primarily perform well under low-dynamic and illumination-stable conditions, while exhibiting substantial performance degradation in scenarios involving fast motion, drastic illumination changes, and complex dynamic backgrounds. Moreover, most approaches focus on large, high-contrast shadows, often neglecting subtle, fragmented, or low-contrast shadow regions, resulting in both false positives and false negatives. These limitations largely stem from the inherent constraints of conventional frame-based cameras. Under high-speed motion or low-light conditions, motion blur and uneven illumination often obscure shadow boundaries and fine details, severely hindering the model’s ability to capture dynamic shadow variations.

To overcome these limitations, sensing modalities capable of capturing rapid shadow dynamics and adapting to complex illumination conditions are required. Event cameras [30, 39], which asynchronously record pixel-wise brightness changes instead of capturing full image frames, offer high temporal resolution, low latency, and a wide dynamic range. These properties enable event cameras to effectively capture fast motion and abrupt illumination variations that challenge conventional sensors. Consequently, event cameras serve as a promising complementary modality for modeling fine-grained shadow dynamics in complex environments. However, event data inherently lacks color and texture cues, and directly extracting features from raw event streams often yields limited or unstable gains, or even degrades performance. Therefore, choosing an appropriate event representation and designing effective cross-modal fusion are key challenges.

To address these challenges, we propose an event-guided diffusion framework for video shadow detection. To the best of our knowledge, this is the first work to introduce event cameras into the VSD task. Our key idea is to leverage motion and edge cues embedded in event streams to guide temporal modeling in RGB videos, thereby enhancing the diffusion model’s ability to capture dynamic shadow variations.

Specifically, (1) we design a Motion Trajectory Attention (MTA) module, which extracts motion trajectories from event streams and constructs trajectory-aware cross-frame attention to model the spatiotemporal evolution of shadows. (2) To provide finer-grained temporal constraints for diffusion, we introduce a Bidirectional Mask Generation (BMG) module, which generates preliminary masks in both forward and backward temporal directions from event streams. This process forms a longer RGB preliminary mask sequence within the same temporal window, providing denser and more structured temporal supervision

for the diffusion model. (3) Finally, we reconstruct event streams into intensity images via ESI and employ a lightweight network with a prediction head to produce event-based preliminary mask sequences. During prediction for the target frame, the Dual-Modal Temporal Guidance (DMTG) module injects the preliminary mask sequences from both modalities across neighboring frames as conditional guidance into the diffusion model, enabling EDM to progressively refine representations through the reverse diffusion process and directly decode the final shadow mask.

Our main contributions are summarized as follows:

- We propose an **Event-guided Diffusion Model (EDM)** for video shadow detection in complex dynamic scenes. To the best of our knowledge, this is the first work to introduce event cameras into the VSD task, effectively leveraging event streams to significantly enhance detection performance under challenging conditions.
- We construct a real-world multimodal dataset, **EVSD**, combining RGB frames and event streams under complex illumination and high-dynamic scenarios. In addition, we build **Event-ViSha**, a multimodal extension of the ViSha dataset, providing new benchmarks for multimodal shadow detection research.
- We conduct comprehensive comparisons with state-of-the-art methods on both EVSD and Event-ViSha datasets. Experimental results demonstrate that our approach achieves significant improvements across all evaluation metrics and establishes new state-of-the-art performance.

2 Related Works

2.1 Shadow Detection

Early research on shadow detection primarily focused on single images [8, 18, 21, 22, 24, 26]. With technological advancements, video shadow detection has emerged as a more challenging direction. Chen et al. [7] introduced ViSha, the first video shadow detection dataset, and proposed TVSD-Net, which employs a triple-parallel network to aggregate features from multiple frames. Subsequently, STICT [32] adopted a Teacher-Student paradigm with spatio-temporal interpolation consistency training to enhance temporal modeling. More recently, approaches such as SCOTCH and SODA [31] have proposed novel video shadow detection frameworks: SODA leverages trajectory attention to track shadows with complex deformations, while SCOTCH employs a contrastive loss to improve shadow feature representation. TBGDiff [43] was the first to introduce diffusion models to video shadow detection, proposing a lightweight diffusion paradigm and integrating Spatio-Temporal Encoding Embedding (STEE) to capture temporal information. In [17], a two-stage detection paradigm was proposed, incorporating novel temporal and spatial adaptation modules to effectively extend image-based models to video-level detection. Despite these advances, existing

methods are mainly evaluated and optimized under stable illumination and low-dynamic scenarios, and video shadow detection under complex lighting and high-dynamic conditions remains a challenging problem.

2.2 Event-based Segmentation

Event-based segmentation leverages the sparse and asynchronous data streams produced by event cameras to achieve pixel- or region-level scene understanding, showing particular advantages under high dynamic range and low-light conditions. EV-IMO [34] introduced the first event dataset with instance-level annotations, while Ev-SegNet [2] preserved spatiotemporal event information through a six-channel encoding scheme. To address cross-modal adaptation under poor illumination, CMDA [37] enabled knowledge transfer from daytime RGB images to nighttime event data. More recently, several works [38, 41, 42] have explored dedicated fusion strategies for RGB and event modalities, further improving segmentation performance.

3 Event-based Video Shadow Dataset (EVSD)

Datasets play a critical role in algorithm development and performance evaluation. Given the lack of multimodal video shadow detection datasets that jointly incorporate event cameras and RGB imagery, we construct the Event-ViSha dataset based on ViSha using the V2E simulator [23].

However, Event-ViSha still has several inherent limitations. Specifically, it lacks high-dynamic scenes and contains no motion blur. Moreover, the illumination conditions are relatively homogeneous, with limited samples under extreme lighting environments. In addition, the simulated event streams may suffer from excessive noise, sparse distributions, and unstable event quality in certain scenarios, which collectively limit the diversity and realism of the dataset.

To address these limitations, we further construct a real-world video shadow detection dataset, termed EVSD (Event-based Video Shadow Dataset). This dataset substantially improves the quality of event data while systematically enhancing scene diversity. As illustrated in Fig. 1, EVSD covers a wide range of challenging environments, including low-light scenes, high dynamic range (HDR) conditions, and high-speed scenarios. The videos contain shadow instances with diverse object categories, materials, and scales, enabling more realistic modeling of the complex interactions between shadows and objects in real-world environments. Compared with ViSha, EVSD provides richer samples with motion blur and extreme illumination conditions, allowing for a more comprehensive evaluation of algorithm robustness and generalization in complex dynamic scenes.

The scale of the EVSD, the video acquisition frame rate, and the split between training and validation sets are designed following the benchmark video shadow detection dataset ViSha. All videos were captured using a DAVIS346 event camera, which provides synchronized APS frames and event streams. The recordings were conducted at approximately 30 fps. Shadow annotations were

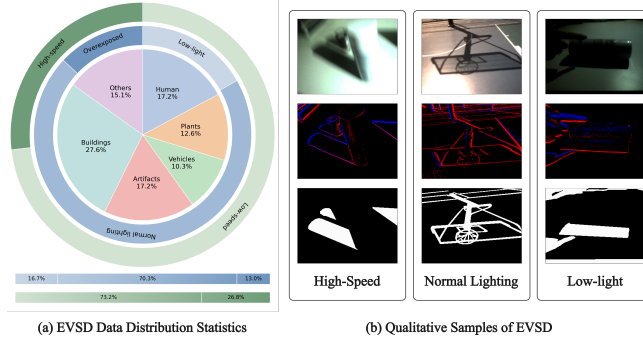


Fig. 1: Statistical overview of the EVSD. (a) illustrates the category composition and scene distribution of the dataset. (b) presents representative visual examples from EVSD.

generated based on both the corresponding APS images and the accumulated event images. The annotation process was conducted by a professional labeling team and further verified through multiple rounds of manual review to ensure the accuracy and consistency of the annotations.

The EVSD contains 108 scenes, comprising a total of 9,708 images. According to object motion speed, the dataset includes 29 high-speed scenes and 79 low-speed scenes. In terms of illumination conditions, the dataset consists of 18 low-light scenes, 76 normal-light scenes, and 14 overexposed scenes. To ensure a balanced distribution across different scene types and lighting conditions, the dataset is divided into training and validation sets, with 49 video sequences (4,435 images) used for training and 59 video sequences (5,273 images) used for validation. More details are provided in the **Supplementary Material**.

4 Proposed Method

4.1 Overview

In this section, we systematically present the overall pipeline of EDM. As illustrated in Figure 2, the entire architecture comprises two main stages, shifting from coarse preliminary mask generation to fine-grained diffusion-based mask prediction. Below, we first provide a holistic overview of these two stages, followed by detailed elaborations on each core component (MTA, BMG, and DMTG) in subsequent subsections.

Stage 1: Preliminary Mask Prediction. Given $2i + 1$ consecutive RGB frames $\{I_{T+k}\}_{k=-i}^i$ and the corresponding $2i$ inter-frame event voxel segments $\{E_{T+k+\frac{1}{2}}\}_{k=-i}^{i-1}$, this stage aims to generate coarse spatial-temporal shadow predictions. Specifically, we employ an RGB encoder and a motion encoder to extract multi-scale visual features $\{F_{T+k}\}_{k=-i}^i$ and explicit motion field representations from $\{I_{T+k}\}_{k=-i}^i$ and $\{E_{T+k+\frac{1}{2}}\}_{k=-i}^{i-1}$, respectively. Based on the extracted

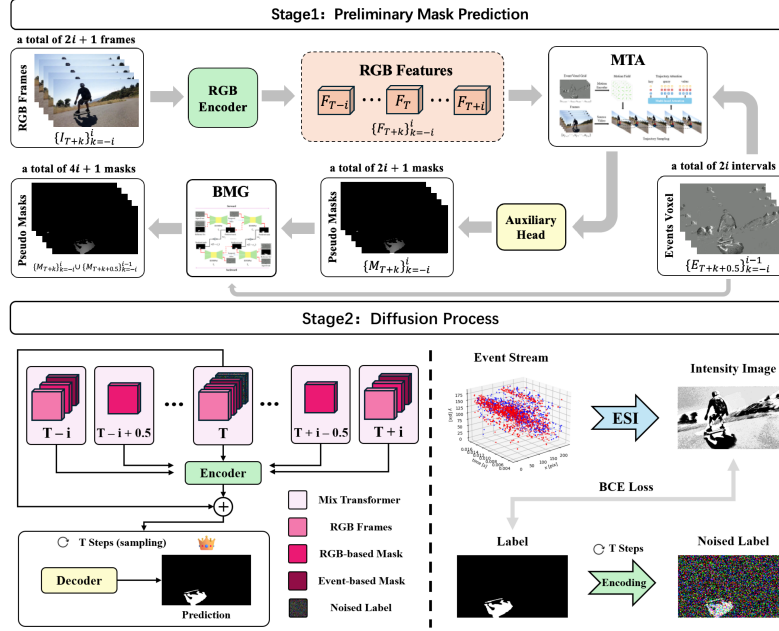


Fig. 2: Overview of the pipeline. EDM consists of two stages: Preliminary Mask prediction and Diffusion Process. The former employs the MTA and BMG modules to fuse dual-modal inputs and construct event-based temporal guidance, while the latter injects this guidance as a conditional embedding into the diffusion denoising process to produce the final shadow prediction.

motion fields, the MTA module temporally enhances the RGB features to yield $\{F'_{T+k}\}_{k=-i}^i$. These enhanced features are then mapped by a prediction head to produce the final preliminary shadow masks $\{M'_{T+k}\}_{k=-i}^i$.

Building upon this, the BMG module leverages $\{M'_{T+k}\}_{k=-i}^i$ together with the corresponding event streams $\{E_{T+k+\frac{1}{2}}\}_{k=-i}^{i-1}$ to estimate intermediate preliminary masks $\{M'_{T+k+\frac{1}{2}}\}_{k=-i}^{i-1}$. The original and intermediate masks are then combined to form a denser temporal guidance sequence $\{M'_{T+k}\}_{k=-i}^i \cup \{M'_{T+k+\frac{1}{2}}\}_{k=-i}^{i-1}$.

Stage 2: Diffusion Process. In this stage, the event streams $\{E_{T+k}\}_{k=-i}^i$ are first reconstructed into intensity image sequences $\{E'_{T+k}\}_{k=-i}^i$ via an event-to-intensity conversion. These reconstructed intensity images are then processed by an event encoder to predict the corresponding event-modality preliminary mask sequence $\{M'_{T+k}\}_{k=-i}^i$. Subsequently, the DMTG module fuses the RGB frames $\{I_{T+k}\}_{k=-i}^i$ with the preliminary mask sequences from both modalities—specifically, the dense sequence $\{M'_{T+k}\}_{k=-i}^i \cup \{M'_{T+k+\frac{1}{2}}\}_{k=-i}^{i-1}$ from the first stage and the event-based sequence $\{M'_{T+k}\}_{k=-i}^i$. The unified dual-modal representations are then injected into the diffusion model as conditional guid-

ance. During the diffusion process, an analog bits strategy is employed for noise injection, allowing the network to directly predict the final refined shadow mask.

4.2 Motion Trajectory Attention

The output of an event camera essentially reflects the temporal dynamics of pixel-level luminance changes within a scene, thereby implicitly encoding the local motion trajectories of objects. In contrast to conventional optical flow methods that rely on the brightness constancy assumption, event streams can directly capture instantaneous object displacements with higher temporal resolution and dynamic responsiveness. Building upon this property, we leverage the event stream between two consecutive RGB frames and encode the event sequence using the MotionEncoder module to predict the corresponding motion field. Specifically, the MotionEncoder is composed of 5 stacked 3×3 convolutional blocks with a stride of 2, which efficiently extract spatial displacement features induced by the events. The encoded features are then mapped to x- and y-direction displacement components via a multi-layer perceptron (MLP), thereby generating a pixel-wise two-dimensional motion field. This process can be formally expressed as follows:

$$(\Delta x, \Delta y) = \text{MLP}(\text{MotionEncoder}(E_{t \rightarrow t+1})), \quad (1)$$

where $E_{t \rightarrow t+1}$ represents the event stream recorded during the time interval $[t, t+1]$.

Next, based on the estimated displacement fields $(\Delta x, \Delta y)$, patch-level trajectory sampling is performed on the extracted RGB features F . Each trajectory is iteratively traced starting from the first frame. Using the displacement fields derived from the event stream, the position of a patch at (x_0, y_0) in the first frame can be propagated to its corresponding locations in subsequent frames. Consequently, for N frames of extracted feature maps with spatial dimensions $H \times W$, a complete set of patch trajectories can be obtained, capturing the temporal evolution of each spatial location. The formulation is presented as follows:

$$\mathcal{T} = \{\tau_{x_0, y_0} \mid \tau_{x_0, y_0} = ((x_0, y_0), \dots, (x_{N-1}, y_{N-1})), (x_0, y_0) \in [1, H] \times [1, W]\}, \quad (2)$$

where each trajectory τ_{x_0, y_0} represents the positions of a patch across N frames, and (x_{t+1}, y_{t+1}) is computed iteratively as

$$x_{t+1} = x_t + \Delta x_t(x_t, y_t), \quad y_{t+1} = y_t + \Delta y_t(x_t, y_t), \quad t = 0, \dots, N-2. \quad (3)$$

After sampling the patch trajectories, we collect the patch embeddings from the extracted RGB feature map F that belong to the same trajectory. Multi-head attention is then performed among the patch embeddings along each trajectory. For a query vector $\mathbf{z}(x_k, y_k)$, the corresponding keys and values are taken from the other patch embeddings on the same trajectory, excluding the current query

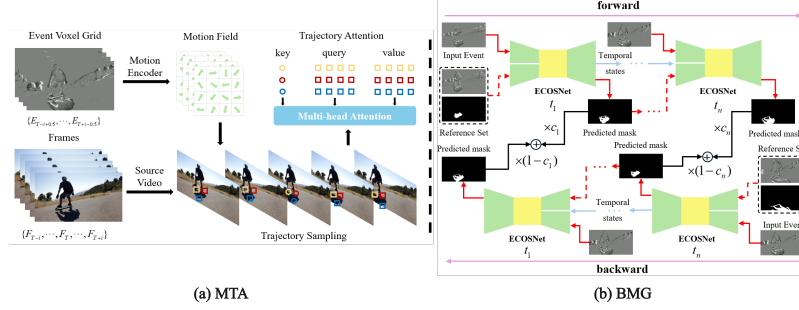


Fig. 3: Illustration of the MTA and BMG modules. (a) depicts the architecture of the MTA module, and (b) presents the structure of the BMG module.

patch. No additional positional encoding is introduced. The computation can be formally expressed as follows:

$$\begin{aligned}
 \mathbf{Q} &= \mathbf{z}(x_k, y_k), \\
 \mathbf{K} &= \mathbf{V} = \mathbf{z}_{\text{traj}}^{\setminus k}, \\
 \text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &= \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}.
 \end{aligned} \tag{4}$$

where $\mathbf{z}_{\text{traj}}$ denotes the set of patch embeddings along the same trajectory, d is the dimension of the query and key vectors, and $\mathbf{z}_{\text{traj}}^{\setminus k}$ represents the set of remaining patch embeddings along the trajectory after excluding the current query patch $\mathbf{z}(x_k, y_k)$.

4.3 Bidirectional Mask Generation

In the final diffusion-based mask prediction stage, we use preliminary masks predicted from adjacent frames as temporal guidance to model shadow evolution. However, under high-speed motion, large displacements between neighboring masks may weaken the modeling of inter-frame dynamics. To address this issue, we exploit the asynchronous triggering property of event cameras and leverage the inter-frame event streams $\{E_{T+k+\frac{1}{2}}\}_{k=-i}^{i-1}$ to capture instantaneous shadow displacements, yielding finer and temporally continuous motion cues.

Inspired by the semi-supervised segmentation framework in [44], where ECOSNet predicts the current mask using historical event streams and masks, we adopt a similar strategy. Specifically, intermediate preliminary masks between adjacent frames are estimated using the predicted masks and the inter-frame event streams, and inserted into the original mask sequence to provide refined spatio-temporal guidance.

To avoid introducing additional losses that may interfere with the optimization of the main segmentation model, we first pre-train ECOSNet independently using only the training split of the current dataset, without relying on extra data

or supervision. The pre-trained weights are then loaded during the main training stage to enable implicit knowledge transfer while maintaining a fair training protocol and stable convergence. Note that the initial masks predicted by BMG serve solely as temporal guidance and are not subject to additional supervision.

Since ECOSNet adopts an LSTM-based recurrent prediction scheme, prediction errors may accumulate over time. To mitigate this issue, we employ a bidirectional prediction and fusion strategy. Intermediate masks are predicted in both forward (from the first frame) and backward (from the last frame) directions, and fused as

$$M_t^{\text{fusion}} = W_t \otimes M_t^{\text{forward}} + (1 - W_t) \otimes M_t^{\text{backward}}, \quad (5)$$

where the fusion weight decreases linearly with time step t :

$$W_t = 1 - \frac{t}{T}, \quad (6)$$

and T denotes the total number of frames.

Finally, the intermediate masks are concatenated with the original preliminary masks to form a denser temporal guidance sequence

$$\{M_{T+k}\}_{k=-i}^i \cup \{M_{T+k+\frac{1}{2}}\}_{k=-i}^{i-1}.$$

This design provides denser masks within a short temporal window, offering finer and more continuous temporal guidance for the diffusion process.

4.4 Dual-Modal Temporal Guidance

To further strengthen the temporal modeling within the diffusion process prior to the final shadow mask prediction, we introduce a Dual-Modal Temporal Guidance (DMTG) module, as illustrated in Fig. 2, which supplies the diffusion model with stronger temporal constraints by aggregating cues from both modalities.

The RGB branch produces the preliminary mask set $\mathcal{M}_{\text{rgb}} = \{M_{T+k}\}_{k=-i}^i \cup \{M_{T+k+\frac{1}{2}}\}_{k=-i}^{i-1}$. We further exploit the event modality to obtain complementary guidance masks, since event streams capture fine-grained shadow boundaries and intensity variations that offer dynamic cues complementary to RGB. The choice of event representation, however, has a non-trivial impact on performance: a naive temporal accumulation tends to yield low-quality masks and may interfere with backbone optimization. Inspired by [16], we therefore adopt Event-based Single Integration (ESI) to reconstruct intensity images from the event stream, thereby preserving the high dynamic range and high temporal resolution inherent to event cameras.

Based on the reconstructed intensity images, a lightweight ResNet extracts event features and, under ground-truth supervision, predicts an event-based preliminary mask sequence $\{M'_{T+k}\}_{k=-i}^i$. Finally, DMTG jointly encodes the RGB preliminary masks $\mathcal{M}_{\text{rgb}}$, the event-based masks $\{M'_{T+k}\}_{k=-i}^i$, and the RGB frames $\{I_{T+k}\}_{k=-i}^i$, and injects the fused representation as conditional guidance into the diffusion model, thereby improving both segmentation accuracy and temporal consistency. Further details are provided in the **Supplementary material**.

4.5 Diffusion-based Mask Prediction

Traditional diffusion models typically rely on computationally expensive U-Net architectures to predict noise, which leads to high computational overhead in high-resolution segmentation tasks. To address this issue, we adopt a diffusion framework that directly predicts segmentation masks and progressively denoises the latent variable z_t .

Specifically, we construct a diffusion process that transforms the ground-truth label y into a noisy variable z_t . To enable diffusion modeling on discrete segmentation labels, we employ analog bits strategy [6], which converts discrete labels into binary strings and maps them into continuous representations. Gaussian noise is then gradually added to the encoded label to obtain the noisy variable z_t . The forward diffusion process is defined as

$$q(z_t | z_0) = \mathcal{N}(z_t; \sqrt{\bar{\alpha}_t}z_0, (1 - \bar{\alpha}_t)I), \quad (7)$$

where z_0 denotes the encoded ground-truth label and $\bar{\alpha}_t$ is controlled by a pre-defined noise scheduler.

In the reverse process, a network parameterized by θ , denoted as $p_\theta(y_t, g, x)$, progressively predicts y_0 , where g and x represent the conditional guidance and image input, respectively. The reverse Markov process can be written as

$$p_\theta(z_{0:T} | x) = p(z_T) \prod_{t=T}^1 p_\theta(z_{t-1} | z_t, x), \quad (8)$$

which reconstructs the clean sample z_0 from noise z_T through iterative denoising.

During inference, EDM starts from a random Gaussian noise sample z_T and progressively reconstructs the final mask. At each sampling step, the noisy variable z_t is fused with conditional features and fed into the decoder to predict the denoised output. The next latent variable z_{t-1} is then obtained through reparameterization. To accelerate inference, we adopt the DDIM sampling strategy.

4.6 Loss Function

We supervise the final shadow mask using a combination of binary cross-entropy (BCE) loss and Lovász-Hinge loss. In addition, BCE loss is applied to the preliminary mask predictions from both RGB frames and ESI-reconstructed event intensity images as auxiliary supervision. The overall loss is defined as

$$\mathcal{L} = \mathcal{L}_{\text{BCE}}(y, \hat{y}) + \mathcal{L}_{\text{hinge}}(y, \hat{y}) + \lambda_1 \mathcal{L}_{\text{BCE}}(r, \hat{r}) + \lambda_2 \mathcal{L}_{\text{BCE}}(e, \hat{e}), \quad (9)$$

where y denotes the final mask prediction, while r and e represent the preliminary masks predicted from RGB frames and event intensity images, respectively. In all experiments, $\lambda_1 = \lambda_2 = 0.5$.

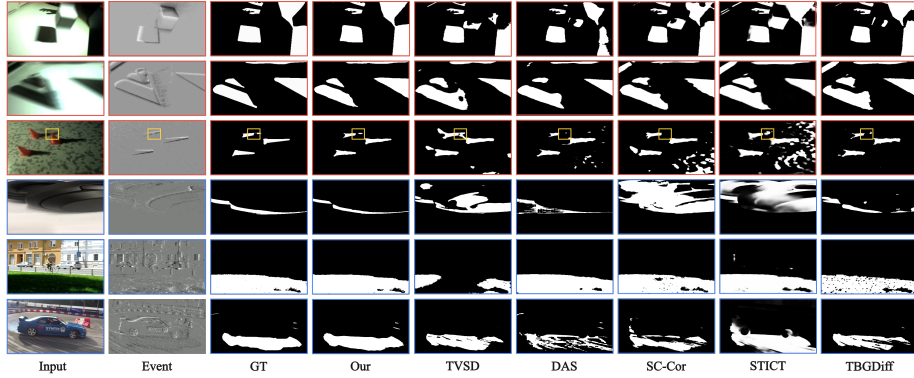


Fig. 4: Qualitative comparison of EDM with SOTA methods. The first three rows (red borders) illustrate mask predictions on the EVSD, while the last three rows (blue borders) correspond to results on the Event-ViSha dataset.

5 Experiments

5.1 Implementation Details

We adopt MixTransformer-B3 as the backbone for RGB feature extraction. The model is optimized using AdamW with a learning rate of 1×10^{-5} and a batch size of 2. Each input clip consists of three consecutive RGB frames and the event stream accumulated within the same temporal interval, with all images resized to 512×512 . Random seeds are fixed for reproducibility. All experiments are conducted on 8 A6000 GPUs for 18 epochs. For the diffusion model, we use a cosine noise scheduler, 15 sampling steps, and a scale weight of 0.03.

5.2 Datasets and Evaluation Metrics

We performed training and evaluation on both EVSD and Event-ViSha datasets. For quantitative evaluation, we employ four commonly used metrics: Mean Absolute Error (MAE), Intersection over Union (IoU), F-measure (F_β), and Balanced Error Rate (BER).

5.3 Comparison with SOTA Methods

Comparison Methods. We compare our method with 16 state-of-the-art approaches across four related tasks, including Image Shadow Detection (ISD), Video Object Segmentation (VOS), Event-RGB Semantic Segmentation (ERSS), and Video Shadow Detection (VSD). Semi-supervised VOS methods [4, 11, 36] typically require ground-truth annotations of the first frame during inference. To ensure a fair comparison, we initialize these methods using the first-frame mask bounding box generated by the ISD model, aligning their evaluation protocol

Table 1: Performance comparison on EVSD and Event-ViSha dataset.

Methods		Event-ViSha				EVSD			
Tasks	Models	MAE↓	F_β ↑	IoU↑	BER↓	MAE↓	F_β ↑	IoU↑	BER↓
ISD	MTMT [9]	0.043	0.729	0.517	20.28	0.085	0.591	0.512	21.07
	SDDNet [12]	0.040	0.754	0.548	14.05	0.076	0.638	0.522	17.78
	SILT [40]	0.031	0.796	0.606	12.80	0.080	0.611	0.537	18.97
VOS	STCN [11]	0.048	0.684	0.528	12.42	0.080	0.611	0.536	22.16
	Pix2Seq [5]	0.034	0.775	0.618	10.63	0.055	0.644	0.555	17.94
	SAM3 [4]	0.031	0.799	0.632	13.47	0.046	0.661	0.594	14.97
ERSS	BRENet [41]	0.043	0.729	0.517	9.01	0.042	0.679	0.612	15.72
	EISNet [38]	0.038	0.688	0.565	16.77	0.051	0.688	0.623	14.96
	CMNeXt [42]	0.043	0.773	0.623	9.87	0.043	0.673	0.625	13.78
VSD	TVSD [7]	0.033	0.757	0.567	17.70	0.076	0.658	0.584	18.97
	STICT [33]	0.046	0.702	0.545	16.60	0.109	0.578	0.539	22.34
	SC-Cor [15]	0.042	0.762	0.615	13.61	0.074	0.626	0.576	19.16
	DTTNet [29]	0.016	0.849	0.718	6.45	-	-	-	-
	TBGDiff [43]	0.023	0.797	0.667	8.58	0.054	0.644	0.612	16.13
	DAS [36]	0.034	0.754	0.575	12.58	0.066	0.673	0.601	15.99
	Scotch&SODA [31]	0.029	0.793	0.640	9.07	0.073	0.625	0.586	18.05
Our		0.015	0.861	0.739	5.76	0.035	0.724	0.664	11.05

Table 2: Performance on stratified EVSD scenarios.

Models	EDM		DAS	
	IoU↑	BER↓	IoU↑	BER↓
Low-light	0.682	13.15	0.573	19.98
Normal Lighting	0.654	10.67	0.622	14.73
Overexposed	0.663	11.26	0.545	17.84
Low-Speed	0.668	10.89	0.621	15.08
High-Speed	0.647	11.57	0.532	18.86
Overall	0.664	11.05	0.601	15.99

with existing VSD methods and our approach. For ERSS methods, we strictly follow the feature representation and input settings described in their original papers. Since the official implementation of DTTNet is not publicly available, it cannot be reproduced on EVSD.

Qualitative Comparison. Figure 4 presents qualitative comparisons between our method and several state-of-the-art video shadow detection approaches. The first three rows correspond to EVSD (red borders), and the last three to Event-ViSha (blue borders). Our method consistently outperforms competing approaches under challenging conditions, including complex backgrounds and motion blur. In the first two rows, although the RGB frames suffer from severe motion blur, the event streams preserve clear shadow boundaries, leading to more accurate predictions. In the third row, the granite texture confuses other methods and the shadow of the upper bucket is partially occluded (yellow region). Nevertheless, the event streams reliably capture the boundary, enabling

Table 3: Ablation study on the contributions of the MTA, BMG, and DMTG modules, as well as the event input.

Setting	MTA	BMG	DMTG	Event-ViSha				EVSD			
				MAE↓	F_β ↑	IoU↑	BER↓	MAE↓	F_β ↑	IoU↑	BER↓
Baseline	×	×	×	0.030	0.788	0.605	11.74	0.074	0.596	0.563	19.12
EDM w/o MTA	×	✓	✓	0.024	0.816	0.692	7.35	0.047	0.682	0.633	13.21
EDM w/o BMG	✓	×	✓	0.019	0.834	0.706	6.64	0.039	0.701	0.645	11.98
EDM w/o DMTG	✓	✓	×	0.027	0.799	0.663	8.72	0.057	0.651	0.602	15.64
EDM w/o event	-	-	-	0.030	0.784	0.621	9.88	0.066	0.617	0.585	17.32
EDM	✓	✓	✓	0.015	0.861	0.739	5.76	0.035	0.724	0.664	11.05

Table 4: Ablation Study on Different Motion-Field Extraction Strategies in MTA.

Setting	Event-ViSha				EVSD			
	MAE↓	F_β ↑	IoU↑	BER↓	MAE↓	F_β ↑	IoU↑	BER↓
RGB-based Optical Flow	0.018	0.842	0.716	6.34	0.048	0.689	0.615	13.97
Event-based Optical Flow	0.023	0.827	0.718	6.16	0.047	0.712	0.632	12.86
Event-based Motion (ours)	0.015	0.861	0.739	5.76	0.035	0.724	0.664	11.05

EDM to correctly identify the shadow region. These results indicate that EDM maintains stable and accurate shadow detection in complex dynamic scenes.

Quantitative Comparison. Table 1 reports quantitative comparisons with several state-of-the-art approaches on the EVSD and Event-ViSha datasets. Our method achieves the best performance across all evaluation metrics on both datasets. EVSD contains numerous high-dynamic scenes and challenging illumination conditions, which pose significant difficulties for video shadow detection. As shown in the table, methods that incorporate event information (ERSS) generally exhibit stronger robustness on EVSD, highlighting the benefit of complementary event cues in complex dynamic environments. The official implementation of DTTNet is not publicly available and therefore cannot be evaluated on EVSD. Nevertheless, the results on both datasets indicate that EDM achieves consistently strong performance. These observations suggest that integrating event information provides valuable cues for improving video shadow detection in challenging scenarios.

Stratified Dataset Analysis. To further dissect the factors underlying the performance improvement of event cameras on the VSD task, we stratify the dataset by scene types and compare EDM against DAS, the state-of-the-art model on EVSD, under these stratified scenarios, with results reported in Table 2. Compared with RGB-based methods, EDM demonstrates more pronounced advantages in challenging scenarios, including high-speed motion, low-light conditions, and overexposure. Low-light scenes remain difficult due to the presence of shadow-like appearance cues that complicate discrimination. In contrast, in high-speed scenarios, EDM effectively leverages fine-grained event motion information and explicitly models shadow trajectories via MTA, leading to substantial performance gains.

Table 5: Patch-size ablation study on Event-ViSha.

Patch Size	MAE↓	F_β ↑	IoU↑	BER↓
16x16	0.020	0.794	0.751	6.89
64x64	0.017	0.826	0.713	6.23
32x32 (Ours)	0.015	0.861	0.739	5.76

Table 6: Ablation of Unidirectional vs Bidirectional Strategies in BMG.

Setting	Event-ViSha				EVSD			
	MAE↓	F_β ↑	IoU↑	BER↓	MAE↓	F_β ↑	IoU↑	BER↓
unidirectional	0.019	0.857	0.728	6.24	0.046	0.707	0.643	11.47
Bidirectional (ours)	0.015	0.861	0.739	5.76	0.035	0.724	0.664	11.05

5.4 Ablation Study

In Table 3, we present a systematic ablation study of the key components in the proposed framework to quantify the contribution of each module. Specifically, we remove MTA, BMG, DMTG, and the event-stream input individually to measure their impact on overall performance. The baseline is constructed by removing all these components.

MTA Module. The results in Table 3 demonstrate the importance of MTA for temporal modeling. Furthermore, we replace the motion encoder with event-based optical flow (E-RAFT [20]) and RGB-based optical flow (RAFT [35]) to extract motion information, and report the results in Table 4. The results show that inter-frame event streams inherently provide cleaner motion cues, enabling a lightweight encoder to extract motion representations that outperform those estimated by optical flow methods.

Furthermore, since MTA establishes cross-frame motion trajectories at the patch level, the choice of patch size inevitably affects model performance. To investigate this, we conduct an ablation study on patch sizes, with results reported in Table 5. Among the configurations evaluated, a patch size of 32×32 achieves the best performance in terms of both MAE and BER, striking a favorable balance between quantization error and trajectory stability.

BMG Module. Compared with the baseline, the BMG module improves performance by generating finer-grained intermediate masks. We further compare uni-directional and bi-directional weighting strategies (Table 6). The results indicate that bi-directional weighting effectively mitigates error accumulation in BMG during long-horizon prediction.

DMTG Module. Table 3 demonstrates the importance of the DMTG module in guiding the denoising process. To further analyze the role of event inputs and their representations, we conduct additional ablations in Table 7. Specifically, we consider three settings: removing the event-based preliminary mask sequence in DMTG, and replacing the event inputs with voxel grid and time surface representations, respectively. For fair comparison, the temporal windows

Table 7: Ablation Study on Different Event-Representation Embedding Strategies in the DMTG Module.

Setting	Event-ViSha				EVSD			
	MAE↓	F_β ↑	IoU↑	BER↓	MAE↓	F_β ↑	IoU↑	BER↓
w/o Event Mask	0.025	0.835	0.705	7.52	0.064	0.596	0.618	14.69
Voxel-based Mask	0.034	0.767	0.654	10.89	0.057	0.632	0.582	17.94
TimeSurface-based Mask	0.021	0.814	0.706	6.97	0.035	0.693	0.652	13.24
ESI-based Mask (ours)	0.015	0.861	0.739	5.76	0.035	0.724	0.664	11.05

of both representations are kept consistent with ESI (33 ms). The results indicate that the event-based preliminary mask sequence provides complementary cues from an alternative modality, improving performance. Moreover, the choice of event representation significantly affects the results, and inappropriate representations may even degrade performance.

Additional ablations on diffusion-related hyperparameters are provided in the **Supplementary Material**.

6 Conclusions

We propose an Event-guided Diffusion Model (EDM) for video shadow detection in complex dynamic scenes. To the best of our knowledge, this is the first work to introduce event cameras into video shadow detection (VSD). By leveraging event streams as a complementary modality to RGB, EDM improves robustness under challenging motion and illumination conditions. Specifically, the MTA module captures shadow trajectories from event streams to enhance temporal modeling, while BMG generates denser mask sequences within short temporal windows to improve temporal consistency. Through ESI, event streams are reconstructed into intensity images to predict event-based preliminary masks, which are fused with RGB-based masks via DMTG and injected into the diffusion process as conditional guidance for direct mask prediction.

In addition, we construct EVSD, the first dual-modality dataset combining event streams and RGB videos for complex video shadow detection. Extensive experiments demonstrate that our method achieves superior performance on both public benchmarks and the proposed EVSD.

7 Acknowledgments

This work was supported by the Science and Technology Talent and Platform Program of Yunnan Province, China (Grant No. 202505AT350001), and the Reward Funds for Research Projects of Tianmushan Laboratory (Grant No. TK-2024-D-010).

References

1. Abrams, A., Schillebeeckx, I., Pless, R.: Structure from shadow motion. In: 2014 IEEE International Conference on Computational Photography (ICCP). pp. 1–8 (2014). <https://doi.org/10.1109/ICCPHOT.2014.6831802>
2. Alonso, I., Murillo, A.C.: Ev-segnet: Semantic segmentation for event-based cameras (2018), <https://arxiv.org/abs/1811.12039>
3. Babenko, B., Yang, M.H., Belongie, S.: Robust object tracking with online multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33**(8), 1619–1632 (2011)
4. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollár, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: Sam 3: Segment anything with concepts (2025), <https://arxiv.org/abs/2511.16719>
5. Chen, T., Li, L., Saxena, S., Hinton, G., Fleet, D.J.: A generalist framework for panoptic segmentation of images and videos. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 909–919 (2023)
6. Chen, T., Zhang, R., Hinton, G.: Analog bits: Generating discrete data using diffusion models with self-conditioning (2023), <https://arxiv.org/abs/2208.04202>
7. Chen, Z., Wan, L., Zhu, L., Shen, J., Fu, H., Liu, W., Qin, J.: Triple-cooperative video shadow detection (2021), <https://arxiv.org/abs/2103.06533>
8. Chen, Z., Zhu, L., Wan, L., Wang, S., Feng, W., Heng, P.A.: A multi-task mean teacher for semi-supervised shadow detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5610–5619 (2020). <https://doi.org/10.1109/CVPR42600.2020.00565>
9. Chen, Z., Zhu, L., Wan, L., Wang, S., Feng, W., Heng, P.A.: A multi-task mean teacher for semi-supervised shadow detection. In: CVPR (2020)
10. Chen, Z., Gao, T., Sheng, B., Li, P., Chen, C.P.: Outdoor shadow estimating using multiclass geometric decomposition based on bls. *IEEE transactions on cybernetics* **50**(5), 2152–2165 (2018)
11. Cheng, H.K., Tai, Y.W., Tang, C.K.: Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in neural information processing systems* **34**, 11781–11794 (2021)
12. Cong, R., Guan, Y., Chen, J., Zhang, W., Zhao, Y., Kwong, S.: Sddnet: Style-guided dual-layer disentanglement network for shadow detection. In: Proceedings of the 31st ACM international conference on multimedia. pp. 1202–1211 (2023)
13. Cucchiara, R., Grana, C., Piccardi, M., Prati, A.: Detecting moving objects, ghosts, and shadows in video streams. *IEEE Transactions on Pattern Analysis Machine Intelligence* **25**(10), 1337–1342 (2003)
14. Cui, Y., Gao, B., Zhang, Y., Dong, X., Xiang, J., Li, D., Tu, Z.: Wd-detr: Wavelet denoising-enhanced real-time object detection transformer for robot perception with event cameras (2025), <https://arxiv.org/abs/2506.09098>
15. Ding, X., Yang, J., Hu, X., Li, X.: Learning shadow correspondence for video shadow detection. In: European Conference on Computer Vision. pp. 705–722. Springer (2022)
16. Dong, X., Zhang, Y., Cui, Y., Xiang, J., Li, D., Tu, Z.: An event-based fast intensity reconstruction scheme for uav real-time perception (2025), <https://arxiv.org/abs/2508.02238>

17. Duan, X., Cao, Y., Zhu, L., Fu, G., Wang, X., Zhang, R., Li, P.: Two-stage video shadow detection via temporal-spatial adaption. In: *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XLVIII*. p. 196–214. Springer-Verlag, Berlin, Heidelberg (2024). https://doi.org/10.1007/978-3-031-73195-2_12, https://doi.org/10.1007/978-3-031-73195-2_12
18. Finlayson, G.D., Drew, M.S., Lu, C.: Intrinsic images by entropy minimization. In: *European Conference on Computer Vision (ECCV)*. Lecture Notes in Computer Science, vol. 3953, pp. 582–595. Springer (2006). https://doi.org/10.1007/11744078_45
19. Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Martinez-Gonzalez, P., Garcia-Rodriguez, J.: A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing* p. S1568494618302813 (2018)
20. Gehrig, M., Millh  usler, M., Gehrig, D., Scaramuzza, D.: E-raft: Dense optical flow from event cameras. In: *2021 International Conference on 3D Vision (3DV)*. pp. 197–206. IEEE (2021)
21. Guo, R., Dai, Q., Hoiem, D.: Single-image shadow detection and removal using paired regions. In: *CVPR 2011*. pp. 2033–2040 (2011). <https://doi.org/10.1109/CVPR.2011.5995725>
22. Hu, X., Wang, T., Fu, C.W., Jiang, Y., Wang, Q., Heng, P.A.: Revisiting shadow detection: A new benchmark dataset for complex world. *IEEE Transactions on Image Processing* **30**, 1925–1934 (2021). <https://doi.org/10.1109/tip.2021.3049331>, <http://dx.doi.org/10.1109/TIP.2021.3049331>
23. Hu, Y., Liu, S.C., Delbruck, T.: v2e: From video frames to realistic dvs events (2021), <https://arxiv.org/abs/2006.07722>
24. Khan, S.H., Bennamoun, M., Sohel, F., Togneri, R.: Automatic feature learning for robust shadow detection. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1939–1946 (June 2014). <https://doi.org/10.1109/CVPR.2014.249>
25. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Doll  r, P., Girshick, R.: Segment anything (2023), <https://arxiv.org/abs/2304.02643>
26. Lalonde, J.F., Efros, A.A., Narasimhan, S.G.: Detecting ground shadows in outdoor consumer photographs. In: *Proceedings of the 11th European Conference on Computer Vision (ECCV)*. Lecture Notes in Computer Science, vol. 6312, pp. 322–335. Springer (2010). https://doi.org/10.1007/978-3-642-15555-0_24
27. Lee, Y.S., Chung, W.Y.: Visual sensor based abnormal event detection with moving shadow removal in home healthcare applications. *Sensors* **12**(1), 573–584 (2012)
28. Li, J., Li, H.: Neural reflectance for shape recovery with shadow handling. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 16221–16230 (2022)
29. Li, Z., Sun, K., Yao, R., Zhu, H., Hu, F., Zhao, J., Shao, Z., Zhou, Y.: Dtnet: Improving video shadow detection via dark-aware guidance and tokenized temporal modeling (2025), <https://arxiv.org/abs/2511.06925>
30. Lichtsteiner, P., Posch, C., Delbruck, T.: A 128 128 120 db 15 s latency asynchronous temporal contrast vision sensor (2006)
31. Liu, L., Prost, J., Zhu, L., Papadakis, N., Li  , P., Sch  nlieb, C.B., Aviles-Rivero, A.I.: Scotch and soda: A transformer video shadow detection framework (2023), <https://arxiv.org/abs/2211.06885>

32. Lu, X., Cao, Y., Liu, S., Long, C., Chen, Z., Zhou, X., Yang, Y., Xiao, C.: Video shadow detection via spatio-temporal interpolation consistency training (2022), <https://arxiv.org/abs/2206.08801>
33. Lu, X., Cao, Y., Liu, S., Long, C., Chen, Z., Zhou, X., Yang, Y., Xiao, C.: Video shadow detection via spatio-temporal interpolation consistency training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3116–3125 (2022)
34. Mitrokhin, A., Ye, C., Fermuller, C., Aloimonos, Y., Delbruck, T.: Ev-imo: Motion segmentation dataset and learning pipeline for event cameras (2020), <https://arxiv.org/abs/1903.07520>
35. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European conference on computer vision. pp. 402–419. Springer (2020)
36. Wang, Y., Zhou, W., Mao, Y., Li, H.: Detect any shadow: Segment anything for video shadow detection. *IEEE Transactions on Circuits and Systems for Video Technology* **34**(5), 3782–3794 (2024). <https://doi.org/10.1109/TCSVT.2023.3320688>
37. Xia, R., Zhao, C., Zheng, M., Wu, Z., Sun, Q., Tang, Y.: Cmda: Cross-modality domain adaptation for nighttime semantic segmentation (2023), <https://arxiv.org/abs/2307.15942>
38. Xie, B., Deng, Y., Shao, Z., Li, Y.: EISNet: A multi-modal fusion network for semantic segmentation with events and images. *IEEE Trans. Multimedia* **26**, 8639–8650 (2024)
39. Yang, M., Delbruck, T., Brandli, C., Liu, S.-C., Berner, R.: A 240×180 130 db 3 s latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits* **49**(10), 2333–2341 (2014)
40. Yang, H., Wang, T., Hu, X., Fu, C.W.: Silt: Shadow-aware iterative label tuning for learning to detect shadows from noisy labels. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12687–12698 (2023)
41. Yao, Z., Ying, X., Zhu, Z., Chuah, M.C.: Learning flow-guided registration for rgb-event semantic segmentation (2025), <https://arxiv.org/abs/2505.01548>
42. Zhang, J., Liu, R., Shi, H., Yang, K., Reiß, S., Peng, K., Fu, H., Wang, K., Stiefelhagen, R.: Delivering arbitrary-modal semantic segmentation. In: CVPR (2023)
43. Zhou, H., Wang, H., Ye, T., Xing, Z., Ma, J., Li, P., Wang, Q., Zhu, L.: Timeline and boundary guided diffusion network for video shadow detection. In: Proceedings of the 32nd ACM International Conference on Multimedia. p. 166–175. MM '24, ACM (Oct 2024). <https://doi.org/10.1145/3664647.3681236>, <http://dx.doi.org/10.1145/3664647.3681236>
44. Zhu, L., Chen, X., Wang, L., Wang, X., Tian, Y., Huang, H.: Continuous-time object segmentation using high temporal resolution event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024)