DIMOS: Disentangling Instance-level Moving Object Segmentation

Hongxiang Huang Hongwei Ren Xiaopeng Lin Yulong Huang Zeke Xie Bojun Cheng
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
hhuang516@connect.hkust-gz.edu.cn bocheng@hkust-gz.edu.cn

Abstract

Moving instance segmentation (MIS) attracts increasing attention due to its broad applications in traffic surveillance, autonomous driving, and animal tracking. Event cameras record asynchronous brightness changes, providing high temporal resolution and dynamic range, which makes them highly sensitive to motion information. By fusing event and image features, motion cues from events can complement spatial details from images, enhancing the performance of MIS. However, current multimodal MIS methods still struggle to segment small moving instances, as event cameras often yield sparse features under limited resolution. In addition, event features entangle appearance attributes with motion cues, which further restricts effective cross-modal fusion. To address these challenges, we first propose a dual-disentangling feature extraction framework that separates and extracts appearance and motion information within both image and event modalities, thereby improving feature density. Subsequently, a multi-granularity cross-modal alignment is introduced to align distributionally and semantically consistent features across modalities, enabling more effective fusion with rich spatial and temporal details. The experiment results demonstrate that our method achieves state-of-the-art performance in multimodal MIS, especially for small instances under challenging conditions such as fast motion and low-light settings. [1]

1. Introduction

Moving instance segmentation has gained increasing attention owing to its wide applications in traffic surveillance [54], autonomous driving [53], and animal tracking [15]. This task is inherently more challenging than conventional semantic segmentation, as it requires not only distinguishing object categories but also segmenting individual instances and identifying their independent motions. Recent advances in moving instance segmentation algorithms have achieved impressive results [5, 46]. However, the performance of image-based methods still shows limitations in extreme environments, such as low illumination, backlighting, and high-speed motion scenarios [15, 23].

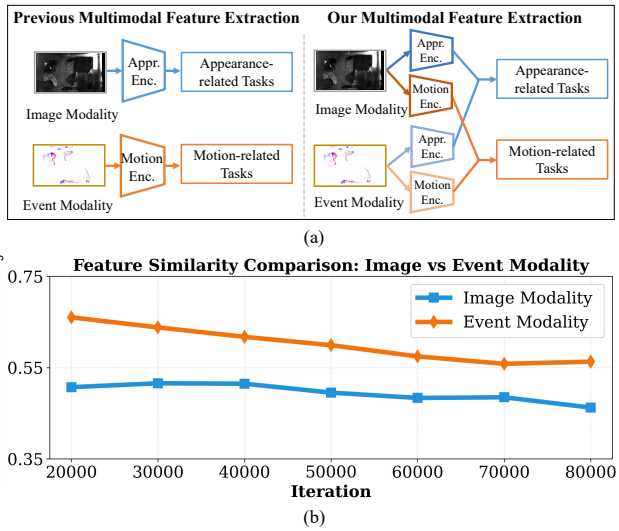

Figure 1. Feature Extraction Comparison. (a) Our method extracts both appearance and motion features from each modality. (b) Model checkpoints at different iterations are sampled to compute the cosine similarity between appearance and motion features extracted from encoders of the same modality on MouseSIS dataset.

Event cameras, with their unique advantages in low latency, high dynamic range, and low power consumption, have shown potential to overcome the limitations of image-based methods [11, 19]. Built on more complex pixel circuits, event cameras achieve extremely high temporal resolution but relatively low spatial resolution [25]. In addition, they generate sparse and asynchronous events rather than dense frame-based signals. Although these characteristics are advantageous for capturing rapid motion in extreme environments [18, 28, 36, 37, 50], they also introduce challenges for dense prediction tasks such as instance segmentation. In particular, the sparsity of event streams leads

[1] Project page: https://github.com/Neuromorphic-Electronics-Photonics-Lab/DIMOS-Moving-Instance-Segmentation-CVPR2026.

to incomplete spatial context, making it difficult to capture clear object boundaries and regions. As a result, although event data are valuable for mitigating limitations under extreme conditions, models relying solely on event inputs still struggle to achieve competitive performance in moving instance segmentation.

Recent research has increasingly explored multimodal fusion [24, 27, 47], in which each modality provides complementary strengths. Image frames provide appearance cues such as texture and structural details, while event streams provide motion information. This paradigm shows clear performance improvements. However, event cameras still suffer from limited resolution due to their larger pixel pitch, and current fusion methods often require maintaining consistent resolution between image and event sensors [16, 41]. These factors make small instance segmentation very challenging. Since small objects occupy only a limited number of pixels, both appearance and motion information become constrained. This sparsity leads to insufficient feature density and degraded segmentation quality. In practice, each modality naturally contains both types of cues. Image frames contain motion cues, as widely utilized in optical flow estimation [10, 34]. The density or distribution of event streams provides implicit clues about appearance patterns related to shape, texture, and material, as these attributes determine the surface reflectance [12, 13, 31, 51]. Such inherent properties have not been fully explored or exploited in moving instance segmentation.

A promising solution is to extract both appearance and motion cues from each modality instead of relying on a strict separation between them, which increases feature density and enables more effective use of the available pixels, as shown in Figure 1(a). However, achieving such joint extraction is not equally straightforward for different modalities. For image data, the separation between appearance and motion cues is well-understood. Appearance details are naturally captured by the camera, while motion information can be derived from temporal differences or motion blur [14]. In contrast, event data exhibits strong entanglement between appearance and motion cues because both motion and appearance characteristics can induce variations in event density and distribution that are difficult to distinguish. This coupling complicates the extraction of clean appearance and motion features from the event modality and ultimately weakens the effectiveness of cross-modal fusion. As shown in Figure 1(b), features extracted from the event modality exhibit higher similarity across different types compared to those from the image modality. To solve the entanglement problem, feature disentanglement has been explored in other dense prediction tasks, such as image segmentation [33, 44], super resolution [22, 42], and image generation [6, 9], where it helps isolate different semantic factors and improve feature interpretability. These observations motivate an intra-modal disentanglement strategy to effectively separate appearance and motion cues, especially for event data.

In this work, we focus on disentangling and extracting both appearance and motion features within each modality. Specifically, we design a Disentangling Instance-level Moving Object Segmentation (DIMOS) framework, using dual-disentangling encoders with intra-modal contrastive learning and task-specific supervision to disentangle appearance and motion features from both image and event modalities. Intra-modal contrastive learning enhances the discriminability between appearance and motion information, while task-specific supervision constrains the disentangled features to learn appearance and motion cues, respectively. Since disentanglement produces two types of features per modality, effective multimodal learning requires aligning appearance and motion features across modalities. To this end, we introduce a multi-granularity cross-modal alignment mechanism that combines adversarial domain adaptation and modality translation. This design facilitates effective feature fusion by jointly enforcing distributional and semantic alignment between cross-modal features.

- We propose a dual-disentangling mechanism that extracts both appearance and motion features from each modality, enhancing feature density and representation quality.
- We design a multi-granularity cross-modal alignment to enforce distributional and semantic consistency for effective feature fusion.
- Experiments demonstrate that our approach achieves state-of-the-art performance, validating the effectiveness of the proposed disentanglement framework.

2. Related Work

2.1. Video Object Segmentation

Video object segmentation (VOS) lays the foundation for most moving object and instance segmentation tasks. In the semi-supervised setting, an initial annotation in the first frame is propagated over time. Early methods focused on efficient temporal propagation, while later methods emphasized robust feature matching. Representative methods such as FEELVOS [40], CFBI [48], and CFBI+ [49] significantly improved temporal consistency and stability, forming the basis for many subsequent frameworks.

Beyond, unsupervised or zero-shot VOS methods further relax need for an initial annotation by relying only on the intrinsic visual and motion cues in videos. MAT-Net [56] combines appearance and motion for foreground discovery, while Isomer [52] leverages transformer-based architectures for long-range temporal modeling. Another approach [55] explores multi-source fusion to improve segmentation accuracy in more diverse scenarios. Although these approaches improve generalization, they remain sen-

sitive to complex camera motion and adverse conditions. To address this limitation, recent works integrate event data into VOS. Event cameras offer high temporal resolution and clean motion cues under challenging lighting conditions. ELVOS [23] demonstrates that fusing event streams with images significantly improves temporal correspondence and segmentation reliability, highlighting cross-modal fusion as a promising strategy for robust VOS.

2.2. Moving Instance Segmentation

Moving instance segmentation (MIS) extends VOS from foreground localization to distinguishing multiple independently moving objects. Early motion segmentation approaches, such as FgSegNet [2], primarily focus on identifying moving regions through background subtraction, optical flow estimation, or motion clustering. Event-based variants, including EVIMO [30], GConv [31], and UnEVIMO [43], further leverage the high temporal resolution of event cameras to separate independent motions. However, due to the sparse and asynchronous nature of event data, these approaches often produce coarse object boundaries and incomplete contours, limiting their performance in fine-grained instance-level segmentation. To achieve more precise instance discrimination, image-based MIS methods adapt video instance segmentation (VIS) architectures that exploit rich appearance cues. IDOL [46] introduces an interaction mechanism between detection and segmentation to maintain temporal consistency. Although it performs well in structured scenes, its reliance on single-modality inputs restricts robustness under motion blur, occlusion, and other degraded conditions.

To overcome these limitations, multimodal pipelines emerge to integrate complementary cues from both images and events. ModelMixSort [15] combines YOLO [35] detectors on RGB and event-derived grayscale frames with a SAM [8, 21] model, whereas EvInsMOS [41] explicitly fuses image texture and event-based motion cues through contrastive learning [7, 32] and cross-modal masked attention. Despite these advances, most multimodal MIS frameworks still follow a simplified paradigm that extracts appearance information from images and motion information from events. Such designs often result in insufficient feature density for small objects and weak semantic correspondence across modalities, highlighting the necessity for a unified framework that jointly disentangles and aligns appearance–motion representations to achieve more robust moving instance segmentation.

3. Method

In this section, we present the **Disentangling Instance-Level Moving Object Segmentation (DIMOS)** framework. It extracts both appearance and motion features from each modality and introduces feature disentanglement with multi-granularity cross-modal alignment to enhance feature quality and fusion robustness.

3.1. Problem Definition

Given a sequence of image frames $\mathbf{I}_t$ and the corresponding event stream $\mathbf{E}_{[t,t+\Delta t]}$ recorded by an event camera over the same time interval, our goal is to perform pixel-level segmentation of all independently moving objects in the scene and to predict their motion states at the current time step. Specifically, the model is required to predict an instance mask $\hat{m}_k$ and a binary motion label $\hat{y}_k \in \{0, 1\}$ for each instance. The overall output can be expressed as follows,

$$\hat{\mathbf{M}} = \{\hat{m}_k\}_{k=1}^{K}, \quad \hat{\mathbf{Y}} = \{\hat{y}_k\}_{k=1}^{K}, \tag{1}$$

where K denotes the number of instances in the current frame.

Event cameras produce asynchronous event streams $\mathbf{E} = \{e_i\}_{i=1}^{N}$, where each event $e_i = (x_i, y_i, t_i, p_i)$ encodes a brightness change at time t_i and pixel (x_i, y_i) with polarity p_i. While event data offers ultra-high temporal resolution, it is inherently sparse and irregular, making it less suitable for dense prediction tasks. To address this, we discretize the event stream into B temporal bins and accumulate events spatially and temporally to obtain a voxel representation as

$$\mathbf{V}_t(x, y, b) = \sum_{\substack{i \\ (x_i, y_i)=(x,y)}} p_i \, \max\Big(0,\, 1 - \Big|b - \tfrac{t_i - t}{\Delta t}(B-1)\Big|\Big), \tag{2}$$

where $b \in \{0, 1, \ldots, B-1\}$ is the bin index, B is the total number of bins, and Δt denotes the duration of the event slice between the current frame and the next frame.

The two modalities are combined as $\mathbf{X} = \{\mathbf{I}_t, \mathbf{V_t}\}$ and assigned to the segmentation output through a mapping function with learnable parameters ϕ as follows,

$$f_\phi(\mathbf{X}) = (\hat{\mathbf{M}}, \hat{\mathbf{Y}}). \tag{3}$$

3.2. Method Overview

Our framework consists of four main components: a dual-disentangling module, a cross-modal alignment and fusion module, a cross-type interaction module, and a task-specific module. Compared to previous works in VIS [4, 45], we make several key modifications. First, both image and event are encoded independently with a dual-branch encoder to extract appearance and motion features, as shown inside the orange dashed box in Figure 2. To ensure a clear separation of appearance and motion features, we adopt intra-modal contrastive learning to enhance disentanglement within each branch. Second, the two inputs yield four feature vectors that are subsequently fused across the image and event modalities into appearance and motion features. This is accomplished by a multi-granularity cross-modal

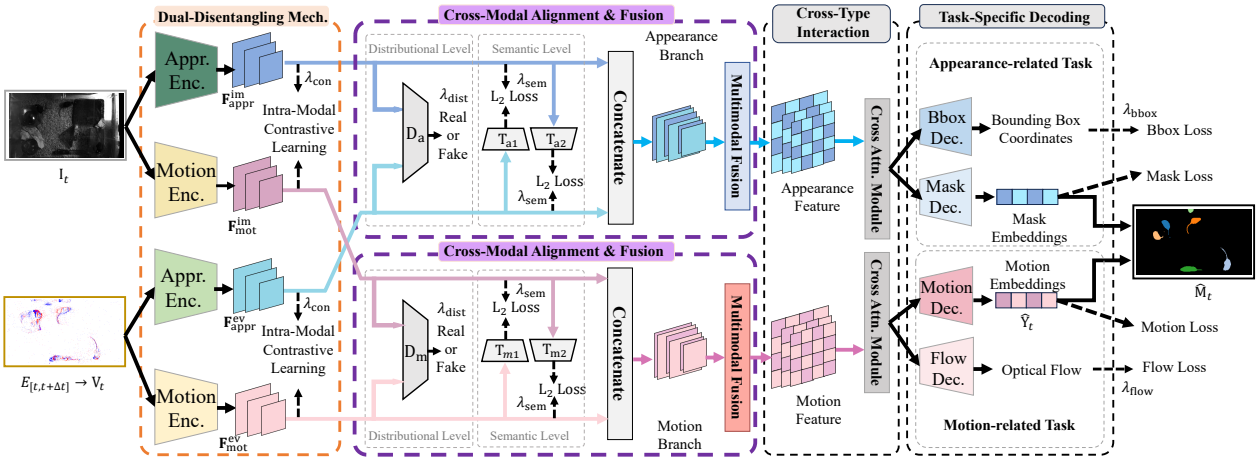

Figure 2. **Overview of the proposed DIMOS framework.** The pipeline consists of four major components: (1) **Dual-Disentangling Mechanism** with appearance and motion encoders for each modality. (2) **Multi-Granularity Cross-Modal Alignment & Fusion** that enforces consistency at both distributional and semantical levels. (3) **Cross-Type Interaction** for joint reasoning between appearance and motion cues via cross attention. and (4) **Task-Specific Decoders** for appearance-related and motion-related predictions. D_a and D_m denote the domain discriminators for appearance and motion branches used in adversarial distribution alignment, while T_{a1}, T_{a2}, T_{m1}, and T_{m2} represent modality translation modules that perform bidirectional feature reconstruction for semantic-level alignment.

alignment module and a lightweight CNN fusion layer for the image and event modalities, as shown inside the purple dashed boxes in Figure 2. Third, to preserve the semantics of disentangled features, we employ not only instance segmentation and bounding box regression for decoders of appearance-related tasks, but also motion classification and optical flow estimation for decoders of motion-related tasks.

We employ different strategies for inference and training. During inference, we utilize the mask fusion procedure proposed in [41]. The model first upsamples the predicted mask embeddings to full resolution and then fuses the motion classification results by applying a confidence threshold θ to the motion scores. Only masks whose confidence exceeds θ are retained as moving instances, producing the final instance-level segmentation map. In contrast, during training, the supervision is assigned by Hungarian matching between predicted masks and ground-truth instances [3], ensuring a one-to-one correspondence for both mask prediction and motion classification without thresholding.

3.3. Dual-Disentangling Mechanism

To ensure sufficient information for small instance segmentation, we aim to jointly exploit appearance and motion information extracted from each modality. Providing complementary information from both the image and the event modalities can alleviate the lack of dense information within a single modality. Therefore, we adopt a dual-branch encoder for each modality to simultaneously extract appearance and motion information from image and event inputs.

The dual-branch encoders of the two modalities share the same input but are parameterized independently and trained with different task-specific supervision signals to learn distinct semantic representations. We denote the appearance and motion features extracted from images as $\mathbf{F}^{\text{im}}_{\text{appr}}$ and $\mathbf{F}^{\text{im}}_{\text{mot}}$, and those from events as $\mathbf{F}^{\text{ev}}_{\text{appr}}$ and $\mathbf{F}^{\text{ev}}_{\text{mot}}$. To further enhance the disentanglement between the appearance and motion branches, we incorporate **intra-modal contrastive learning**. Unlike previous works that apply contrastive learning for improving cross-modal feature discriminability, we focus on intra-modal separation of appearance and motion features. This encourages the network to emphasize differences between appearance and motion semantics rather than across modalities, avoiding redundant or mixed representations across branches. For each modality, positive samples $\mathbf{F}^{+}$ are selected from the same type (appearance or motion) and consecutive frames, while negative samples $\mathbf{F}^{-}$ are sampled from different types or non-consecutive frames. The InfoNCE loss [32] for intra-modal contrastive learning is defined as

$$\mathcal{L}_{\text{con}} = -\log \frac{\exp\left(\mathbf{F}\cdot\mathbf{F}^{+}/\tau\right)}{\exp\left(\mathbf{F}\cdot\mathbf{F}^{+}/\tau\right) + \sum_{\mathbf{F}^{-}} \exp\left(\mathbf{F}\cdot\mathbf{F}^{-}/\tau\right)}, \tag{4}$$

where $\cdot$ denotes dot product between two ℓ_2-normalized features and τ is a temperature factor. The specific construction of positive and negative samples for appearance and motion features is detailed in the supplementary material.

3.4. Multi-Granularity Cross-Modal Alignment

Previous multimodal fusion methods often combine features from different modalities through concatenation and simple linear or convolutional operations. Without proper

alignment before fusion, these methods fail to ensure semantic consistency or fully exploit the complementarity between image and event data. This limitation arises from the inherent distributional and semantic gap between the two modalities. Therefore, we propose a **multi-granularity cross-modal alignment** that aligns appearance and motion features across modalities by enforcing consistency at both distributional and semantic levels during the feature fusion.

Distribution-Level Alignment via Domain Adaptation. At the distributional level, the two modalities can be regarded as two "domains" of the same underlying scene. The distribution gap between them makes feature alignment difficult, even if they share the same semantic category (e.g., appearance or motion). To bridge this gap, we employ **adversarial domain adaptation** [39] to learn domain-invariant representations.

We introduce two discriminators for the appearance and motion branches. The discriminators classify feature modality, while the encoders, through a gradient reversal layer, learn to minimize this gap. An asymmetric strategy is adopted, where image features serve as the reference domain for appearance alignment and event features for motion alignment. This design leverages the fact that images provide more explicit appearance cues, while events contain clearer motion cues. The adversarial loss is as follows:

$$\mathcal{L}_{\text{adv}} = \mathcal{L}_{\text{adv}}^{\text{appr}} + \mathcal{L}_{\text{adv}}^{\text{mot}}, \tag{5}$$

where each component corresponds to an adversarial training objective between the encoder and its corresponding domain discriminator. The adversarial loss can be formulated as a min–max optimization as below,

$$\min_{G}\max_{D}\ \mathbb{E}_{x\sim p_{\text{ref}}}\big[\log D(x)\big]+\mathbb{E}_{z\sim p_{\text{src}}}\big[\log(1-D(G(z)))\big], \tag{6}$$

where G denotes the feature encoder, D the domain discriminator, p_{ref} the reference domain distribution, and p_{src} the source domain distribution. For the appearance branch, $x = \mathbf{F}_{\text{appr}}^{\text{im}}$ and $G(z) = \mathbf{F}_{\text{appr}}^{\text{ev}}$. For the motion branch, $x = \mathbf{F}_{\text{mot}}^{\text{ev}}$ and $G(z) = \mathbf{F}_{\text{mot}}^{\text{im}}$. Detailed formulations are provided in the supplementary material.

Semantic-Level Alignment via Modality Translation. Distribution-level alignment alone cannot fully ensure semantic consistency across modalities. To bridge this gap, we introduce two lightweight convolutional **modality transform modules** that translate appearance and motion features between the image and event spaces, enforcing bidirectional semantic consistency during training. A reconstruction loss regularizes this process as below,

$$\begin{aligned}\mathcal{L}_{\text{trans}} =&\|T_{a1}(\mathbf{F}_{\text{appr}}^{\text{im}}) - \mathbf{F}_{\text{appr}}^{\text{ev}}\|_2^2 + \|T_{a2}(\mathbf{F}_{\text{appr}}^{\text{ev}}) - \mathbf{F}_{\text{appr}}^{\text{im}}\|_2^2 \\ &+ \|T_{m1}(\mathbf{F}_{\text{mot}}^{\text{im}}) - \mathbf{F}_{\text{mot}}^{\text{ev}}\|_2^2 + \|T_{m2}(\mathbf{F}_{\text{mot}}^{\text{ev}}) - \mathbf{F}_{\text{mot}}^{\text{im}}\|_2^2,\end{aligned} \tag{7}$$

where T is the translation module. This design ensures that features of the same semantic type are mutually translatable, strengthening cross-modal alignment and providing a more stable foundation for fusion.

Importantly, both distributional and semantic alignments are entirely unsupervised and only applied during training, introducing no extra cost at inference time.

3.5. Optimizing Objectives

The overall training objective of DIMOS integrates task-specific supervision, intra-modal contrastive learning, and cross-modal alignment.

Main Task Loss. Following Sec. 3.1, the model predicts both instance masks and motion states. The instance segmentation loss is defined as

$$\mathcal{L}_{\text{mov\_seg}} = \frac{1}{K}\sum_{k=1}^{K}\big[\mathcal{L}_{\text{cls}}(\hat{y}_k, y_k) + \mathcal{L}_{\text{mask}}(\hat{m}_k, m_k)\big], \tag{8}$$

where y_k and $\hat{y}_k$ are the ground truth and predicted motion labels, and m_k and $\hat{m}_k$ denote the ground truth and predicted masks. $\mathcal{L}_{\text{cls}}$ is a standard cross-entropy loss used for class prediction, and $\mathcal{L}_{\text{mask}}$ is a binary cross-entropy loss for mask supervision.

Extra Task-Specific Loss. To enhance appearance and motion perception, we introduce two extra objectives. For motion modeling, an unsupervised optical flow estimation loss [41] is defined as

$$\mathcal{L}_{\text{flow}} = \sum_{\mathbf{c}} \psi\Big(\mathbf{I}_t(\mathbf{c}) - \mathbf{I}_{t+\Delta}\big(\mathbf{c} + \hat{\mathbf{F}}_{t\to t+\Delta}(\mathbf{c})\big)\Big), \tag{9}$$

where $\hat{\mathbf{F}}_{t\to t+\Delta}$ denotes the predicted optical flow between two adjacent frames through an FPN-based flow decoder. Here, $\mathbf{I}_t(\mathbf{c})$ and $\mathbf{I}_{t+\Delta}(\mathbf{c})$ represent the pixel intensities of two consecutive frames viewed as continuous functions of spatial coordinate $\mathbf{c}$. The term $\mathbf{I}_{t+\Delta}(\mathbf{c} + \hat{\mathbf{F}}_{t\to t+\Delta}(\mathbf{c}))$ samples the next frame at a displaced location determined by the estimated flow, effectively warping it toward the current frame. The robust function $\psi(\cdot)$ follows $(|u| + \epsilon)^q$ with $\epsilon = 0.01$ and $q = 0.4$ [29]. For appearance modeling, a bounding box regression loss is given by

$$\mathcal{L}_{\text{bbox}} = \|\hat{\mathbf{c}_{\text{b}}} - \mathbf{c}_{\text{b}}\|_1, \tag{10}$$

where $\mathbf{c}_{\text{b}}$ denotes the reference coordinates of bounding boxes.

Finally, we combine all losses, including the intra-modal contrastive loss $\mathcal{L}_{\text{con}}$ (Sec. 3.3) and the cross-modal alignment losses $\mathcal{L}_{\text{adv}}$ and $\mathcal{L}_{\text{trans}}$ (Sec. 3.4), into the total objective as follows,

$$\begin{aligned}\mathcal{L}_{\text{total}} = &\mathcal{L}_{\text{mov\_seg}} + \lambda_{\text{flow}}\mathcal{L}_{\text{flow}} + \lambda_{\text{bbox}}\mathcal{L}_{\text{bbox}} \\ &+ \lambda_{\text{con}}\mathcal{L}_{\text{con}} + \lambda_{\text{dist}}\mathcal{L}_{\text{adv}} + \lambda_{\text{sem}}\mathcal{L}_{\text{trans}},\end{aligned} \tag{11}$$

where λ_{flow}, λ_{bbox}, λ_{con}, λ_{dist}, and λ_{sem} are balancing coefficients.

4. Experiments

In this section, we evaluate our proposed method (DIMOS) on three challenging datasets that contain both image and event modalities. We compare our approach with frame-based and event-assisted methods. We further provide ablation studies to analyze the contribution of each component in our architecture.

Table 1. Summary of the three datasets

Dataset	Avg. Inst. per Frame	Avg. Inst. Mask Area	Avg. Foreg. Mask Area
MouseSIS	4.10	0.73%	3.01%
SEVD-Fixed	7.68	0.15%	1.12%
EVIMO	1.34	3.74%	5.03%

4.1. Datasets

We conduct extensive experiments on three benchmarks: MouseSIS [15], SEVD-Fixed [1], and EVIMO [30].

MouseSIS [15] contains synchronized grayscale frames and event streams of interacting mice with over 75000 temporally consistent instance masks. The targets are small and frequently occluded, making it suitable for evaluating fine-grained segmentation of objects with low foreground ratio.

SEVD-Fixed [1] is a synthetic traffic surveillance dataset with RGB, event, depth, and semantic labels captured under diverse lighting and weather conditions. Foreground objects like vehicles and pedestrians are often small, challenging precise instance segmentation.

EVIMO [30] provides indoor event streams with ground-truth motion masks and depth for up to three moving objects, serving as a standard benchmark for motion segmentation.

As shown in Table 1, MouseSIS and SEVD-Fixed exhibit lower average instance mask area ratio (0.73% and 0.15%) compared to EVIMO (3.74%), and lower foreground coverage. This highlights their particular challenge on small instance segmentation. More dataset details are provided in the supplementary material.

4.2. Implementation Details

We implement our framework in PyTorch. For the three datasets, we train the network for 400K iterations on MouseSIS, 500K on EVIMO, and 800k on SEVD-Fixed with a batch size of 16. We use the Adam optimizer [20] with a weight decay of 1×10^{-6} and employ a one-cycle learning rate schedule, with the peak learning rate set to 1×10^{-4}. The number of event bins is set to $B = 10$ and the moving confidence threshold is set to $\theta = 0.1$ across all experiments. The loss weights are set to $\lambda_{\text{flow}} = 10.0$, $\lambda_{\text{con}} = 0.5$, $\lambda_{\text{bbox}} = 0.01$, $\lambda_{\text{dist}} = 0.1$, and $\lambda_{\text{sem}} = 10.0$. All experiments, including ablations and comparisons with prior methods, are conducted on the same evaluation machine. Training is performed on dual A40 GPUs, and inference is conducted on a single RTX 5090 GPU to ensure consistent evaluation settings. More architectural and training details are provided in the supplementary material.

Due to the relatively low spatial resolution of most existing DVS sensors and the requirement to maintain consistent resolution across modalities, we downsample different datasets to specific target resolutions. The input resolutions are resized to 320×180 for MouseSIS and 512×384 for SEVD-Fixed. This downsampling strategy significantly reduces computational overhead while preserving instance-level discriminability, particularly for small objects. For EVIMO, we use the original resolution of 346×260. Importantly, the lower resolution setting further highlights the challenge of small instance segmentation.

Following previous works [26, 41, 57], we adopt three primary metrics for moving instance segmentation: $\mathbf{mIoU}_{\text{ins}}$, $\mathbf{mIoU}_{01}$, and $\mathbf{mAP}$. Specifically, $\mathbf{mIoU}_{\text{ins}}$ evaluates instance-level segmentation accuracy for each moving object, while $\mathbf{mIoU}_{01}$ measures the 0–1 binary foreground mask accuracy. We further report $\mathbf{mAP}$ to account for false positives and overall detection precision.

4.3. Quantitative Results

We quantitatively evaluate the proposed framework against representative frame-based (IDOL [46]) and event-assisted (ModelMixSort [15] and EvInsMos [41]) methods on three challenging benchmarks: MouseSIS, SEVD-Fixed, and EVIMO. The results are summarized in Table 2.

On the MouseSIS dataset, our method achieves the best instance-level segmentation accuracy with the $\mathbf{mIoU}_{\text{ins}}$ of 70.25%, outperforming both classical frame-based method (IDOL [46]) and event-assisted baselines. Notably, ModelMixSort [15] and EvInsMOS [41] already yield improvements over image-only methods, which confirms the benefits of leveraging event data under the challenging illumination condition. Our method further boosts performance through explicit disentanglement and alignment strategies. On the SEVD-Fixed dataset, which presents more challenging scenarios than MouseSIS due to its complex outdoor environments, diverse weather conditions, and a larger number of smaller instances, our framework achieves 62.05% $\mathbf{mIoU}_{\text{ins}}$, outperforming EvInsMOS [41] by 5.55%. This consistent gain highlights the superior robustness of our approach under extreme conditions, where event signals effectively complement degraded image data. On the EVIMO dataset, although EvInsMOS [41] and ModelMixSort [15] already perform strongly by leveraging both modalities, our method achieves the highest $\mathbf{mIoU}_{\text{ins}}$ of 72.08%, indicating the effectiveness of our disentanglement strategy.

Overall, these results demonstrate that our method achieves consistent improvements across different bench-

Table 2. Quantitative comparison on MouseSIS, SEVD-Fixed, and EVIMO.

Dataset	Method	Backbone	$\mathbf{mIoU}_{\text{ins}}$ (%)	$\mathbf{mIoU}_{01}$ (%)	mAP (%)	FLOPs
MouseSIS	IDOL [46]	ResNet-50	60.66	66.96	26.73	68.13G
	ModelMixSort [15]	YOLO-SAM	63.72	75.79	23.11	5.49T
	EvInsMOS [41]	ResNet-50	62.54	75.34	30.94	26.08G
	DIMOS (ours)	ResNet-50	**70.25**	**77.30**	**45.18**	60.42G
SEVD-Fixed	IDOL [46]	ResNet-50	45.49	52.17	16.13	195.85G
	ModelMixSort [15]	YOLO-SAM	49.56	61.43	18.47	5.55T
	EvInsMOS [41]	ResNet-50	56.50	58.45	20.24	87.52G
	DIMOS (ours)	ResNet-50	**62.05**	**61.53**	**23.29**	201.26G
EVIMO	IDOL [46]	ResNet-50	69.35	72.01	33.08	106.12G
	ModelMixSort [15]	YOLO-SAM	71.67	**78.33**	33.99	5.50T
	EvInsMOS [41]	ResNet-50	71.26	75.19	35.97	40.95G
	DIMOS (ours)	ResNet-50	**72.08**	75.74	**36.44**	94.81G

marks with diverse illumination conditions, motion patterns, and scene complexities. Importantly, both MouseSIS and SEVD-Fixed contain a large number of small instances, where accurate segmentation strongly relies on dense appearance and motion features. Our disentanglement framework effectively enhances the segmentation of such small instances by simultaneously extracting appearance and motion information from both modalities. In contrast, conventional frame-based or simple fusion methods do not explicitly perform such dual-modality disentangling, which highlights the advantage of our approach in jointly leveraging complementary appearance and motion cues.

4.4. Qualitative Results

Figure 3 presents qualitative comparisons among representative methods on MouseSIS datasets. The results consistently demonstrate that our proposed DIMOS framework achieves more accurate and temporally consistent segmentation, particularly for small moving instances.

Compared to IDOL, our model produces cleaner object boundaries and avoids missing detections under motion blur or low-illumination conditions. The superiority becomes evident where image-only models often fail to distinguish object contours or confuse overlapping instances. Event-assisted baselines, including ModelMixSort [15] and EvInsMOS [41], perform better by leveraging event information; however, they still exhibit fragmented masks or inaccurate separations when multiple objects move closely or interact. Overall, the visual comparisons align well with the quantitative results, confirming that the proposed framework achieves superior robustness for small instance segmentation under limited resolution.

4.5. Ablation Study

To investigate the contribution of each component, we perform ablation experiments on MouseSIS by progressively enabling task-specific supervision, dual-disentangling mechanism, alignment modules, and encoder backbones. Results are shown in Table 3 and Table 4.

Table 3. Ablation study on MouseSIS. We incrementally add dual-disentangling mechanism (Dual. Mech.), extra task-specific losses (UnFlow and BBox), semantic alignment loss (Sem. Align.), and distributional alignment loss (Dist. Align.).

UnFlow	BBox	Dual. Mech.	Sem. Align.	Dist. Align.	$\mathbf{mIoU}_{\text{ins}}$ (%)
×	×	×	×	×	60.47
✓	×	×	×	×	62.54
✓	✓	×	×	×	63.46
✓	✓	✓	×	×	68.11
✓	✓	✓	✓	×	69.23
✓	✓	✓	✓	✓	70.25

Baseline without extra modules. As shown in Table 3, without any additional modules, the model reduces to a simple multimodal interaction pipeline, achieving only 60.47% $\mathbf{mIoU}_{\text{ins}}$, which confirms the weakness of insufficient feature extraction in handling small instance segmentation under challenging conditions.

Effect of extra task-specific supervision. As shown in Table 3, introducing unsupervised flow estimation improves $\mathbf{mIoU}_{\text{ins}}$ from 60.47% to 62.54%, demonstrating that additional motion guidance helps capture motion cues. Incorporating bounding box supervision further raises it to 63.46%, providing a spatial prior that enhances localization, particularly for small or overlapping objects. For sequences without box annotations, pseudo boxes are generated from the outer boundaries of instance masks.

Effect of dual-disentangling mechanism. Table 3 also shows that the dual-disentangling mechanism brings a significant performance increase to 68.11% $\mathbf{mIoU}_{\text{ins}}$. This

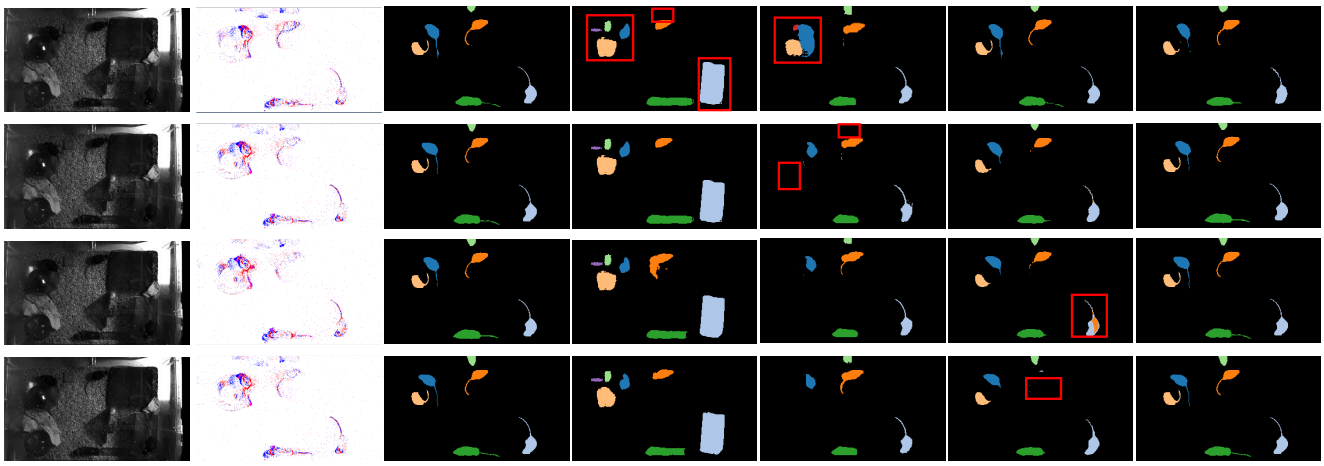

Figure 3. Visual comparisons show consecutive frames sampled from a video sequence of MouseSIS [15], arranged from top to bottom. Red boxes highlight regions with missed segmentation or unclear boundaries.

module explicitly separates appearance and motion information within both modalities, resulting in denser and more discriminative features after cross-modal fusion, which are particularly beneficial for small instance segmentation. By introducing intra-modal contrastive learning for explicit disentanglement, the network prevents the mutual interference between appearance and motion semantics and ensures that each branch focuses on its corresponding representation.

Effect of semantic and distributional alignment. Adding the semantic-level alignment further improves the performance to 69.23% $\mathbf{mIoU}_{\text{ins}}$, indicating that transforming corresponding appearance and motion features across modalities enhances fusion effectiveness. Moreover, enabling distribution alignment pushes the performance to 70.25% $\mathbf{mIoU}_{\text{ins}}$ since features extracted from image and event streams naturally follow different distributions that may degrade performance. The semantic reconstruction and adversarial domain adaptation jointly ensure that representations from different modalities remain coherent.

Table 4. Backbone ablation on MouseSIS dataset.

Backbone	Param.	FLOPs	$\mathbf{mIoU}_{\text{ins}}$ (%)
MobileNetV2 [38]	$\sim$ 3.4M	12.24G	68.62
ResNet18 [17]	$\sim$ 11.7M	20.10G	69.32
ResNet-50 [17]	$\sim$ 25.6M	60.42G	**70.25**

Effect of different encoder backbone. Table 4 shows the results using different backbones on MouseSIS, confirming that our disentangling and alignment modules effectively leverage the representational capacity of deeper networks. Importantly, using lightweight backbones such as MobileNetV2 [38] and ResNet-18 [17] results in only marginal drops of 1.63% and 0.93%, demonstrating the strong backbone-agnostic generalization of our framework. This shows that performance gains come from the proposed modules rather than large encoders alone. This is particularly advantageous given that our disentangling framework involves multiple encoder branches. By adopting dual lightweight backbone networks, we can reduce overall parameters while maintaining or even surpassing the performance of conventional methods that rely on a single, large-capacity backbone (e.g., dual MobileNetV2 with around 7.0M parameters vs. single ResNet-50 [17] with around 25.6M parameters), achieving a favorable performance–efficiency trade-off.

5. Conclusion

In this work, we addressed the challenge of small moving instance segmentation by proposing the DIMOS framework. Our method disentangles and extracts appearance and motion representations within each modality and aligns them through a multi-granularity cross-modal alignment strategy. This design enhances feature density and fusion effectiveness, leading to consistent improvements across multiple datasets. Experimental results demonstrate the effectiveness and robustness of our approach, particularly for small instances under challenging conditions.

Limitation. Our framework, like most multimodal segmentation systems, still relies on paired inputs. However, such synchronized dual-modality inputs are not always available. Existing methods often experience severe performance degradation or even fail completely in such single-modality settings. Therefore, enhancing the single-modality compatibility of multimodal systems represents an important and meaningful future research direction.

6. Acknowledgements

This work was partially supported by the Young Scientists Fund of the National Natural Science Foundation of China under Grant (62305278), as well as the Youth S&T Talent Support Program of GDSTA under Grant (SKXRC2025460). The authors gratefully acknowledge the financial support that made this research possible.

References

[1] Manideep Reddy Aliminati, Bharatesh Chakravarthi, Aayush Atul Verma, Arpitsinh Vaghela, Hua Wei, Xuesong Zhou, and Yezhou Yang. Sevd: Synthetic event-based vision dataset for ego and fixed traffic perception. *arXiv preprint arXiv:2404.10540*, 2024.

[2] Long Ang Lim and Hacer Yalim Keles. Foreground segmentation using a triplet convolutional neural network for multiscale feature encoding. *arXiv e-prints*, pages arXiv–1801, 2018.

[3] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021.

[4] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022.

[5] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European conference on computer vision*, pages 640–658. Springer, 2022.

[6] Gang Dai, Yifan Zhang, Qingfeng Wang, Qing Du, Zhuliang Yu, Zhuoman Liu, and Shuangping Huang. Disentangling writer and character styles for handwriting generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5977–5986, 2023.

[7] Gang Dai, Yifan Zhang, Quhui Ke, Qiangya Guo, and Shuangping Huang. One-dm: One-shot diffusion mimicker for handwritten text generation. In *European Conference on Computer Vision*, pages 410–427. Springer, 2024.

[8] Gang Dai, Qingfeng Wang, Yutao Qin, Gang Wei, and Shuangping Huang. Vg-sam: Visual in-context guided sam for universal medical image segmentation. *Fractal and Fractional*, 9(11):722, 2025.

[9] Gang Dai, Yifan Zhang, Yutao Qin, Qiangya Guo, Shuangping Huang, and Shuicheng Yan. Beyond isolated words: Diffusion brush for handwritten text-line generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19054–19064, 2025.

[10] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[11] Haotian Fu, Yue Zhou, Zhuo Zhang, Hongzhao Zheng, Renxu Yang, Yulong Huang, Dezhen Yang, Yannan Xing, Tugba Demirci, Ning Qiao, et al. Spikeram: A 48.1 pw/synapse/bit event-driven spiking compute-near/in-memory processor with neuromorphic sensor enabling lifelong on-chip learning. In *2026 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 314–316. IEEE, 2026.

[12] Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3867–3876, 2018.

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

[14] Dong Gong, Jie Yang, Lingqiao Liu, Yanning Zhang, Ian Reid, Chunhua Shen, Anton Van Den Hengel, and Qinfeng Shi. From motion blur to motion flow: A deep learning solution for removing heterogeneous motion blur. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2319–2328, 2017.

[15] Friedhelm Hamann, Hanxiong Li, Paul Mieske, Lars Lewejohann, and Guillermo Gallego. Mousesis: A frames-and-events dataset for space-time instance segmentation of mice. In *European Conference on Computer Vision*, pages 156–173. Springer, 2024.

[16] Friedhelm Hamann, Emil Mededovic, Fabian Gülhan, Yuli Wu, Johannes Stegmaier, Jing He, Yiqing Wang, Kexin Zhang, Lingling Li, Licheng Jiao, et al. Sis-challenge: Event-based spatio-temporal instance segmentation challenge at the cvpr 2025 event-based vision workshop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4675–4683, 2025.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] Hongxiang Huang, Xiaopeng Lin, Hongwei Ren, Yue Zhou, and Bojun Cheng. Exploring temporal dynamics in event-based eye tracker. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5145–5154, 2025.

[19] Yulong Huang, Xiaopeng Lin, Hongwei Ren, Haotian Fu, Yue Zhou, Zunchang Liu, Biao Pan, and Bojun Cheng. Clif: Complementary leaky integrate-and-fire neuron for spiking neural networks. In *International Conference on Machine Learning*, pages 19949–19972. PMLR, 2024.

[20] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[22] Zhe Kong, Le Li, Yong Zhang, Feng Gao, Shaoshu Yang, Tao Wang, Kaihao Zhang, Zhuoliang Kang, Xiaoming Wei, Guanying Chen, et al. Dam-vsr: Disentanglement of appearance and motion for video super-resolution. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–11, 2025.

[23] Hebei Li, Jin Wang, Jiahui Yuan, Yue Li, Wenming Weng, Yansong Peng, Yueyi Zhang, Zhiwei Xiong, and Xiaoyan Sun. Event-assisted low-light video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3250–3259, 2024.

[24] Hebei Li, Yansong Peng, Jiahui Yuan, Peixi Wu, Jin Wang, Yueyi Zhang, and Xiaoyan Sun. Efficient event-based semantic segmentation via exploiting frame-event fusion: A hybrid neural network approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18296–18304, 2025.

[25] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 dB 15 μs latency asynchronous temporal contrast vision sensor. *IEEE journal of solid-state circuits*, 43 (2):566–576, 2008.

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[27] Xiaopeng Lin, Yulong Huang, Hongwei Ren, Zunchang Liu, Hongxiang Huang, Yue Zhou, Haotian Fu, and Bojun Cheng. Clearsight: Human vision-inspired solutions for event-based motion deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7462–7471, 2025.

[28] Xiaopeng Lin, Hongwei Ren, Yulong Huang, Zunchang Liu, Yue Zhou, Haotian Fu, Biao Pan, and Bojun Cheng. Event-based motion deblurring via multi-temporal granularity fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

[29] Pengpeng Liu, Irwin King, Michael R Lyu, and Jia Xu. Ddflow: Learning optical flow with unlabeled data distillation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8770–8777, 2019.

[30] Anton Mitrokhin, Chengxi Ye, Cornelia Fermüller, Yiannis Aloimonos, and Tobi Delbruck. Ev-imo: Motion segmentation dataset and learning pipeline for event cameras. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6105–6112. IEEE, 2019.

[31] Anton Mitrokhin, Zhiyuan Hua, Cornelia Fermuller, and Yiannis Aloimonos. Learning visual motion segmentation using event surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14414–14423, 2020.

[32] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[33] Chenhao Pei, Fuping Wu, Liqin Huang, and Xiahai Zhuang. Disentangle domain features for cross-modality cardiac image segmentation. *Medical Image Analysis*, 71:102078, 2021.

[34] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170, 2017.

[35] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[36] Hongwei Ren, Zhuo Li, Aiersi Tuerhong, Haobo Liu, Fei Liang, Yongxiang Feng, Wenhui Wang, Yaoyuan Wang, Ziyang Zhang, Weihua He, et al. E2b: A single modality point-based tracker with event cameras. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6461–6468. IEEE, 2025.

[37] Hongwei Ren, Yue Zhou, Jiadong Zhu, Xiaopeng Lin, Haotian Fu, Yulong Huang, Yuetong Fang, Fei Ma, Hao Yu, and Bojun Cheng. Rethinking efficient and effective point-based networks for event camera classification and regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[38] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[39] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[40] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9481–9490, 2019.

[41] Zhexiong Wan, Bin Fan, Le Hui, Yuchao Dai, and Gim Hee Lee. Instance-level moving object segmentation from a single image with events. *International Journal of Computer Vision*, pages 1–22, 2025.

[42] Yingqian Wang, Longguang Wang, Gaochang Wu, Jungang Yang, Wei An, Jingyi Yu, and Yulan Guo. Disentangling light fields for super-resolution and disparity estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):425–443, 2022.

[43] Ziyun Wang, Jinyuan Guo, and Kostas Daniilidis. Unevimo: Unsupervised event-based independent motion segmentation. In *European Conference on Computer Vision*, pages 228–245. Springer, 2024.

[44] Zhixiang Wei, Lin Chen, Tao Tu, Pengyang Ling, Huaian Chen, and Yi Jin. Disentangle then parse: Night-time semantic segmentation with illumination disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21593–21603, 2023.

[45] Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. Seqformer: Sequential transformer for video instance

segmentation. In *European Conference on Computer Vision*, pages 553–569. Springer, 2022.

[46] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *European Conference on Computer Vision*, pages 588–605. Springer, 2022.

[47] Bochen Xie, Yongjian Deng, Zhanpeng Shao, and Youfu Li. Eisnet: A multi-modal fusion network for semantic segmentation with events and images. *IEEE Transactions on Multimedia*, 26:8639–8650, 2024.

[48] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *European Conference on Computer Vision*, pages 332–348. Springer, 2020.

[49] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by multi-scale foreground-background integration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4701–4712, 2021.

[50] Man Yao, Huanhuan Gao, Guangshe Zhao, Dingheng Wang, Yihan Lin, Zhaoxu Yang, and Guoqi Li. Temporal-wise attention spiking neural networks for event streams classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10221–10230, 2021.

[51] Bohan Yu, Jin Han, Boxin Shi, and Imari Sato. Eventpsr: Surface normal and reflectance estimation from photometric stereo using an event camera. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11427–11436, 2025.

[52] Yichen Yuan, Yifan Wang, Lijun Wang, Xiaoqi Zhao, Huchuan Lu, Yu Wang, Weibo Su, and Lei Zhang. Isomer: Isomerous transformer for zero-shot video object segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 966–976, 2023.

[53] Matthias Zeller, Vardeep S Sandhu, Benedikt Mersch, Jens Behley, Michael Heidingsfeld, and Cyrill Stachniss. Radar instance transformer: Reliable moving instance segmentation in sparse radar point clouds. *IEEE Transactions on Robotics*, 40:2357–2372, 2023.

[54] Bo Zhang and Jian Zhang. A traffic surveillance system for obtaining comprehensive information of the passing vehicles based on instance segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 22(11):7040–7055, 2020.

[55] Xiaoqi Zhao, Shijie Chang, Youwei Pang, Jiaxing Yang, Lihe Zhang, and Huchuan Lu. Adaptive multi-source predictor for zero-shot video object segmentation. *International Journal of Computer Vision*, 132(8):3232–3250, 2024.

[56] Tianfei Zhou, Jianwu Li, Shunzhou Wang, Ran Tao, and Jianbing Shen. Matnet: Motion-attentive transition network for zero-shot video object segmentation. *IEEE transactions on image processing*, 29:8326–8338, 2020.

[57] Yi Zhou, Guillermo Gallego, Xiuyuan Lu, Siqi Liu, and Shaojie Shen. Event-based motion segmentation with spatio-temporal graph cuts. *IEEE transactions on neural networks and learning systems*, 34(8):4868–4880, 2021.