Injecting Frame-Event Complementary Fusion into Diffusion for Optical Flow in Challenging Scenes

Haonan Wang[1], **Hanyu Zhou**[2*], **Haoyue Liu**[1], **Luxin Yan**[1]
[1]National Key Lab of Multispectral Information Intelligent Processing Technology,
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
[2]School of Computing, National University of Singapore
{whn_aurora,yanluxin}@hust.edu.cn, hy.zhou@nus.edu.sg

Abstract

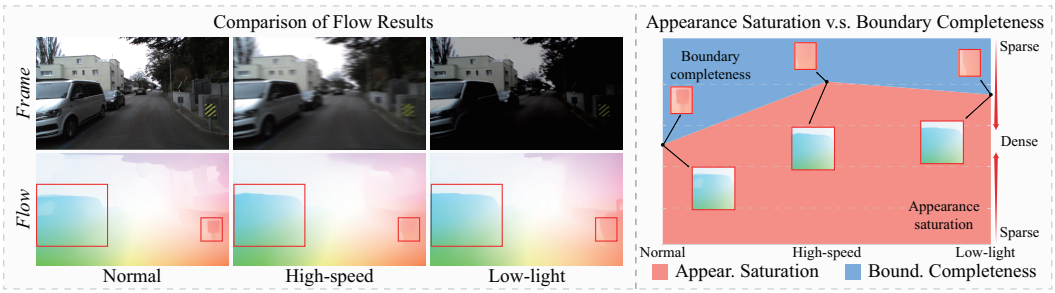
Optical flow estimation has achieved promising results in conventional scenes but faces challenges in high-speed and low-light scenes, which suffer from motion blur and insufficient illumination. These conditions lead to weakened texture and amplified noise and deteriorate the appearance saturation and boundary completeness of frame cameras, which are necessary for motion feature matching. In degraded scenes, the frame camera provides dense appearance saturation but sparse boundary completeness due to its long imaging time and low dynamic range. In contrast, the event camera offers sparse appearance saturation, while its short imaging time and high dynamic range gives rise to dense boundary completeness. Traditionally, existing methods utilize feature fusion or domain adaptation to introduce event to improve boundary completeness. However, the appearance features are still deteriorated, which severely affects the mostly adopted discriminative models that learn the mapping from visual features to motion fields and generative models that generate motion fields based on given visual features. So we introduce diffusion models that learn the mapping from noising flow to clear flow, which is not affected by the deteriorated visual features. Therefore, we propose a novel optical flow estimation framework Diff-ABFlow based on diffusion models with frame-event appearance-boundary fusion. Inspired by the appearance-boundary complementarity of frame and event, we propose an Attention-Guided Appearance-Boundary Fusion module to fuse frame and event. Based on diffusion models, we propose a Multi-Condition Iterative Denoising Decoder. Our proposed method can effectively utilize the respective advantages of frame and event, and shows great robustness to degraded input. In addition, we propose a dual-modal optical flow dataset for generalization experiments. Extensive experiments have verified the superiority of our proposed method. The code is released at https://github.com/Haonan-Wang-aurora/Diff-ABFlow.

1 Introduction

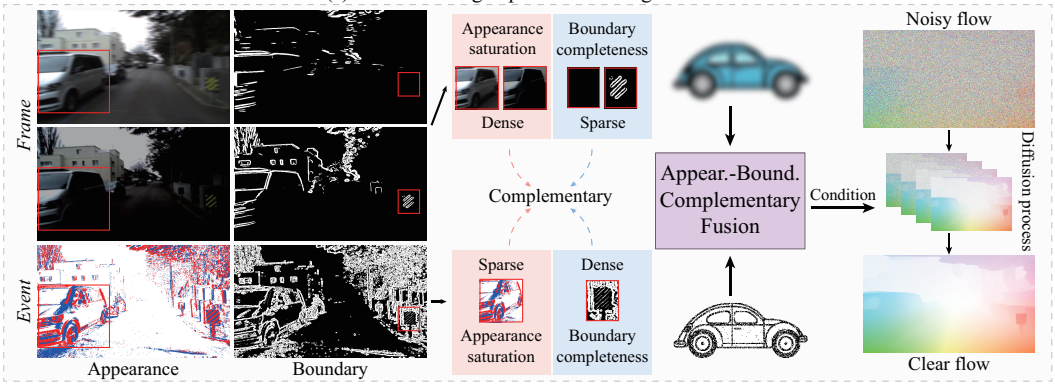
Optical flow estimation is a visual task that models pixel-wise displacements between adjacent frames. Existing methods [11, 17, 28] focus on conventional scenes, while challenging degraded scenes such as high-speed and low-light scenes remain to be further explored. The motion blur of high-speed scenes and the insufficient illumination of low-light scenes both lead to weakened texture and amplified noise. These degradations severely deteriorate visual features and violate the photometric consistency assumption, which further brings about invalid motion feature matching.

*Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



(a) Problem of high-speed and low-light scenes



(b) Illustration of our main idea

Figure 1: **Illustration of problem and idea.** Motion blur in high-speed scenes and insufficient illumination in low-light scenes reduce the boundary completeness of frame images, resulting in unclear boundary in optical flow. In this work, we explore the appearance-boundary complementarity of frame and event to guide the fusion of these two modalities. In addition, we introduce diffusion models to reconstruct the paradigm of optical flow estimation as a denoising process from noisy optical flow to clear optical flow conditioned on fused visual features.

Typically, existing methods adopt either uni-modal visual enhancement or dual-modal motion fusion to deal with high-speed and low-light conditions [8, 33, 38, 41]. The former, including deblurring and low-light enhancement improves the apparent visual effect, but the inner features remain deteriorated, contributing nothing to the photometric constancy assumption and motion feature matching. The latter utilizes event cameras to improve boundary completeness while the appearance features are still degraded and unqualified for motion feature matching. Specifically, appearance saturation refers to the abundance of appearance texture information within a visual modality. It reflects the degree of spatial variation in pixel intensity caused by fine-grained textures, shading, and color details. Boundary completeness denotes the continuity and integrity of object boundaries within a modality. It evaluates how well the modality captures clear, coherent, and complete boundary structures.

To solve these problems, we mainly explore in two aspects: sensors that can improve visual features and models that are robust to degraded input features. As shown in Fig. 1, on the one hand, we utilize the appearance-boundary complementarity of frame and event to obtain better visual features. The frame camera captures appearance with dense saturation, but due to its long imaging time and low dynamic range, the boundary captured under high-speed and low-light conditions shows sparse completeness. Thus, we introduce the event camera, a neuromorphic visual sensor [4], which offers dense boundary completeness because of its short imaging time and high dynamic range, despite its sparse appearance saturation. The appearance-boundary complementarity enables us to estimate optical flow with saturated appearance and complete boundary. On the other hand, we utilize the paradigm of diffusion models to adapt to degraded input features. Discriminative and generative models both rely on high-quality visual feature input. The former learns the mapping from visual features to motion fields, and the latter learns the process of generating motion fields based on given visual features. Both are severely affected by the degradation of visual features. Different from those two, the paradigm proposed by DDPM [9] models the denoising process from noisy data to clear data. We apply it to optical flow estimation and learn the mapping from noisy optical flow to clear optical flow, which demonstrates strong robustness to the degradation of visual features.

Based on these motivations, we propose Diff-ABFlow, a novel diffusion-based optical flow estimation framework guided by frame and event modalities. To exploit the appearance-boundary complementarity of frame and event, we propose an Attention-based Appearance-Boundary Fusion (Attention-ABF) module, which effectively combines the appearance feature of the frame and the boundary feature of the event to obtain fusion features with saturated appearance and complete boundary. Based on diffusion models, we propose a Multi-Condition Iterative Denoising Decoder (MC-IDD) as the optical flow backbone, including a Time-Visual-Motion Multi-way Cross-Attention (TVM-MCA) module and a Memory-GRU Denoising Decoder (MGDD) module. TVM-MCA integrates the fused visual feature, motion feature and temporal embedding to obtain the comprehensive feature including information from three aspects. MGDD is an optical flow inference module that combines the denoising paradigm proposed by DDIM and the iterative refinement method in optical flow estimation, which retains the generalization and robustness of diffusion models while improving efficiency. In summary, our contributions are as follows:

- We propose a novel framework, Diff-ABFlow, which leverages diffusion models with a frameevent complementary fusion strategy for accurate optical flow estimation in high-speed and low-light scenes. To the best of our knowledge, this is the first work that utilizes dual-modal data input to guide diffusion models for optical flow estimation.
- We propose the Attention-ABF module. Attention-ABF effectively utilizes the appearance-boundary complementarity of frame cameras and event cameras to obtain fusion features with high-quality appearance and boundary information.
- We propose the MC-IDD module. MC-IDD is an innovative optical flow backbone based on the DDIM paradigm and improved for optical flow estimation tasks, which combines visual features, motion features, and temporal embeddings to guide the denoising process.
- We conduct extensive experiments on both synthetic and real-world datasets to comprehensively demonstrate that our proposed Diff-ABFlow achieves state-of-the-art performance in optical flow estimation under high-speed and low-light conditions.

2 Related Work

Optical Flow Estimation. Optical flow estimation methods have developed rapidly with the advancement of deep neural networks. Earlier optical flow methods used a simple U-Net structure [3, 12]. Subsequent research gradually integrated modules such as feature pyramid and cost volume into optical flow estimation [23, 28, 35]. In addition, powerful techniques such as GRU [14, 29] and Transformer [11, 25, 34, 37] have been incorporated as the backbone for optical flow estimation. However, these frame-based methods often suffer from the motion blur in high-speed scenes and insufficient illumination in low-light scenes. In contrast, the event camera with short imaging time and high dynamic range captures high-quality visual signals especially in boundary areas. Event-based approaches [5, 7, 16, 21, 42] mainly follow the frame-based framework and reconstruct the event stream into event frame as input. In this work, we utilize the appearance-boundary complementarity of frame and event to obtain better visual features for optical flow estimation in degraded scenes.

Degraded Scenes Optical Flow. To deal with the motion blur of high-speed scenes and the insufficient illumination of low-light scenes, some researches directly perform visual enhancement such as deblurring and low-light enhancement. However, visual enhancement destroys the visual features and leads to invalid motion feature matching. Besides, a few methods use techniques such as feature fusion [8, 33] and domain adaptation [38, 39, 40, 41] to introduce event cameras to improve visual features. These approaches can indeed utilize event cameras to improve boundary completeness, but the appearance features provided by frame cameras are still degraded and unqualified for motion feature matching. The degradation of features severely affects discriminative models, which map from visual features to motion fields, and generative models, which generate motion fields given visual features. Therefore, we introduce diffusion models to reconstruct the optical flow estimation paradigm and reduce the impact of input visual feature degradation.

Diffusion Model. Diffusion models were originally proposed for image generation [9]. In contrast to previous discriminative and generative models, they model the denoising process from noisy samples to clear samples. The paradigm of diffusion models has been widely used in various fields

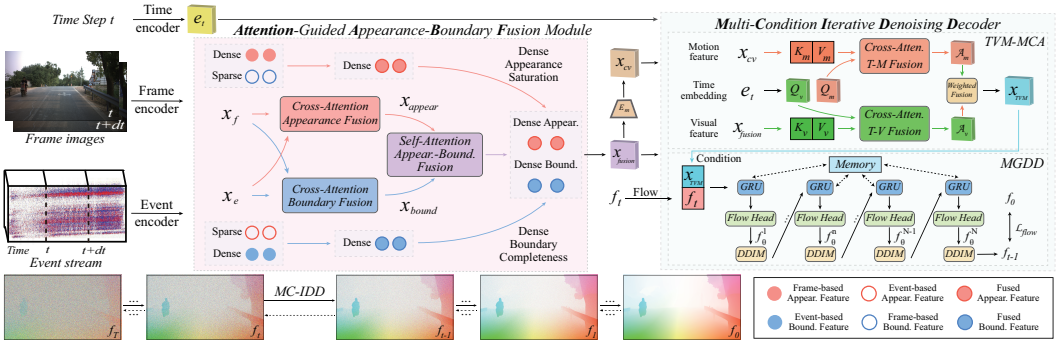


Figure 2: **Overall framework of Diff-ABFlow.** Diff-ABFlow mainly contains two parts: Attention-ABF for feature fusion and MC-IDD for denoising. In Attention-ABF, we utilize the appearance-boundary complementarity to fuse frame and event. In MC-IDD, we first integrate time embedding, visual feature and motion feature in the TVM-MCA module based on multi-way cross-attention mechanism. Then in MGDD, we input the comprehensive feature and the optical flow of the current time step into multiple GRUs with memory slots for iterative denoising. We repeatedly run MC-IDD a certain number of times on the noisy optical flow to obtain the clear optical flow.

of computer vision, such as semantic segmentation [1, 22, 30, 31, 36], depth estimation [15], trajectory prediction [13, 18]. The paradigm of these tasks has been reconstructed into the denoising process from noisy information to clear information with visual conditions. The practice in these fields has confirmed the strong robustness and generalization of diffusion models. Previous work has used diffusion models for optical flow estimation [17, 24], which reconstructed the task into a denoising process from noisy optical flow field to clear optical flow field, and achieved promising results. Therefore, we combine the appearance-boundary complementarity of frame and event, and the paradigm of diffusion models to propose a novel optical flow estimation framework that achieves state-of-the-art performance with strong generalization and robustness in degraded scenes.

3 Our Diff-ABFlow

3.1 Overall Framework

We propose a framework based on diffusion models with frame-event appearance-boundary fusion, which reformulates optical flow estimation as a denoising process from noisy flow to clear flow conditioned on frame and event. As shown in Fig. 2, the whole framework can be divided into two parts, one for feature fusion, and the other for iterative denoising based on multi-condition inputs. Based on the image pair and event stream, we extract the frame feature x_f and event feature x_e respectively, and input them into the Attention-ABF module to obtain the fused feature x_{fusion}, which is then used to construct a 4D cost volume x_{cv}. The time step t is encoded into the time embedding e_t through Sinusoidal Embedding [32] and MLP, and is then input into the TVM-MCA module together with the fused visual feature x_{fusion} and motion feature x_{cv} to obtain the enhanced feature x_{TVM}, which is finally input into the MGDD module together with the current optical flow f_t for iterative denoising. The MC-IDD module is repeatedly executed to obtain a clear optical flow.

3.2 Attention-Guided Appearance-Boundary Fusion Module

To verify the appearance-boundary complementarity of frame and event, we design an analysis experiment to study the complementarity between frame and event at the feature level. We use the Sobel operator to extract the boundary of frame and event respectively and use K-means clustering to analyze the distribution of appearance and boundary features of frame and event. As shown in Fig. 3, the frame provides dense appearance saturation and sparse boundary completeness, while the event is the opposite. This verifies the appearance-boundary complementarity of frame and event, which motivates us to design the Attention-Guided Appearance-Boundary Fusion Module.

In the Attention-ABF module, for the input frame features x_f and event features x_e, we first utilize appearance and boundary extractors to obtain appearance and boundary representations: $[x_{fa}, x_{fb}]$ and $[x_{ea}, x_{eb}]$. Then we obtain features with dense appearance saturation x_{appear} and features with

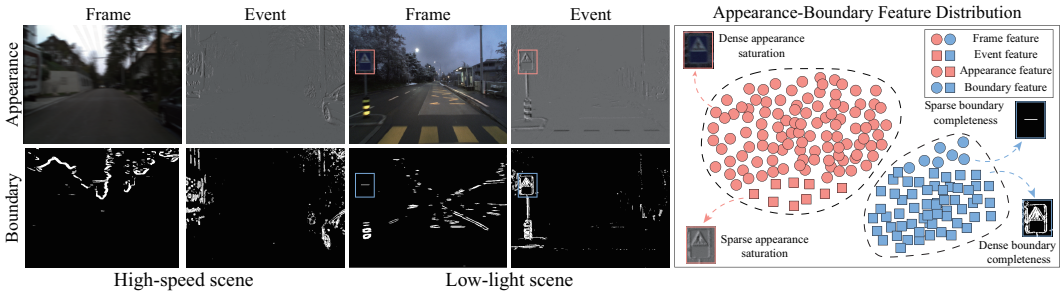


Figure 3: **Appearance-boundary feature distribution of frame and event in high-speed and low-light scenes.** We use K-means clustering to analyze the distribution of appearance and boundary features from frame and event features. The frame image has dense appearance saturation but sparse boundary completeness due to the motion blur of high-speed scenes and the insufficient illumination of low-light scenes. On the contrary, the event stream provides complete boundary in such degraded scenes while its appearance saturation is sparse. This motivates us to design a feature fusion module to fuse the two modalities utilizing the appearance-boundary complementarity.

dense boundary completeness x_{bound} based on two cross-attention modules:

$$x_{appear} = \mathrm{CAtten}(x_{fa}, x_{ea}), x_{bound} = \mathrm{CAtten}(x_{fb}, x_{eb}). \tag{1}$$

Subsequently, we utilize the self-attention mechanism to fuse appearance and boundary information from two features: $x_{fusion} = \mathrm{SAtten}(x_{appear}, x_{bound})$.

3.3 Multi-Condition Iterative Denoising Decoder

To intuitively demonstrate the superiority of diffusion models with deteriorated input features, we select a Transformer-based discriminative method and a GAN-based generative method to study the robustness to degraded inputs. Given degraded visual inputs, we utilize t-SNE to analyze the visual features and corresponding motion labels from three models. As shown in Fig. 4, both discriminative models and generative models have a certain degree of deviation when accepting degraded inputs, while the denoising process of diffusion models is almost unaffected, which motivates us to design an optical flow estimation backbone based on diffusion models, called MC-IDD.

MC-IDD utilizes the fused visual feature x_{fusion}, motion feature x_{cv} and time embedding e_t as conditions to denoise the optical flow field $\mathbf{f}_t$ at the current time step. MC-IDD includes two main parts, where TVM-MCA is used to integrate three conditions to obtain the feature x_{TVM} that contains temporal, visual, and motion information, while MGDD uses the comprehensive feature x_{TVM} to guide the denoising process based on GRU with memory slot.

Time-Visual-Motion Multi-Way Cross-Attention Module. The TVM-MCA module mainly uses two-way cross attention and gated fusion to effectively fuse time, vision, and motion features. Based on time embedding, we split it into visual query vector Q_v and motion query vector Q_m, each of which is fused with the visual features and motion features using cross-attention, to obtain time-visual attention features $\mathcal{A}_v$ and time-motion attention features $\mathcal{A}_m$:

$$\mathcal{A}_v = Softmax\left(\frac{Q_v K_v^T}{\sqrt{d}}\right) V_v, \mathcal{A}_m = Softmax\left(\frac{Q_m K_m^T}{\sqrt{d}}\right) V_m, \tag{2}$$

where K_v, V_v and K_m, V_m are obtained by flattening the visual feature x_{fusion} and the motion feature x_{cv}, and projecting them through linear layers, respectively. d denotes the number of feature dimensions. Then we utilize the learnable MLP to calculate the weights g of the two attention features and perform weighted fusion to obtain x_{TVM}:

$$x_{TVM} = g \cdot \mathcal{A}_v + (1 - g) \cdot \mathcal{A}_m, \tag{3}$$

which will be used to guide optical flow denoising later.

Memory-GRU Denoising Decoder Based on the paradigm of diffusion models, we first perform a forward diffusion process on the ground truth to obtain noisy optical flow, which gradually adds

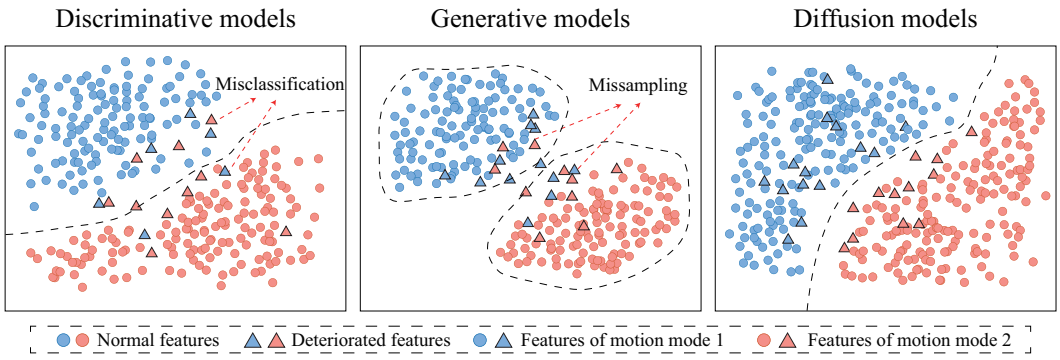


Figure 4: **t-SNE of visual features and corresponding motion labels from three different models.** Obviously, when inputting degraded features into those models, there exist misclassifications in discriminative models and missamplings in traditional generative models, while diffusion models demonstrate strong robustness to degraded inputs. This motivates us to introduce the paradigm of diffusion models and design a denoising decoding module for optical flow estimation.

Gaussian noise to the flow T times using a Markovian chain. The process is formulated as:

$$q(\mathbf{f}_t \mid \mathbf{f}_0) = \mathcal{N}\left(\mathbf{f}_t \mid \sqrt{\bar{\alpha}_t}\mathbf{f}_0, (1-\bar{\alpha}_t)\mathbf{I}\right), \quad t \in \{0, 1, \ldots, T\}, \tag{4}$$

where $\mathbf{f}_0$ indicates the ground truth of optical flow and $\mathbf{f}_t$ denotes the noisy flow. $\bar{\alpha}_t$ is defined as $\bar{\alpha}_t := \prod_{s=0}^{t} \alpha_s = \prod_{s=0}^{t}(1-\beta_s)$, where β_s is the pre-defined noise variance schedule, indicating the degree of Gaussian noise applied at each step.

When it comes to the reverse denoising process, our proposed MGDD module utilizes Gated Recurrent Unit (GRU) with a memory slot to iteratively denoise the optical flow $\mathbf{f}_t$. First, x_{TVM} and the stored memory are jointly input into GRU as conditions and are encoded into latent features together with the optical flow. The latent features are then used to update the memory slot and input into the flow head to obtain the coarse flow prediction $\mathbf{f}_\theta^n$. The memory slot is used to store hidden features in the current iteration, which helps retain feature details at each noise level. Then we follow the DDIM [26] paradigm to calculate the denoised optical flow as Eq. 5. After N iterations, the optical flow $\mathbf{f}_{t-1}$ of the next time step is obtained as:

$$\mathbf{f}_{t-1} = \sqrt{\alpha_{t-1}}\mathbf{f}_\theta^N + \sqrt{1-\alpha_{t-1}}\,\tilde{\boldsymbol{\epsilon}}_t, \tag{5}$$

where $\tilde{\epsilon}_t$ denotes the predicted noise at time step t: $\tilde{\boldsymbol{\epsilon}}_t = \frac{\mathbf{f}_t - \sqrt{\alpha_t}\,\mathbf{f}_\theta^N}{\sqrt{1-\alpha_t}}$. For the noisy optical flow $\mathbf{f}_T$, we run the MGDD module K times to obtain the denoising sequence $\{\mathbf{f}_T, \mathbf{f}_{T-1}, , \mathbf{f}_{T-K}\}$.

3.4 Optimization

For each prediction of optical flow in the denoising sequence $\{\mathbf{f}_T, \mathbf{f}_{T-1}, , \mathbf{f}_{T-K}\}$, we introduce three loss functions to supervise the learning of the network. To supervise the optical flow prediction with the ground-truth data, we adopt an L1 loss between the predicted flow $\hat{\mathbf{f}}$ and the ground-truth flow $\mathbf{f}_0$, which is formulated as:

$$\mathcal{L}_{flow} = \|\mathbf{f}_0 - \hat{\mathbf{f}}\|, \tag{6}$$

which directly minimizes the endpoint displacement error. Then we utilize the frame to add a smoothness loss with a boundary-aware term, which encourages the predicted optical flow to be spatially smooth while preserving the boundary of optical flow:

$$\mathcal{L}_{\text{smooth}} = \sum_{x,y}\left(\left|\nabla_x \hat{\mathbf{f}}(x,y)\right| \cdot e^{-\alpha|\nabla_x I(x,y)|} + \left|\nabla_y \hat{\mathbf{f}}(x,y)\right| \cdot e^{-\alpha|\nabla_y I(x,y)|}\right) \tag{7}$$

where $I(x,y)$ denotes the first input frame and α is the weight of boundary-aware term. Finally we utilize event data $E_t(x,y)$ to introduce an event consistency loss, which encourages the consistency between flow and event thus improving the accuracy of optical flow in the boundary area:

$$\mathcal{L}_{event} = \sum_{x,y} \|E_t(x,y) - E_{t+1}(x + \hat{\mathbf{f}}_u(x,y), y + \hat{\mathbf{f}}_v(x,y))\| \tag{8}$$

Table 1: Quantitative results on synthetic and real datasets, where VE denotes Visual Enhancement.

Method		Discriminative model							Generative model	
		GMA [14]		FF [25]	E-RAFT [7]	BFlow [8]		ABDA-Flow [38]	FD [17]	**Ours**
		w/o VE	w/. VE	-	-	w/o VE	w/. VE	-	-	-
Input		Frame		Frame	Event	Frame-event		Frame-event	Frame	Frame-event
HS-KITTI	EPE ↓	1.71	1.73	0.69	2.49	0.55	0.53	1.02	0.62	**0.46**
	Fl-all ↓	11.44	12.08	2.18	16.99	1.90	1.81	3.27	1.94	**1.12**
LL-KITTI	EPE ↓	1.98	1.83	0.71	3.08	0.68	0.69	0.64	0.67	**0.59**
	Fl-all ↓	12.36	11.76	2.85	19.21	2.54	2.53	2.46	2.43	**2.23**
HS-DSEC	EPE ↓	2.21	2.25	1.61	2.72	1.15	1.25	1.85	1.17	**1.09**
	Fl-all ↓	9.65	10.45	7.32	13.87	4.13	4.93	10.43	4.78	**3.83**
LL-DSEC	EPE ↓	2.43	2.41	1.70	3.49	1.73	1.76	1.62	1.69	**1.50**
	Fl-all ↓	12.78	12.02	9.65	18.56	6.48	6.97	5.69	6.03	**4.39**

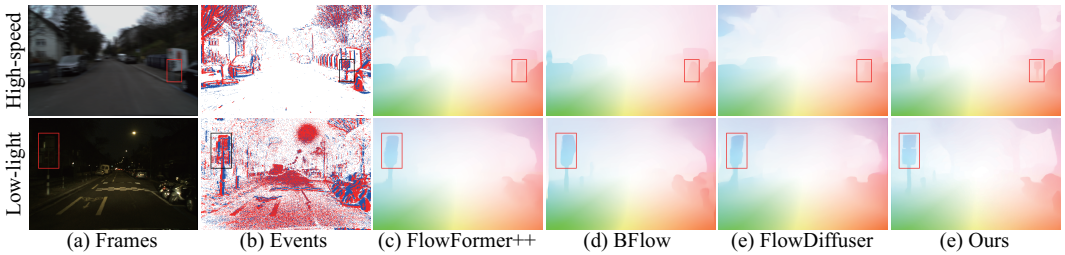


Figure 5: Visualization results on real high-speed and nighttime images of HS-DSEC and LL-DSEC.

where $\hat{\mathbf{f}}_u(x, y)$ and $\hat{\mathbf{f}}_v(x, y)$ respectively denotes the horizontal and vertical components of the flow. In summary, the total loss function is formulated as:

$$\mathcal{L} = \mathcal{L}_{flow} + \lambda_{smooth} \cdot \mathcal{L}_{smooth} + \lambda_{event} \cdot \mathcal{L}_{event}, \tag{9}$$

where λ_{smooth} and λ_{event} are the weights for corresponding losses.

4 Experiments

4.1 Experiment Setup

Dataset We conducted extensive experiments on both synthetic and real datasets. The synthetic datasets, HS-KITTI and LL-KITTI, are derived from the KITTI2015 dataset [20] by applying motion blur and low-light processing, respectively, where the v2e model [10] is used to generate corresponding event streams. The real datasets include HS-DSEC, obtained by applying motion blur to the DSEC dataset [6], and LL-DSEC, which consists of nighttime segments from the original DSEC dataset. In addition, we propose a **H**igh-**S**peed **F**rame-**E**vent **F**low **D**ataset (HS-FEFD) and a **L**ow-**L**ight **F**rame-**E**vent **F**low **D**ataset (LL-FEFD), which are collected by our custom-built frame-event co-optical axis imaging device in various scenes. Note that our proposed datasets are intended for generalization evaluation and are not used for training.

Implementation Details For model parameters, we set the diffusion step number T as 50 for forward diffusion following DDIM [26], the iterative decoding number N in the MGDD module as 6, and the denoising step number K as 4 for reverse denoising. During the training phase, we first pre-train the model on AutoFlow [27], FlyingChairs [3], FlyingThings [19], and MPI-Sintel [2]. Then we fine-tune it on the training sets of HS-KITTI, LL-KITTI, HS-DSEC, and LL-DSEC respectively. Finally, we conduct comparison and generalization experiments with the trained models on these datasets. All training and evaluation are performed on a single RTX 3090 GPU.

Table 2: Quantitative results on the proposed unseen HS-FEFD and LL-FEFD datasets.

Method		Discriminative model					Generative model	
		GMA [14]	FF [25]	E-RAFT [7]	BFlow [8]	ABDA-Flow [38]	FD [17]	**Ours**
Input		Frame	Frame	Event	Frame-event	Frame-event	Frame	Frame-event
HS-FEFD	EPE ↓	8.99	7.82	16.85	6.18	8.35	6.72	**4.69**
	Fl-all ↓	58.11	57.24	72.35	45.84	57.96	49.51	**28.77**
LL-FEFD	EPE ↓	11.07	9.41	15.79	7.53	6.90	7.45	**5.23**
	Fl-all ↓	65.82	59.36	75.89	55.73	43.31	52.04	**31.49**

(a) Frames (b) Events (c) FlowFormer++ (d) BFlow (e) FlowDiffuser (e) Ours

Figure 6: Visualization results on the proposed unseen HS-FEFD and LL-FEFD datasets.

Comparison Methods We select multiple methods with different input settings and paradigms for comparison. For methods based on frame, we choose GMA [14] and FlowFormer++ (FF) [25] that use discriminative models and FlowDiffuser (FD) [17] that uses generative models. For methods based on event, we choose E-RAFT [7] and for methods based on frame-event, BFlow [8] is chosen. These methods all use the same training process as ours to ensure fairness. In addition, we add deblurring and low-light enhancement to some of these methods to test the impact of visual enhancement methods on optical flow estimation. For evaluation, we choose End-Point-Error (EPE) and the percentage of erroneous pixels (Fl-all) as metrics for quantitative evaluation.

4.2 Comparison Experiments

Comparison on Synthetic Datasets. In Table 1, we list the evaluation metrics of the proposed method and all the comparison methods on synthetic datasets. Obviously, our proposed method significantly outperforms all competing methods with different input data and different paradigms. In addition, the visual enhancement method does not significantly improve the metrics of optical flow estimation, and sometimes even makes the results worse.

Comparison on Real Datasets. In Table 1 and Fig. 5, we compare the proposed method with competing methods in real high-speed and nighttime scenes. First, we can conclude that the method with frame-event input performs much better than those with only frame or event input, which confirms that the complementarity of frame and event is beneficial for optical flow estimation. Second, for the methods with the same frame input, the method using diffusion models is significantly better than those using discriminative models, which verifies the excellent performance of diffusion models. Finally, the metrics and the visualization results have demonstrated the superiority of our proposed method based on diffusion models with frame-event complementarity fusion.

Generalization for Unseen Datasets. In Fig. 6, we compare the generalization on our proposed datasets. Under unseen real high-speed and low-light conditions, the discriminative method with frame input fails to estimate optical flow while the frame-event method performs slightly better. The generative method performs well overall, but the blurry boundary still exist. On the contrary, our proposed method works well for both appearance and boundary, reflecting strong generalization.

4.3 Ablation Study

Effectiveness of Input Modality. In Table 3, we conduct an ablation study on the input modalities. Obviously, utilizing the complementarity of frame and event data significantly improves the accuracy

Table 3: Ablation study on modalities.

Modality	EPE ↓	FI-all ↓
Frame	0.58	1.93
Event	0.65	2.13
Frame+Event	**0.46**	**1.12**

Table 4: Discussion on fusion strategies.

Fusion Strategy	EPE ↓	Fl-all ↓
w/ Concatenating	0.57	1.84
w/ Weighting	0.55	1.79
w/ **Attention Guided**	**0.46**	**1.12**

Table 5: Discussion on flow estimation backbones.

Flow Backbone		EPE ↓	FI-all ↓
Discriminative	FlowFormer	0.59	1.89
Generative	GAN	0.65	2.97
	Diffusion Models	**0.46**	**1.12**

Table 6: Ablation experiments on proposed modules.

Attention-ABF	TVM-MCA	Memory Slot	EPE ↓	Fl-all ↓
✗	✗	✗	0.93	4.73
✗	✓	✗	0.65	2.59
✗	✗	✓	0.78	3.87
✗	✓	✓	0.59	2.09
✓	✗	✗	0.73	2.98
✓	✓	✗	0.54	1.63
✓	✗	✓	0.61	2.07
✓	✓	✓	0.46	1.12

Table 7: Discussion on diffusion settings.

Method		EPE ↓	Fl-all ↓	Inference Time/ms ↓
Diffusion Module	U-Net	0.63	2.75	97.5
	Conditional-RDD	0.57	1.74	**63.2**
	MGDD	**0.46**	**1.12**	64.6
Denoising Steps K	1	0.61	2.05	**37.5**
	2	0.52	1.61	46.4
	3	0.49	1.35	55.6
	4	**0.46**	**1.12**	64.6
	5	0.46	1.13	73.9

of the flow estimation results, achieving much better performance than using a single modality alone. This demonstrates that the two modalities provide mutually beneficial information and lead to more robust and precise flow estimation under challenging scenes.

Influence of Proposed Modules. In Table 6, we conduct ablation experiments on the proposed modules to reveal the effects of each module. The frame-event fusion module Attention-ABF plays the most important role in improving the results and TVM-MCA follows closely behind. The memory slot of GRU also makes a positive contribution.

4.4 Discussion

How does Feature Fusion Module work? In Fig. 7, in order to reveal the role of the feature fusion module Attention-ABF, we construct 4D cost volumes from frame features, event features, and fusion features respectively and analyze the response intensity histograms corresponding to different gradients to reflect the feature distribution in appearance and boundary areas. Moreover, we provide the flow results from the three cost volumes. On the one hand, the responses of the cost volumes from frame and event are concentrated in low-gradient and high-gradient intervals, i.e., the appearance and boundary regions, respectively, while the cost volume from the fusion features is evenly distributed in different gradient intervals. On the other hand, the flow inferred from frame has dense appearance saturation but sparse boundary completeness, and the flow inferred from event is the opposite, while the flow obtained by the fusion cost volume has dense appearance saturation and boundary completeness.This shows that our proposed fusion module effectively combines the appearance saturation and boundary completeness from frame and event.

Impact of Feature Fusion Strategies. In Table 4, we discuss the impact of various feature fusion strategies, including simple concatenation, weighted fusion, and our proposed attention-guided fusion. The results clearly demonstrate that the attention-guided fusion strategy significantly outperforms the other two. This superiority arises from its ability to effectively utilize the characteristics of frame and event features in appearance and boundary respectively.

Analysis on Optical Flow Backbone. In Table 5, we analyze the impact of different optical flow backbone architectures, including discriminative models (e.g., FlowFormer), traditional generative models (e.g., GAN-based methods), and the diffusion-based models adopted in our framework. From the results, we observe that discriminative and traditional generative models exhibit comparable performance, showing no significant differences. In contrast, diffusion models achieve substantially better accuracy and generalization in optical flow estimation.

Choices of Diffusion Settings. In Table 7, we conduct experiments to select the best diffusion module and denoising step. The results demonstrate that our proposed module MGDD outperforms other modules with similar inference time. In addition, the inference time increases linearly with the number of denoising steps. Thus, we set the denoising step number K to 4 to achieve the best possible results without causing excessively long inference time.

Table 8: Discussion on computational efficency.

Methods	Modules	Parameters(M)	Memory Consumption (GB)	Inference Time (ms)
FlowFormer++ [25]	Overall	17.6	13.2	141.2
BFlow [8]	Overall	5.9	10.9	148.7
FlowDiffuser [17]	Overall	16.3	15.4	186.9
Ours	Attention-ABF	4.5	3.9	52.3
	TVM-MCA	3.8	3.5	46.8
	MGDD	8.9	8.4	99.4
	Overall	17.2	15.8	198.5

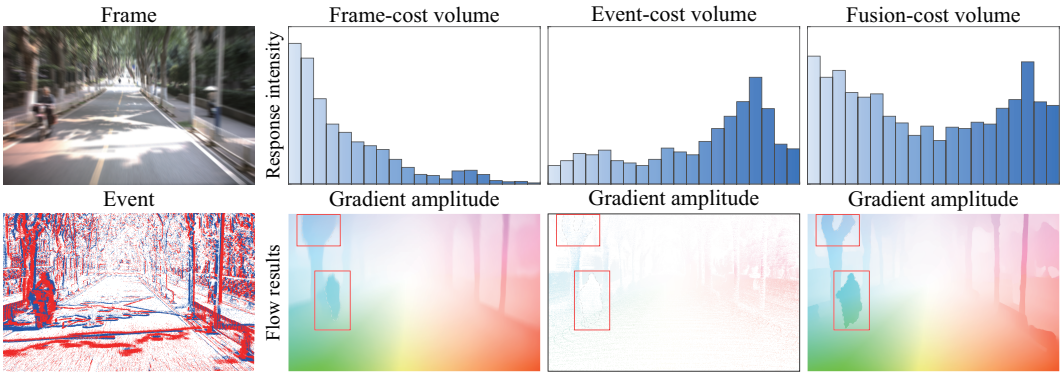


Figure 7: **Analysis on feature fusion module.** To analyze the role of the feature fusion module, we count the response intensity at different gradients of three cost volumes constructed from frame, event, and fusion features and compare the flow results from those cost volumes. The results demonstrate that the Attention-ABF module effectively utilizes the appearance-boundary complementarity of frame and event, and obtains fusion features with saturated appearance and complete boundary.

Discussion on Computational efficiency. For the purpose of verifying the computational efficiency of each module in our proposed method, we conduct some additional experiments on images from HS-DSEC and LL-DSEC with a resolution of 640×480 to test the number of parameters, memory consumption, and inference time of the modules and other optical flow methods, using a single RTX 3090 GPU as the inference platform. As shown in Table 8, the computational cost of each module in our proposed method remains within a reasonable range. Moreover, our approach achieves substantial performance gains with only a minor increase in computational cost.

Limitations Our proposed model performs well under both high-speed and low-light conditions, but fails to estimate optical flow when facing textureless planes. Neither frame nor event cameras can capture discriminative visual signals for the textureless planes since there exist no spatial brightness changes. In future work, we plan to incorporate another visual sensor, LiDAR, to perceive the distance from sensors to the planes thus obtaining the optical flow.

5 Conclusion

In this work, we propose a novel diffusion-based framework with event-frame appearance-boundary fusion for optical flow under both high-speed and low-light conditions. We are the first to utilize the paradigm of diffusion models with fused frame and event to solve the problem of optical flow in degraded scenes. We design the effective appearance-boundary fusion module Attention-ABF to lead the fusion of frame and event, taking advantage of their respective characteristics. In addition, we propose the innovative diffusion-based optical flow backbone MC-IDD, which aggregates information from multi-aspects including time step, visual features, and motion features, to guide the denoising process. Ours proposed method Diff-ABFlow achieves state-of-the-art performance far ahead previous methods. I believe that the multi-condition guided denoising diffusion paradigm we proposed can be used not only in the field of optical flow estimation, but also in many other fields of computer vision such as depth estimation and semantic segmentation.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Natural Science Foundation of China under Grant U24B20139. The computation is completed in the HPC Platform of Huazhong University of Science and Technology. We also thank the reviewers for their constructive comments and suggestions.

References

[1] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

[2] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, pages 611–625. Springer, 2012.

[3] 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.

[4] 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.

[5] 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.

[6] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021.

[7] Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza. E-raft: Dense optical flow from event cameras. In *2021 International Conference on 3D Vision (3DV)*, pages 197–206. IEEE, 2021.

[8] Mathias Gehrig, Manasi Muglikar, and Davide Scaramuzza. Dense continuous-time optical flow from event cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4736–4746, 2024.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[10] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1312–1321, 2021.

[11] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer: A transformer architecture for optical flow. In *European conference on computer vision*, pages 668–685. Springer, 2022.

[12] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.

[13] Chiyu Jiang, Andre Cornman, Cheolho Park, Benjamin Sapp, Yin Zhou, Dragomir Anguelov, et al. Motiondiffuser: Controllable multi-agent motion prediction using diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9644–9653, 2023.

[14] Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. Learning to estimate hidden motions with global motion aggregation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9772–9781, 2021.

[15] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024.

[16] Haotian Liu, Guang Chen, Sanqing Qu, Yanping Zhang, Zhijun Li, Alois Knoll, and Changjun Jiang. Tma: Temporal motion aggregation for event-based optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9685–9694, 2023.

[17] Ao Luo, Xin Li, Fan Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Flowdiffuser: Advancing optical flow estimation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19167–19176, 2024.

[18] Weibo Mao, Chenxin Xu, Qi Zhu, Siheng Chen, and Yanfeng Wang. Leapfrog diffusion model for stochastic trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5517–5526, 2023.

[19] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016.

[20] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3061–3070, 2015.

[21] Federico Paredes-Vallés and Guido CHE De Croon. Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3446–3455, 2021.

[22] Koutilya Pnvr, Bharat Singh, Pallabi Ghosh, Behjat Siddiquie, and David Jacobs. Ld-znet: A latent diffusion approach for text-based image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4157–4168, 2023.

[23] 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.

[24] Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, and David J Fleet. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. *Advances in Neural Information Processing Systems*, 36:39443–39469, 2023.

[25] Xiaoyu Shi, Zhaoyang Huang, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer++: Masked cost volume autoencoding for pretraining optical flow estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1599–1610, 2023.

[26] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[27] Deqing Sun, Daniel Vlasic, Charles Herrmann, Varun Jampani, Michael Krainin, Huiwen Chang, Ramin Zabih, William T Freeman, and Ce Liu. Autoflow: Learning a better training set for optical flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10093–10102, 2021.

[28] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018.

[29] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.

[30] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3554–3563, 2024.

[31] Aysim Toker, Marvin Eisenberger, Daniel Cremers, and Laura Leal-Taixé. Satsynth: Augmenting image-mask pairs through diffusion models for aerial semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27695–27705, 2024.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[33] Zhexiong Wan, Yuchao Dai, and Yuxin Mao. Learning dense and continuous optical flow from an event camera. *IEEE Transactions on Image Processing*, 31:7237–7251, 2022.

[34] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8121–8130, 2022.

[35] Jia Xu, René Ranftl, and Vladlen Koltun. Accurate optical flow via direct cost volume processing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1289–1297, 2017.

[36] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023.

[37] Shiyu Zhao, Long Zhao, Zhixing Zhang, Enyu Zhou, and Dimitris Metaxas. Global matching with overlapping attention for optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17592–17601, 2022.

[38] Hanyu Zhou, Yi Chang, Haoyue Liu, Wending Yan, Yuxing Duan, Zhiwei Shi, and Luxin Yan. Exploring the common appearance-boundary adaptation for nighttime optical flow. In *The Twelfth International Conference on Learning Representations*.

[39] Hanyu Zhou, Yi Chang, and Zhiwei Shi. Bring event into rgb and lidar: Hierarchical visual-motion fusion for scene flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26477–26486, 2024.

[40] Hanyu Zhou, Yi Chang, Zhiwei Shi, Wending Yan, Gang Chen, Yonghong Tian, and Luxin Yan. Adverse weather optical flow: Cumulative homogeneous-heterogeneous adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[41] Hanyu Zhou, Haonan Wang, Haoyue Liu, Yuxing Duan, Yi Chang, and Luxin Yan. Bridge frame and event: Common spatiotemporal fusion for high-dynamic scene optical flow. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27904–27913, 2025.

[42] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Ev-flownet: Self-supervised optical flow estimation for event-based cameras. *arXiv preprint arXiv:1802.06898*, 2018.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We claim the contributions of this work at the end of the Introduction and elaborate how we achieve them in the Method section. Moreover, we provide sufficient experimental results in the Experiments section.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in the 9th page.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We provide the full set of assumptions and proof for each theoretical result in the Method section.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide complete model implementation details, experimental settings, and training data description in Experiments section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will provide the code project once the paper is accepted.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have presented the detailed experimental settings in the Experiments section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We use the statistical significance in the evaluation metrics.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We describe the computer resources in the Experiments section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: There is no concern about ethics involved in this work.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no concern about broader impacts involved in this work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: There is no concern about safeguards involved in this work.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We credit all works cited including dataset, code and model in the references.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Our paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The paper does not involve the usage of LLMs in the core methods

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.