FlowFM: Advancing Dark Optical Flow Estimation with Flow Matching

Anonymous CVPR submission

Paper ID

1. Feature Encoding

1.1. Feature and Context Backbone

We follow a similar design to the well-known RAFT [22], incorporating a feature backbone and a context feature backbone. The latter is introduced to extract the original features at 1/8 resolution, mapping $\mathrm{F}_{1\&2} \in \mathbb{R}^{3\times H\times W}$ to $\mathbf{x}' \in \mathbb{R}^{C\times H/8\times W/8}$ where we set C is set to 256. Both the feature and context backbones consist of 6 residual blocks: 2 at 1/2 resolution, 2 at 1/4 resolution, and 2 at 1/8 resolution. Furthermore, the structure of the context backbone is identical to that of the feature backbone, except that BatchNorm regularization is used in the context branch and InstanceNorm in the feature backbone, respectively. **The code will be made publicly available upon the paper's acceptance, and we kindly ask for your understanding.**

1.2. Gated Attention Encoder

In DOFE computation, dark degradations critically undermine the accuracy of long-range feature matching. In particular, the inherent sinks of the attention mechanism may cause key features to be overlooked in a single frame during the feature affinity process. This phenomenon becomes especially pronounced in conditions characterized by noise and darkness. To counteract this, we introduce a novel gated attention encoder (GAE) that synergistically combines a gated modulation module (GMM) and a masked feature encoder (MFE) to achieve more discriminative modeling of large motions. To effectively ensure feature interaction, our GMM first modulates MFE to capture spatially coherent features, background, and object details. Currently, our MFE uses a gated strategy to capture fine-grained features, and h distinct k values to dynamically control the magnitude of mask sparsity, enabling parallel encoding of salient top-k global features while suppressing noise and interference.

Gated Modulation Module. To begin with, GMM is designed to generate multiple components, including offset priors for basic feature enhancement, and spatial and channel convolutional attention that provide dynamic properties for long-term information aggregation. Given the basic feature $\mathbf{x} \in \mathbb{R}^{C\times H/8\times W/8}$ from the feature backbone, the GMM is,

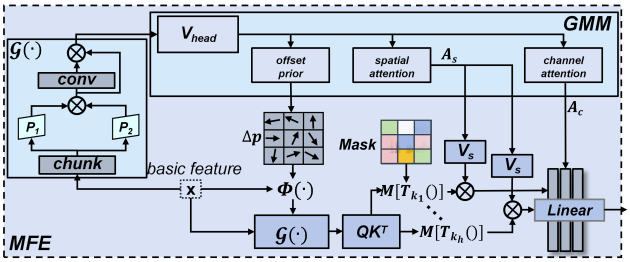

Figure 1. Overview of the proposed Gated Attention Encoder.

$$\begin{aligned} \mathrm{V} &= \mathrm{V}_{\mathrm{head}}(\mathcal{G}(\mathbf{x}')), \Delta p = r \cdot f_{\mathrm{offsets}}(\mathbf{V}), \\ A_s &= f_{\mathrm{spatial}}(\mathbf{V}), A_c = f_{\mathrm{channel}}(\mathbf{V}), \end{aligned} \tag{1}$$

where $\mathcal{G}(\cdot)$ denotes our proposed gated block. Practically, input features are split into two equal-sized chunks, with matrix multiplication computed between them to derive attention weights. As shown in **Fig.1**, a 3×3 convolution block is then introduced to extract fine-grained features, followed by fusing the convolved result with the attention weights. This strategy anchors attention weight calculation on local and global interactions of input features, rather than over-binding to fixed feature positions, mitigating attention concentration on a single sink position. $\mathrm{V}_{\mathrm{head}}(\cdot)$ projects the feature $\mathbf{x}$ to the Value $\mathbf{V}$. $\Delta p \in \mathbb{R}^{2\times h\times w}$ denotes the offset that warps the basic characteristic $\mathbf{x}$ to enhance the expression of the object. r is a scalar that controls the offset range via the tanh function. Besides, $f_{\mathrm{spatial}}(\cdot)$ denotes spatial attention implemented with a function $mean$ and convolutions, while $f_{\mathrm{channel}}(\cdot)$ is channel attention that includes avgpool, sigmoid, and convolutions.

Masked Feature Encoder. After that, the offset Δp is integrated into the basic feature $\mathbf{x}$ using bilinear interpolation $\phi(\cdot)$, enabling the MFE to achieve more flexible and content-adaptive learning,

$$\mathbf{x}_\Delta = \phi(\mathbf{x}', \Delta p), \tag{2}$$

After that, we project $\mathbf{x}_\Delta$ and $\mathbf{x}$ through a depth-wise layer $\mathrm{QK}_{\mathrm{head}}(\cdot)$ to obtain the query and key maps. In the

Table 1. EPE comparisons by switch components within RAFT and GMA, and FlowFM frameworks (trained on FCDN).

Task	Method	Para.(M)	EPE↓	
			FCDN	VBOF
Feature	RAFT	5.26	1.23	21.84
	+DBME	7.61	1.14	21.43
	+Def.	7.67	1.17	21.52
	+SpE	8.28	1.15	21.40
	+GAE	7.53	1.08	20.95
Context	GMA	5.80	1.18	21.77
	+DBME	8.21	1.14	20.75
	+Trans.	6.56	1.17	21.23
	+SpE	8.93	1.14	20.82
	+MCE	8.21	1.12	20.69
Exchange	FlowFM	12.22	0.87	13.28
	all GAE	11.93	0.92	13.79
	all MCE	12.64	0.96	13.92
	GAE ⇌ MCE	12.22	0.94	13.87
	FlowFM	12.22	**0.87**	**13.28**

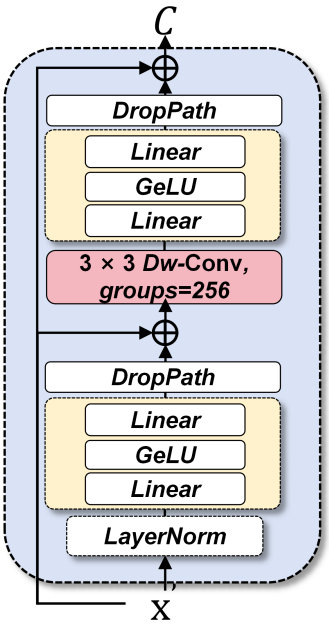

Figure 2. Overview of the proposed MLP-based Context Encoder.

encoding phase, treating all pixel features equally has long been regarded as unsuitable and inefficient; a refinement strategy is applied to QK^T using the top-k function and mask matrix. The process is as follows,

$$\begin{aligned} \mathrm{Q},\mathrm{K} &= \mathrm{QK}_{\mathrm{head}}(\mathcal{G}(\mathrm{C}[\mathbf{x}_\Delta,\mathbf{x}'])); \mathrm{V}_s = \mathrm{V}\cdot A_s, \\ \mathbf{x}'_h &= \mathrm{softmax}(\mathcal{M}[\mathcal{T}_{k_h}(\mathrm{QK}^T/\sqrt{d})])\cdot \mathrm{V}_s, \\ \mathbf{x} &= f_{\mathrm{linear}}(\mathrm{C}[\mathbf{x}'_1,...,\mathbf{x}'_h])\cdot A_c, \end{aligned} \tag{3}$$

where $\mathcal{T}_k(\cdot)$ performs a top-k selection, identifying the significant feature pairs of k. A mask matrix $\mathcal{M}\in\mathbb{R}^{C_m\times C_m}$ sets the scores of non-top-k elements to negative infinity, thereby filtering out deceptive affinities caused by interference. Yet, relying on a single top-k may erroneously shield important features. Consequently, we adopt h distinct top-k values to dynamically adjust the mask sparsity magnitude in parallel, denoted as $\mathcal{T}_{k_h}(\cdot)$. Notably, h is not multi-head, but the number of parallel branches with distinct top-k settings, and using these top-k selections effectively suppresses stubborn interference and reduces motion asymmetry-related uncertainty. Lastly, $f_{\mathrm{linear}}(\cdot)$ is a linear block that serves as an adjustment strategy, recalibrating the importance of encoded channel-wise expressions through attention A_c.

1.3. MLP-based Context Encoder

After the context backbone computation, we introduce an MLP-based context encoder (MCE) for reliable scene and background motion analysis. As illustrated in **Fig.2**, we first use a LayerNorm layer and a Linear module to refine low-level contextual representations. These features are then fed into a depth-wise convolution layer to precisely highlight spatial consistency and make them more expressive, precisely. Subsequently, the features are fed into another MLP module to capture global features and improve the understanding of contextual information. The resulting outputs from the MCE can facilitate global connections and interactions, capturing holistic scene patterns crucial for better performance.

1.4. Comparison with Different Encoders.

Task Feature. We conducted a comparison among our Gated Attention Encoder (GAE), dual-branch motion encoder (DBME) [31], deformable attention methods (Def.) [30], and sparse attention encoder (SpE) [32] within the RAFT framework. As shown in **Tab. 1** (Task Feature), the RAFT equipped with GAE delivers outstanding EPE values of 1.08 and 20.95 on the FCDN and VBOF datasets, markedly outperforming recent methods like DBME by 5.26% and 2.24%; and Def. by 7.69% and 2.65% margins. This is because DBME's effectiveness is constrained by its reliance on a rigid convolutional space, rendering it susceptible to neighboring noise and visual degradation due to the lack of robustness to long-range pixel connections. Def. is hindered by its oversimplified object modeling, resulting in ineffective global understanding and inadequate interference suppression in dark scenarios. Although SpE incorporates an anti-interference mechanism, the inherent flaw of attention sinking induced by darkness hinders the extraction of more effective motion-related information. In contrast, our GAE starts with interference suppression, dynamic perception, and mitigating attention sinking, which significantly enhances feature learning in dark scenarios.

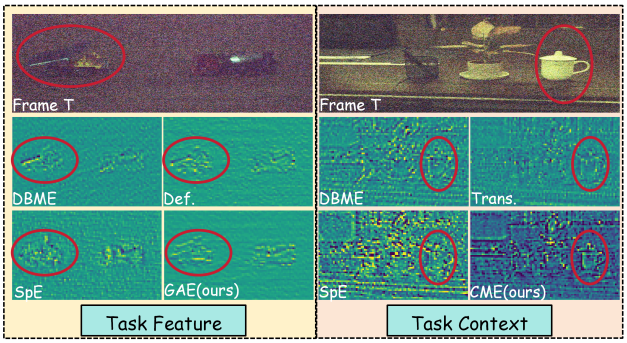

Figure 3. Visualization of the feature maps of different components. Two sets of input frames and corresponding feature maps at different stages are depicted in the feature pipeline and the context pipeline. Zoom in for details.

Meanwhile, our GAE, with its integrated MFE, offers fewer parameters than DBME, Def., and SpE while providing substantial improvements in the global feature encoding performance of DOFE tasks.

Task Context. We compare our MLP-based context encoder (MCE) with the DBME, SpE, and Transformers (Trans) blocks [25] on a common baseline GMA [9] to ensure fair evaluation. As indicated in **Tab.** 1 (Task Context), our MCE's contextual representations with linear layers realize notable enhancements, reducing the EPE value from 1.18 to 1.12 on FCDN and from 21.77 to 20.69 on VBOF. Compared with DBME and SpE methods, our MCE is specifically designed for contextual feature learning. It enables cross-dimensional context extraction and dynamic enhancement through the interaction between MLP-based transformations in the channel dimension and depthwise convolutions in the spatial dimension.

Task Exchange. Under Task Exchange, both GAE and MCE modules, are employed following the backbone layer. It is evident that there is a significant performance drop, and the two modules exhibit distinct strengths: in contrast to MCE, which excels at extracting more robust contextual information for scene understanding through cross-dimensional interaction; our GAE demonstrates superior capability in perceiving the underlying feature representations in the dark. Finally, the interchangeability of GAE and MCE proves to be unsatisfactory, further confirming that both modules are task-oriented by design and play a crucial role in DOFE.

Visual Experiment. We provide a visualization of features between task features and task contexts. As shown in **Fig.** 3 (Task Feature), the red boxes highlight that the car produced by GAE exhibits smoother features and more complete target information, whereas the car's details from other methods are more distorted. In **Fig.** 3 (Task Context), our MCE enables feature extraction that retains richer scene information, with more distinct cup contours and appearance details within the red-boxed regions.

Table 2. EPE comparisons by switch different components within GAE and MCE (trained on FCDN).

Task	Method	Para.(M)	EPE↓	
			FCDN	VBOF
	FlowFM	**12.22**	**0.87**	**13.28**
GAE	w/o $\mathcal{G}(\cdot)$	11.99	0.91	13.68
	w/o GMM	11.27	0.93	13.95
	w/o MFE	11.08	0.93	14.16
MCE	w/o MLP	10.44	0.92	14.35
	w/o Dw-Conv	11.59	0.91	14.10
	w/o Dropout	12.22	0.88	13.40
	w/o Norm	12.22	0.94	14.08

Table 3. EPE comparisons by adjust parameters h and k in MFE (trained on FCDN). C denotes the channel dimensions.

h	k	EPE↓	
		FCDN	VBOF
0	-	0.95	13.80
1	$\frac{C}{2}$	0.90	13.77
2	$\frac{C}{2}, \frac{2C}{3}$	0.88	13.60
3	$\frac{C}{2}, \frac{2C}{3}, \frac{3C}{4}$	0.87	13.45
4	$\frac{C}{2}, \frac{2C}{3}, \frac{3C}{4}, \frac{4C}{5}$	**0.87**	**13.28**
5	$\frac{C}{2}, \frac{2C}{3}, \frac{3C}{4}, \frac{4C}{5}, \frac{5C}{6}$	0.92	13.63

1.5. Ablation on GAE and MCE modules.

To validate the effectiveness of GAE, we performed an ablation experiment by systematically removing its components. As detailed in row 1 of **Tab.** 2, we eliminated the gated strategy $\mathcal{G}(\cdot)$ and directly used the basic features for value projection, which led to an increase of 4.60% in EPE (0.87 $\rightarrow$ 0.91) in FCDN and of 3.01% in EPE (13.28 $\rightarrow$ 13.68) in VBOF. Subsequently, we removed GMM and replaced MFE with standard self-attention, as indicated in rows 2 and 3. These experiments demonstrate that GMM and MFE play a crucial role in useful feature induction and important long-range information aggregation.

In the 4th to 7th rows of **Tab.** 2, we validated the effectiveness of each component within the MCE module, respectively. Specifically, the experimental results show that removing any single component, including MLP, Dw-Conv, Dropout, and Norm, leads to noticeable performance degradation across key metrics. The corresponding EPE scores show a significant decrease from the full-model, which fully demonstrates that each component of our proposed MCE is indispensable for the DOFE task, as each contributes uniquely to the module's overall performance.

1.6. Analysis on Parameters h and k.

As presented in **Tab.** 3, we empirically assessed the impact of parameter adjustments on performance by tuning h and

Table 4. Ablation analysis for other attention methods on our model (trained on FCDN).

Method	EPE↓	
	FCDN	VBOF
Gating Unite	0.97	14.59
Soft Attention	1.03	14.94
Top-k	**0.87**	**13.28**

Table 5. EPE comparisons by switch different components within IFDD (trained on FCDN).

Task	Method	Para.(M)	EPE↓	
			FCDN	VBOF
	FlowFM	**12.22**	**0.87**	**13.28**
IFDD	w/o GRU unite	10.82	1.02	14.81
	w/o FMR(+CGA)	11.57	0.92	13.95
	w/o FMR(+CDD)	14.24	0.90	13.66
FMR	SAM ⇌ FreEnc	12.22	0.87	13.40
	w/o SAM(Attn)	13.46	0.87	13.26
	w/o FreEnc(wavelet)	12.75	1.14	15.29
	w/o FreEnc(SpDe.)	13.18	0.95	14.57

k within the MFE. When the top-k settings were adjusted to $\frac{C}{2}$, $\frac{2C}{3}$, $\frac{3C}{4}$ and $\frac{4C}{5}$, the EPE in FCDN improved from 0.88 to 0.85, and in VBOF from 13.79 to 13.28, indicating a gradual performance improvement. Nevertheless, continuing to increase top-k led to a decline in performance. This phenomenon arises because a larger top-k value induces more redundant affinities and ambiguities, which fail to provide reliable guidance for refining global motion.

1.7. Analysis on Different Attentions.

As shown in **Tab.** 4, we performed an ablation study on alternative suppression methods within the MFE module. Specifically, we substituted the top-k strategy module of our MFE with a gating unit [1] and a soft attention unit [12]. Notably, these alternative methods lack dedicated interference suppression mechanisms, resulting in subpar performance relative to our top-k masking approach when applied to the challenging DOFE task. This further confirms the superiority of our top-k masking strategy in targeted interference suppression for DOFE.

2. Motion Denoising Decoder

2.1. Implicit Fourier Denoising Decoder.

In our FlowFM framework, our Implicit Fourier Denoising Decoder (IFDD) is equipped with gated recurrent units (GRUs) for accurate motion hidden state inference. To further verify the effectiveness of IFDD, we conducted additional ablation experiments in the supplementary materials,

Figure 4. Visualization of the dark optical flow between IFDD and FMR components.

as shown in the **Tab.**5. To begin with, we removed the GRU unit and directly input the noisy flow into the FMR module for motion decoding. Clearly, performance decreased significantly. This shows that the latent state inference strategy of GRU is crucial for exploring degenerate motion patterns. Subsequently, we adopted contour-guided attention (CGA) [32] and conditional denoising decoder (CDD) [16] to replace our FMR module while retaining the GRU module. Nonetheless, CGA strengthens motion boundaries via long-range feature connections guided by contour priors, yet its limited noise robustness prevents effective suppression of noise present in dark environments. CDD adapts to noise and data distribution through a simple convolutional network, while FMR substantially enhances object discriminability in the frequency domain by leveraging the spatial positional relationships of the target, thus achieving state-of-the-art performance in DOFE tasks. These results demonstrate that our IFDD is currently the most well-suited motion denoising decoder for noisy flow.

Fig. 4 presents a set of optical flow comparisons. It can be seen that the performance of 'w/o GRU unit' is the worst, demonstrating that GRU plays a crucial role in hidden state inference for dark optical flow estimation. Meanwhile, 'w/o FMR(+CGA)' and 'w/o FMR(+CDD)' can improve motion discriminability, but there remains a significant gap compared with our proposed IFDD method.

2.2. Fourier Motion Refactor

As a composite module integrating spatial and frequency-domain optimization, the Fourier Motion Refactor (FMR) processes the input features as follows. The input feature map undergoes LayerNorm2d normalization, 1×1 depth-wise convolution to extract refined feature, then is weighted by the Spatial Attention Module (SAM) to emphasize key spatial regions; the resulting weight fused with the original input, normalized again, fed into Fourier-based Frequency Enhancer (FreEnc) for frequency-domain enhancement, and finally residual-fused with the intermediate feature to yield the deeply enhanced output.

2.2.1. Spatial Attention Module.

SAM is a spatial-channel modulation module, comprising adaptive avgpool that compresses $H \times W$ to 1×1 to capture per-channel global context, and a 1×1 convolution for channel-wise attention weight generation (shape (B, c, 1, 1)). These weights are element-wise multiplied with original spatial features via broadcasting to achieve dynamic spatial importance modulation.

2.2.2. Fourier-based Frequency Enhancer.

The FreEnc serves as the core component of the proposed FMR method. In practice, we first apply real-valued Fourier transform on the input feature map to separate its magnitude and phase components; then optimize the magnitude via a sequential module consisting of 1×1 convolutions and LeakyReLU activation. While retaining the original phase, the model retains the original phase, then reconstructs complex values by multiplying the optimized magnitude with trigonometric functions (cosine for the real part, sine for the imaginary part). This is the transformation of complex numbers from **polar coordinates** to Cartesian coordinates (amplitude + phase $\rightarrow$ real part + imaginary part). Lastly, the optimized "real part + imaginary part" is recombined into a complex frequency domain signal using the tensor function (*e.g.*, torch.complex (real, imag) in PyTorch), which is then fed into an inverse Fourier transform to map the feature back to the spatial domain for output.

2.2.3. Analysis on SAM and FreEnc component.

As shown in **Tab.** 5, the effectiveness of the spatial attention module and fourier-based frequency enhancer components in the Fourier motion refactor module is further validated.

In row 4th of **Tab.** 5, swapping the order of the SAM and FreEnc resulted in suboptimal performance. In dark optical flow estimation, performing the SAM first to focus on valid motion features and target regions in the spatial domain enables the FreEnc to conduct a targeted frequency-domain refinement of key information afterward, while avoiding the synchronous amplification of noise and invalid data that occurs when FreEnc is applied first, thus achieving superior performance. As shown in row 5th of **Tab.** 5, replacing SAM with a more complex self-attention module added 1.1M parameters but only yielded a 0.02 EPE improvement on VBOF, providing minimal gain for model practicality.

Additionally, we adopted the wavelet-based decomposition and enhancement method [8] and the spatial frequency-based enhancement method [31] to replace our proposed FreEnc. Obviously, they exhibit varying degrees of performance degradation. The wavelet method does not provide features that anchor the target's spatial positional relationships, such as magnitude (mag), resulting in inadequate modeling of complex motion patterns with spatial components under challenging low-light conditions. Essentially, in contrast to our Fourier-based method, the spatial frequency method primarily operates on distinct attribute-specific features within the spatial (image) domain, whereas our Fourier-based method transforms image data into the frequency domain to decompose and manipulate signal components based on their frequency characteristics. By contrast, the Fourier method offers unparalleled efficiency in isolating global frequency components (*e.g.*, low-frequency smooth regions, and high-frequency texture details), thereby facilitating tasks, such as large-scale noise suppression and motion patterns prediction. As shown in row 7 of **Tab.** 5, the spatial frequency-based method is confined to image-domain processing.

Fig. 4 presents the optical flow results of the FMR ablation experiments. The wavelet-based method fails to perform adequately for dark optical flow estimation, leading to distortions in large-scale optical flow fields. By contrast, other methods perform better, yet still exhibit limitations compared to our proposed FMR method. This demonstrates that FMR provides a pioneering frequency-domain approach to dark optical flow estimation, marking an advancement over spatial-domain-only processing.

3. Why Do Neural Models with Explicit Flow Field Regression Perform Better Than Conventional Velocity Field Regression?

Today, diffusion and flow matching models in practice commonly predict noise or noised quantities (*e.g.*, velocity field v). Given this paradigm, less attention has been paid to what the optical flow network should directly predict, with the implicit assumption that the network is capable of performing the assigned task. However, the roles of clean optical flow and noisy quantities (including noise itself) are not equal. Optical flow is subject to strong constraints from physical laws, such as object motion, occlusion, and geometric principles, thereby giving rise to inherent structural properties such as smoothness and consistency. Importantly, the superposition of noise ε onto degraded features and unreliable motion affinities can introduce a domain shift, causing the learned transformation path to deviate from the ideal velocity field trajectory. Intuitively, in dark optical flow estimation, predicting clean optical flow versus estimating noise magnitudes exhibits distinct levels of complexity, whereas directly predicting optical flow represents an inherently more straightforward endeavor.

According to the hypothesis proposed in [3, 4] that "*(high-dimensional) data lie (roughly) in a low-dimensional manifold*", optical flow-though residing in a high-dimensional pixel space-is essentially distributed over such a low-dimensional "manifold". By contrast, noise is purely a random perturbation in the high-dimensional space, devoid of any adherence to manifold structures. Consequently, as a mixture of image-derived signals and noise,

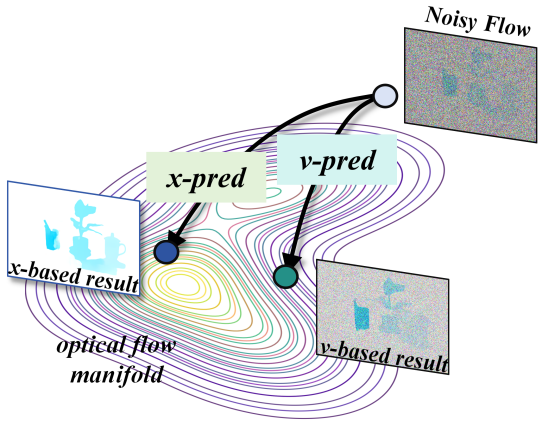

Figure 5. The **Manifold Assumption** [4] posits that natural images lie on a low-dimensional manifold embedded within the high-dimensional pixel space. In the flow matching paradigm, an optical flow x can be modeled as lying on the manifold, whereas the velocity field v (*e.g.*, $v = gt\text{-}\epsilon$) is inherently off-manifold. Training a neural network to predict a clean optical flow (*i.e.*, x-prediction) yields fundamentally distinct performance relative to training it to predict noisy quantities (*e.g.*, ϵ, v-prediction).

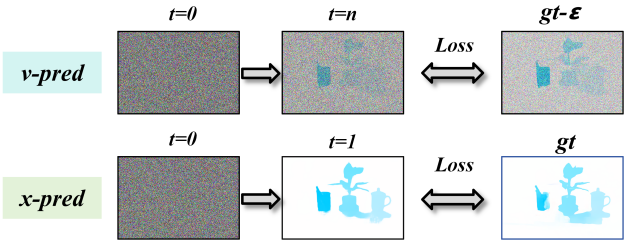

Figure 6. Predictions exhibit distinct discrepancies across different training pipelines. Specifically, v-pred imposes a highly volatile constraint term during training, compelling the model to adapt to unstructured noise in the high-dimensional space, which is incompatible with the inherent structure attributes of the dark optical flow. As with our proposed FlowFM, directly learning the optical flow field (x-pred) while conducting denoising enables the model to comprehensively capture the intrinsic characteristics of the data.

the velocity field in the flow matching paradigm also falls outside the bounds of this "manifold".

As illustrated in **Fig. 5**, in the flow matching paradigm, we provide two pipelines: one is x, or optical flow prediction (x-prediction); the other is velocity prediction (v-prediction). It can be seen that it is more effective to allocate the model to directly model optical flow data than to learn the noise distribution. Therefore, having a neural network predict a target on the manifold (clean optical flow) and a target outside the manifold (noise or flow velocity) are two tasks with completely different levels of difficulty. We believe that predicting clean optical flow is essentially simpler because it allows the network to focus on learning the intrinsic low-dimensional structure of the data, while naturally filtering out high-dimensional noise and dark degeneration. As shown in **Fig. 6**, in our FlowFM, by minimizing ground truth distance using loss function $\mathcal{L}_1$, enforces the learned vector field to satisfy the intrinsic low-dimensional structural representation of optical flow, while naturally filtering out high-dimensional noise and suppressing interference.

To further demonstrate the effectiveness of our method for directly predicting optical flow sample, we evaluated the performance of different flow matching strategies on FCDN and VBOF datasets. **Tab. 6** and **Tab. 7** present the EPE metrics on the FCDN and Mix datasets, respectively. It is evident that directly predicting optical flow outperforms predicting high-dimensional noise and mixed data, as low-dimensional optical flow data embodies more prominent structural regularity and smoothness constraints—both of which are more conducive to dark optical flow.

4. Additional Experiment

4.1. Experiment Setup

Datasets. Similar to previous successful work [31, 32], we trained on all FCDN data [28], testing the model's fitting ability on FCDN and generalization on VBOF [26]. For Mix (FCDN + VBOF) dataset, we used all for training, tested fitting on both datasets, and provided generalization visualizations on real-world dark image/video datasets (FLIR ADAS [7], SDSD [23], GOF [11]). We use the end-point error (EPE) [6] as an evaluation metric to assess the predicted dark optical flow.

Implementation Details The proposed FlowFM was implemented on the PyTorch platform, utilizing an Intel (R) i5 2.59GHz CPU and an NVIDIA RTX 4080 GPU. During training, we followed the previous work CEDFlow [31] and CEDFlow++ [32], equipping our FlowFM with the AdamW optimizer and clipping gradients in the range of [-1, 1]. We used the one-cycle learning rate schedule to minimize loss using the distance $\mathcal{L} - 1$. All models were initialized from scratch with random weights, and the resolution of the clipped input frames was set to 368×496.

For fair comparison, 350k iterations were run on FCDN dataset and 400k on the Mix dataset during training. The batch size was set to 4 (training phase) and 1 (testing phase), while the learning rate was 2.5e-4 for both FCDN and Mix dataset, respectively. For large models like FlowDiffuser [16] and Flowformer++ [19], we reduced the batch size under hardware constraints but increased training steps by 150K for fairness. The pre-trained encoders were re-trained in FCDN.

Comparison Methods. To our knowledge, few studies have tackled the challenging DOFE. We evaluated FlowFM against 15 top performing methods from the Sintel [2] and KITTI [17] rankings, including PWC-Net [21], RAFT,

Table 6. Analysis on different flow matching strategies evaluations on FCDN and VBOF datasets (trained on FCDN).

Method	Reference	EPE(trained on FCDN)↓									
		FCDN	Canon	Fuji	Fuji2	Nikon	Nikon2	Sony	Sony2	Sony3	VBOF
w/o FM(Flow Matching)	-	1.23	20.76	37.58	14.20	26.54	28.77	16.46	32.97	19.52	20.47
w/o v-pred (**main paper's Eq.11**)	-	1.07	14.25	28.99	11.54	21.37	22.70	12.47	20.90	14.86	16.77
FlowFM(x-pred)	-	**0.87**	**12.09**	**23.96**	**8.99**	**17.38**	**16.59**	**9.82**	**16.65**	**11.97**	**13.28**

Table 7. Analysis on different flow matching strategies evaluations on FCDN and VBOF datasets (trained on Mix).

Method	Reference	EPE(trained on Mix)↓									
		FCDN	Canon	Fuji	Fuji2	Nikon	Nikon2	Sony	Sony2	Sony3	VBOF
w/o FM(Flow Matching)	-	1.27	6.26	8.25	4.74	8.68	8.12	4.34	5.77	6.15	6.10
w/o v-pred (**main paper's Eq.11**)	-	1.10	5.97	8.15	3.96	7.21	8.72	4.14	5.40	6.03	5.82
FlowFM(x-pred)	-	**1.06**	**5.83**	**7.03**	**3.30**	**6.75**	**7.85**	**3.74**	**4.37**	**5.98**	**5.24**

GMA, SCV [10], Flow1D [24], GyroFlow [11], GMFlowNet [27], GMFlow [25], AGFlow, KPAFlow [13], C-RAFT [20], GAFlow [15], Flowformer++, FlowDiffuser, and DPFlow [18]. And DOFE-oriented CEDFlow [31] and CEDFlow++ [32], excluding code-unreleased ABDA-Flow [29] and multi-modal [5].

4.2. Comparison with State-of-the-Arts

Trained on FCDN. As illustrated in **Tab. 8**, we provide additional quantitative evaluations on 17 state-of-the-art methods. The proposed FlowFM achieves an EPE score of 0.87 on the FCDN dataset. It outperforms well-known methods including GMFlow, C-RAFT, and GAFlow and even exceeds the recent CEDFlow by 19.4% (1.08 → 0.87) and CEDFlow++ by 13.8% (1.01 → 0.87). From columns 4 to 12 of **Tab. 8**, generalization evaluations on Canon, Fuji, Fuji2, Nikon, Nikon2, Sony, Sony2, Sony3, and VBOF (all parts) show that our FlowFM achieves remarkable robustness in dark scenes. These results demonstrate that FlowFM outperforms other state-of-the-art models in effectively tackling DOFE tasks, thanks to its masked global encoder and graph-based motion reconstruction.

Trained on Mix. Tab. 9 depicts a quantitative evaluation using FlowFM trained on the Mix (FCDN + VBOF) dataset. Compared with existing SOTA methods, FlowFM achieves significantly better performance, with superior EPE scores on VBOF and its subsets. While optical flow methods generally exhibit strong performance, most advanced approaches still face notable challenges under substantial illumination variations, with AGFlow and GAFlow yielding particularly poor results.

4.3. Why Do Models Trained on Mix Yield Higher EPE on FCDN Than FCDN-Only?

Notably, models trained on the more complex Mix dataset typically yield higher EPE scores on the FCDN dataset than those trained solely on the FCDN. This is because the noise distribution of FCDN differs from that of VBOF, which VBOF incorporates camera-specific sensor noise, dark noise, and color distortion. During Mix training, the models must adapt to these two distinct noise modes, which weakens their specialization in the FCDN noise pattern. Additionally, the ground truth of the VBOF is estimated from ideally exposed images rather than directly rendered optical flow, thus introducing minor inaccuracies.

The two key aspects are summarized:

Experiment 1 (noise distribution isolation evaluation): We randomly selected 100 illumination-ideal image pairs from FCDN. The Mix (original) consists of this FCDN subset and 100 illumination-ideal image pairs from VBOF. Mix (noise corrected) comprises the FCDN subset and 100 illumination-ideal image pairs from VBOF without dark noise and color distortion injection. In the Mix training phase, we used the same batch size and learning rate as those in the original manuscript, with a total of 100k epochs. As shwn in Tab.10, it can be observed that the EPE values of Mix (noise corrected) are lower and closer to those of FCDN, which indirectly confirms that noise distribution mismatch is the key factor causing performance degradation.

Experiment 2 (flow label accuracy analysis): To verify the impact of the accuracy of VBOF's optical flow labels on FCDN performance, we trained the model using 'FCDN+VBOF' with corrected optical flow labels (where the labels were corrected using the SOTA method DPFlow [18]). Due to constraints, we only tested 100 image pairs from Experiment 1 to observe whether the performance on the FCDN dataset was restored, thus verifying the impact of VBOF label bias on model training. As shown in **Fig. 7**, we also supplemented the visualization results for reference. The vast majority of optical flow labels in VBOF are accurate; we only identified a small number of images with

Table 8. EPE comparison of different optical flow evaluations on FCDN and VBOF datasets (trained on FCDN dataset). Bold indicates best performance, while underlining denotes second rank.

Method	Reference	EPE(trained on FCDN)↓									
		FCDN	Canon	Fuji	Fuji2	Nikon	Nikon2	Sony	Sony2	Sony3	VBOF
PWC-Net[21]	CVPR-18	1.81	24.10	54.38	15.77	33.65	31.57	18.49	38.71	22.83	26.77
RAFT[22]	ECCV-20	1.23	20.25	39.52	14.20	26.40	25.45	15.61	32.43	19.16	21.84
GMA[9]	CVPR-21	1.18	20.37	39.64	14.40	27.54	26.07	15.83	30.23	19.04	21.77
SCV[10]	CVPR-21	1.29	22.47	42.08	14.96	29.43	27.50	17.29	37.95	22.01	24.13
Flow1D[24]	ICCV-21	1.22	21.01	49.33	14.25	28.93	27.48	16.72	37.63	21.86	21.79
GyroFlow[11]	ICCV-21	1.79	23.38	50.07	15.52	31.54	28.82	17.10	39.53	22.37	24.76
GMFlowNet[27]	CVPR-22	1.56	21.80	41.78	14.51	26.76	26.71	15.52	34.77	20.03	22.87
GMFlow[25]	CVPR-22	1.18	19.79	39.83	14.56	27.96	25.30	15.29	30.33	18.85	22.72
AGFlow[14]	AAAI-22	1.15	20.13	39.79	14.16	26.30	25.15	16.31	30.67	18.78	21.05
KPAFlow[13]	CVPR-22	1.24	22.19	47.21	14.71	29.97	27.58	16.15	36.65	21.61	23.10
C-RAFT[20]	CVPR-22	1.20	21.55	39.78	14.49	27.55	26.24	15.46	30.91	19.72	22.04
GAFlow[15]	ICCV-23	1.10	19.53	37.31	13.95	25.37	24.70	15.21	30.37	19.13	20.84
Flowformer++[19]	CVPR-23	1.12	20.09	39.34	14.03	25.49	25.46	15.10	30.32	18.87	20.94
FlowDiffuser[16]	CVPR-24	1.09	19.83	36.74	14.00	25.32	24.26	15.17	30.57	19.14	20.77
CEDFlow[31]	AAAI-24	1.08	19.69	39.21	13.94	25.24	25.09	15.16	30.30	18.73	20.89
DPFlow[18]	CVPR-25	1.17	19.88	37.20	14.32	26.07	25.56	15.20	27.73	19.29	20.92
CEDFlow++[32]	IJCV-25	1.01	20.07	34.82	13.93	25.85	24.34	15.05	28.58	19.00	20.56
FlowFM(Ours)	-	**0.87**	**12.09**	**23.96**	**8.99**	**17.38**	**16.59**	**9.82**	**16.65**	**11.97**	**13.28**

Table 9. EPE comparison of different flow evaluations on FCDN and VBOF datasets (trained on Mix dataset). Bold indicates best performance, while underlining denotes second rank.

Method	Reference	EPE(trained on Mix)↓									
		FCDN	Canon	Fuji	Fuji2	Nikon	Nikon2	Sony	Sony2	Sony3	VBOF
PWC-Net[21]	CVPR-18	1.92	10.64	18.48	8.06	13.45	13.77	7.97	13.15	12.92	9.90
RAFT[22]	ECCV-20	1.38	8.25	16.77	7.34	11.43	12.06	6.00	11.56	11.29	8.89
GMA[9]	CVPR-21	1.26	6.33	10.11	4.90	11.56	11.96	5.16	7.04	6.63	6.81
SCV[10]	CVPR-21	1.27	7.41	15.78	6.48	11.79	12.37	5.84	11.83	11.21	7.76
Flow1D[24]	ICCV-21	1.30	6.60	10.34	5.13	10.93	13.44	6.07	7.63	7.19	6.93
GyroFlow[11]	ICCV-21	1.83	8.77	16.49	7.92	12.32	12.75	7.17	10.44	12.36	9.40
GMFlowNet[27]	CVPR-22	1.70	9.09	16.16	7.71	12.74	12.82	7.26	10.71	11.20	8.66
GMFlow[25]	CVPR-22	1.31	6.26	10.69	5.76	10.90	10.76	5.69	6.74	6.88	7.23
AGFlow[14]	AAAI-22	1.27	6.25	10.34	4.97	8.61	10.01	5.15	6.51	6.76	6.75
KPAFlow[13]	CVPR-22	1.39	7.36	12.70	6.11	10.14	10.51	6.22	8.71	8.37	7.47
C-RAFT[20]	CVPR-22	1.27	6.87	10.79	5.45	9.74	10.21	5.93	6.88	7.02	7.04
GAFlow[15]	ICCV-23	1.26	6.27	9.87	4.76	8.24	9.90	6.03	6.72	6.84	6.79
Flowformer++[19]	CVPR-23	1.24	6.22	9.73	4.77	8.38	9.72	4.64	6.74	6.82	6.46
FlowDiffuser[16]	CVPR-24	1.20	6.38	9.74	4.40	9.32	9.63	5.11	6.57	6.89	6.57
CEDFlow[31]	AAAI-24	1.23	6.22	9.41	4.69	8.23	9.87	4.66	6.50	6.10	6.52
DPFlow[18]	CVPR-25	1.24	6.31	9.55	4.56	7.80	9.50	4.74	6.23	6.43	6.21
CEDFlow++[32]	IJCV-25	1.14	5.85	9.66	4.38	7.45	9.29	4.63	6.18	6.77	6.25
FlowFM(Ours)	-	**1.06**	**5.83**	**7.03**	**3.30**	**6.75**	**7.85**	**3.74**	**4.37**	**5.98**	**5.24**

minor label errors, which slightly interfere with the model's learning of motion patterns.

Notably, due to the small test sample size, we do not intend to include these results in the main paper. However, they can indirectly confirm that the 'synthesis-to-real domain shift' resulting from Mix training is the main factor that causes the performance degradation on the FCDN dataset. Going forward, we will substantially refine the VBOF labels.

4.4. Visual Comparison with State-of-the-Arts

Trained on FCDN. Fig. 8 presents a qualitative comparison of representative methods on the challenging FCDN dataset. Obviously, in **Fig. 8**, our FlowFM demonstrates advantages in terms of high optical flow accuracy (closest to the ground truth in shape and position within the red-circled area) and strong effectiveness in low-light FCDN scenarios, significantly outperforming methods such as CEDFlow, DPFlow, and CEDFlow++. **Fig. 9** provides a generalization performance evaluation on the challenging VBOF

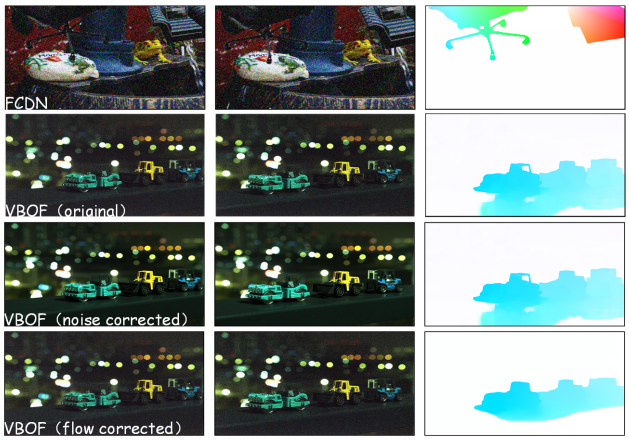

Figure 7. Flow label analysis of Mix dataset domain shift.

Table 10. Noise distribution analysis of Mix dataset domain shift.

Method	EPE(trained on FCDN or Mix)			
	FCDN	Canon	Nikon	VBOF
FCDN	0.85	11.79	16.28	13.15
Mix(original)	1.01	4.34	6.46	5.33
Mix(noise corrected)	0.99	4.32	6.45	5.27
Mix(flow corrected)	1.00	4.34	6.45	5.29

dataset (featuring heavy noise and dark environments); in this dataset, the optical flow results of FlowDiffuser, CEDFlow, and CEDFlow++ exhibit blurred regions, unclear boundaries, and insufficient restoration of shape details for background and car contours. By contrast, our FlowFM delineates much sharper contours and scenes that are highly consistent with the ground truth, thanks to our proposed specific methods, *e.g.*, flow matching with explicit flow field regression and the frequency-domain enhancement.

Trained on Mix. Fig. 10 presents a qualitative comparison of representative methods on the FCDN dataset. All of these methods are trained on the Mix dataset. It can be seen that our FlowFM achieves clear and accurate optical flow results for the legs of the bench and the outlines of the chair within the purple box, significantly outperforming methods such as CEDFlow, DPFlow, and CEDFlow++. **Fig. 11** presents an evaluation of the challenging VBOF dataset with strong heavy noise; in this scene, the flow results of FlowDiffuser, CEDFlow, and CEDFlow++ for background and car contours exhibit blurred regions, unclear boundaries, and insufficient restoration of shape details. In contrast, FlowFM delineates much sharper contours and scenes that are highly consistent with the ground truth.

Real-world flow evaluation. To further explore FlowFM's generalization and robustness in real-world dark scenarios. **Fig. 12** shows its optical flow results on the FLIR ADAS [7], SDSD [23], and GOF [11] datasets. The FLIR ADAS dataset encompasses a variety of night driving scenarios with rapid object motion, profound darkness, and multiple light sources. The SDSD comprises 150 high-resolution spatially aligned video pairs that show the same scenes in both low and normal lighting. The GOF contains real-world scenes in 4 different categories with synchronized gyro readings, including a regular scene and three challenging cases such as low-light, foggy, and rainy scenes. **Despite the fact that FLIR ADAS, SDSD, and GOF lack optical flow ground truth**, FlowFM exhibits superior performance in these real-world conditions characterized by fast-moving objects, dynamic scenes, and varied lighting.

To be specific, **Fig. 12** presents three challenging dark scenarios: GOF (dark, low-illuminated, blurry, and rapidly moving targets), where motion targets lack significant discriminative attributes, FLIR ADAS (dark, noisy with multi-light sources, with dense shadows); and SDSD (dark, with scarce textures and prominent noise). In these challenging cases, Flowformer++, FlowDiffuser, CEDFlow, and DPFlow all produce severely blurred optical flow results. For example, vehicle contours are severely blurred in GOF, and details of the circular ring on the table disappear in SDSD. These challenges pose a significant issue, stemming from poor feature discriminability and low visibility conditions. CEDFlow++ shows partial effectiveness but has fragmented details (*e.g.*, fractured vehicle edges in GOF, blurred and distorted shapes in SDSD). Our FlowFM, leveraging the noise robustness of the flow matching strategy and implicit Fourier-based enhancement, retains flow field details more comprehensively and outperforms CEDFlow and CEDFlow++ in the majority of scenarios.

4.5. Limitation Analysis

In particular, our approach has limitations in the dark. As shown in **Fig. 13**, in dark and high-dynamic range scenes that contain large same-attribute motion, extremely small targets, such as GOF data where discriminative features are almost absent (*e.g.*, DID's bottom region), accurate motion estimation is impaired by a critical visual cue deficit. Future work will inject priors or multi-modal data to alleviate these extreme limitations, and adapt the FlowFM module to other vision tasks with higher efficiency and accuracy.

References

[1] Wasi Uddin Ahmad, Nanyun Peng, and Kai-Wei Chang. Gate: graph attention transformer encoder for cross-lingual relation and event extraction. In *AAAI*, pages 12462–12470, 2021. 4

[2] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, pages 611–625. Springer, 2012. 6

[3] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009. 5

[4] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009. 5, 6

[5] Weichen Dai, Hexing Wu, Xiaoyang Weng, Yuxin Zheng, Yuhang Ming, and Wanzeng Kong. Multi-modal synergistic implicit image enhancement for efficient optical flow estimation. In *CVPR*, pages 2173–2182, 2025. 7

[6] 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 *ICCV*, pages 2758–2766, 2015. 6

[7] F. A. Group. Flir thermal dataset for algorithm training. In *https://www.flir.in/oem/adas/adas-dataset-form/*, 2018. 6, 9

[8] Hai Jiang, Ao Luo, Haoqiang Fan, Songchen Han, and Shuaicheng Liu. Low-light image enhancement with wavelet-based diffusion models. *ACM TOG*, 42(6):1–14, 2023. 5

[9] Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. Learning to estimate hidden motions with global motion aggregation. In *ICCV*, pages 9772–9781, 2021. 3, 8

[10] Shihao Jiang, Yao Lu, Hongdong Li, and Richard Hartley. Learning optical flow from a few matches. In *CVPR*, pages 16592–16600, 2021. 7, 8

[11] Haipeng Li, Kunming Luo, and Shuaicheng Liu. Gyroflow: Gyroscope-guided unsupervised optical flow learning. In *ICCV*, pages 12869–12878, 2021. 6, 7, 8, 9

[12] Jiachen Lu, Jinghan Yao, Junge Zhang, Xiatian Zhu, Hang Xu, Weiguo Gao, Chunjing Xu, Tao Xiang, and Li Zhang. Soft: Softmax-free transformer with linear complexity. *NeurIPS*, 34:21297–21309, 2021. 4

[13] Ao Luo, Fan Yang, Xin Li, and Shuaicheng Liu. Learning optical flow with kernel patch attention. In *CVPR*, pages 8906–8915, 2022. 7, 8

[14] Ao Luo, Fan Yang, Kunming Luo, Xin Li, Haoqiang Fan, and Shuaicheng Liu. Learning optical flow with adaptive graph reasoning. In *AAAI*, pages 1890–1898, 2022. 8

[15] Ao Luo, Fan Yang, Xin Li, Lang Nie, Chunyu Lin, Haoqiang Fan, and Shuaicheng Liu. Gaflow: Incorporating gaussian attention into optical flow. In *CVPR*, pages 9642–9651, 2023. 7, 8

[16] Ao Luo, Xin Li, Fan Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Flowdiffuser: Advancing optical flow estimation with diffusion models. In *CVPR*, pages 19167–19176, 2024. 4, 6, 8

[17] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *CVPR*, pages 3061–3070, 2015. 6

[18] Henrique Morimitsu, Xiaobin Zhu, Roberto M Cesar, Xiangyang Ji, and Xu-Cheng Yin. Dpflow: Adaptive optical flow estimation with a dual-pyramid framework. In *CVPR*, pages 17810–17820, 2025. 7, 8

[19] 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 *CVPR*, pages 1599–1610, 2023. 6, 8

[20] Xiuchao Sui, Shaohua Li, Xue Geng, Yan Wu, Xinxing Xu, Yong Liu, Rick Goh, and Hongyuan Zhu. Craft: Cross-attentional flow transformer for robust optical flow. In *CVPR*, pages 17602–17611, 2022. 7, 8

[21] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*, pages 8934–8943, 2018. 6, 8

[22] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pages 402–419. Springer, 2020. 1, 8

[23] Ruixing Wang, Xiaogang Xu, Chi-Wing Fu, Jiangbo Lu, Bei Yu, and Jiaya Jia. Seeing dynamic scene in the dark: A high-quality video dataset with mechatronic alignment. In *ICCV*, pages 9700–9709, 2021. 6, 9

[24] Haofei Xu, Jiaolong Yang, Jianfei Cai, Juyong Zhang, and Xin Tong. High-resolution optical flow from 1d attention and correlation. In *ICCV*, pages 10498–10507, 2021. 7, 8

[25] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *CVPR*, pages 8121–8130, 2022. 3, 7, 8

[26] Mingfang Zhang, Yinqiang Zheng, and Feng Lu. Optical flow in the dark. *TPAMI*, 2021. 6

[27] Shiyu Zhao, Long Zhao, Zhixing Zhang, Enyu Zhou, and Dimitris Metaxas. Global matching with overlapping attention for optical flow estimation. In *CVPR*, pages 17592–17601, 2022. 7, 8

[28] Yinqiang Zheng, Mingfang Zhang, and Feng Lu. Optical flow in the dark. In *CVPR*, pages 6749–6757, 2020. 6

[29] 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. *arXiv preprint arXiv:2401.17642*, 2024. 7

[30] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2

[31] Fengyuan Zuo, Zhaolin Xiao, Haiyan Jin, and Haonan Su. Cedflow: Latent contour enhancement for dark optical flow estimation. In *AAAI*, pages 7909–7916, 2024. 2, 5, 6, 7, 8

[32] Fengyuan Zuo, Haiyan Jin, Zhaolin Xiao, and Haonan Su. Cedflow++: Latent contour enhancement for dark optical flow estimation. In *IJCV*, pages 1–20, 2025. 2, 4, 6, 7, 8

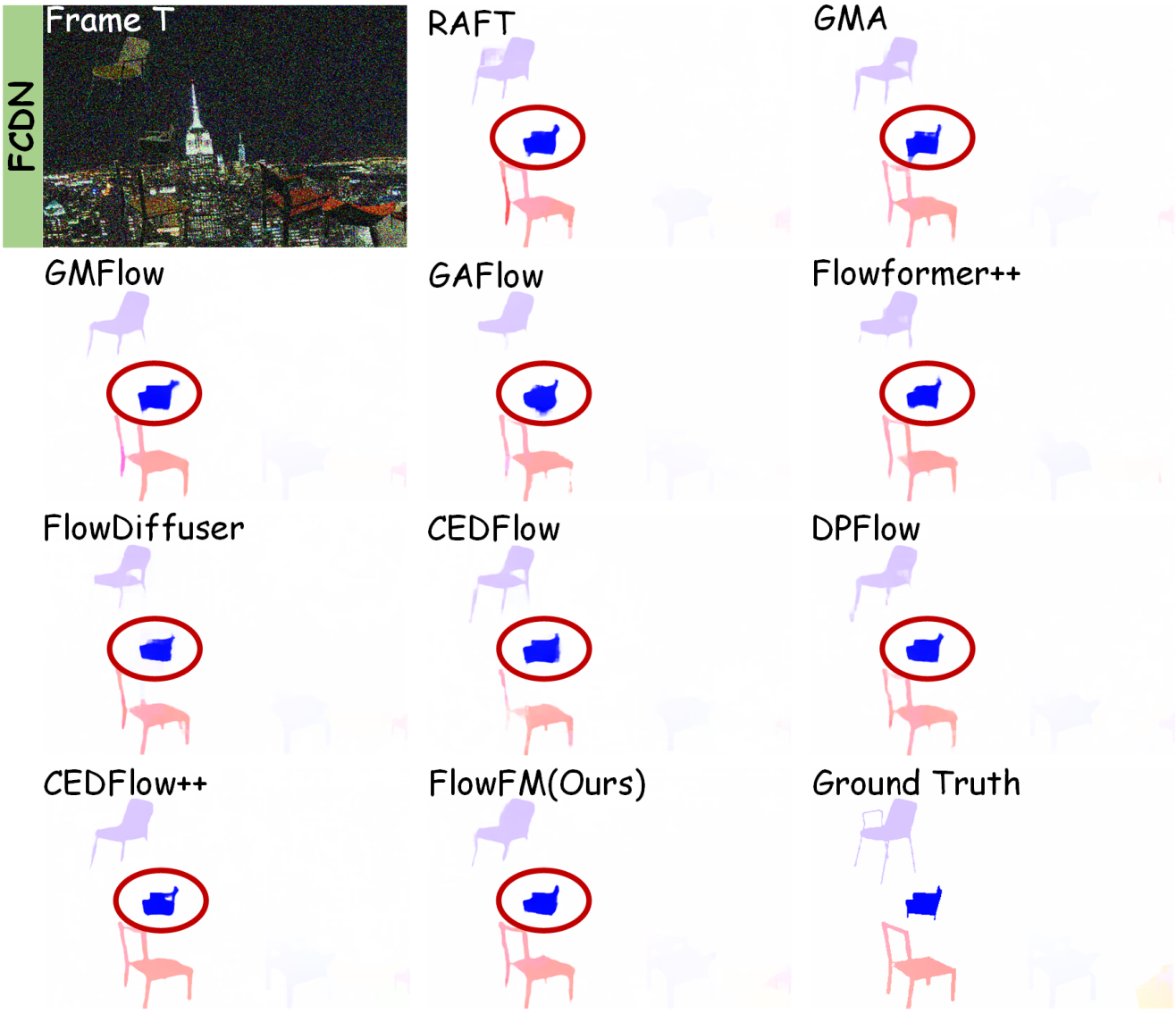

Figure 8. Qualitative comparisons of FlowFM with RAFT, GMA, GMFlow, GAFlow, Flowformer++, FlowDiffuser, CEDFlow, DPFlow, and CEDFlow++ are presented on FCDN dataset (trained on FCDN). Red box represents the difference area of dark optical flow.

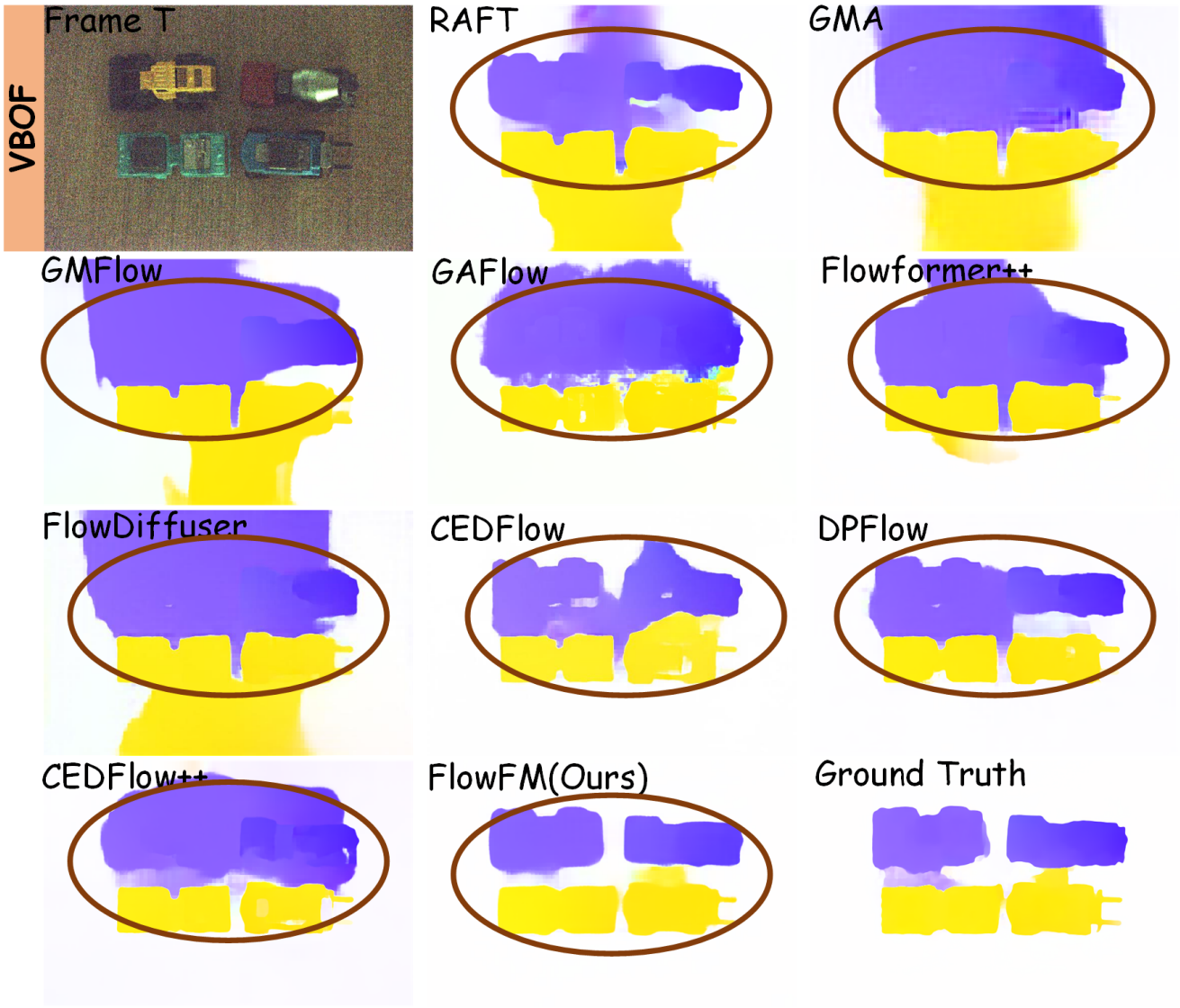

Figure 9. Generalization comparisons of FlowFM with RAFT, GMA, GMFlow, GAFlow, Flowformer++, FlowDiffuser, CEDFlow, DPFlow, and CEDFlow++ are presented on the challenging VBOF dataset (trained on FCDN). Color box denotes the difference area of dark optical flow.

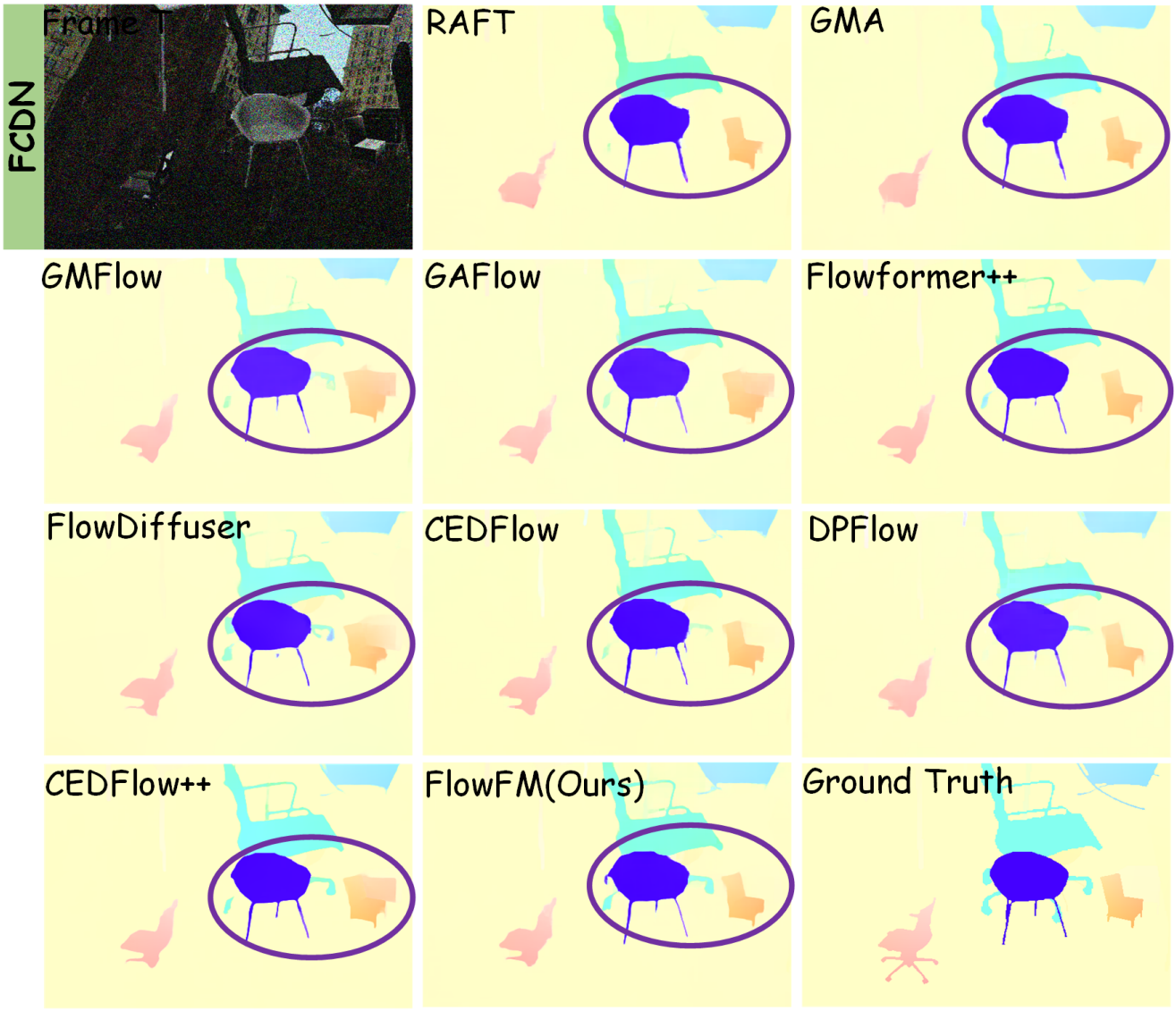

Figure 10. Qualitative comparisons of FlowFM with RAFT, GMA, GMFlow, GAFlow, Flowformer++, FlowDiffuser, CEDFlow, DPFlow, and CEDFlow++ are presented on FCDN dataset (trained on Mix). Purple box denotes the difference area of dark optical flow.

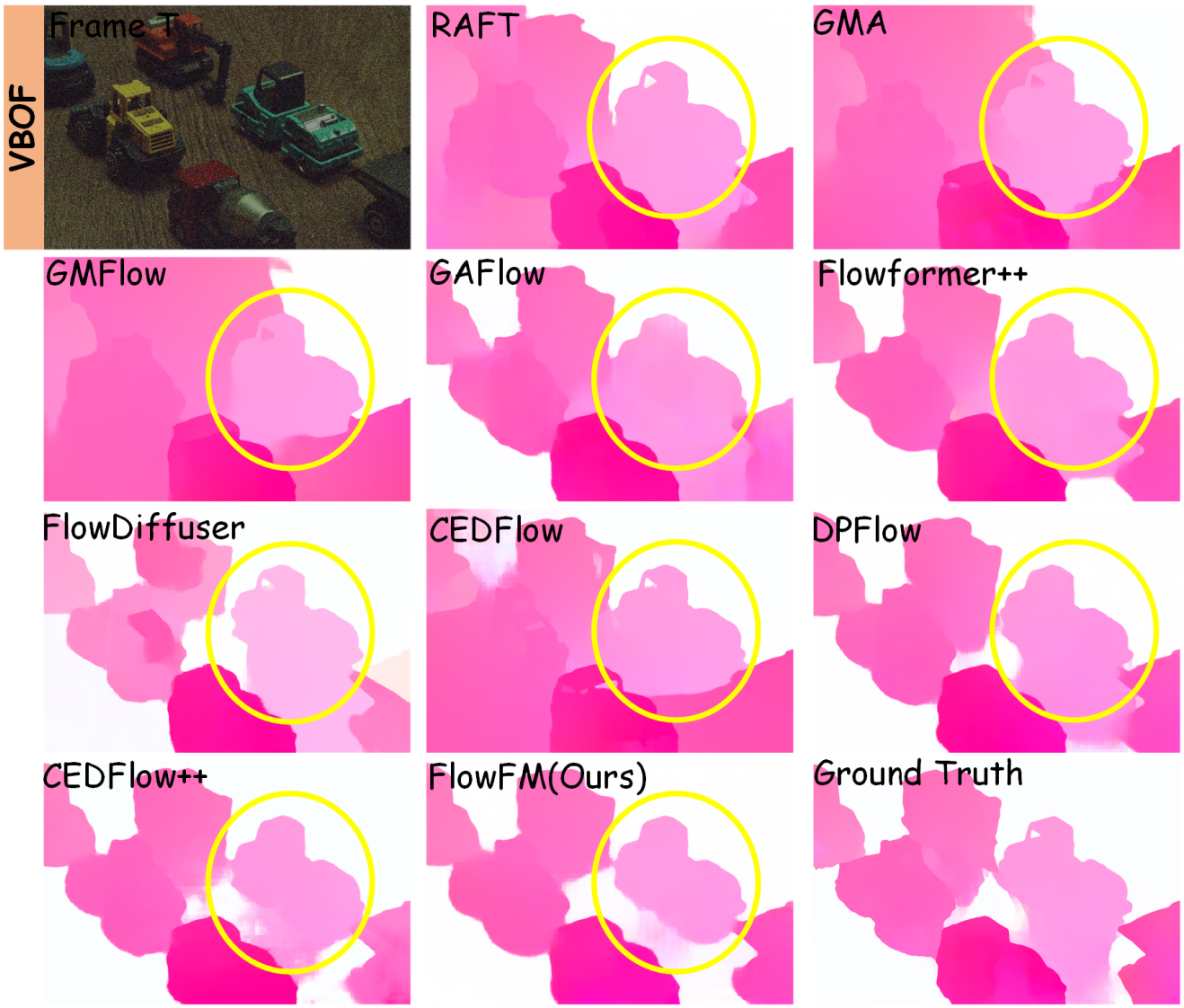

Figure 11. Qualitative comparisons of FlowFM with RAFT, GMA, GMFlow, GAFlow, Flowformer++, FlowDiffuser, CEDFlow, DPFlow, and CEDFlow++ are presented on the challenging VBOF dataset (trained on Mix). Yellow box denotes the difference area of dark optical flow.

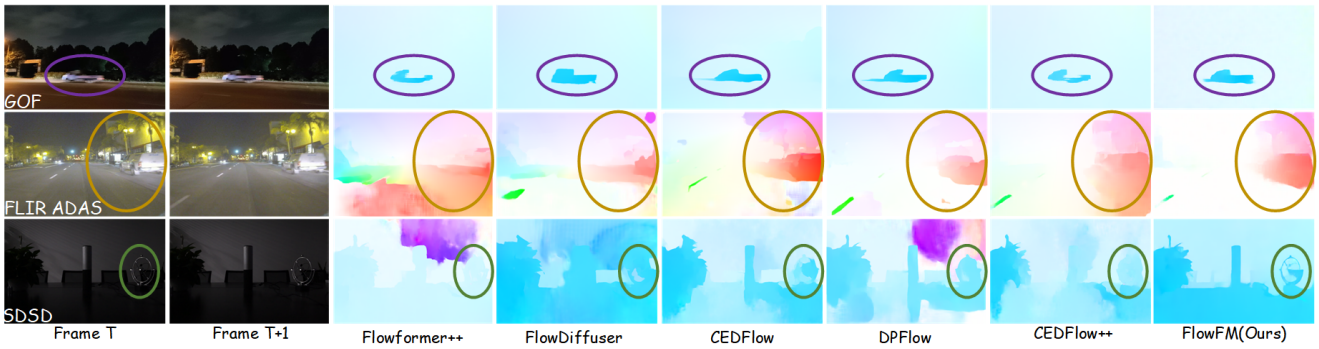

Figure 12. The DOFE results on the FLIR ADAS (a), SDSD (b), and GOF (c) datasets demonstrate FlowFM's ability to effectively manage rapid motions, complex visual elements, and exhibit significant generalizability in real-world driving scenarios and low-light conditions (Best viewed with Zoom).

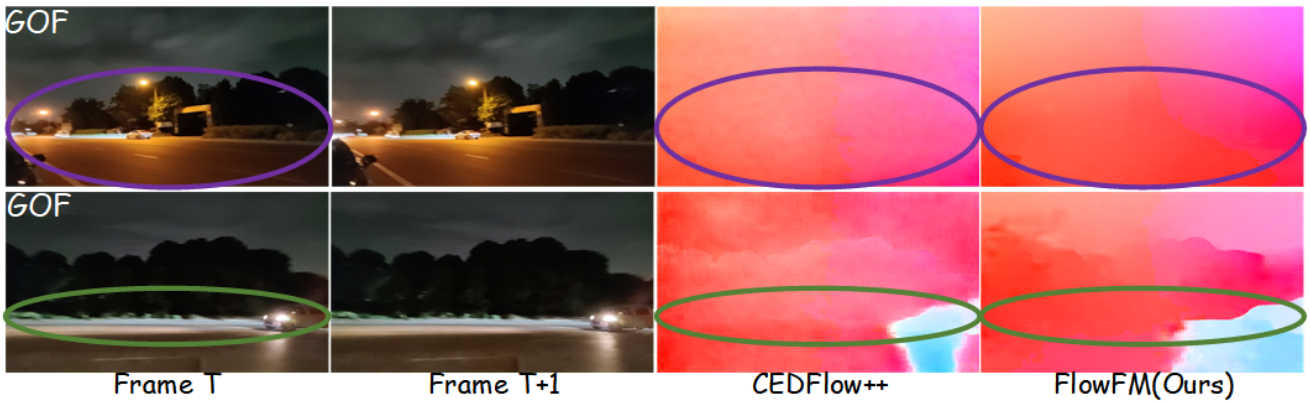

Figure 13. Failure cases.