FlowAnyTime: Efficient Fine-tuning with Intra-Inter Frame Distillation for All-Weather Optical Flow Estimation

Zixu Wang[1], Hongye Chen[2], Xiaochun Zou[3*], Congxuan Zhang[2], Zhen Chen[1,2*], Xinbo Zhao[1*]

[1]School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2]School of Instrument Science and Optoelectronic Engineering, Nanchang Hangkong University, Nanchang, China
[3]School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China
wangzixu0827@163.com, xczou@nwpu.edu.cn

Abstract

Motion estimation in degraded scenes has long been a significant challenge, primarily attributed to substantial scene variations and insufficient training data. Existing approaches typically address this limitation by incorporating additional training strategies or modifying network architectures within conventional frameworks. However, these solutions not only require cumbersome training procedures or additional modal inputs, but also lack generalization capabilities. To address this problem, we propose a unified optical flow estimation framework specifically designed for degraded scenes. In this work, we employ large-scale pre-trained optical flow foundation models as both teacher and student networks. Our objective is to compensate for feature incompleteness during image degradation through pre-trained large models. Subsequently, we leverage supervised signals for fine-tuning and introduce an intra-inter frame distillation method to enable the student network to adapt to diverse cross-domain scenarios. Our proposed methodology provides deeper insights into learning style-invariant features from these learnable fine-tuning layers. Extensive experiments demonstrate that our approach achieves superior generalization performance and state-of-the-art results in degraded scenes (including low-light, rain, fog and other conditions) while requiring minimal training resources.

Introduction

Optical flow estimation is a fundamental task in computer vision that aims to accurately determine pixel-level motion correspondences between adjacent frames. This high-precision motion information provides essential cues for downstream tasks including video segmentation (Chen et al. 2023), video frame interpolation (Park et al. 2021), object tracking (Qin et al. 2023) and other fields (Liu et al. 2025). However, optical flow estimation performs poorly in degraded scenarios because such conditions violate the brightness and gradient constancy assumptions upon which most optical flow methods rely, resulting in significant performance degradation.

Current optical flow estimation methods typically decompose degraded scenarios into multiple specific subtasks, such as nighttime optical flow and adverse weather optical flow (e.g., rain and fog scenarios). Many existing approaches are tailored to specific types of degradation and applications. Among these, some methods commonly treat these problems as a two-stage process: the first stage preprocesses images to restore them to a clean state, followed by network-based computation for motion estimation (Zheng, Zhang, and Lu 2020; Li et al. 2019). However, this framework fails to establish connections between the preprocessing and motion estimation processes, resulting in suboptimal estimation performance. Alternative approaches introduce additional auxiliary knowledge and training strategies to compensate for missing visual motion information (Li, Luo, and Liu 2021; Zhou et al. 2024a; Dai et al. 2025). Nevertheless, estimation systems for degraded scenarios require the integration of multiple specialized models and specific training procedures, leading to complex and computationally expensive frameworks. Furthermore, since unsupervised loss functions for optical flow are typically designed for brightness constancy scenarios, unsupervised and semi-supervised optical flow estimation methods exhibit poor performance in degraded conditions. Recently, CEDFlow (Zuo et al. 2024) has improved nighttime optical flow estimation accuracy by incorporating Transformers and contour enhancement modules into the encoder component of RAFT-based (Teed and Deng 2020; Wang et al. 2024, 2025) frameworks. However, these carefully designed modules focus on individual degraded scenarios, resulting in limited generalization capability and introducing additional computational overhead.

In this paper, we formulate optical flow estimation in degraded scenarios (including low-light, rain, fog, and etc.) as a unified task. While existing methods exhibit poor performance under these conditions, we observe that networks trained on normal images can extract more robust geometric representations. To leverage the strong generalization capabilities of cross-domain pre-trained models, we propose transferring the rich knowledge from pre-trained optical flow foundation models to degraded domains. Specifically, we explore a fine-tuning paradigm that operates on two fronts. On one hand, we selectively unfreeze certain parameters in the pre-trained model to learn domain-specific features in degraded scenarios through supervised signals. On the other hand, we construct an intra-inter frame distillation strategy that suppresses degraded feature learning through normal features, thereby encouraging consistent se-

*Corresponding Authors

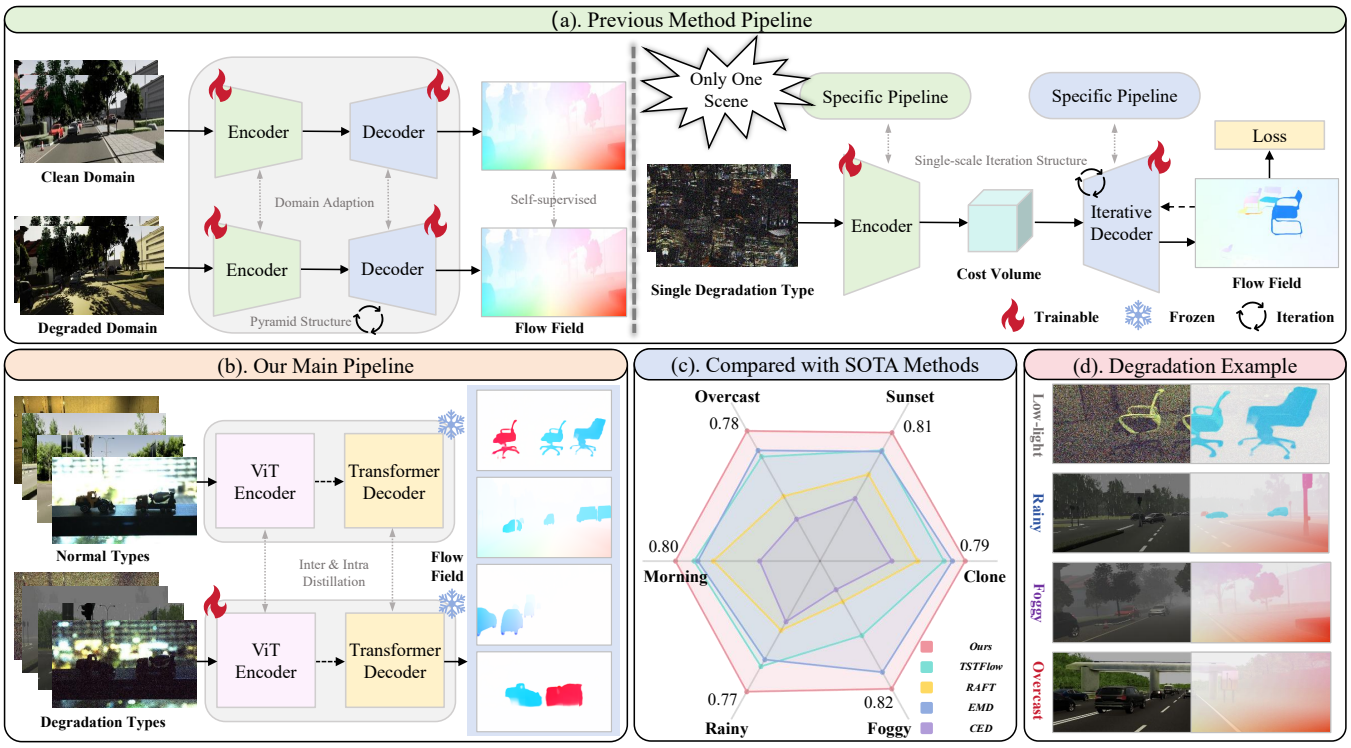

Figure 1: **(a)** and **(b)**: Pipeline comparison between our method and previous approaches. **(c)**: Performance comparison of our method against other representative state-of-the-art (SOTA) approaches on the VKITTI2 dataset; **(d)**: Illustration of several representative degradation types.

mantic and geometric relationships within the feature domain. Additionally, given the inherent style differences between normal and degraded domains, completely eliminating cross-domain style disparities is not only challenging but also practically impossible. Therefore, we propose to focus model optimization on several key layers, utilizing the intra-inter frame distillation strategy to minimize the gap between layer embeddings from the normal teacher and degraded student networks. Overall, our contributions are summarized as follows:

- We propose FlowAnyTime, a unified framework for optical flow estimation in degraded scenarios. By leveraging large-scale pre-trained optical flow foundation models, we eliminate redundant frameworks and cumbersome training procedures, treating optical flow estimation across different degraded scenarios as a unified task.
- We introduce a novel feature fine-tuning strategy for adapting to diverse cross-domain features. Through selective unfreezing of a small number of learnable parameters, we model both variant and invariant motion information from intra and inter-frame perspectives, enabling the model to extract robust motion features and establish accurate correspondences, thereby enhancing optical flow estimation accuracy in degraded scenarios.
- Extensive experiments demonstrate that our proposed method achieves state-of-the-art optical flow estimation performance across various degradation scenarios while maintaining computational efficiency. Compared to previous state-of-the-art approaches, our method improves the average optical flow estimation accuracy by 20% across diverse degradation conditions.

Related Work

Optical Flow Estimation in Degraded Scenes. Degraded scene scenarios have long posed significant challenges for optical flow estimation, as they fundamentally violate the brightness and gradient constancy assumptions upon which most optical flow methods rely. Since optical flow estimation seeks motion correspondences through visual imagery, several approaches have attempted to address this by either extracting degradation-specific features or directly feeding image-restored outputs into optical flow networks. For instance, DarkFlow (Zheng, Zhang, and Lu 2020), RainFlow (Li et al. 2019), and FogFlow (Yan, Sharma, and Tan 2020) propose data-driven optical flow estimation methods specifically tailored to low-light, rainy, and foggy conditions, respectively. However, existing image enhancement techniques are not necessarily well-suited for two-frame optical flow estimation tasks, as the enhanced images may lose critical visual representations, leading to erroneous motion feature extraction.

Fundamentally, optical flow estimation in degraded scenarios still requires establishing pixel correspondences be-

tween consecutive frames. Consequently, subsequent methods have attempted to improve cross-domain optical flow estimation by incorporating additional data modalities (Zou et al. 2026, 2024) or developing novel training strategies. Several methods enhance optical flow estimation through additional data modalities. GyroFlow (Li, Luo, and Liu 2021) combines gyroscope data with images for unsupervised learning. ABDAFlow (Zhou et al. 2024a) and CHDAFlow (Zhou et al. 2024b) use auxiliary domains (event cameras and intermediate bridges respectively) to transfer motion knowledge to nighttime scenarios. MSIFlow (Dai et al. 2025) incorporates depth point clouds for nighttime enhancement. However, these multi-modal approaches require additional training/inference processes or complex loss functions (Du et al. 2024), making training cumbersome and data acquisition challenging. CEDFlow (Zuo et al. 2024) similarly integrates Transformers and contour constraints into RAFT-based encoders for low-light motion feature enhancement. In this work, we leverage solely visual information, building upon large-scale pre-trained optical flow estimation models as our foundation. By incorporating fine-tuning and intra-frame to inter-frame distillation strategies, we address the challenge of diminished accuracy in optical flow estimation networks under degraded scene conditions.

Methodology

Preliminaries. Optical flow estimation aims to establish pixel-wise displacement correspondences between consecutive RGB frames. Given input frames I_1 and I_2, the network $\Theta(\cdot)$ produces a two-dimensional optical flow field f:

$$f = \Theta(I_1, I_2), \tag{1}$$

In this work, we employ CroCo v2 (Weinzaepfel et al. 2023) as our backbone model. CroCo v2 is a Vision Transformer (ViT)-based encoder-decoder architecture for optical flow estimation. The entire CroCo v2 model is pre-trained on a large scale using 5.3 million real-world image pairs, enabling accurate capture of pixel correspondences in general scenes for optical flow estimation. For detailed explanations of CroCo v2, please refer to the supplementary material.

Distillation Fine-turning with Paired Inputs. Our overall network framework structure is illustrated in Fig. 2. As shown in Fig. 2, we deploy the CroCo v2 large model as our baseline, where the teacher model freezes all parameters during training.

To adapt to tokens in degraded domains and ensure accurate motion feature extraction, we propose the following two regularization strategies: (1) Due to the inherent differences in information captured between degraded and normal domain visible light, we employ a fine-tuning strategy that unfreezes partial parameters, enabling the network to learn features specific to degraded domains. (2) To ensure model robustness, we design an innovative learning strategy. **The core idea of this strategy is to leverage paired high-quality image sequences as additional auxiliary information sources during the training phase.** Since paired normal and degraded scene images in our dataset share identical content, we believe that the discrepancy between their predictions will be largely influenced by their stylistic differences. The teacher provides the neural style of its intermediate feature maps as additional supervision for the student, enabling the student to learn to bridge the stylistic gap between normal and degraded images in the feature space. This approach ultimately leads to learned representations that are insensitive to varying degradation conditions.

Given that the strong generalization and zero-shot capabilities of CroCo v2 may stem from their large-scale pre-training datasets, we do not modify the model architecture to reuse the pre-trained model. Furthermore, rather than re-training the entire model, we fine-tune several key layers in the network to avoid overfitting while preserving the inherent learned patterns. Based on these techniques, an optical flow estimation model for degraded scenarios can be trained. In the next section, we will discuss enhancing the robustness of flow estimation in degraded scenarios through intra-frame and inter-frame distillation methods.

Intra-frame Distillation. As illustrated in Fig. 2, the teacher and student networks take consecutive frame pairs from normal and degraded types as inputs, respectively. For intra-frame motion feature extraction, if normal and degraded image pairs share identical content, the model should exhibit consistent motion patterns to both. While CroCo v2 demonstrates excellent performance on clean images, its robustness to degraded inputs remains limited. Direct training on degraded images often leads to overfitting and fails to leverage the rich geometric understanding learned from clean data. To address this challenge, our key idea is to exploit large-scale pre-trained knowledge to obtain invariant information across varying image quality conditions.

We argue that optical flow fields between consecutive frames should satisfy local smoothness and global geometric consistency constraints, with motion patterns encoded across different channels reflecting these temporal geometric relationships. Furthermore, in optical flow estimation tasks, tokens at different spatial locations contribute variably to the final motion estimation. Critical regions such as motion boundaries and occlusion areas should receive higher attention weights. Additionally, when the encoder processes dual-frame inputs, its internal representations implicitly model inter-frame motion correspondences, where the quality of these correspondences directly impacts the final optical flow prediction accuracy.

Therefore, we design an intra-frame distillation strategy to constrain the student network in acquiring salient intra-frame features. We aim to preserve the teacher network understanding of scene geometry and motion patterns while prioritizing key tokens that contribute to flow estimation. Through this approach, the student learns to handle degradation while maintaining optical flow estimation quality.

Let $\mathcal{I}^c = \{(I_1^c, I_2^c)\}$ denote clean image pairs and $\mathcal{I}^d = \{(I_1^d, I_2^d)\}$ denote their degraded counterparts, where degradation is modeled as:

$$I^d = \mathcal{D}(I^c) = I^c + \eta \tag{2}$$

where η represents various degradation factors (e.g., Gaussian noise, motion blur, rainy, foggy). The teacher network $\mathcal{T}$ with parameters θ processes normal pairs to estimate op-

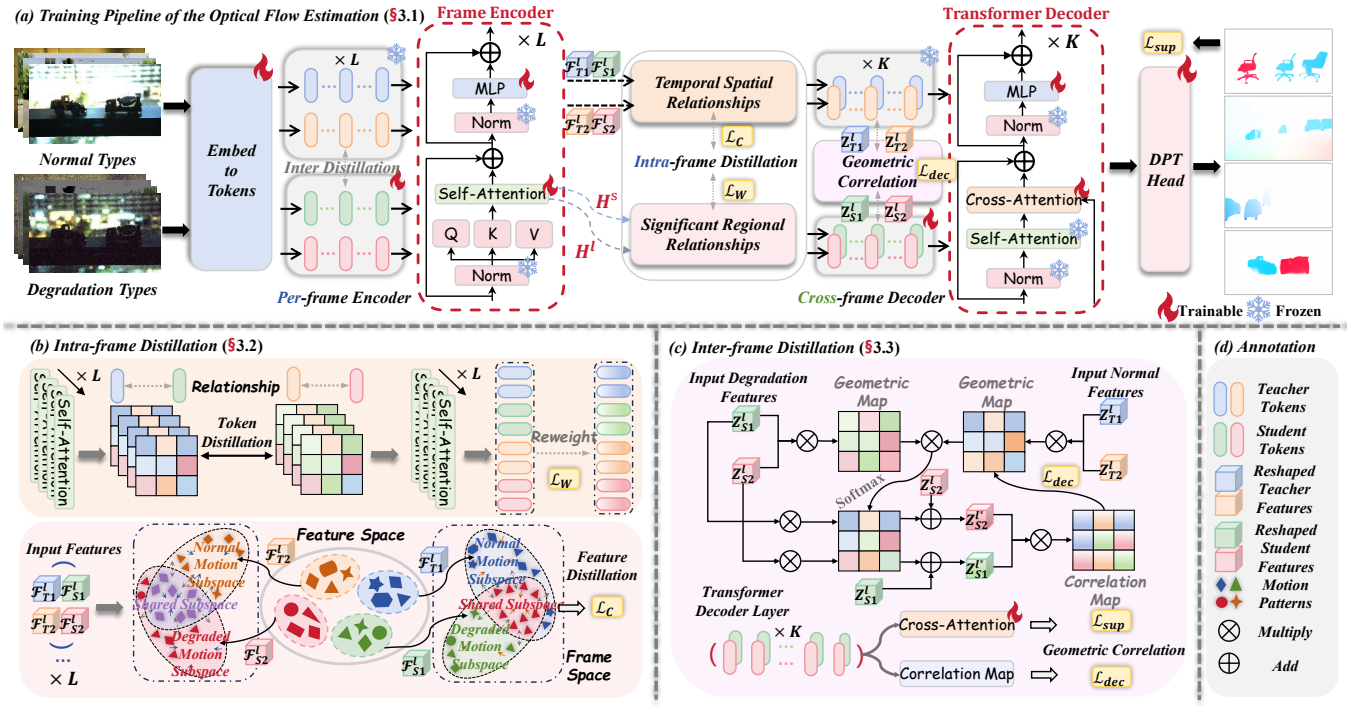

Figure 2: Overview of the proposed network architecture.

tical flow $\mathcal{T}(I_1^c, I_2^c; \theta_T)$, while the student network $\mathcal{S}$ with parameters θ processes degraded pairs: $\mathcal{S}(I_1^d, I_2^d; \theta_S)$.

For encoder layer l, let $\mathcal{F}_\mathcal{T}^l \in \mathbb{R}^{N\times C}$ and $\mathcal{F}_\mathcal{S}^l \in \mathbb{R}^{N\times C}$ denote teacher and student features respectively, where N is the number of tokens and c is the channel dimension. Each channel acts as a specialized motion detector. To compare these detectors across clean and degraded domains, we need a normalization scheme that preserves intra-channel spatial relationships while removing inter-domain magnitude differences. Following the insight that different channels capture distinct motion patterns and geometric features, we apply channel normalization to intermediate feature maps. For each layer l and channel c, we compute spatial attention maps:

$$\mathbf{A}_{\mathcal{T},\mathcal{S}}^{(l,c)} = \frac{\exp(\mathcal{F}_{\mathcal{T},\mathcal{S}}^{(l,c)}/\tau)}{\sum_{i,j}\exp(\mathcal{F}_{\mathcal{T},\mathcal{S}}^{(l,c)}[i,j]/\tau)} \in \mathbb{R}^{N_l}, \tag{3}$$

where $c \in \{1, ..., C\}$ indexes channels, and τ is the temperature parameter. N_l are the token dimensions at layer l. Traditional spatial distillation treats all channels equally, ignoring their specialized roles. Therefore, we construct a distillation loss that measures distribution differences while respecting channel structures:

$$\mathcal{L}_C^{(l)} = \frac{\tau^2}{C}\sum_{c=1}^{C}\mathrm{KL}(\mathbf{A}_{\mathcal{T},c}^{(l)} \| \mathbf{A}_{\mathcal{S},c}^{(l)}), \tag{4}$$

The Kullback-Leibler (KL) divergence heavily penalizes when the student fails to activate regions that the teacher considers important, while being more forgiving when the student activates additional regions. This is crucial because degradation may cause spurious activations. This asymmetric formulation ensures that regions with high teacher attention (salient motion) are strongly supervised while those with low teacher attention (static/occluded) have relaxed constraints.

Notably, not all spatial tokens contribute equally to optical flow estimation: motion boundaries are critical for accurate flow discontinuities, texture-rich regions provide reliable correspondence cues, and homogeneous areas are less informative for flow estimation. The self-attention in encoder process naturally captures these dependencies, which we can exploit to guide distillation. For layer l, the attention matrix $\mathbf{P}^{(l)} \in \mathbb{R}^{N\times N}$ encodes pairwise token relationships. We compute the importance weight for each token based on its influence on subsequent layers:

$$\mathbf{P}^{(l)} = \frac{1}{N}\sum_{h=1}^{N}\mathrm{Softmax}\left(\frac{Q^{(l)}K^{(l)\top}}{\sqrt{d_k}}\right), \tag{5}$$

where N is the number of attention heads, and $Q^{(l)}, K^{(l)}$ are query and key matrices. However, single-layer attention is myopic. We need to capture how information flows through the network hierarchy. The cumulative importance from layer l to the final layer L:

$$\mathbf{E}^{(l)} = (\beta_l\mathbf{P}^{(l)} + (1-\beta_l)\mathbf{T})\cdots(\beta_L\mathbf{P}^{(L)} + (1-\beta_L)\mathbf{T}), \tag{6}$$

where $\mathbf{T}$ is the identity matrix, and $\beta \in [0, 1]$ balances attention and residual connections. The attention flow matrix $\mathbf{E}^{(l)}$ formulates how much each token at layer l influences the final representation. We aggregate this influence to obtain

Method	Trained on FCDN+VBOF			Trained on VKITTI 2					
	FCDN	VBOF(Fuji2)	VBOF(All)	Clone	Fog	Rain	Morning	Overcast	Sunset
RAFT[ECCV-20]	1.38	7.34	8.89	1.53	2.34	1.84	1.39	1.92	1.54
Flow1D[CVPR-21]	1.30	5.13	6.93	1.49	2.95	2.41	1.95	2.31	2.06
GMA[ICCV-21]	1.26	4.90	6.81	1.65	2.31	1.88	1.46	1.99	1.42
AGFlow[AAAI-22]	1.27	4.97	6.75	1.38	2.24	1.80	1.35	1.78	1.35
GMFlowNet[CVPR-22]	1.70	7.71	8.66	1.67	3.09	2.56	1.41	1.86	1.44
GMFlow[CVPR-22]	1.31	5.76	7.23	1.93	2.55	1.99	2.11	2.32	1.96
KPA-Flow[CVPR-22]	1.39	6.11	7.47	1.33	2.01	1.55	1.21	1.45	1.28
EMD-Flow[ICCV-23]	1.18	4.86	7.09	0.99	1.11	1.33	1.15	1.12	1.14
CEDFlow[AAAI-24]	1.23	4.69	6.52	-	-	-	-	-	-
TSTFlow[WACV-24]	-	-	-	1.12	1.75	1.21	1.09	1.23	1.12
FlowAnyTime[Ours]	**0.85**	**4.23**	**4.57**	**0.79**	**0.82**	**0.77**	**0.80**	**0.78**	**0.81**

Table 1: Comparison with state-of-the-art methods on FCDN, VBOF and VKITTI 2 datasets. The best results are in **bold**.

scalar importance scores and acquire the weighted token-wise distillation loss:

$$\mathcal{L}_{\text{W}}^{(l)} = \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbf{E}_{ij}^{(l)} \cdot \|\mathcal{F}_{\mathcal{T},ij}^{(l)} - \mathcal{F}_{\mathcal{S},ij}^{(l)}\|_1, \tag{7}$$

The total inter-frame distillation loss is:

$$\mathcal{L}_{\text{enc}} = \sum_{l\in L} \left[\mathcal{L}_{\text{C}}^{(l)} + \mathcal{L}_{\text{W}}^{(l)}\right], \tag{8}$$

The proposed method provides robustness through an information bottleneck, where temporal spatial distillation acts as an information bottleneck to filter noise while preserving motion-relevant features. Additionally, attention regularization is enforced via token weighting, preventing overfitting to degradation artifacts in less important regions. Moreover, Multi-scale consistency is ensured by hierarchical distillation, maintaining alignment across feature scales.

Inter-frame Distillation. We propose a novel framework that integrates inter-frame interaction distillation into optical flow estimation to enhance robustness under degraded imaging conditions. The key insight is that optical flow estimation under degraded conditions can benefit from knowledge learned from clean image pairs, as the geometric correlation between image pairs remain consistent regardless of degradation. In flow decoder, cross-attention capture the relationships between different motion cues. By computing these correlations separately for each frame, we can understand how features from different time points relate to each other, which is essential for motion correspondence. The geometric map encode abundant information about geometric correlation that remains relatively stable across domains.

Let $\mathbf{Z}_{\mathcal{T}_1}^{(k)}, \mathbf{Z}_{\mathcal{T}_2}^{(k)} \in \mathbb{R}^{N\times C}$ denote the teacher's token features and $\mathbf{Z}_{\mathcal{S}_1}^{(k)}, \mathbf{Z}_{\mathcal{S}_2}^{(k)} \in \mathbb{R}^{N\times C}$ denote the student's token features in the flow decoder, where N is the number of tokens, C is the channel dimension, and k is the number of decoder layer K. To capture cross-frame token relationships, we compute the token-wise correlation matrices:

$$\mathbf{A}_{\mathcal{T},\mathcal{S}}^{(k)} = \mathbf{Z}_{\mathcal{T}_1,\mathcal{S}_1}^{(k)} \cdot (\mathbf{Z}_{\mathcal{T}_2,\mathcal{S}_2}^{(k)})^{\top} \in \mathbb{R}^{N\times N}, \tag{9}$$

Subsequently, the dot product of $\mathbf{A}^{\mathcal{T}}$ and $\mathbf{A}^{\mathcal{S}}$ is computed to obtain the geometric similarity map between the teacher and student:

$$\mathbf{M}^{geo} = \text{Softmax}\left(\mathbf{A}^{\mathcal{T}} \cdot \mathbf{A}^{\mathcal{S}}\right) \in \mathbb{R}^{N\times N}, \tag{10}$$

Next, the two-frame features $\mathbf{Z}_{\mathcal{S}_1}^{(k)}, \mathbf{Z}_{\mathcal{S}_2}^{(k)}$ from the student network are respectively multiplied by $\mathbf{M}^{geo}$ to incorporate domain adaptation knowledge transfer:

$$\mathbf{Z}_{\mathcal{S}_1,\mathcal{S}_2}^{(k)^*} = \mathbf{Z}_{\mathcal{S}_1,\mathcal{S}_2}^{(k)} + \mathbf{M}^{geo} \cdot \mathbf{Z}_{\mathcal{S}_1,\mathcal{S}_2}^{(k)}, \tag{11}$$

Finally, we compute the transfer correlation matrices $\mathbf{A}_{tra}^{\mathcal{S}}$ to calculate the geometric consistency distillation loss:

$$\mathcal{L}_{\text{dec}} = \sum_{k\in K} \frac{1}{C} \cdot \|\mathbf{A}^{\mathcal{T}_{(k)}} - \mathbf{A}_{tra}^{\mathcal{S}_{(k)}}\|_2^2, \tag{12}$$

Training Targets. By integrating the previously mentioned distillation training strategy with the encoder-decoder framework for inter-frame and intra-frame distillation, we optimize the network via the following objective function:

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda_e \mathcal{L}_{\text{enc}} + \lambda_d \mathcal{L}_{\text{dec}}, \tag{13}$$

Here, $\mathcal{L}_{\text{sup}}$ is the Laplacian loss function in CroCo v2 (Weinzaepfel et al. 2023). Empirically, we apply intra-frame distillation to the tokens using the corresponding encoder layers $L = \{3, 6, 12, 18, 20, 24\}$, followed by inter-frame distillation applied to the tokens with the corresponding decoder layers $K = \{4, 8, 12\}$. In the experiment, we set λ_e and λ_d to 0.5 respectively.

Experiment

Training Details. The training procedure consists of two distinct stages: a pre-training phase on large-scale datasets followed by a fine-tuning phase for model adaptation. We adopt the training strategy from CroCo v2 (Weinzaepfel et al. 2023) as our pre-training foundation.

In the fine-tuning stage, we utilize paired normal-degraded scenario datasets as benchmarks for training and

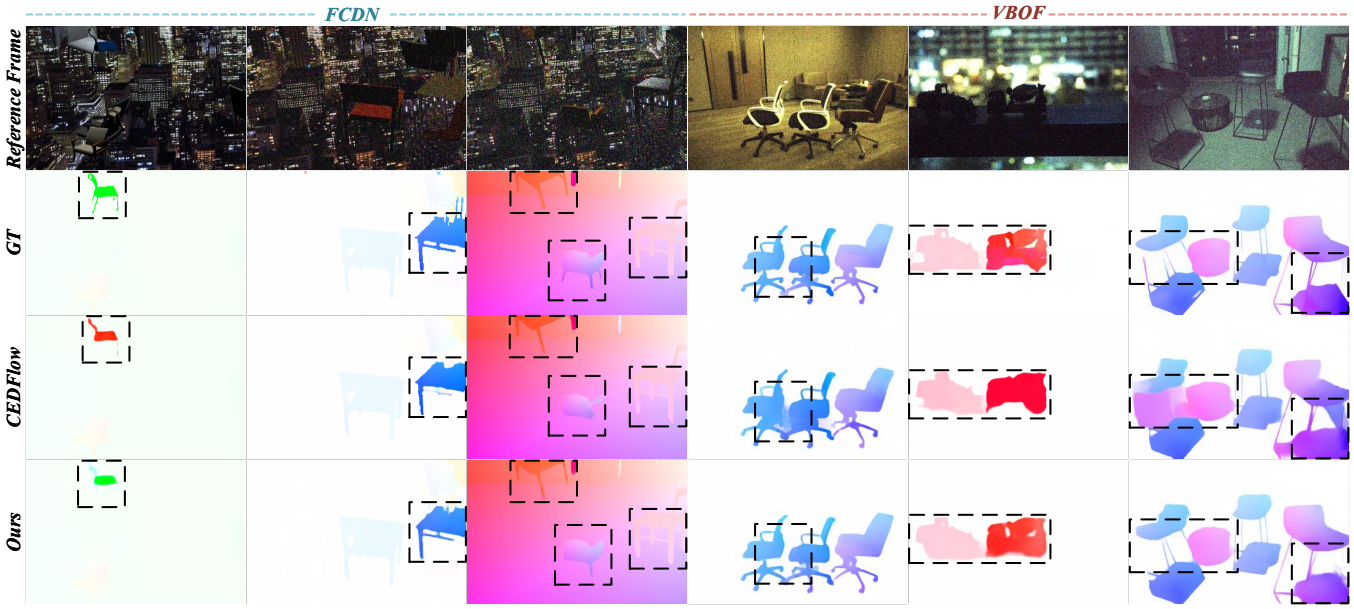

Figure 3: Flow visualization comparison under challenging low-light and high-noise conditions.

evaluation of our pipeline. Specifically, following (Zuo et al. 2024), we employ the FCDN (Zheng, Zhang, and Lu 2020) and VBOF (Zhang, Zheng, and Lu 2022) datasets as benchmarks for optical flow estimation in low-light scenarios. Additionally, following (Yoon et al. 2024), we use the VKITTI2 (Cabon, Murray, and Humenberger 2020) dataset to evaluate performance under various weather conditions: Clone, Fog, Morning, Overcast, Rain, and Sunset. In our experiments, we designate Clone as the normal scenario and treat the remaining weather conditions as degraded scenarios. During the fine-tuning stage, we train on all training datasets with a batch size of 4, a learning rate of 2×10^{-5} and 150k iterations. All other settings remain consistent with the aforementioned training procedure. All fine-tuning experiments are conducted on a single RTX 4090 GPU. Throughout the training process, we use FlyingChairs as the teacher model input for FCDN. For VBOF, we extract the first pair of images from each camera image sequence as the teacher model input. For VKITTI2, we employ the Clone scenario as the teacher model input.

Standard Evaluation. We conducted a comprehensive evaluation of our model by comparing it against several state-of-the-art approaches (RAFT (Teed and Deng 2020), Flow1D (Xu et al. 2021), GMA (Jiang et al. 2021), AGFlow (Luo et al. 2022b), GMFlowNet (Zhao et al. 2022), GMFlow (Xu et al. 2022), KPA-Flow (Luo et al. 2022a), EMD-Flow (Deng et al. 2023), CEDFlow (Zuo et al. 2024) and TSTFlow (Yoon et al. 2024)) that have demonstrated superior performance in optical flow estimation across both standard and degraded imaging conditions. To ensure fair comparison, all models were trained using identical training strategys on the FCDN + VBOF and VKITTI 2 datasets. Detailed implementation specifications and experimental configurations are provided in the supplementary materials.

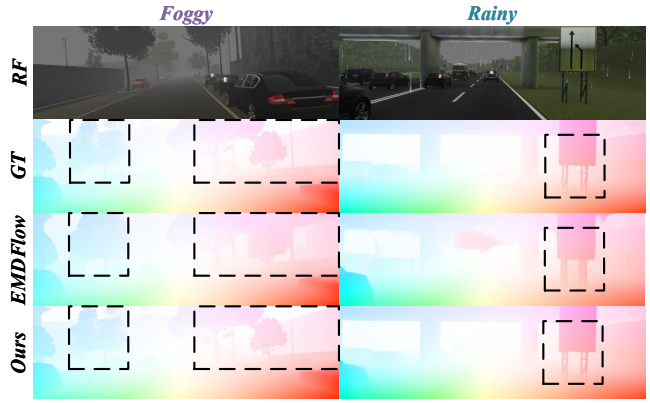

Figure 4: Comparison of multiple degraded scenarios on VKITTI2.

Training on the FCDN and VBOF Datasets. We evaluate the optical flow estimation accuracy of our model under low-light scenarios using FCDN and VBOF benchmarks. To ensure fair comparison, we adopt the same experimental strategy as CEDFlow (Zuo et al. 2024). As shown in Tab. 1, our method achieves the best end-point-error (EPE) among all competing methods. On FCDN, our method attains an EPE of 0.85, outperforming the second-best method CEDFlow by 31% (1.23 *vs* 0.85). Additionally, our model requires significantly fewer training iterations (150k *vs* 350k) to achieve superior performance. These results demonstrate that, benefiting from our training strategy and robust foundational framework, our model exhibits enhanced robustness to noise and low-light conditions. On the VBOF dataset, our method surpasses other models on both the Fuji2 subset and across all scenes. The proposed FlowAnyTime achieved

Fine-turning Layers	Dataset	
	FCDN	VBOF(All)
Only PE	4.37	9.97
PE + All Encoder + All Decoder	0.87	4.75
PE + All Encoder	2.21	6.75
PE + All Decoder	1.98	6.41
PE + 6 Encoder + All Decoder	1.09	5.21
PE + All Encoder + 3 Decoder	1.31	5.36
PE + 6 Encoder + 3 Decoder	**0.85**	**4.57**

Table 2: Ablation study results of fine-tuning different layers. All models are trained on a mixed dataset of FCDN and VBOF, then evaluated separately on FCDN and VBOF test datasets. PE indicates patch embedding layer; All Encoder denotes fine-tuning all frame encoder layers; All Decoder represents fine-tuning all transformer decoder layers and DPT head; 6 Encoder indicates fine-tuning frame encoder layers $L = \{3, 6, 12, 18, 20, 24\}$; 3 Decoder represents fine-tuning transformer decoder layers $K = \{4, 8, 12\}$ and DPT head. Underline for the adopt strategy.

an EPE of 4.23 on the Fuji2 subset and 4.57 across all VBOF scenes, outperforming other state-of-the-art methods. Specifically, on the VBOF Fuji2 subset, our method leads the second-best approach CEDFlow by 10% (4.23 vs 4.69). Across all scenes of VBOF dataset, our method surpasses CEDFlow by 30%. Furthermore, based on the Fuji2 and all metrics, our method demonstrates robust generalization capability on the VBOF dataset. These results further validate that our optical flow estimation approach can better mitigate the adverse effects of illumination and noise under low-light conditions. Furthermore, Fig. 3 presents flow visualization results for low-light scenarios, demonstrating that our method exhibits superior optical flow estimation performance in noise-corrupted low-light regions.

Training on the VKITTI2 Datasets. Similar to TSTFlow (Yoon et al. 2024), we utilize the VKITTI2 dataset to evaluate our model's optical flow estimation accuracy under various degraded scenarios. Our training strategy involves training on one degraded scenario and then validating on all other scenarios, with the average EPE across all validation scenarios serving as the performance metric for that particular scenario. Detailed experimental results are provided in the supplementary material. According to the results presented in Table 1, our method achieves the best EPE performance among all compared approaches. In common degraded scenarios such as rain and fog, our method outperforms the second-best methods EMD-Flow and TSTFlow by 26% (0.82 vs 1.11) and 36% (0.77 vs 1.21), respectively. Furthermore, our method demonstrates excellent performance under different lighting conditions. In morning, overcast, and sunset scenarios, our approach achieves improvements of 27%, 30%, and 28% respectively compared to the second-best methods. Additionally, our method exhibits consistent performance across all scenarios, which further validates the effectiveness of our approach and its superior generalization capability to different degraded conditions. Moreover, we present visualization results for different VKITTI2 scenarios in Fig. 4. The visualization results in Fig. 4 demonstrate that our method maintains good robustness when confronted with adverse weather conditions and extreme lighting situations.

Exp.	$\mathcal{L}_{sup}$	$\mathcal{L}_C$	$\mathcal{L}_W$	$\mathcal{L}_{Dec}$	Dataset	
					FCDN	VBOF(All)
1.	✓	✗	✗	✗	1.01	5.65
2.	✓	✓	✗	✗	0.97	5.12
3.	✓	✓	✓	✗	0.94	4.91
4.	✓	✗	✗	✓	0.97	5.69
5.	✓	✓	✓	✓	**0.85**	**4.57**

Table 3: Ablation study of different distillation schemes.

Ablation Study for Fine-tuning Different Layers. As aforementioned, preserving the original architecture and weights is crucial for managing the robust flow estimation capabilities of CroCo in degraded scenario optical flow estimation tasks. To validate this principle, we conduct an ablation study demonstrating the network performance after retraining with different numbers of fine-tuning layers unfrozen using our training strategy. As shown in Tab. 2, we present the results with an increasing number of encoder-decoder fine-tuning layers in the CroCo framework. We observe a gradual performance improvement in the initial stages, indicating that fine-tuning certain layers indeed enhances performance. However, as the number of retraining parameters continues to increase, the optical flow estimation capability of the network for degraded scenarios progressively deteriorates.

Ablation Study for Adding Different Distillation. We validate the effectiveness of each distillation strategy through ablation experiments. First, through supervised learning, the network can effectively acquire degradation knowledge. Second, by introducing the intra-frame distillation strategy, the optical flow estimation accuracy demonstrates notable improvement. This validates that our strategy can effectively learn invariant motion patterns during the per-frame encoding stage. Finally, with the application of inter-frame distillation, the student model further learns invariant geometric correlation knowledge.

Conclusion

In this work, we present FlowAnytime, a novel approach for optical flow estimation in degraded scenarios. By fine-tuning well-pretrained optical flow foundation models and incorporating intra-frame and inter-frame distillation strategies, our method achieves robust optical flow estimation across arbitrary degraded conditions. Extensive and comprehensive experiments demonstrate the effectiveness of our proposed approach. Through this framework, we aim to establish a simple yet effective paradigm for optical flow estimation in degraded scenarios.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (61871326, 62222206, 62272209 and U2441241), the Key Research and Development Program of Jiangxi Province (20232BBE50006) and the Major Research and Development Project of Jiangxi Province (20232ACC01007).

References

Cabon, Y.; Murray, N.; and Humenberger, M. 2020. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*.

Chen, T.; Li, L.; Saxena, S.; Hinton, G.; and Fleed, D. J. 2023. A Generalist Framework for Panoptic Segmentation of Images and Videos. In *IEEE/CVF International Conference on Computer Vision*, 909–919.

Dai, W.; Wu, H.; Weng, X.; Zheng, Y.; Ming, Y.; and Kong, W. 2025. Multi-Modal Synergistic Implicit Image Enhancement for Efficient Optical Flow Estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2173–2182.

Deng, C.; Luo, A.; Huang, H.; Ma, S.; Liu, J.; and Liu, S. 2023. Explicit motion disentangling for efficient optical flow estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9521–9530.

Du, S.; Zou, Y.; Wang, Z.; Li, X.; Li, Y.; Shang, C.; and Shen, Q. 2024. Unsupervised Hyperspectral and Multispectral Image Fusion via Self-Supervised Modality Decoupling. *arXiv preprint arXiv:2412.04802*.

Jiang, S.; Campbell, D.; Lu, Y.; Li, H.; and Hartley, R. 2021. Learning to Estimate Hidden Motions with Global Motion Aggregation. *IEEE/CVF International Conference on Computer Vision*, 9752–9761.

Li, H.; Luo, K.; and Liu, S. 2021. Gyroflow: Gyroscope-guided unsupervised optical flow learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12869–12878.

Li, R.; Tan, R. T.; Cheong, L.-F.; Aviles-Rivero, A. I.; Fan, Q.; and Schonlieb, C.-B. 2019. Rainflow: Optical flow under rain streaks and rain veiling effect. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7304–7313.

Liu, Y.; Zou, Y.; Li, X.; Zhu, X.; Han, K.; Jiang, Z.; Ma, L.; and Liu, J. 2025. Toward a Training-Free Plug-and-Play Refinement Framework for Infrared and Visible Image Registration and Fusion. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 1268–1277.

Luo, A.; Yang, F.; Li, X.; and Liu, S. 2022a. Learning optical flow with kernel patch attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8906–8915.

Luo, A.; Yang, F.; Luo, K.; Li, X.; Fan, H.; and Liu, S. 2022b. Learning optical flow with adaptive graph reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 1890–1898.

Park, M.; Kim, H. G.; Lee, S.; and Ro, Y. M. 2021. Robust Video Frame Interpolation With Exceptional Motion Map. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(2): 754–764.

Qin, Z.; Zhou, S.; Wang, L.; Duan, J.; Hua, G.; and Tang, W. 2023. MotionTrack: Learning Robust Short-Term and Long-Term Motions for Multi-Object Tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17939–17948.

Teed, Z.; and Deng, J. 2020. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. *European Conference on Computer Vision (ECCV)*, 402–419.

Wang, Z.; Zhang, C.; Chen, Z.; Chen, H.; Ge, L.; and Lu, K. 2025. MotionFlow: Joint Motion Priors and Appearance Enhancement for High-Accuracy Optical Flow Estimation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Wang, Z.; Zhang, C.; Chen, Z.; Hu, W.; Lu, K.; Ge, L.; and Wang, Z. 2024. ACR-Net: Learning High-Accuracy Optical Flow via Adaptive-Aware Correlation Recurrent Network. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10): 9064–9077.

Weinzaepfel, P.; Lucas, T.; Leroy, V.; Cabon, Y.; Arora, V.; Brégier, R.; Csurka, G.; Antsfeld, L.; Chidlovskii, B.; and Revaud, J. 2023. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17969–17980.

Xu, H.; Yang, J.; Cai, J.; Zhang, J.; and Tong, X. 2021. High-resolution optical flow from 1d attention and correlation. In *IEEE/CVF International Conference on Computer Vision*, 10498–10507.

Xu, H.; Zhang, J.; Cai, J.; Rezatofighi, H.; and Tao, D. 2022. Gmflow: Learning optical flow via global matching. In *IEEE/CVF conference on computer vision and pattern recognition*, 8121–8130.

Yan, W.; Sharma, A.; and Tan, R. T. 2020. Optical flow in dense foggy scenes using semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13259–13268.

Yoon, J.; Kim, S.; Kwak, S.; and Cho, M. 2024. Optical flow domain adaptation via target style transfer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2111–2121.

Zhang, M.; Zheng, Y.; and Lu, F. 2022. Optical Flow in the Dark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9464–9476.

Zhao, S.; Zhao, L.; Zhang, Z.; Zhou, E.; and Metaxas, D. 2022. Global matching with overlapping attention for optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17592–17601.

Zheng, Y.; Zhang, M.; and Lu, F. 2020. Optical flow in the dark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6749–6757.

Zhou, H.; Chang, Y.; Liu, H.; Yan, W.; Duan, Y.; Shi, Z.; and Yan, L. 2024a. Exploring the Common Appearance-Boundary Adaptation for Nighttime Optical Flow. *arXiv preprint arXiv:2401.17642*.

Zhou, H.; Chang, Y.; Shi, Z.; Yan, W.; Chen, G.; Tian, Y.; and Yan, L. 2024b. Adverse Weather Optical Flow: Cumulative Homogeneous-Heterogeneous Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zou, Y.; Chen, Z.; Zhang, Z.; Li, X.; Ma, L.; Liu, J.; Wang, P.; and Zhang, Y. 2026. Contourlet refinement gate framework for thermal spectrum distribution regularized infrared image super-resolution. *International Journal of Computer Vision*.

Zou, Y.; Li, X.; Jiang, Z.; and Liu, J. 2024. Enhancing neural radiance fields with adaptive multi-exposure fusion: A bilevel optimization approach for novel view synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7882–7890.

Zuo, F.; Xiao, Z.; Jin, H.; and Su, H. 2024. CEDFlow: latent contour enhancement for dark optical flow estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7909–7916.