This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Multi-Modal Synergistic Implicit Image Enhancement for Efficient Optical Flow Estimation

Weichen Dai[1] Hexing Wu[1], Xiaoyang Weng[1], Yuxin Zheng[2], Yuhang Ming[1], and Wanzeng Kong[1]*
[1]Hangzhou Dianzi University [2]Tsinghua University *

Abstract

As a fundamental visual task, optical flow estimation has widespread applications in computer vision. However, it faces significant challenges under adverse lighting conditions, where low texture and noise make accurate optical flow estimation particularly difficult. In this paper, we propose an optical flow method that employs implicit image enhancement through multi-modal synergistic training. To supplement the scene information missing in the original low-quality image, we utilize a high-low frequency feature enhancement network. The enhancement network is implicitly guided by multi-modal data and the specific subsequent tasks, enabling the model to learn multi-modal knowledge that enhances feature information suitable for optical flow estimation during inference. By using RGBD multi-modal data, the proposed method avoids the reliance on the images captured from the same view, a common limitation in traditional image enhancement methods. During training, the encoded features extracted from the enhanced images are synergistically supervised by features from the RGBD fusion as well as by the optical flow task. Experiments conducted on both synthetic and real datasets demonstrate that the proposed method significantly improves performance on public datasets.

1. Introduction

In the field of computer vision, optical flow estimation [1] is an important foundational task, widely applied in areas such as pose estimation [31, 38], and object tracking [26]. Traditional optical flow estimation methods typically rely on assumptions such as photometric consistency [1, 2]. These methods have made significant progress in both computational efficiency and processing accuracy, and they were once the mainstream methods for early optical flow estimation. With the rapid development of deep learning, optical flow estimation techniques based on deep learning have gradually become a research focus [7, 15, 18, 44]. Recent methods such as FlowNet [10], PWC-Net [29], and RAFT [30] have demonstrated outstanding performance in optical flow estimation tasks [6].

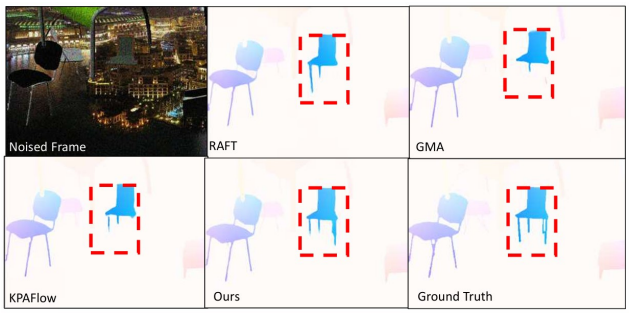


Figure 1. Qualitative assessment of samples from the noised Flychair dataset[42]. Regions with significant improvements are highlighted in red dashed boxes.

In practice, varying lighting conditions, particularly low-light environments, present substantial challenges. Under these adverse conditions, image quality deteriorates rapidly [34], violating the photometric consistency assumption crucial to traditional optical flow methods [42]. Additionally, images captured in adverse illumination often suffer from increased noise and a loss of texture details [4], which reduces texture consistency across frames and complicates optical flow estimation.

To address the challenges posed by adverse illumination and enhance performance, most methods focus on developing advanced feature extraction networks to capture distinctive features for optical flow computation [42, 46]. However, the information available in images captured under poor lighting conditions is inherently limited, imposing an upper bound on the quality of features these networks can extract [43]. Consequently, achieving feature quality sufficient for subsequent steps, such as correlation calculations, becomes challenging, creating a bottleneck in the optical flow process. Thus, a crucial step toward enhancing optical flow performance lies in supplementing the data with high-quality information to support more accurate optical flow estimation.

*Corresponding author

To enhance low-quality images for better optical flow results, many methods focus on using image enhancement techniques. However, these methods require the images captured from the same view, as it is challenging to obtain large quantities of high- and low-quality images from identical viewpoints. Therefore, some methods attempt to improve images using style transfer techniques based on GAN [11, 45], which do not require the image pairs with strict geometric alignment. However, GAN-generated images often lack geometric consistency [14]. Other methods attempt to simulate low-light conditions by artificially generating low-light data from normal-light datasets [42]. However, this synthetic data still differs significantly from real low-light data. Additionally, so-called high-quality data is typically defined based on human visual perception, and a brighter image does not necessarily yield better results for optical flow estimation.

Depth images, which are easily obtainable and can be aligned with RGB images, not only provide more high-quality scene information but also help address the challenge of acquiring sufficient image enhancement data. In this paper, we propose a novel multi-modal synergistic implicit image enhancement method for efficient optical flow estimation. The image enhancement model is implicitly optimized by high-quality data synergistically learned from the fusion of RGB and depth images, thereby improving the image enhancement ability to boost image quality. We first implement a high-low frequency feature enhancement network. Following the output of the enhanced images, iterative optical flow estimation is performed to calculate the correlation matrix and estimate the flow iteratively. The proposed method is trained end-to-end from two perspectives. First, the encoded image and context features based on the enhanced images are supervised by high-quality features resulting from the synergistic learning of RGB and depth images. Meanwhile, the optical flow estimation loss is also backpropagated to the enhancement network. As a result, the image enhancement network is implicitly guided to supplement the image quality with multi-modal knowledge, improving the quality of the extracted features during inference, and enhancing the performance of RGB optical flow estimation.

The main contributions are as follows:

- We propose a novel multi-modal synergistic implicit image enhancement method for efficient optical flow estimation, which leverages multi-modal synergistic learning to implicit supervise the network for RGB images, improving the performance of the optical flow estimation.
- We introduce a multi-modal synergistic learning method that utilizes RGBD fusion with 3D to 2D projection, ensuring that the enhanced features retain geometric interpretability. This method also makes the image enhancement feasible for tasks that require geometric consistency.
- We demonstrate through extensive experiments on synthetic and real datasets that the proposed proposed method significantly improves the performance of optical flow estimation.

2. RELATED WORK

2.1. Optical Flow

Optical flow focuses on estimating the pixel-wise 2D motion between consecutive frames, describing the motion field in the 2D space. Traditional optical flow algorithms typically rely on the assumption of photometric consistency and derive the motion by optimization [1]. Deep learning-based optical flow methods [21, 22, 36] initially introduced FlowNet, which uses two convolutional neural networks to estimate optical flow from paired images, eliminating the manual feature extraction step found in traditional methods and demonstrating the feasibility of learning optical flow directly from images. With the development of deep learning, the design of optical flow estimation networks has become increasingly diverse [5]. PWC-Net [29] employs a spatial pyramid structure to estimate optical flow progressively across multiple levels, introducing a feature pyramid and cost volume-based matching, which improves the accuracy of optical flow estimation while maintaining computational efficiency. RAFT [30] utilizes a novel network structure that iteratively updates optical flow through a multi-scale 4D correlation volume lookup with GRU, while GMA [13] enhances the global motion matching capability of feature extraction and encoding through global motion information aggregation and a local attention mechanism for optical flow calculation.

Existing optical flow methods are primarily designed for image data captured under normal environmental conditions. However, in complex environments, image data often suffer from information loss. For example, in low-light conditions, increased noise and reduced texture prominence degrade image quality, disrupting the photometric consistency assumption that optical flow estimation relies on. The increase in noise and the lack of texture detail reduce the texture consistency between frames [40], resulting in decreased performance of subsequent optical flow estimation [42] networks.

2.2. Image Restoration

Image restoration has long been an important topic in visual tasks, aiming to recover a clean underlying image from degraded observations. Most conventional image restoration (IR) methods can be classified within the Maximum a Posteriori (MAP) framework, a Bayesian paradigm that includes a likelihood (loss) term and a prior (regularization) term. Commonly used L2 or L1 losses in this framework correspond to Gaussian or Laplacian assumptions about im-

age noise. To depict noise distributions more accurately, several more robust alternatives have been introduced, such as the Poisson-Gaussian model[8] and Gaussian Mixture Model (GMM)[24]. Simultaneously, in recent years, there has been growing interest in leveraging image priors to address challenges in image restoration. Typical image priors include total variation[27], wavelet filtering[28], non-local similarity[3], sparsity, and low-rankness, among others. These traditional methods are mainly limited by the model capacity and the subjectivity inherited from manually designed assumptions regarding noise and priors.

In recent years, the field of image restoration has gradually been dominated by deep learning methods. Early pioneering research proposed using convolutional neural networks to address image restoration problems, outperforming traditional model-based methods in tasks such as image denoising[41], super-resolution[32], and deblurring[25]. Subsequently, many studies have focused on designing more refined network architectures to further improve restoration performance, such as the transformer architecture[23, 33]. At the same time, methods have been proposed that solve the problem using learnable image priors, such as generating network priors based on non-local similarity[39] or using transfer learning with networks trained on large datasets. The introduction of these methods has substantially expanded the boundaries of image restoration tasks.

2.3. Optical flow in Challenge Environments

There are already many methods for optical flow estimation in challenging environments, such as those under adverse weather conditions or low-light environments. For example, in rainy scenes, RobustNet[16] estimates optical flow by utilizing a rain-free residual channel, while RainFlow[17] generates rain streak and rain-fog invariant features to reduce the impact of rain streaks and fog on image features. For foggy scenes, [37] performs both image dehazing and optical flow estimation simultaneously by synthesizing foggy data and applying style transfer.

In low-light environments, Zheng et al. [42] proposed a synthetic dark noise optical flow benchmark dataset, called the FlyingChairs Dark&Noise (FCDN) dataset, which addresses the issue of missing datasets by adding dark image noise to a normally lit dataset. They also introduced the Various Brightness Optical Flow (VBOF) dataset, which includes images at different exposure levels along with optical flow pseudo labels. Additionally, CEDFlow [46] introduces high- and low-frequency adaptive enhancements, as well as edge enhancement of feature dimensions, designing a structure specifically for image feature extraction in low-light environments to improve optical flow estimation performance. These methods train optical flow estimation networks using synthetic images with dark noise. However, relying solely on the limited information in low-light images to extract image features restricts the performance of optical flow estimation in such environments. The reduced light intake in low-light conditions makes it difficult to extract essential information [20], which hinders the extraction of high-quality features required for subsequent optical flow calculations. Furthermore, since the network is only trained using low-quality images, this reduces the performance that would normally be achieved under normal lighting conditions. Therefore, effectively utilizing known knowledge priors to guide the enhancement of image features and supplement the information needed for subsequent optical flow calculations is crucial for achieving efficient optical flow estimation.

3. Method

As shown in Fig. 2, the proposed method consists of three main components: a high-low frequency feature enhancement network, a 2D-3D feature fusion network, and an iterative optical flow estimation module. First, the high-low frequency feature enhancement network decomposes the image into high and low frequency features, which are then processed separately by a dense convolutional network and a multi-scale feature enhancement network. These features are then fused via a channel attention mechanism to produce enhanced features, which are further processed by a convolutional neural network to compute a 4D correlation matrix between two RGB images, aiding optical flow estimation. High-low frequency feature enhancement network independently addresses damaged image texture details and edge contours, improving the feature extraction capabilities.

To ensure that the high-low frequency feature enhancement network significantly improves feature quality, a pretrained 2D-3D feature fusion network extracts features from RGB images and point clouds(depth images), using learnable interpolation to fuse point cloud features into image features. This setup supervises the feature extraction process of the enhancement network and implicitly guides the high-low frequency feature enhancement network during training. This method enhances the the images with point cloud information to aid optical flow estimation for RGB images.

3.1. High-Low Frequency Feature Enhancement Network

The high-frequency feature enhancement network first decomposes the image into high-frequency and low-frequency information. These components are then adaptively enhanced using a dense convolutional network for high-frequency information and a multi scale feature enhancement network for low-frequency information, allowing for better handling of details and contours. Subsequently, the enhanced high-frequency and low-frequency features are

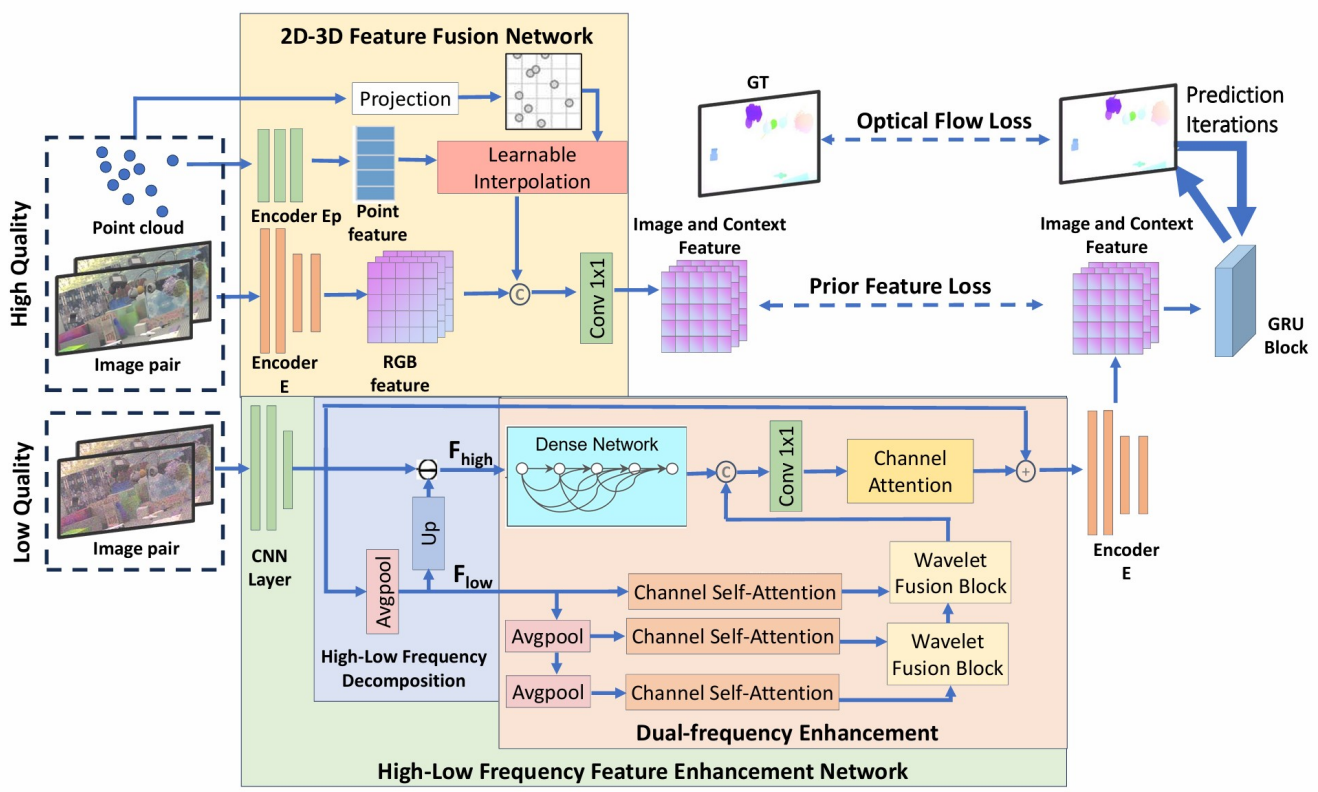


Figure 2. Overview. The proposed method mainly includes a high-low frequency feature enhancement network, 2D-3D feature fusion network and optical flow estimation of a GRU block and supervised by dual types of loss.

fused through channel attention, leveraging inter-channel attention mechanisms to enhance the network representation capability.

3.1.1. High-Low Frequency Decomposition

We first use convolution to extract image features $\mathbf{F}$. To separate the high-frequency and low-frequency information of the features, we first use AvgPooling to obtain the low-frequency information. Then, by subtracting the low-frequency information from the overall features, we obtain the high-frequency information:

$$\mathbf{F}_{low} = \text{Avgpool}(\mathbf{F}) \tag{1}$$

$$\mathbf{I}_{high} = \mathbf{F} - \text{Upsampling}(\mathbf{F}_{low}) \tag{2}$$

where Avgpool$(\cdot)$ represents average pooling, and Upsampling$(\cdot)$ denotes bilinear interpolation upsampling.$\mathbf{F}_{low}$ represents the separated low-frequency information, while $\mathbf{F}_{high}$ represents the separated high-frequency information.

This process ensures that we can handle the high-frequency and low-frequency information separately, applying the most suitable enhancement method for each.

3.1.2. Dual-frequency Enhancement

We separately process the extracted high-frequency and low-frequency features. Since high-frequency information represents the details of the image, we need to use smaller receptive fields to better focus on the local image information, enabling more precise enhancement of the details. For the high-frequency feature $\mathbf{F}_{high}$, we obtain enhanced feature $\mathbf{F}'_{high}$ through a dense network with multiple small convolutional kernels:

$$\mathbf{F}''_{high} = \text{Dense}(\mathbf{F}_{high}) \tag{3}$$

where Dense$(\cdot)$ represents dense convolutional network.

Small convolutional kernels help better focus on detailed areas, while residual connections excel at exploring high-frequency information. Therefore, this combination is highly effective for extracting high-frequency information.

For low-frequency information, which includes the background and contours of the image, it benefits from considering long-range dependencies. We use a multi-scale feature enhancement network to capture these long-range interactions. The input low-frequency feature $\mathbf{F}_{low}$ is enhanced through the channel self-attention transformer block to obtain $\mathbf{F}'_{low}$:

$$\mathbf{F}'_{low} = \text{CA}(\mathbf{F}_{low}) \tag{4}$$

where CA(·) represents channel self-attention transformer block.

The processing is similarly performed across different scales,and finally, the multi-scale information is effectively fused through a wavelet fusion network:

$$\mathbf{F}''_{low} = \mathrm{WF}(\mathbf{F}'_{low}, \mathrm{WF}(\mathbf{F}'_{low1}, \mathbf{F}'_{low2})) \tag{5}$$

where WF(·) represents the wavelet fusion block, and $\mathbf{F}'_{\text{low1}}$ and $\mathbf{F}'_{\text{low2}}$ denote the features obtained by downsampling $\mathbf{F}_{\text{low}}$ once and twice, respectively, followed by processing through CA(·).

Specifically, the wavelet fusion (WF) module is implemented using Discrete Wavelet Transform (DWT) and Inverse Discrete Wavelet Transform (IDWT). The DWT decomposes the large-scale feature information into {HH, HL, LH, LL}, where the original small-scale information is first fused with LL. Then, the merged information is combined with {HH, HL, LH} through the IDWT. This approach helps prevent the loss of structural information that may occur when directly upsampling small-scale feature maps and merging them with large-scale feature maps.

The obtained high and low frequency information features are adaptively weighted and fused using channel attention through residual connections to obtain enhanced features Fen:

$$\mathbf{F}_{en} = \mathrm{Conv}(\mathrm{CA}(\mathrm{Conv}([\mathbf{F}''_{low}, \mathbf{F}''_{low}]))) + \mathbf{F} \tag{6}$$

where Conv(·) represents convolutional layer, and CA(·) denotes channel attention block.

3.2. Iterative Optical Flow Estimation

The 4D correlation volume is computed for establishing feature correspondences between pixels (4D-CV) [30], as shown in Equ. (7). Given extracted features by encoder E, the correlation volume is constructed by computing the dot product between all pairs of feature vectors. The correlation volume $\mathbf{C}$ can be efficiently computed as a single matrix multiplication.

$$\mathbf{C}_{ijkl} = \sum_{h} g_\theta(\mathbf{I}_1)_{ijh} \cdot g_\theta(\mathbf{I}_2)_{klh} \tag{7}$$

where $g_\theta(\cdot) \in \mathbb{R}^{H\times W\times D}$ is the image encoder. ij and kl denote the image position in image 1 and 2, respectively.

Using the obtained multi-scale 4D correlation volume, a GRU-based update operator is employed to iteratively update the flow estimation result by looking up values from it. Let the initial flow field $\mathbf{p}_0$ be zero, and in the k-th iteration, it produces an updated flow $\mathbf{p}_k^{\Delta}$, which is added to the current estimate. After that, we retrieve correlation features from the correlation pyramid $\mathbf{C}$ to compute the updated flow $\mathbf{p}_k^{\Delta}$:

$$\begin{cases} \mathbf{p}_k = \mathbf{p}_k^{\Delta} + \mathbf{p}_{k-1}, \\ \mathbf{p}_k^{\Delta} = \mathrm{GRU}(\mathbf{f}_c, \mathbf{C}(\mathbf{p}^{k-1})) \end{cases} \tag{8}$$

where $\mathbf{f}_c$ is the context feature of the image.

Finally, the network is supervised by the L_1 distance between the predicted flow and the ground truth flow, covering the entire prediction sequence $\{\mathbf{p}_1, \ldots, \mathbf{p}_N\}$, with exponentially increasing weights $\gamma < 1$. Given the ground truth flow $\mathbf{p}_{gt}$, the optical flow estimation loss function is

$$L_{flow} = \sum_{i=1}^{N} \gamma^{N-i} \|\mathbf{p}_{gt} - \mathbf{p}_i\|_1 \tag{9}$$

3.3. Implicit Guidance for Feature Enhancement

3.3.1. 2D-3D Feature Fusion Network

The 2D-3D feature fusion network first extracts the 2D image features $\mathbf{F_{2D}} \in \mathbb{R}^{H\times W\times C_{2D}}$ by using convolutional network, as well as the 3D point features $\mathbf{G_{3D}} = \{g_i | i = 1, \ldots, N\} \in \mathbb{R}^{N\times C_{3D}}$ and point positions $\mathbf{P} = \{p_i | i = 1, \ldots, N\} \in \mathbb{R}^{N\times 3}$, where $\mathbf{N}$ denotes the number of points. Since image features are dense and point cloud features are sparse, alignment is required before fusion. Learnable interpolation [19] is applied to the point cloud features to make them dense and result in the features $\mathbf{F'_{3D}} \in \mathbb{R}^{H\times W\times C_{2D}}$. As illustrated in Fig. 2, for each target pixel in the dense image, the k nearest projected points are found in the image plane, and the neighboring features are weighted using ScoreNet based on the coordinate offsets.

$$\mathbf{F'_{3D_{ij}}} = \mathrm{KNN}_{ij}(\mathbf{F_{3D}} \circ \mathrm{ScoreNet}(\mathbf{P})) \tag{10}$$

After that, the 1x1 convolutional network fuses $\mathbf{F_{2D}}$ with $\mathbf{F'_{3D}}$ to obtain the features $\mathbf{f}^{ig}$ used for implicit guidance.

$$\mathbf{f^{ig}} = \mathrm{Conv}(\mathbf{F_{2D}} + \mathbf{F'_{3D}}) \tag{11}$$

3.3.2. Implicit Guidance Loss

The enhanced feature $\mathbf{F}_{\text{en}}$ obtained through the high-low frequency feature enhancement network, as introduced in Sec. 3.1, is fed through the feature extraction module of the optical flow estimation network to extract the image features $\mathbf{f}_i^l$ and context features $\mathbf{f}_c^l$, respectively. The corresponding RGB image and point cloud are passed through the point cloud feature fusion network with pre-trained parameters, to extract higher quality image features $\mathbf{f}_i^{ig}$ and context features $\mathbf{f}_c^{ig}$. Subsequently, the prior feature loss is calculated as follows:

$$L_{prior} = \left\|\mathbf{f}_i^l - \mathbf{f}_i^{ig}\right\|_2^2 + \left\|\mathbf{f}_c^l - \mathbf{f}_c^{ig}\right\|_2^2 \tag{12}$$

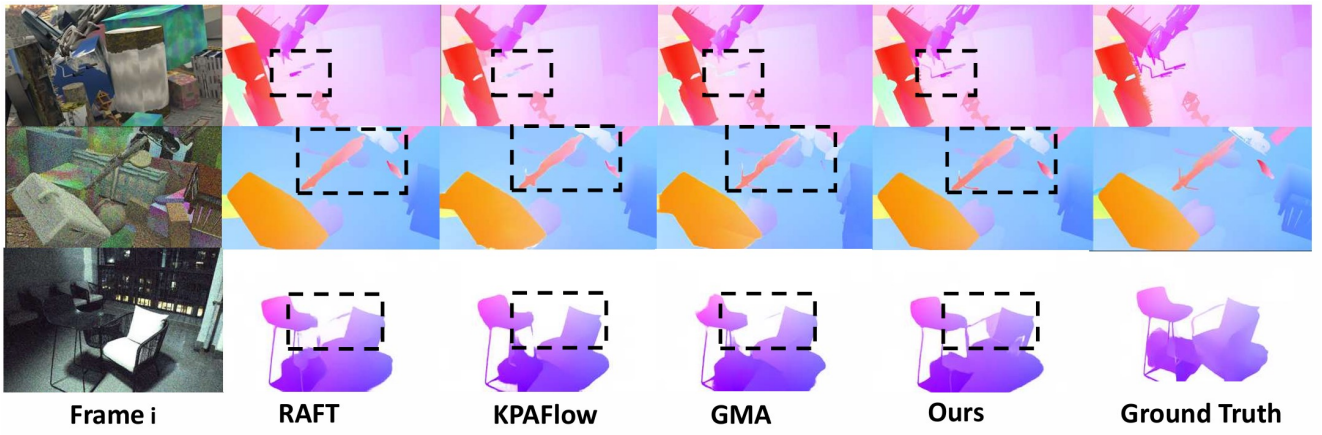


Figure 3. Qualitative comparison of different flow estimations. In difficult environments with noise, the proposed method can provide better results compared to other methods. Regions with significant improvements are highlighted in black dashed boxes.

Table 1. Performance comparison on the "val" split of the FlyingThings3D subset and the noised FlyingThings3D subset. "C" and "N" denote the clean and noised images of the FlyingThings3D subset training set, respectively.

Method	Training Data	Things(clean)		Training Data	Things(clean)		Things(noise)	
		$EPE_{2D} \downarrow$	$ACC_{1px} \uparrow$		$EPE_{2D} \downarrow$	$ACC_{1px} \uparrow$	$EPE_{2D} \downarrow$	$ACC_{1px} \uparrow$
RAFT [30]	C	3.61	81.02%	C+N	3.77	80.34%	4.41	73.11%
GMA [13]	C	3.35	83.29%	C+N	3.54	81.06%	4.22	73.77%
KPAFlow [21]	C	3.19	84.53%	C+N	3.81	80.07%	4.48	72.96%
GMFlow [36]	C	3.21	84.24%	C+N	3.50	**83.85%**	4.18	74.02%
Ours	C	**2.91**	**86.54%**	C+N	**3.25**	83.49%	**3.85**	**76.66%**

3.4. Implementation Details

In the implementation, we set γ to 0.9, combined with multiple optical flow prediction iterations ($N = 12$), to achieve a better coarse-to-fine optical flow update.

The final loss is set as the sum of the prior feature loss and the optical flow estimation loss:

$$\mathcal{L} = \sum_{b=1}^{\mathbf{B}} (L_{\text{flow},b} + \lambda L_{\text{prior},b}) \tag{13}$$

where $\mathbf{B}$ is the number of batches, and λ is set to 0.2

We implemented the model in this paper using PyTorch. The model is based on modifications to RAFT and is trained using the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) for a total of 200k epochs. The initial learning rate is set to 2×10^{-4} and is gradually reduced to 1×10^{-6} during training through a cosine annealing strategy.

4. Experiments

4.1. Dataset

We trained on a subset of the FlyingThings3D dataset with injected noise and tested on both this dataset, Sintel dataset and the VBOF (Variable Brightness Optical Flow) dataset. FlyingThings3D is a synthetic dataset specifically designed for optical flow estimation in 3D scenes, comprising pairs of images with dynamic 3D objects in complex backgrounds, along with their corresponding ground-truth optical flow data. The images in this dataset are generated from synthetic scenes featuring flying objects, simulating various camera viewpoints and lighting conditions. The noise-injected dataset is created based on FlyingThings3D by generating noisy optical flow images that simulate adverse illumination conditions. This is done by introducing uncorrected white balance effects and adding noise models, resulting in synthetic images with low-light noise characteristics alongside their ground-truth optical flow data. The VBOF dataset, on the other hand, includes multiple sets of images under varying exposure conditions for the same scenes, captured by four different cameras (Canon, Fuji, Nikon, and Sony), and provides ground-truth optical flow data for each exposure condition.

The FlyingThings3D dataset does not contain point cloud files by default. Following the unprojection method[9, 35], we construct the corresponding point cloud data using camera parameters and depth image files.

Table 2. EPE comparisons on VBOF

Method	Sony	fuji	Canon	Nikon	All
RAFT[30]	20.64	14.40	19.82	24.37	20.23
GMA[13]	20.81	15.24	20.41	25.30	21.14
KPAFlow[21]	21.73	15.51	21.76	25.74	22.48
GMFlow [36]	20.37	13.87	**19.48**	**23.49**	20.11
Ours	**20.05**	**13.72**	19.73	23.68	**19.99**

4.2. Comparison of the Accuracy

4.2.1. Baseline

These methods include RAFT, GMA, KPAFlow, and GMFlow. All models were fairly trained on the train split of clean and noised FlyingThings3D subset.

4.2.2. Experiment on clean and noised FlyingThings3D subset

Table 1 presents the evaluation results of five models trained on the clean and noised FlyingThings3D subset, assessed on their valset in terms of End-Point Error (EPE). In columns 3 and 4, our model achieves an EPE of 2.91 and an ACC_1px of 86.54% on Things (clean), outperforming all other methods. Compared to the second-ranked GMFlow, our model improves EPE by approximately 9.3% (3.21 $\rightarrow$ 2.91), and by about 19.3% compared to RAFT (3.61 $\rightarrow$ 2.91). These results indicate that introducing implicit feature supervision during training can indirectly guide the model to learn effective feature extraction, significantly enhancing subsequent optical flow estimation.

In columns 6, 7, 8, and 9 of Table 1, the models are trained and tested on both the Things(clean) and Things(noise) datasets. Our model achieves the best performance across all metrics. Compared to GMFlow, our model improves the EPE metric by 7.1% (3.50 $\rightarrow$ 3.25) and 7.8% (4.18 $\rightarrow$ 3.85). These results demonstrate that, with the support of the enhanced network and prior feature guidance, our model excels in addressing optical flow estimation challenges compared to models trained solely on RGB images.

4.2.3. Experiment on VBOF Datasets

To compare the performance of our model in low-light optical flow estimation using synthetic data, we conducted evaluations on the VBOF dataset. Table 2 shows the EPE metrics for various methods on the VBOF dataset and its SONY subset. Our model achieved the best results on the Sony and Fuji subsets, with an EPE of 20.05 in the Sony scenario, representing an improvement of approximately 1.5% over the second-best GMFlow (20.37 $\rightarrow$ 20.05). In the Fuji scenario, our model achieved an EPE of 13.72, outperforming GMFlow by 1.08%. On the Canon and Nikon subsets, our model obtained the second-best results. The varying degrees of improvement across different subsets can be attributed to the differences in noise distribution among the Sony, Fuji, Canon, and Nikon subsets within the VBOF dataset. Following Zheng's data synthesis [42], the noise distribution can be parameterized using Poisson and Gaussian noise models. The synthetic low-light noise dataset used in our method was generated based on the noise distribution of the Sony subset. As a result, our model showed the most significant improvement on the Sony subset, while the Fuji subset, whose noise distribution is relatively similar to Sony, as shown in Figure 4 of the paper [42], also achieved a competitive improvement. This suggests that performance in real-world low-light conditions is influenced by the way synthetic noise is generated. In other words, the generalization should be improved by increasing the diversity of noise in the synthetic dataset.

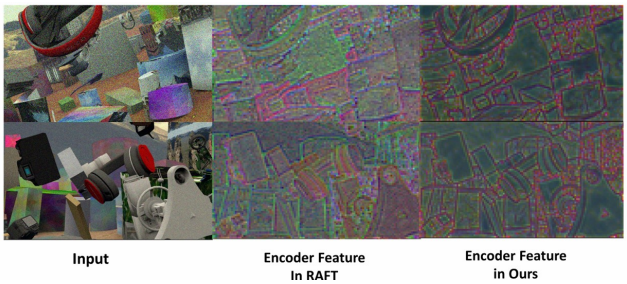


Figure 4. Feature visualization, presented from left to right: noised RGB input images,features extracted from RAFT's Encoder, and features extracted from the enhanced feature through Encoder E

4.3. Qualitative Comparison

Fig. 3 presents a visual comparison of our optical flow estimation method with the RAFT, KPAFlow, and GMA algorithms on the clean and noised Flythings subset dataset, and VBOF dataset(trained on C+N). The first row shows the output for a clean RGB image. As seen in the figure, under the guidance of implicit features, the proposed method can effectively track the contours of small objects when estimating optical flow for moving objects. The second row presents an RGB image with injected noise. Despite the added noise, our method demonstrates resilience by accurately recovering information from the corrupted image, maintaining high precision in optical flow estimation. The third row displays images in a real low-light scenario. Even with real environmental noise, our method effectively handles optical flow estimation. In contrast, the compared methods struggle in regions with noisy, elongated object contours, highlighting the limited capability of noise-compromised image features when lacking additional prior guidance.

Fig. 4 visualizes the features extracted by ours method and RAFT. It is evident that ours method can still extract rel-

Table 3. EPE comparisons by change component on RAFT framework

Framework	EPE	1px
RAFT[30]	4.41	73.11%
Enhance Network + RAFT	4.23	73.83%
Implicit Guidance + RAFT	4.06	75.51%
Ours	3.85	76.66%

atively stable features and more prominent object contours from noise-affected images. In contrast, the feature extraction of RAFT is significantly affected by noise, reflecting the robustness of our method in addressing the optical flow estimation challenges posed by noise.

4.4. Ablation Study

To validate the effectiveness of the feature enhance network and prior feature Loss used in our model, we conducted ablation experiments by replacing components in the RAFT framework. All tested models were trained on the mixed dataset.

As shown in Table 3, we pre-processed the input images using the enhance network (2nd row), and added implicit guidance to RAFT (3rd row). Table 3 shows that the Prior Feature Loss results in an improvement of about 7.9% (4.41 $\rightarrow$ 4.06), while the enhance network results in an improvement of about 4.1% (4.41 $\rightarrow$ 4.23). With both components combined, the performance improves by approximately 12.7% (4.41 $\rightarrow$ 3.85). This indicates that both the enhance network and Prior Feature Loss contribute to enhancing the performance of the optical flow estimation network, with the best results achieved when both are used together.

4.5. Computation Analysis

Parameters: In the 2nd column of Table 4, we compare the parameter capacities of different models. Our model has 6.2M parameters, the third largest, which is smaller than CEDFlow. This is because, compared to RAFT, the only additional learnable parameters are from the lightweight enhancement network. The introduction of prior features does not add extra computation during model inference.

Memory: We also show the memory requirements of different models in Table 4. When inputting images with a resolution of 736×480, our model requires 3.1GB of memory. Considering the significant improvement in accuracy, its computational cost is acceptable when addressing the challenging problem of optical flow estimation in low-light conditions.

Table 4. Comparison of Model Parameters, and Memory Usage

Models	Param(M)	Memory(GB)
RAFT[30]	5.3	1.7
GMFlowNet [36]	9.3	3.4
GMA[13]	5.9	1.8
KPAFlow[21]	5.8	2.6
AGFlow[12]	5.6	1.9
SCV	5.3	1.6
ours	6.2	3.1

5. CONCLUSIONS

This paper proposes a novel multi-modal synergistic implicit image enhancement method aimed at improving optical flow estimation performance in challenging environments. By introducing multi-modal (RGBD) synergistic training, the proposed method effectively enhances the quality of input images through the enhancement network, enabling more accurate optical flow estimation under difficult conditions. The synergistic learning of RGB and depth images at the feature level allows the enhancement network to implicitly acquire multi-modal knowledge with geometric consistency, thereby improving the feature extraction capability for optical flow computation. Extensive experiments on both synthetic and real datasets validate the effectiveness of the proposed method, with results showing a improvement in optical flow estimation performance.

In the future, we will integrate multi-spectral information for further implicit multi-modal implicit image enhancement, guided by three-dimensional depth maps, to further improve optical flow estimation. Additionally, we will optimize the fusion loss design to allow the RGB model to gain more knowledge from multiple modalities.

Acknowledgement

This work was supported by National Science and Technology Innovation(2022ZD0208800), National Natural Science Foundation of China (62306096, U20B2074, 62471169), Zhejiang Provincial Natural Science Foundation of China (LQ22F030022), Key Research and Development Project of Zhejiang Province (2023C03026), and Key Laboratory of Brain Machine Collaborative Intelligence of Zhejiang Province (2020E10010).

References

[1] Simon Baker and Iain Matthews. Lucas-kanade 20 years on: A unifying framework. *International journal of computer vision*, 56:221–255, 2004. 1, 2

[2] Andrés Bruhn, Joachim Weickert, and Christoph Schnörr. Lucas/kanade meets horn/schunck: Combining local and

global optic flow methods. *International journal of computer vision*, 61:211–231, 2005. 1

[3] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, pages 60–65. Ieee, 2005. 3

[4] Linwei Chen, Ying Fu, Kaixuan Wei, Dezhi Zheng, and Felix Heide. Instance segmentation in the dark. *International Journal of Computer Vision*, 131(8):2198–2218, 2023. 1

[5] Gabriele Costante and Thomas Alessandro Ciarfuglia. Ls-vo: Learning dense optical subspace for robust visual odometry estimation. *IEEE Robotics and Automation Letters*, 3 (3):1735–1742, 2018. 2

[6] Qiaole Dong, Chenjie Cao, and Yanwei Fu. Rethinking optical flow from geometric matching consistent perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1337–1347, 2023. 1

[7] 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. 1

[8] Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. *IEEE transactions on image processing*, 17(10):1737–1754, 2008. 3

[9] Xiuye Gu, Yijie Wang, Chongruo Wu, Yong Jae Lee, and Panqu Wang. Hplflownet: Hierarchical permutohedral lattice flownet for scene flow estimation on large-scale point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3254–3263, 2019. 6

[10] 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. 1

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[12] Haiyan Jiang, Haoyi Xiong, Dongrui Wu, Ji Liu, and Dejing Dou. Agflow: fast model selection of penalized pca via implicit regularization effects of gradient flow. *Machine Learning*, 110(8):2131–2150, 2021. 8

[13] 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. 2, 6, 7, 8

[14] Eunah Jung, Nan Yang, and Daniel Cremers. Multi-frame gan: Image enhancement for stereo visual odometry in low light. In *Conference on Robot Learning*, pages 651–660. PMLR, 2020. 2

[15] Taewoo Kim, Kwonyoung Ryu, Kyeongseob Song, and Kuk-Jin Yoon. Loop-net: Joint unsupervised disparity and optical flow estimation of stereo videos with spatiotemporal loop consistency. *IEEE Robotics and Automation Letters*, 5(4): 5597–5604, 2020. 1

[16] Ruoteng Li, Robby T Tan, and Loong-Fah Cheong. Robust optical flow in rainy scenes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 288–304, 2018. 3

[17] Ruoteng Li, Robby T Tan, Loong-Fah Cheong, Angelica I Aviles-Rivero, Qingnan Fan, and Carola-Bibiane Schonlieb. Rainflow: Optical flow under rain streaks and rain veiling effect. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7304–7313, 2019. 3

[18] Yicheng Lin, Shuo Wang, Yunlong Jiang, and Bin Han. Breaking of brightness consistency in optical flow with a lightweight cnn network. *IEEE Robotics and Automation Letters*, 2024. 1

[19] Haisong Liu, Tao Lu, Yihui Xu, Jia Liu, Wenjie Li, and Lijun Chen. Camliflow: bidirectional camera-lidar fusion for joint optical flow and scene flow estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5791–5801, 2022. 5

[20] Xiaoning Liu, Zongwei Wu, Ao Li, Florin-Alexandru Vasluianu, Yulun Zhang, Shuhang Gu, Le Zhang, Ce Zhu, Radu Timofte, Zhi Jin, et al. Ntire 2024 challenge on low light image enhancement: Methods and results. *arXiv preprint arXiv:2404.14248*, 2024. 3

[21] Ao Luo, Fan Yang, Xin Li, and Shuaicheng Liu. Learning optical flow with kernel patch attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8906–8915, 2022. 2, 6, 7, 8

[22] Ao Luo, Fan Yang, Kunming Luo, Xin Li, Haoqiang Fan, and Shuaicheng Liu. Learning optical flow with adaptive graph reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1890–1898, 2022. 2

[23] Achleshwar Luthra, Harsh Sulakhe, Tanish Mittal, Abhishek Iyer, and Santosh Yadav. Eformer: Edge enhancement based transformer for medical image denoising. *arXiv preprint arXiv:2109.08044*, 2021. 3

[24] Deyu Meng and Fernando De La Torre. Robust matrix factorization with unknown noise. In *Proceedings of the IEEE international conference on computer vision*, pages 1337–1344, 2013. 3

[25] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017. 3

[26] Nicola A Piga, Yuriy Onyshchuk, Giulia Pasquale, Ugo Pattacini, and Lorenzo Natale. Roft: Real-time optical flow-aided 6d object pose and velocity tracking. *IEEE Robotics and Automation Letters*, 7(1):159–166, 2021. 1

[27] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992. 3

[28] Eero P Simoncelli and Edward H Adelson. Noise removal via bayesian wavelet coring. In *Proceedings of 3rd IEEE international conference on image processing*, pages 379–382. IEEE, 1996. 3

[29] 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. 1, 2

[30] 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. 1, 2, 5, 6, 7, 8

[31] Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[32] Chunwei Tian, Yong Xu, Wangmeng Zuo, Bob Zhang, Lunke Fei, and Chia-Wen Lin. Coarse-to-fine cnn for image super-resolution. *IEEE Transactions on Multimedia*, 23: 1489–1502, 2020. 3

[33] Chunwei Tian, Menghua Zheng, Wangmeng Zuo, Shichao Zhang, Yanning Zhang, and Chia-Wen Lin. A cross transformer for image denoising. *Information Fusion*, 102: 102043, 2024. 3

[34] Wencheng Wang, Xiaojin Wu, Xiaohui Yuan, and Zairui Gao. An experiment-based review of low-light image enhancement methods. *Ieee Access*, 8:87884–87917, 2020. 1

[35] Wenxuan Wu, Zhiyuan Wang, Zhuwen Li, Wei Liu, and Li Fuxin. Pointpwc-net: A coarse-to-fine network for supervised and self-supervised scene flow estimation on 3d point clouds. *arXiv preprint arXiv:1911.12408*, 2019. 6

[36] 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. 2, 6, 7, 8

[37] Wending Yan, Aashish Sharma, and Robby T Tan. Optical flow in dense foggy scenes using semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13259–13268, 2020. 3

[38] Michela Zaccaria, Fabian Manhardt, Yan Di, Federico Tombari, Jacopo Aleotti, and Mikhail Giorgini. Self-supervised category-level 6d object pose estimation with optical flow consistency. *IEEE Robotics and Automation Letters*, 8(5):2510–2517, 2023. 1

[39] Zhiyuan Zha, Xin Yuan, Jiantao Zhou, Ce Zhu, and Bihan Wen. Image restoration via simultaneous nonlocal self-similarity priors. *IEEE Transactions on Image Processing*, 29:8561–8576, 2020. 3

[40] Fan Zhang, Yu Li, Shaodi You, and Ying Fu. Learning temporal consistency for low light video enhancement from single images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4967–4976, 2021. 2

[41] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017. 3

[42] Yinqiang Zheng, Mingfang Zhang, and Feng Lu. Optical flow in the dark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6749–6757, 2020. 1, 2, 3, 7

[43] Han Zhou, Wei Dong, Xiaohong Liu, Shuaicheng Liu, Xiongkuo Min, Guangtao Zhai, and Jun Chen. Glare: Low light image enhancement via generative latent feature based codebook retrieval. *arXiv preprint arXiv:2407.12431*, 2024. 1

[44] Youjie Zhou, Guofeng Mei, Yiming Wang, Fabio Poiesi, and Yi Wan. Attentive multimodal fusion for optical and scene flow. *IEEE Robotics and Automation Letters*, 2023. 1

[45] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2

[46] Fengyuan Zuo, Zhaolin Xiao, Haiyan Jin, and Haonan Su. Cedflow: Latent contour enhancement for dark optical flow estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7909–7916, 2024. 1, 3