CFFlow: An Optical Flow Estimation Hinging on Cross-Frequency Attention

Wang Mengfei[1,2]
https://orcid.org/0009-0005-5586-4820
Zhu Dongchen[1,2*]
https://orcid.org/0000-0002-1579-3942
Wang Lei[1,2]
https://orcid.org/0009-0001-6655-188X
Li Jiamao[1,2]
https://orcid.org/0000-0002-7478-4544

[1] Bionic Vision System Laboratory, State Key Laboratory of Transducer Technology, Science and Technology on Micro-system Laboratory, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China.

[2] University of Chinese Academy of Sciences, Beijing 100049, China.

Abstract

Due to issues in training datasets and motion estimation in optical flow, existing optical flow models overly emphasize surface-level information features, leading to inaccurate optical flow estimations. Through theoretical analysis, we reveal that multi-frame approaches require more attention to interconnections among diverse information than two-frame methods. Consequently, we propose **CFFlow**, which integrates cross-frequency information into optical flow estimation networks. To address challenges posed by large displacements and small object motion in real-world scenarios, we introduce two specialized modules: Large Displacement Attention (LDA) and Small Object Displacement Attention (SODA), designed to handle distinct motion patterns. CFFlow effectively resolves small target mismatches and large displacement challenges, achieving superior optical flow estimation accuracy. On Sintel and KITTI benchmarks, our CFFlow attains an Average Endpoint Error (AEPE) of **0.99** (clean pass) and **1.65** (final pass), with an F1-all error of **4.21%**, ranking first among all three-frame and two-frame methods.

1 Introduction

Optical flow estimation is a fundamental computer vision task that estimates pixel-level correspondence between consecutive frames. As a crucial paradigm of utilizing video temporal continuity, it is widely used in a variety of downstream video-related tasks, such as video interpolation [15], motion recognition [46], object detection [10], video in-painting [42], video understanding and analysis [40].

With the advancement of neural network architectures, many optical flow methods have emerged [9, 25, 28]. However, existing frameworks face two key limitations: (1) **Limited utilization of existing datasets.** Many methods [1, 21, 38] primarily rely on appearance features and visual similarity to match, often overlooking the underlying motion structure

* Indicates corresponding author.

(a) Input Image (b) Image High-Frequency Feature (c) Image Low-Frequency Feature

Figure 1: **(a)** Input image from the KITTI dataset. **(b)** High-frequency feature preserves fine motion details and highlights fast-moving small objects. **(c)** Low-frequency feature retains global motion structure and captures large displacements more effectively.

within scenes. This limits robustness under complex motion conditions. Although some work has explored the integration of segmentation information to aid in optical flow estimation [36, 45], our paper focuses on improving optical flow estimation efficiently by leveraging frequency-domain features.

(2) **Overreliance on surface information while overlooking the potential of cross-frequency features.** Many existing methods emphasize pre-trained image features [7, 16], yet optical flow estimation often benefits more from abundant low-level cues [43, 44]. Although frequency-aware features have been applied in dense image prediction and object detection tasks [3, 32], their potential remains underexplored in optical flow. Surface features primarily capture basic visual patterns, whereas cross-frequency features encode rich multi-scale structural information, offering greater promise for motion estimation. This paper aims to bridge that gap by fully leveraging cross-frequency features in flow estimation.

In this paper, we focus on leveraging cross-frequency features to enhance optical flow estimation. Existing methods often ignore multi-band frequency information. **We define cross-frequency features as the integration of motion patterns across frequency bands and gradient variations at image edges**, capturing rich multi-scale motion and structural details. These features enhance localized motion and edge-aware flow estimation, addressing limitations of prior approaches.

To enrich cross-frequency features beyond the limitations of two-frame designs, we adopt a multi-frame framework. We decomposed input images into low-frequency (smooth structural) and high-frequency (detailed) components. As illustrated in Figure 1, low-frequency features help estimate global large displacements, while high-frequency components capture fine details, improving small-object motion estimation and reducing mismatches.

To exploit the distinct roles of cross-frequency components, we design specialized modules for each. The Image Frequency Separation (IFS) Module first decomposes inputs into low- and high-frequency components. The Large Displacement Attention (LDA) Module captures long-range motion in the low-frequency domain, while the Small Object Displacement Attention (SODA) Module focuses on fast-moving small objects in the high-frequency domain. Both modules extract motion cues in their respective domains, contributing jointly to the 4D cost volume.

In summary, the contributions of this paper are as follows:

- We define cross-frequency information in optical flow tasks for the first time and design the IFS module to analyze these components, offering a complementary perspective to traditional spatial surface information.

- To fully exploit low- and high-frequency components, we introduce the LDA and SODA modules, targeting large displacements and small object motions, respectively,

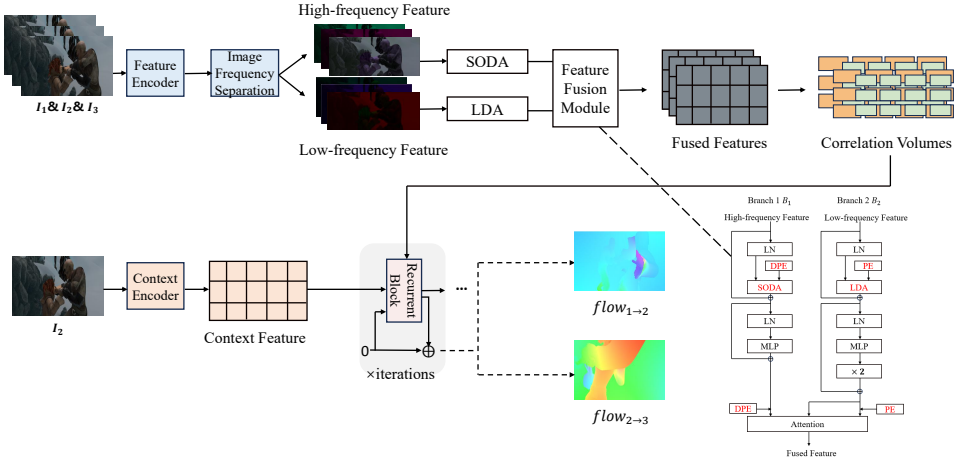


Figure 2: Overview of our LLFlow training pipeline. The IFS module decomposes image features into high- and low-frequency components. Then, SODA models small-object motion from high-frequency features, and LDA captures large displacements from low-frequency features. Finally, the FF module fuses both to produce refined motion estimation.

enhancing motion estimation across scales.

- Our approach achieves state-of-the-art results on Sintel and KITTI-2015 benchmarks, with AEPE of **0.989** and **1.647** on the clean and final passes, and an F1-all error of **4.21**%, ranking 1st among all three-frame methods.

2 Related Work

2.1 Optical Flow Estimation

Optical flow has long been a core vision task. Classical methods like Lucas–Kanade [22] and Horn–Schunck [11] formulated it as an energy optimization problem with handcrafted priors, but struggled with complex real-world motion. With deep learning and synthetic datasets, modern methods now learn optical flow end-to-end, shifting focus from manual priors to model and data design. PWCFusion [24, 29] fuses frame information via GRU-RCN, while MFCFlow [4] enhances motion features using correlation-based compensation. VideoFlow [25] leverages temporal cues for multi-frame bidirectional estimation, and SplatFlow [35] introduces splatting for differentiable motion alignment. Despite these advances, existing multi-frame approaches often emphasize high-level inter-frame features while overlooking cross-frequency motion correlations.

2.2 Cross-frequency Features

Wavelet transform decomposes signals into low-frequency Approximation Coefficients (AC) for global trends and high-frequency Detail Coefficients (DC) for local variations [19, 34, 41]. Similarly, audio processing integrates low-frequency linguistic structures with high-frequency acoustic cues [18], yet many systems rely on single-band features, ignoring their

interplay. This issue extends to optical flow, where high-frequency components capture fine motion (e.g., edges), and low-frequency ones encode global structure. Pretrained models often overemphasize low-frequency semantics [12], missing fine-grained details. To address this, we propose a frequency-aware framework that fuses high-frequency precision with low-frequency consistency for more accurate optical flow.

3 Method

3.1 Theoretical Analysis

Multi-frame optical flow methods aim to match $N : (I_1, I_2, I_3) \longmapsto (f_1, f_2)$, where I_1, I_2, and I_3 are consecutive frames, and (f_1, f_2) are the optical flows for (I_1, I_2) and (I_2, I_3). A multi-frame optical flow network can be formulated as:

$$(f_1, f_2)^* = N_\theta(I_1, I_2.I_3) = \arg\max_{(f_1,f_2)} p(f_1, f_2 \mid I_1, I_2, I_3), \tag{1}$$

where $(f_1, f_2)^*$ are the optimal estimated flows, N_θ is the network with parameters θ, and $p(f_1, f_2 \mid I_1, I_2, I_3)$ is the posterior distribution of multi-frame optical flow. Using Bayes' theorem, we expand $p(f_1, f_2 \mid I_1, I_2, I_3)$:

$$\begin{aligned} p(f_1, f_2 \mid I_1, I_2, I_3) &= \frac{p(f_1, f_2)p(I_1, I_2, I_3 \mid f_1, f_2)}{p(I_1, I_2, I_3)} \\ &= \frac{p(f_1, f_2)p(I_2 \mid f_1, f_2)p(I_1 \mid I_2, f_1, f_2)p(I_3 \mid I_1, I_2, f_1, f_2)}{p(I_1, I_2, I_3)} \\ &= \frac{p(I_2)p(f_1, f_2 \mid I_2)p(I_1 \mid I_2, f_1)p(I_3 \mid I_1, I_2, f_1, f_2)}{p(I_1, I_2, I_3)} \\ &= \frac{p(I_2)p(f_1 \mid I_2)p(f_2 \mid f_1, I_2)p(I_1 \mid I_2, f_1)p(I_3 \mid I_1, I_2, f_1, f_2)}{p(I_1, I_2, I_3)} \end{aligned} \tag{2}$$

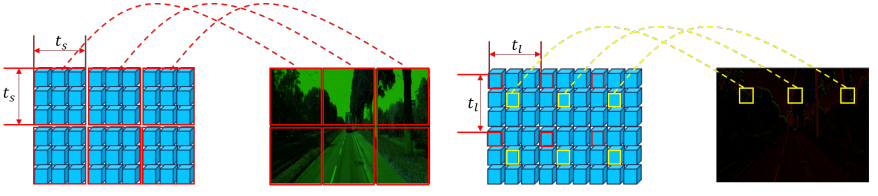
Since $p(f_1, f_2 \mid I_1, I_2, I_3)$ is a probability density of (f_1, f_2), we omit irrelevant terms $p(I_2), p(I_1, I_2, I_3)$ and apply the maximum likelihood estimation to get the following formulation:

$$\begin{aligned} (f_1, f_2)^* = \arg\max_{(f_1,f_2)} [&\underbrace{\{\log(p(f_1 \mid I_2))\}}_{\text{context}} + \underbrace{\{\log(p(f_2 \mid f_1, I_2))\}}_{\text{motion relation}} \\ &+ \underbrace{\{\log(p(I_1 \mid I_2, f_1))\}}_{\text{cost volume 1}} + \underbrace{\{\log(p(I_3 \mid I_1, I_2, f_1, f_2)\}}_{\text{cost volume 2}}]. \end{aligned} \tag{3}$$

Here, according to the definitions of cost volume and context, we get these terms. The motion relation, cost volume 1, and cost volume 2 terms capture high-frequency information, while the context term reflects low-frequency motion data. The context term aids optical flow estimation by enhancing feature understanding.

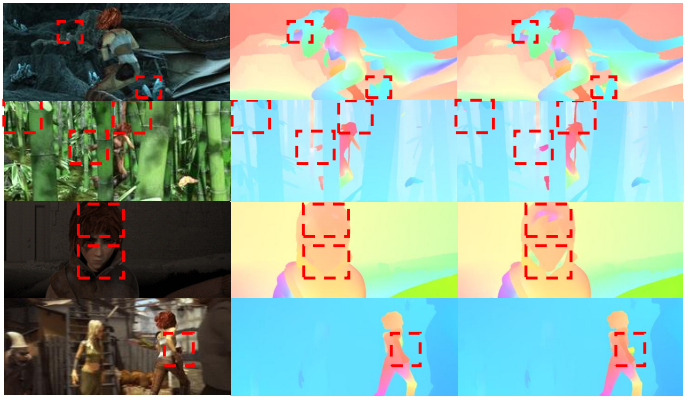
Based on the cost volume terms, it can be observed that image information can affect the estimation of optical flow. Taking into account the different effects of different image features, we consider introducing frequency information to assist in optical flow estimation.

Some methods use pre-trained visual features for high-level understanding [15], while others focus on cost volume representations to capture long-range dependencies [13, 26]. In multi-frame settings, each cost volume encodes pairwise matching cues, with high-level semantics aiding motion reasoning and low-frequency features providing spatial context. Empirical evidence shows that combining cross-frequency features with attention in patch matching preserves motion continuity effectively.



(a) SODA (b) LDA

Figure 3: **(a)** SODA module: High-frequency features (blue cubes) are grouped within red boxes. **(b)** LDA module: Low-frequency features with identical border colors indicate grouping for large displacement modeling.



Reference Frame MemFlow Ours

Figure 4: It can be seen from the red box that CFFlow better preserves the small object motion and the large displacement of the objects.

3.2 Overview

As shown in Figure 2, we redesign the optical flow feature extraction process by introducing the Image Frequency Separation Module (IFS), which decomposes extracted features into low- and high-frequency components. Built on this framework, our model CFFlow incorporates two specialized modules: SODA for small object motion and LDA for large displacement, enabling accurate flow estimation across scales. A Feature Fusion Module (FF) is then used to integrate both frequency branches. The following sections detail the design of IFS, SODA, LDA, and FF.

3.3 Image Frequency Separation Module

We introduce image frequency separation to improve robustness in complex scenes. Traditional methods process spatial features as a whole, making them prone to interference. Our IFS addresses this by decomposing features into high- and low-frequency components, handling large and small displacements. These components differ in spatial resolution. Given input features $f \in R^{H\times W}$, the high-frequency f^h and low-frequency f^l are computed as:

$$\begin{aligned} f_0^l &= \text{Downsample}(f) \in R^{H/s\times W/s}, \\ f_0^h &= f - \text{Upsample}(f_0^l) \in R^{H\times W}, \end{aligned} \tag{4}$$

Table 1: **Experiments on Sintel [2] and KITTI [9] datasets.** '*' denotes the multi-frame methods. We use **bold** and _ to highlight the methods that rank 1st and 2nd.

Training Data	Method	Sintel(train)		KITTI-15(train)		Sintel(test)		KITTI-15(test)
		Clean	Final	EPE	F1-all	Clean	Final	F1-all
C+T	FlowNet2 [14]	2.02	3.54	10.08	30.0	3.96	6.02	-
	SKFlow [30]	1.22	2.46	4.27	15.5	-	-	-
	RAFT [33]	1.43	2.71	5.04	17.4	-	-	-
	FlowFormer [15]	0.94	2.33	4.09	14.72	-	-	-
	FlowFormer++ [16]	0.90	2.30	3.93	**14.13**	-	-	-
	CRAFT [28]	1.27	2.79	4.88	17.5	-	-	-
	GMFlow [34]	1.08	2.48	-	-	-	-	-
	GMA [12]	1.30	2.74	4.69	17.1	-	-	-
	Videoflow-BOF [25]	1.03	2.19	3.96	15.33	-	-	-
	Videoflow-MOF [25]	1.18	2.56	**3.89**	14.20	-	-	-
	Ours	**0.89**	**2.11**	3.88	14.24	-	-	-
C+T+S+K(+H)	PWC-Net [29]	(1.71)	(2.34)	(1.50)	(5.3)	3.44	4.58	7.71
	GMFlow [34]	-	-	-	-	1.74	2.90	9.32
	SKFlow* [30]	(0.52)	(0.78)	(0.51)	(0.94)	1.28	2.27	4.48
	RAFT* [33]	(0.75)	(1.21)	(0.63)	(1.5)	1.94	3.18	5.11
	CRAFT [28]	(0.60)	(1.06)	(0.58)	(1.33)	1.45	2.40	4.81
	GMA [12]	(0.62)	(1.07)	(0.57)	(1.2)	1.39	2.48	5.14
	PWC-Fusion* [27]	-	-	-	-	3.43	4.57	7.17
	FlowFormer [15]	(0.48)	(0.74)	(0.53)	(1.11)	1.16	2.09	4.68
	FlowFormer++ [16]	(0.40)	(0.60)	(0.57)	(1.16)	1.07	1.94	4.52
	MatchFlow(R) [8]	(0.51)	(0.81)	(0.59)	(1.3)	1.33	2.64	4.72
	SAMFlow [35]	-	-	-	-	1.00	2.08	4.49
	MatchFlow(G) [8]	(0.49)	(0.78)	(0.55)	(1.1)	1.16	2.37	4.63
	Videoflow-BOF* [25]	(0.37)	(0.54)	(0.52)	**(0.85)**	1.02	1.84	4.44
	SplatFlow* [32]	(0.53)	(0.91)	(0.80)	(2.4)	1.12	2.07	4.61
	Ours	**(0.35)**	**(0.47)**	**(0.45)**	(0.91)	**0.99**	**1.65**	**4.21**

Table 2: **Results on the Sintel Test Set with Different Evaluation Metrics.**

Method	Sintel Test(clean)						Sintel Test(final)					
	All	Matched	Unmatched	s_{0-10}	s_{10-40}	s_{40+}	All	Matched	Unmatched	s_{0-10}	s_{10-40}	s_{40+}
MemFlow [8]	1.064	0.426	6.091	0.253	0.778	5.623	1.914	0.931	9.928	0.430	1.382	10.556
SAMFlow [35]	1.000	0.384	5.966	0.252	0.760	**5.245**	2.080	1.036	10.60	0.515	1.488	11.278
VideoFlow-BOF [25]	1.005	0.389	6.023	0.229	0.695	5.605	1.713	0.812	9.054	**0.387**	1.242	9.422
StreamFlow [31]	1.041	**0.381**	6.416	**0.212**	**0.690**	5.999	1.874	0.824	10.435	0.409	**1.240**	10.674
FlowFormer++ [16]	1.073	0.390	6.635	0.252	0.796	5.810	1.943	0.878	10.627	0.438	1.404	10.712
Ours(CFFlow)	**0.989**	0.396	**5.819**	0.229	0.694	5.470	**1.647**	**0.788**	**8.650**	0.402	1.242	**8.793**

where s is the stride of upsample and downsample. f_0^l and f_0^h represent initial low-frequency features and high-frequency features of f. In order to fully separate high-frequency and low-frequency information, f_0^l and f_0^h will be separated 4 times:

$$\begin{aligned} f_i^l &= F_{ll}(f_{i-1}^l) + \text{Downsample}(F_{hl}(f_{i-1}^h)) \in R^{H/s \times W/s}, \\ f_i^h &= F_{hh}(f_{i-1}^h) + \text{Upsample}(F_{lh}(f_{i-1}^l)) \in R^{H \times W}, \end{aligned} \tag{5}$$

Here, $i = 1,2,3,4$. F_{ll}, F_{hl}, F_{lh}, and F_{hh} denote convolutions between low/high-frequency input and output channels, respectively. The final high- and low-frequency components of input f are given by f_4^h and f_4^l.

3.4 Large Displacement Attention Module

We divide the self-attention mechanism into two parts: Large Displacement Attention Module (LDA) and Small Object Displacement Attention Module (SODA), which handle f_4^l and f_4^h respectively. LDA is designed to handle the low-frequency component f_4^l, primarily to capture long-range attention. Specifically, feature grouping is required. Since most large displacements require long-range

Table 3: **Results of Different Image Frequency Separation Modules.**

Methods	Sintel(train)		KITTI(train)	
	clean	final	F1-epe	F1-all
OctaveConv [5]	0.39	0.54	**0.48**	1.09
Haar Wavelet [27]	0.51	0.49	0.54	**0.89**
OctaveResent [20]	**0.35**	**0.47**	0.45	0.91

attention, it is considered to group the images in intervals, thereby achieving attention to long-distance features.

As shown in Figure 3, for the input feature f_4^l of size $R^{H/s\times W/s}$, sampling and grouping are performed at a fixed interval t_l. Each sampling group G has a size of $R^{H/(st_l)\times W/(st_l)}$.The f_4^l group form G^l is as follows:

$$G^l = [G_1, G_2, ..., G_{k_l}]^T, \tag{6}$$

where k_l is the number of groups, which is related to the t_l value. After grouping, standard self-attention is computed within each group G in LDA.

$$f_{LDA}^l = \text{Self-attention}(G^l). \tag{7}$$

3.5 Small Object Displacement Attention Module

As described in Sec. 3.4, SODA is designed to process the high-frequency feature f_4^h, focusing on short-range attention for small object motion. Since the displacements of small objects are mostly relatively close in distance and relatively small, they may be overlooked. Therefore, attention must be paid to the features of small neighborhoods.

As illustrated in Figure 3, the input $f_4^h \in \mathbb{R}^{H\times W}$ is divided into adjacent $t_s \times t_s$ pixel groups, forming the group structure G^s used by SODA:

$$G^s = [G_1, G_2, ..., G_{k_s}]^T, \tag{8}$$

where k_s is the number of groups, which is related to the t_s value. After grouping, standard self-attention is performed within each SODA group.

$$f_{SODA}^h = \text{Self-attention}(G^s). \tag{9}$$

3.6 Feature Fusion Module

As shown in Figure 2, high-frequency feature f_{SODA}^h and low-frequency feature f_{LDA}^l are processed in separate branches (B_1, B_2), each with Layer Normalization (LN), positional embedding (PE), dynamic PE (DPE), followed by LDA or SODA, another LN, and an MLP (multi-layer perception). The low-frequency branch is upsampled by 2× for alignment, and both branches are fused via attention to produce the final feature f^u.

$$f_{out}^h = B_1(f_{SODA}^h), f_{out}^l = B_2(f_{LDA}^l), \tag{10}$$

where f_{out}^h and f_{out}^l represent the outputs of two branches. Then we focus on the above results to obtain the final fused features through the following formula:

$$f^u = f + Attention(f_{out}^h, f_{out}^l, f), \tag{11}$$

Where f is the input image feature. Specifically, the features f_1, f_2, and f_3 of the input images I_1, I_2, and I_3 are processed through the IFS, SODA, LDA, and FF modules, ultimately resulting in the fused features f_1^u,f_2^u, and f_3^u.

Table 4: **Results of Different Sizes of LDA and SODA**

Module	Group Interval (t_l, t_s)	Things(val)		KITTI(train)	
		clean	final	F1-epe	F1-all
LDA (t_l)	1	1.48	2.11	4.22	15.07
	3	1.31	**1.78**	**3.99**	14.67
	5	1.22	2.03	4.02	**14.28**
	7	**1.19**	1.84	4.10	14.72
SODA (t_s)	1	1.38	2.08	4.17	15.24
	3	1.12	**1.46**	**3.93**	14.47
	5	**1.06**	1.58	4.08	14.53
	7	1.21	1.49	3.99	14.49

4 Experiments

4.1 Implementation Details

Based on some previous works, we utilize the feature encoder and context encoder as ImageNet Twins-SVT. We divide the training process into two stages: the C+T-stage and the C+T+S+K+H-stage. In the C+T stage, we first train for 300k iterations on the FlyingChairs [8] and FlyingThings [23] datasets. Since the FlyingChairs dataset only contains training data pairs, we achieve the multi-frame effect through grouping. In the C+T+S+K+H stage, we fine-tune the model trained in the C+T stage for 120k iterations on the Sintel [2], KITTI-2015 [9], and HD1K datasets [17]. We use the AdamW optimizer and a one-cycle learning rate scheduler. The batch size for all stages is 6. We use AEPE and F1-all as evaluation metrics.

4.2 Comparison with State-of-the-art Methods

4.2.1 Generalization Performance

In Table 1, the "$C+T$" setting reflects the model's cross-dataset generalization ability. Specifically, on the challenging Sintel final pass, CFFlow ranks first and surpasses the VideoFlow series. It is noteworthy that VideoFlow and FlowFormer++ have 13.4% and 52.9% more parameters than CFFlow (13.5M vs. 18.2M vs. 11.9M). Additionally, FlowFormer++ undergoes pre-training using a masked autoencoder strategy on the YouTube-VOS dataset [39]. And for the clean pass and KITTI-2015, our CFFlow achieves performance on par with the state-of-the-art methods.

4.2.2 Dataset-specific Performance

After the "$C+T+S+K(+H)$" stage, we submit to the Sintel and KITTI benchmarks. As shown in Table 1, our model achieves **0.99** and **1.65** AEPE on the Sintel clean and final passes, outperforming most methods, especially SAMFlow and MatchFlow on the final pass. This confirms that cross-frequency features are more effective than standard image feature enhancements. On KITTI-2015, CFFlow achieves a 4.21% F1-all error, surpassing both FlowFormer++ and VideoFlow-BOF, with consistent gains over the latter under the same three-frame setting. Full results are available on the official benchmark websites.

4.3 Performance Analysis for Different Regions

To comprehensively evaluate CFFlow, we present additional metrics in Table 2, including "**Unmatched**" AEPE for pixels that do not match across frames and s_{0-10}, s_{10-40}, s_{40+} for different ranges of flow

Table 5: **The Parameters and Inference of Different Models.**

Model	Para.	Infer.
GMA [16]	5.9M	74ms
FlowFormer [13]	18.2M	149ms
SAMFlow [25]	-	450ms
CRAFT [23]	6.4M	116ms
MatchFlow(G) [7]	15.4M	126ms
CFFlow(Ours)	11.9M	174ms

magnitudes of the ground truth. Compared with five state-of-the-art methods: MemFlow, SAMFlow, VideoFlow, StreamFlow, and FlowFormer++. CFFlow achieves the **best performance in the "Unmatched" region** on both clean and final passes, demonstrating its strength in handling occlusion and mismatches through cross-frequency features (see Figure 4). Moreover, CFFlow remains competitive in large-displacement scenarios caused by fast motion and severe blur, outperforming SAMFlow and MemFlow on the final pass.

4.4 Visualization

In Figure 4, we present examples of visualization results on a clean pass of the Sintel test sets. From left to right, each row represents input images and predicted flows of the multi-frame method MemFlow [6] and our CFFlow. Compared to MemFlow, CFFlow performs better in regions with small object details (third row, fourth row) and areas prone to unmatched issues, such as occlusions (second row). At the same time, CFFlow also demonstrates advantages regarding object integrity (first row). It shows that cross-frequency is valid in large displacement scenarios.

4.5 Ablation Study

4.5.1 Image Frequency Separation Module

We design the Image Frequency Separation Module using Octave Convolution to reduce spatial redundancy and preserve key frequency information. Inspired by YOLOv9's Haar Wavelet DownSample, we adapt it for frequency separation. As shown in Table 3, OctaveConv provides reasonable results but does not fully optimize separation. Haar Wavelet improves Sintel accuracy, but underperforms on KITTI F1-all. In contrast, OctaveResNet, which replaces standard convolutions in ResNet with OctaveConv, achieves the best balance, improving both Sintel (0.35, 0.47) and KITTI F1-all (0.91). These results show that OctaveConv in ResNet improves frequency separation while maintaining important details.

4.5.2 LDA and SODA Module

We analyze the impact of the interval settings (t_l, t_s) on the model performance through optical flow estimation experiments with LDA and SODA under different intervals of the module group (see Table 4). The results demonstrate that the module group interval is one of the key factors affecting the accuracy and generalization capacity of optical flow estimation. For both SODA and LDA, an interval of 3 proves to be an ideal configuration, where the models achieve a better balance in performance across multiple evaluation metrics. However, the performance variations under different intervals reveal differences between the two methods in feature interaction and optical flow information mining, providing valuable references for further optimization of module design in optical flow estimation models.

4.5.3 Parameters and Runtime Analysis

Table 5 shows that our model achieves a good balance between performance and efficiency. With 11.9M parameters, lower than FlowFormer (18.2M) and close to CRAFT (6.4M), it runs at 174 ms, faster than SAMFlow (450 ms) and CRAFT (422 ms), although slightly slower than MatchFlow (126 ms). Despite added processing over FlowFormer (149ms), it maintains competitive efficiency.

5 Conclusion

This paper investigates the integration of cross-frequency features into optical flow estimation. Current multi-frame methods often focus on surface-level cues, overlooking deeper frequency information. To address this, we propose CFFlow, which incorporates cross-frequency features via two modules: LDA for large displacements and SODA for small object motion. Experiments show that CFFlow achieves state-of-the-art performance on Sintel and KITTI, ranking first among three-frame methods. However, fully embedding cross-frequency knowledge remains a challenge for future work.

6 Acknowledgements

This work was supported by National Science and Technology Major Project from Minister of Science and Technology, China (2021ZD0201403), Youth Innovation Promotion Association, Chinese Academy of Sciences (2021233), Shanghai Academic Research Leader (22XD1424500).

References

[1] Wenbo Bao, Xiaoyun Zhang, Li Chen, and Zhiyong Gao. Kalmanflow 2.0: Efficient video optical flow estimation via context-aware kalman filtering. *IEEE Transactions on Image Processing*, 28 (9):4233–4246, 2019.

[2] D Butler, Jonas Wulff, G Stanley, and M Black. Mpi-sintel optical flow benchmark: Supplemental material. In *MPI-IS-TR-006, MPI for Intelligent Systems (2012*. Citeseer, 2012.

[3] Linwei Chen, Ying Fu, Lin Gu, Chenggang Yan, Tatsuya Harada, and Gao Huang. Frequency-aware feature fusion for dense image prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[4] Yonghu Chen, Dongchen Zhu, Wenjun Shi, Guanghui Zhang, Tianyu Zhang, Xiaolin Zhang, and Jiamao Li. Mfcflow: A motion feature compensated multi-frame recurrent network for optical flow estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5068–5077, 2023.

[5] Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3435–3444, 2019.

[6] Qiaole Dong and Yanwei Fu. Memflow: Optical flow estimation and prediction with memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19068–19078, 2024.

[7] 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.

[8] 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.

[9] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[10] Davar Giveki. Human action recognition using an optical flow-gated recurrent neural network. *International Journal of Multimedia Information Retrieval*, 13(3):29, 2024.

[11] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17 (1-3):185–203, 1981.

[12] Ying Jiao Huang, Xia Fei Ma, Rong Rong Xue, and Hao Tong Ma. Cff-net: cross-frequency fusion network for aero-degraded image restoration. In *AOPC 2024: Optical Sensing, Imaging Technology, and Applications*, volume 13496, pages 350–355. SPIE, 2024.

[13] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer: A transformer architecture for optical flow. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 668–685. Springer, 2022.

[14] 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.

[15] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9000–9008, 2018.

[16] 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.

[17] Daniel Kondermann, Rahul Nair, Katrin Honauer, Karsten Krispin, Jonas Andrulis, Alexander Brock, Burkhard Gussefeld, Mohsen Rahimimoghaddam, Sabine Hofmann, Claus Brenner, et al. The hci benchmark suite: Stereo and flow ground truth with uncertainties for urban autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 19–28, 2016.

[18] Ying-Yee Kong and Louis D Braida. Cross-frequency integration for consonant and vowel identification in bimodal hearing. *Journal of Speech, Language, and Hearing Research*, 54(3):959–980, 2011.

[19] Salim Lahmiri. Wavelet low-and high-frequency components as features for predicting stock prices with backpropagation neural networks. *Journal of King Saud University-Computer and Information Sciences*, 26(2):218–227, 2014.

[20] Pengtao Li, Meihua Wang, Zhun Fan, Han Huang, Guijie Zhu, and Jiafan Zhuang. Our-net: a multi-frequency network with octave max unpooling and octave convolution residual block for pavement crack segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[21] Yawen Lu, Qifan Wang, Siqi Ma, Tong Geng, Yingjie Victor Chen, Huaijin Chen, and Dongfang Liu. Transflow: Transformer as flow learner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18063–18073, 2023.

[22] Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th international joint conference on Artificial intelligence*, volume 2, pages 674–679, 1981.

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

[24] Zhile Ren, Orazio Gallo, Deqing Sun, Ming-Hsuan Yang, Erik B Sudderth, and Jan Kautz. A fusion approach for multi-frame optical flow estimation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2077–2086. IEEE, 2019.

[25] Xiaoyu Shi, Zhaoyang Huang, Weikang Bian, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Videoflow: Exploiting temporal cues for multi-frame optical flow estimation. *arXiv preprint arXiv:2303.08340*, 2023.

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

[27] Radomir S Stanković and Bogdan J Falkowski. The haar wavelet transform: its status and achievements. *Computers & Electrical Engineering*, 29(1):25–44, 2003.

[28] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17602–17611, 2022.

[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.

[30] Shangkun Sun, Yuanqi Chen, Yu Zhu, Guodong Guo, and Ge Li. Skflow: Learning optical flow with super kernels. *Advances in Neural Information Processing Systems*, 35:11313–11326, 2022.

[31] Shangkun Sun, Jiaming Liu, Huaxia Li, Guoqing Liu, Thomas Li, and Wei Gao. Streamflow: streamlined multi-frame optical flow estimation for video sequences. *Advances in neural information processing systems*, 37:9205–9228, 2024.

[32] Yanguang Sun, Chunyan Xu, Jian Yang, Hanyu Xuan, and Lei Luo. Frequency-spatial entanglement learning for camouflaged object detection. In *European Conference on Computer Vision*, pages 343–360. Springer, 2024.

[33] 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.

[34] BCM van Wijk, A Jha, W Penny, and V Litvak. Parametric estimation of cross-frequency coupling. *Journal of neuroscience methods*, 243:94–102, 2015.

[35] Bo Wang, Yifan Zhang, Jian Li, Yang Yu, Zhenping Sun, Li Liu, and Dewen Hu. Splatflow: Learning multi-frame optical flow via splatting. *International Journal of Computer Vision*, 132 (8):3023–3045, 2024.

[36] Mengfei Wang, Wenjun Shi, Dongchen Zhu, Lei Wang, and Jiamao Li. Memoflow: Modifying explicit motion of inconsistency in optical flow. In *International Conference on Pattern Recognition*, pages 219–234. Springer, 2025.

[37] 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.

[38] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[39] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018.

[40] Mingliang Zhai, Xuezhi Xiang, Ning Lv, and Xiangdong Kong. Optical flow and scene flow estimation: A survey. *Pattern Recognition*, 114:107861, 2021.

[41] Dengsheng Zhang. Wavelet transform. In *Fundamentals of image data mining: Analysis, Features, Classification and Retrieval*, pages 35–44. Springer, 2019.

[42] Kaidong Zhang, Jialun Peng, Jingjing Fu, and Dong Liu. Exploiting optical flow guidance for transformer-based video inpainting. *IEEE transactions on pattern analysis and machine intelligence*, 46(7):4977–4992, 2024.

[43] Zhiyong Zhang, Aniket Gupta, Huaizu Jiang, and Hanumant Singh. Neuflow v2: High-efficiency optical flow estimation on edge devices. *arXiv preprint arXiv:2408.10161*, 2024.

[44] Zhiyong Zhang, Huaizu Jiang, and Hanumant Singh. Neuflow: Real-time, high-accuracy optical flow estimation on robots using edge devices. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5048–5055. IEEE, 2024.

[45] Shili Zhou, Ruian He, Weimin Tan, and Bo Yan. Samflow: Eliminating any fragmentation in optical flow with segment anything model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7695–7703, 2024.

[46] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 989–997, 2019.