MFCFlow : A Motion Feature Compensated Multi-Frame Recurrent Network for Optical Flow Estimation

Yonghu Chen[1,2], Dongchen Zhu[1,2], Wenjun Shi[1], Guanghui Zhang[1], Tianyu Zhang[1,2], Xiaolin Zhang[1,2,3,4,5] and Jiamao Li [1,2,3,*]

[1]Bionic Vision System Laboratory, State Key Laboratory of Transducer Technology, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]Xiongan Institute of Innovation, Xiongan, 071700, China
[4]University of Science and Technology of China, Hefei, Anhui, 230027, China
[5]ShanghaiTech University, Shanghai 201210, China
jmli@mail.sim.ac.cn

Abstract

Occlusions have long been a hard nut to crack in optical flow estimation due to ambiguous pixels matching between abutting images. Current methods only take two consecutive images as input, which is challenging to capture temporal coherence and reason about occluded regions. In this paper, we propose a novel optical flow estimation framework, namely MFCFlow, which attempts to compensate for the information of occlusions by mining and transferring motion features between multiple frames. Specifically, we construct a Motion-guided Feature Compensation cell (MFC cell) to enhance the ambiguous motion features according to the correlation of previous features obtained by attention-based structure. Furthermore, a TopK attention strategy is developed and embedded into the MFC cell to improve the subsequent matching quality. Extensive experiments demonstrate that our MFCFlow achieves significant improvements in occluded regions and attains state-of-the-art performances on both Sintel and KITTI benchmarks among other multi-frame optical flow methods.

1. Introduction

Optical Flow (OF) estimation is a fundamental low-level vision task, which describes a 2D displacement field from frame I_t to the next frame I_{t+1}. OF steadily serves as a salutary clue providing dense correspondence for prediction tasks related to motions in videos, providing essential motion clues for high-level practical applications such as autonomous driving [23], action recognition [34, 26], video super-resolution [3, 29, 37], and video interpolation [39, 20, 14].

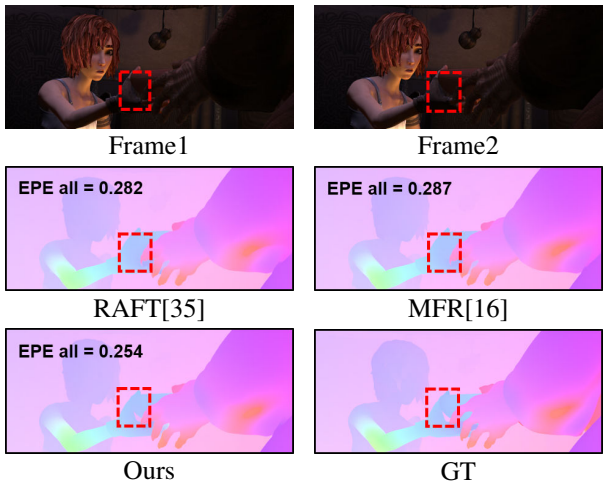


Figure 1: The optical flow field estimated by RAFT, MFR, and our MFCFlow on Sintel test set. Our method can generate better results via motion features compensation, especially for occluded regions, such as the edges of the 'bowl'.

FlowNet (S,C) [4] is the pioneering work to employ convolution neural networks (CNNs) for optical flow. After that, OF estimation methods have made tremendous strides [27, 32, 24, 28, 35, 15]. As Ranjan [27] first proposes a coarse-to-fine spatial pyramid network named SPyNet to

*Corresponding author

estimate OF at multiple resolutions and the cost volume is introduced to discriminatively evaluate the matching similarity in PWC-Net [32] for warped features, the performance of these OF estimation methods have been greatly improved. However, occluded regions are still intractable to deal with for these methods, which spurs us to dig deeper to handle occlusions.

As highlighted in GMA [15], in optical flow estimation, an *occluded point* is defined as a point that toggles between visible and invisible states in the time domain. Here, Figure 1 demonstrates one general case of occlusions, where part of the bowl moves out from behind the fingers. This phenomenon is the mutual occlusion that occurs due to the motion between objects in the shot. In addition, Figure 2 shows another case of occlusions, where part of the blade moves out of the frame view. Both phenomena show that OF estimation is ill-defined in occluded regions, because there is no real corresponding point in the target image. That is, there are no correct matching relationships for these occluded points. As a result, motion features in occluded regions can not be internally consistent, leading to ambiguities in subsequent feature-level matching, especially for two-frame methods.

The ambiguity in cost volume due to occlusions will lead to failure to reason about occluded regions. Aiming at the aforementioned occlusion problem, in classical variational approaches, researchers generally introduce extra objective functions to constrain occlusions [40, 1]. In the deep learning era, they place hopes on CNNs to learn occlusion maps[12, 9, 22, 13, 41], without much success. Nonetheless, given the continuity of motion, the current occluded region may find corresponding points in an earlier frame. Moreover, the "oracle" study about temporal information in MFF [28] confirms that the previous optical flow can provide complementary information, suggesting that leveraging motion features from multiple frames matters.

In view of this, we propose a novel multi-frame framework that leverages motion features concerning time to recover ambiguous features for optical flow estimation. Considering that, for a multi-frame framework, how effectively fuse the features between multiple frames is a crucial challenge. Thus we design an attention-based feature compensation strategy to fuse pair of motion features for recovering the motions from ambiguity. In order to ensure that the subsequent matching is more accurate and reasonable, a TopK Selector is further introduced into the feature fusion as an additional feature filtering on the attention matrix. In summary, the key contributions of our work are as follows.

- For handling ambiguities caused by occlusions in optical flow estimation, we propose a novel multi-frame recurrent framework, namely MFCFlow, that aggregates previous features along the image sequences to recover current ambiguous motions.
- We suggest a novel attention-based feature compensation strategy to exploit the temporal coherence between motion features in the proposed MFC cell. Besides, we deploy the TopK attention to filter the most relevant and effective pixels, significantly reducing redundant information and noisy correlations in feature matching.
- Our MFCFlow outperforms top-performing multi-frame approach MFR[16], and achieves state-of-the-art performances on both Sintel [2] and KITTI [23], especially making remarkable improvements in occluded regions.

2. Related Work

OF Estimation via Deep Learning Variational approaches dominate optical flow estimation since the work of Horn and Schunck [6]. Accompanied by the popularity of computer vision applications, CNNs play a more important role, and optical flow estimation via deep learning is showing an inevitable trend. FlowNet (S, C) [4] is the first deep learning approach for OF estimation, along with a synthetic training dataset, FlyingChairs. However, the accuracy is not as good as classical algorithms. IIg *et al*. [11] combines multiple flownets and proposes a small fusion network for flow refinements, flownet2.0, whose accuracy is comparable to classical methods but inefficient. Some of the follow-up works seek to leverage classical practices, such as warping-based estimation. PWC-Net [32] has then been a baseline for lightweight networks and several top-performing methods [7, 28, 19]. Recently, RAFT [35] achieves new benchmark results via a 4D all-pair cost volume with a recurrent unit. Some modules in our proposed framework are also inspired by the successful RAFT. While these methods obtain good results in most cases, they do not allow actual reasoning in occlusions.

Occlusion Handling As OF is ill-defined in occluded regions due to its violation of the brightness constancy constraints, occlusion handling plays an essential role in precise estimation. Classical methods treat occlusions as outliers and optimize robust objective functions in variational approaches [40, 1]. Other methods jointly estimate OF and occlusions with significant improvements [12, 9]. Occlusion maps matter in unsupervised methods because they need to ignore occluded regions in photometric loss [22, 13]. In the self-supervised method, an occlusion map is also a must for filtering features to avoid ambiguity due to occlusions [41].

Different from previous works, we will not estimate the occlusions map. We target improving performance in occluded regions via the motion feature compensation, not requiring estimating an occlusion map as the optimization

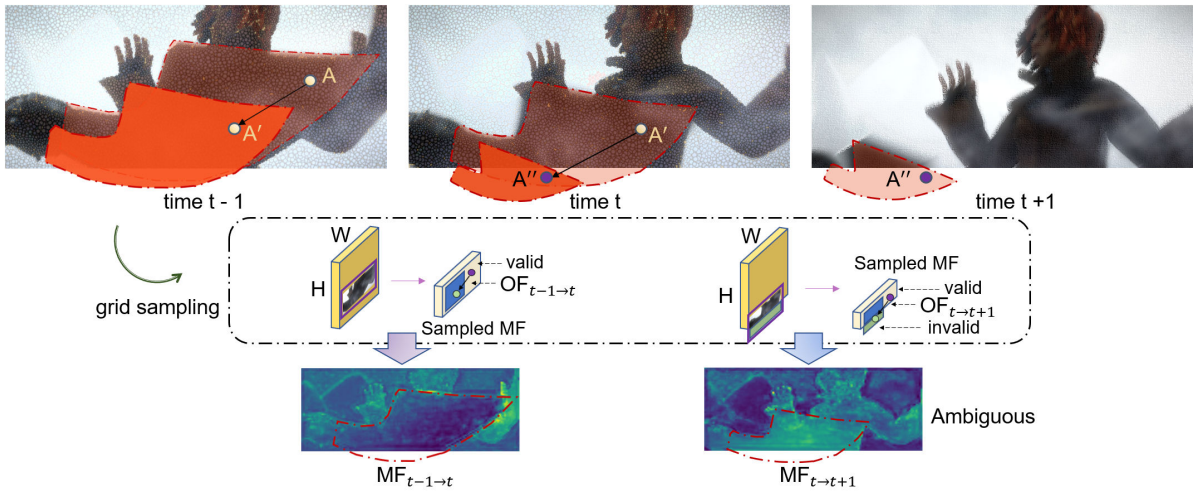


Figure 2: **Illustration of ambiguities caused by occlusions.** Sample frames are selected from Sintel (final) training data. The yellow dot $A^{'}$ at time t moves to the purple dot $A^{''}$ at time $t+1$, and the invisibility of target point $A^{''}$ will lead to the motion feature extracted by grid sampling getting ambiguous. As depicted in the motion feature (MF) visualization, the blade is indistinguishable from the human body. However, at time $t-1$, the yellow dot A is visible, and the $\mathrm{MF}_{t-1\to t}$ is clearly evident. The $\mathrm{MF}_{t-1\to t}$ is highly coherent to our target $\mathrm{MF}_{t\to t+1}$. We can utilize $\mathrm{MF}_{t-1\to t}$ to compensate $\mathrm{MF}_{t\to t+1}$, which in turn improves the final optical flow estimation.

prior. Similar to interpolation approaches, our method models temporal coherence to achieve better interpolation results in occlusions.

Multi-Frame OF Estimation Leveraging temporal coherence has been proven to improve the OF estimation quality that $\mathrm{OF}_{t-1\to t}$ can be utilized to recover ill-defined $\mathrm{OF}_{t\to t+1}$ [28, 24, 5, 22, 19, 16]. Ren *et al*. [28] proposes a multi-frame fusion process to fuse $\mathrm{OF}_{t-1\to t}$ and $\mathrm{OF}_{t\to t+1}$. ContinualFlow [24] introduces a temporal connection to pass $\mathrm{OF}_{t-1\to t}$ to the estimation process to get the target $\mathrm{OF}_{t\to t+1}$. Inspired by ContinualFlow, STaRFlow[5] proposes a STaRFlow cell to pass features in multiple scales, jointly with occlusion maps. Our work is close to the methods mentioned above, but we pass the motion features not OF to subsequent estimation. Moreover, we are surprised to find that unsupervised learning methods using multiple frames also have an improved performance [22, 19]. Very recently, Jiao *et al*. [16] combines multiple frames to leverage motion consistency to obtain a better-performing method named MFR, which is the most relevant work as ours. However, MFR feeds motion features to a two-layer CNN connected by a ReLU activation, which is hard to deal with distracting information, while we develop an attention-based feature fusion strategy to leverage motion features in a more effective multi-frame setting.

3. Methodology

We propose a multi-frame optical flow estimation algorithm for addressing occlusions. To fuse motion features of different moments in the image sequences, we design a simple but effective module, *Motion-guided Feature Compensation cell* (MFC cell), which is applied recurrently concerning the time scale. We first describe the phenomenon of ambiguities caused by occlusions in Section 3.1, and further propose our multi-frame framework and present an unrolled representation of our model in Section 3.2. Finally, we elaborate on the proposed MFC cell, which is used to compensate for ambiguous motion features in Section 3.3.

3.1. Problems Statements

As mentioned in MFR [16], given consecutive image features $g_\theta(I_t), g_\theta(I_{t+1})$, where g_θ is the feature extractor, H and W are the height and width of the feature map respectively, and D is the channel dimension of the feature map. The correlation volume layer can encode the motion feature similarity between $g_\theta(I_t)$ and $g_\theta(I_{t+1})$ to generate the 4D cost volume, $\mathbf{C}$. The center of the sampling grid from $\mathbf{C}$, is determined by the optical flow $\mathrm{OF}_{t\to t+1}$, for examples in Figure 2. The motion feature MF will be able to store the most relevant matching points for each pixel in frame I_t, and the final optical flow can be iteratively updated from MF via CNN blocks[35].

The workflow seems to work perfectly for most of the video scenes. However, when considering the occlusions and large displacement, we can observe that the generated motion feature MF may be sampled from the area that extends beyond the cost volume boundary, resulting in the ambiguity of matching points. In other words, we cannot determine the most relevant points in grid samplings in am-

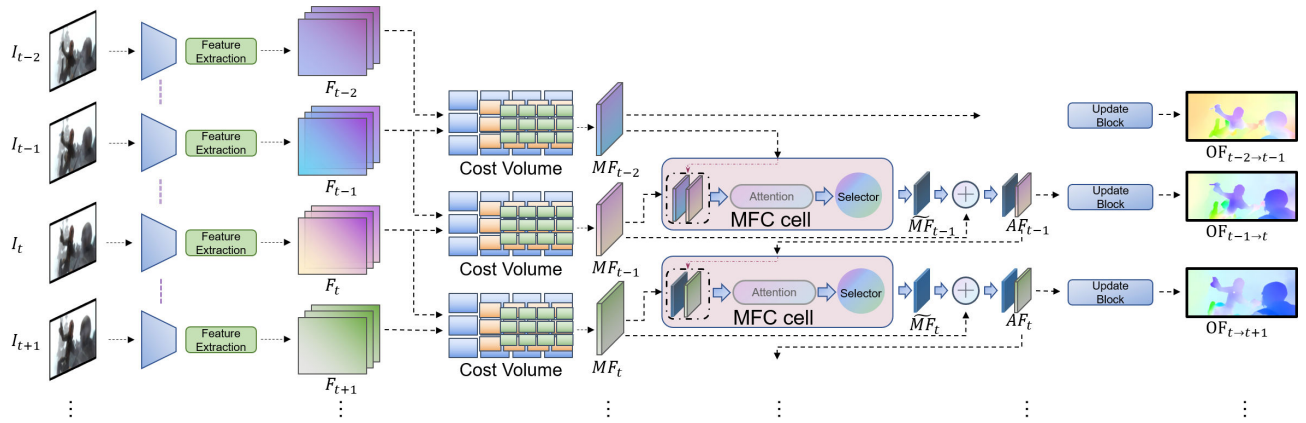


Figure 3: **Unrolled view of the proposed multi-frame recurrent network for optical flow estimation (MFCFlow).** MFCFlow takes $N+1$ frames ($I_{k,k\in t-2,t-1,t,t+1}$) as input and outputs N OFs. Feature extraction blocks that encodes the input frames are identical with shared weights. The cost volume ($\mathbf{C}$) generates motion features (MF_k) from feature pair F_k and F_{k+1}. The Motion-guided Feature Compensation cell (MFC cell) exploits the temporal coherence between consecutive MFs to refine the current MF, which will be passed to update blocks for subsequent OF estimation.

biguous regions. Specifically, as illustrated in Figure 2, the yellow pixel A' from time t will move to the purple dot, which is invisible at time $t+1$ with a large displacement. The invisibility of the target point leads to ambiguous grid sampling during motion feature generations, making it difficult to forward plausible motion features to update blocks. Multiple frames carrying additional information about object motion are conducive to motion feature recovery and temporal coherence constraints.

3.2. Schematic Overview of the Framework

To address the ambiguities caused by occlusions, we propose a multi-frame architecture named MFCFLow to aggregate historical motions and recover ambiguous features. As illustrated in Figure 3, for presentation clarity, we focus on four-frame optical flow estimation. Given four input frames I_{t-2}, I_{t-1}, I_t, and I_{t+1}, we aim to estimate the optical flow from frame I_t to frame I_{t+1}, denoted as $\mathrm{OF}^{I_{t+1}}_{t\to t+1}$. The superscript I indicates that it fuses information from all of the previous frames, as opposed to $\mathrm{OF}_{t\to t+1}$. In the model training phase, our multi-frame estimation method will take $N+1$ frames $(I_{t-N+1},...,I_t,I_{t+1})$ as input and output N optical flows $(\mathrm{OF}^{I_{t-N+2}}_{t-N+1\to t-N+2},...,\mathrm{OF}^{I_t}_{t-1\to t},\mathrm{OF}^{I_{t+1}}_{t\to t+1})$, which will be further supervised. When conducting inferences on 'test' set, we only select the last $\mathrm{OF}^{I_{t+1}}_{t\to t+1}$ as the model output.

MFCFlow takes $N+1$ consecutive frames as input. First, the image features are extracted from the input frames via a shared convolution neural network. The features encoder g_θ, similar to RAFT [35], maps the input frame to dense features maps at a lower resolution. Then we generate the motion features from $g_\theta(I_k)$ and $g_\theta(I_{k+1})$ via a 4D cost volume $\mathbf{C}$. MFC cell will further exploit the temporal coherence between consecutive MFs to refine the current MF_k to provide non-local information for better interpolations in occluded regions. We will give a detailed description of the MFC cell in Section 3.3. MFC cell finally output a motion-augmented feature $\tilde{\mathrm{MF}}_k$. The concatenation of original MF_k and $\tilde{\mathrm{MF}}_k$ is the desired aggregated feature AF_k. The update blocks will iteratively decode the aggregated feature to generate the final optical flow.

3.3. Motion-guided Feature Compensation Cell

Ambiguities caused by occlusions may be difficult to alleviate via only two frames because of insufficient local information in occluded regions. However, temporal coherence through the frame sequences can provide non-local information for motions, which can be viewed as a non-local interpolation for ill-defined optical flow.

Although the motion feature (MF_t) inferred from time t to time $t+1$ is ambiguous, we can model temporal coherence along a sequence of consecutive frames to compensate for MF_t. During a short period of Δt, MF_{t-1} and MF_t should be highly internally coherent. We compensate MF_t with historical $\mathrm{MF}_{i(<t)}$ based on the similarity, which means we should pay more attention to regions with similar motion features through the image sequences. Successful MFR [16] ignores the constraints of relevance in feature selections, which may lead to noisy motions. Inspired by transformers [36], we explore the TopK Cross-attention Selector for our query and value features are the projection of the MF_t, and the key features are the projection of the MF_{t-1}, aiming at fusing temporal motion features. The attention matrix computed from the query and key features is used to augment the value features. The augmented value features will iteratively be decoded to output the final OF.

Figure 4 depicts the detailed structure of the Motion-guided Feature Compensation cell. After previous-

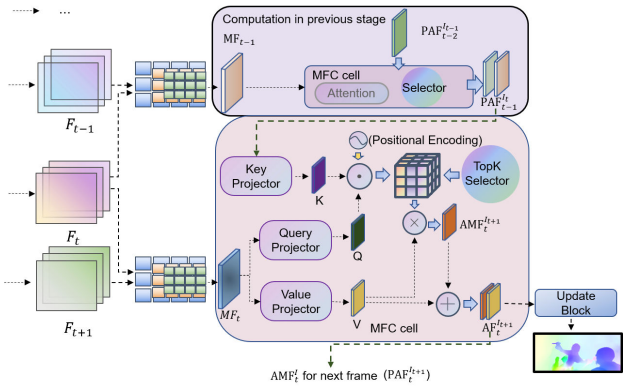


Figure 4: Structure of the Motion-guided Feature Compensation cell (MFC cell).

stage computation, previous aggregated feature $\mathrm{PAF}_{t-1}^{I_t}$ is achieved. We perform a series of feature extraction for frame pair $\{I_t, I_{t+1}\}$ to get the initial ambiguous motion feature MF_t.

Let $x \in \mathbb{R}^{N \times D_m}$ denote the $\mathrm{PAF}_{t-1}^{I_t}$, and $y \in \mathbb{R}^{N \times D_m}$ denote the MF_t, where $N = H \times W$, H and W are the height and width of the feature map respectively, and D_m is the channel dimension of the feature map. The i^{th} feature vector is denoted $x_i \in \mathbb{R}^{D_m}$. The attention part of our MFC cell computes the feature vector as a weighted sum of the projected motion features. The augmented motion feature $\mathrm{AMF}_t^{I_{t+1}}$, denoted as $\tilde{y}$, is defined as :

$$\tilde{y}_i = \sum_{j=1}^{N} f(\mathcal{Q}(y_i), \mathcal{K}(x_j))\mathcal{V}(y_j) \tag{1}$$

where $\mathcal{Q}$, $\mathcal{K}$, $\mathcal{V}$ are the projection function for the query, key, and value vectors, f is the attention function given by:

$$f(x_i, y_j) = \mathrm{softmax}(\frac{\langle x_i, (y_{j_1}, ..., y_{j_l}, ..., y_{j_D})\rangle}{\sqrt{D}}), y_{j_l} \in y_j \tag{2}$$

The aggregated motion features $\mathrm{AF}_t^{I_{t+1}}$, which will be decoded for optical flow, is derived by:

$$\hat{y} = \mathrm{Concat}(y, \tilde{y}) \tag{3}$$

We also explore the use of 2D relative positional embedding [30] like GMA [15], allowing the attention map to depend on both the feature self-similarity and the relative position from the point. For this, the augmented motion vector $\tilde{y}$ will be denoted as :

$$\tilde{y}_i = \sum_{j=1}^{N} f(\mathcal{Q}(y_i), \mathcal{K}(x_j) + \mathrm{Pos}_{j-i})\mathcal{V}(y_j) \tag{4}$$

where Pos_{j-i} denotes the relative positional embedding vector index by the pixel offset $j-i$. Furthermore, the attention map from only the query vector and relative positional embedding vector is investigated, dismissing the feature of self-similarity. Thus, the $\tilde{y}$ will be deduced by:

$$\tilde{y}_i = \sum_{j=1}^{N} f(\mathcal{Q}(y_i), \mathrm{Pos}_{j-i})\mathcal{V}(y_j) \tag{5}$$

The attention matrix demonstrates the point-to-point similarity, which is redundant in Softmax [17] function. Each 2D point in frame t has at most one corresponding location in frame $t+1$. In consideration of brightness, fast motions, motion blur, illumination effects, uniformly colored objects, and other factors, each point will not be relevant to all points in the reference frame yet. We get inspiration from KVT [38], which selects TopK scores for the follow-up Softmax function. We believe TopK reference points according to matching similarity are sufficient for our MFC cell. Therefore, the TopK cross-attention strategy is adopted. To be specific, we only sift out TopK relevant points for subsequent feature matching, and f_{TopK} will be developed by:

$$f(x_i, y_j) = \mathrm{softmax}(\frac{\langle x_i, (y_{j_1}, ..., y_{j_l}, ..., y_{j_D})\rangle_{\mathrm{TopK}}}{\sqrt{D}}) \tag{6}$$

The final output of MFC cell $\hat{y}$ ($\mathrm{AF}_t^{I_{t+1}}$) is the concatenation of the original motion feature and its augmentation, which can be formulated as $\hat{y} = \mathrm{Concat}(y, \tilde{y}) = \mathrm{Concat}(\mathrm{MF}_t, \mathrm{AMF}_t^{I_{t+1}})$.

3.4. Multi-Frame Training Loss

Update Blocks will estimate a sequence of flow estimates $\{f_1, f_2, ..., f_M\}$ for each pair of images, and the final f_M will be viewed as the target refined flow. We use $N+1$ frame training sequences and train our network to estimate the optical flow for a series of frames. Our proposed network takes $N+1$ frames as input and output N flows. Information from previous frames will be transmitted through the feature fusion module. At the end of the sequence, we update the weights to decrease:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_i \tag{7}$$

where $\mathcal{L}_i$ is the l_1 distance between each pairs of predicted and ground truth flow over the entire sequence of predictions $(f_{i1}, ..., f_{iM})$ with exponentially increasing weights. Given the ground truth flow f_{gt}, loss $\mathcal{L}_i$ is defined as:

$$\mathcal{L}_i = \sum_{j=1}^{M}\gamma^{j-M}||f_{gt} - f_{ij}||_1 \tag{8}$$

Training Data	Method	Sintel (train)		Sintel (test)		KITTI-2015 (train)		KITTI-2015 (test)
		Clean	Final	Clean	Final	Epe-all	F1-all	F1-all
C+T+S/K	FlowNet2* [11]	(1.45)	(2.01)	4.16	5.74	(2.30)	(6.8)	11.48
	IRR-PWC* [10]	(1.92)	(2.51)	3.84	4.58	(1.63)	(5.3)	7.65
	MFF* [28]	-	-	3.42	4.57	-	-	7.17
	SelFlow* [19]	(1.68)	(1.77)	3.74	4.26	-	**(1.2)**	8.42
	STaRFlow-ft* [5]	-	-	2.72	3.71	-	-	7.65
	RAFT [35]	(0.77)	(1.20)	2.08	3.41	(0.64)	(1.5)	5.27
	MFR* [16]	(0.65)	**(1.01)**	2.01	3.29	(0.59)	(1.3)	5.17
	Ours*	**(0.58)**	(1.10)	**1.63**	**2.89**	**(0.57)**	**(1.2)**	**5.07**
C+T+S+K+H	LiteFlowNet2 [8]	(1.30)	(1.62)	3.48	4.69	(1.47)	(4.8)	7.74
	PWC-Net+ [33]	(1.71)	(2.34)	3.45	4.60	(1.50)	(5.3)	7.72
	MaskFlowNet [41]	-	-	2.52	4.17	-	-	6.10
	RAFT [35]	(0.76)	(1.22)	1.94	3.18	(0.63)	(1.5)	5.10
	RAFT-warm [35]	(0.77)	(1.27)	1.61	2.86	-	-	-
	MFR* [16]	(0.64)	(1.04)	1.55	2.80	**(0.54)**	**(1.1)**	5.03
	Ours*	**(0.56)**	**(0.89)**	**1.49**	**2.58**	(0.55)	**(1.1)**	**(5.00)**

Table 1: Quantitative comparsions of optical flow with EPE and F1 employed for evalution. Multi-frame methods are marked with *.

4. Experiments

4.1. Implementation Details

We follow the standard optical flow training procedure [32, 35, 11] of first pre-training our model on FlyingChairs [4] and then on FlyingThings [21]. We then finetune on either MPI Sintel [2] or KITTI [23]. Following [5], we first train our multi-frame architecture, except the MFC cell on 2D two-frame data. Then we train the MFC cell on sequences of N images from FlyingThings3D. We use a batch size of 4 for training. We will Finetune on MPI Sintel or KITTI. For Sintel, we can supervise every time step. We only supervise the last-step estimation in KITTI because only the last time step is annotated in the multiview KITTI dataset.

We train our model on four RTX3090 GPUs with Pytorch Library [25]. We adopt the same hyperparameters as RAFT [35] for the base network and utilize the the one-cycle learning rate policy [31], with the highest learning rate set to 2.5×10^{-4} for FlyingChairs and then 1.25×10^{-4} for the rest. For the MFC cell, we empirically choose channel dimensions $D_m = 16$. Other learning settings such as data augmentation are similar to RAFT.

4.2. Comparisons with the State-of-the-Art Works

We evaluate the proposed MFCFlow on standard benchmarks, Sintel and KITTI. Average end-point error (EPE) and percentage of optical flow outliners (F1) are used for evaluation in Table 1. We follow previous methods [35, 16] and train our model on C+T+S/K and C+T+S+K+H, respectively. For a fair comparison, the frame number of our submitted model's input is 4. MFCFlow outperforms RAFT [35] and the top-performing multi-frame method MFR [16] by a large margin especially on Sintel. With C+T+S/K training, we achieve EPE of 1.63 on Clean pass and 2.89 on Final pass in Sintel. MFCFLow improves a lot on KITTI, with F1 of 5.07. Training with more data from HD1k (H)[18] improves the performance of our model and reduces the testing error to 1.49 and 2.58 on Sintel. This phenomenon is probably because MFC cells can extract motion features. Overall, our MFRFlow achieves a new state-of-the-art performance, demonstrating the benefit of exploiting the temporal coherence to address occlusions. Note that, compared to the performance on Sintel, our model achieves minor improvement on KITTI. We believe this is due to the insufficient training data of KITTI (only 200 sequences of images), which is far from enough to train MFC cells. With more training data, MFCFlow may achieve more apparent improvements.

Furthermore, we show the visualization results of our MFCFlow on Sintel and KITTI test set in Figure 5. It can be observed that MFCFlow significantly alleviates the ambiguity caused by occlusions such as the street lamp in front of the car, which is in line with the expectations of the designed MFC cell. Furthermore, thanks to the multi-frame recurrent framework incorporating more matching cues, the reflective regions (car windows) and illumination effects (the human body and clothes) also achieve performance improvements. To conclude, by exploiting the temporal coherence, MFCFlow can give precise estimates and cogent reasoning, and the experimental results demonstrate that recovering ambiguous motion features across temporal coherence is indeed effective.

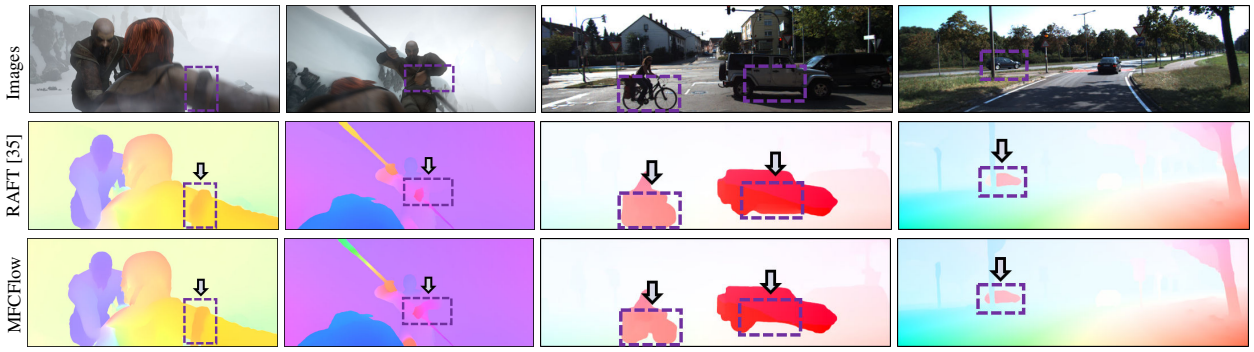


Figure 5: Qualitative comparisons on Sintel and KITTI test datasets. The second row shows the visualization results of the state-of-the-art RAFT[35], and the bottom row depicts ours.

4.3. Occlusions analysis

To explore whether the proposed model has advantages for occlusions, we evaluate the AEPE for different regions ('noc':non-occluded pixels and 'occ':occluded pixels) on Sintel. We take Clean and Final pass as training data and set Albedo pass as the test set with the 'warm-start' strategy [35].

As shown in Table 2, for all Clean, Final, and Albedo, our model is significantly improved in both 'occ' and 'noc' compared with the baseline RAFT. And it is worth noting that the increases of AEPE in 'occ' are much larger than that in 'noc' on all three passes (23.10% vs 10.56%, 20.36% vs 18.66%, and 15.39% vs 4.47%), demonstrating the benefits of MFC cells in handling occlusions. Because of the completely missing features, 'occ-out' occlusions are more challenging than 'occ-in' occlusions for two-frame optical flow estimation methods like RAFT. On the contrary, our method can completely remedy the adverse effects caused by both two kinds of occlusions. As a result, the improvement in 'occ-out' regions is more obvious. Interestingly, for the Final pass, since it adds a lot of blur noise itself, no matter which area it is, there is a big pickup of our model. These results fully demonstrate that, thanks to the MFC cell, the proposed multi-frame recurrent framework is friendly not only to motion-induced occlusions, but also to other noise-induced ambiguous matching.

Sintel Datasets	Type	Pct	RAFT (AEPE)	Ours (AEPE)	Rel.Impr (%)
Clean (Train)	Noc	-	0.341	**0.305**	10.56
	Occ	(8.29%)	6.256	**4.811**	23.10
	Occ-in	(5.35%)	4.974	**4.014**	19.30
	Occ-out	(2.94%)	6.964	**5.127**	26.38
	All	-	0.773	**0.634**	17.98
Final (Train)	Noc	-	0.718	**0.584**	18.66
	Occ	(8.29%)	8.414	**6.701**	20.36
	Occ-in	(5.35%)	7.177	**5.705**	20.51
	Occ-out	(2.94%)	8.619	**7.392**	14.24
	All	-	1.280	**1.031**	19.45
Albedo (Test)	Noc	-	0.358	**0.342**	4.47
	Occ	(8.29%)	7.325	**6.198**	15.39
	Occ-in	(5.35%)	6.084	**5.414**	11.01
	Occ-out	(2.94%)	7.692	**6.247**	18.79
	All	-	0.867	**0.770**	11.19

Table 2: **Occlusions analysis.** The percentage of different regions is denoted as 'Pct'. In(out-of)-frame occlusion is further split and denoted as 'Occ-in(out)'.

4.4. Ablations Results

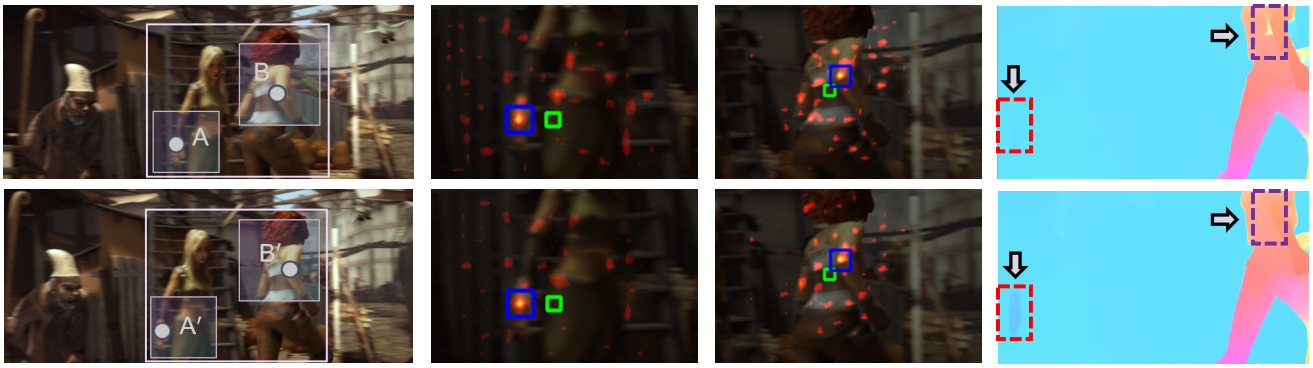
To verify our design, we conducted the following ablation experiments. Albedo pass marked with † is evaluated with the 'warm-start' strategy [35] and settings used in our final model are underlined.

We exploit the temporal coherence in our model and test the inference time in the Sintel dataset on a single RTX 2060 GPU. For presentation clarity, we assume the framework takes N frames as input. The motion feature $\mathrm{MF}_{1\to2}$ extracted from frame 1,2 will pass forward along the time dimension. Aiming to find the temporal relation, thus, some historical motion features should be dropped due to memory cost and low correlation to the current motion feature. Therefore, the number of input frames, N, will inevitably make a difference to our framework. Detailed ablation results are shown below in Table 3. As the number of input frames increases, the performance of the model continues to improve, even in background regions (F1-bg), further demonstrating the effectiveness of the our model. However, the inference time will increase as well. Considering the fairness of comparison with other multi-frame methods and precision, we set $N = 4$ in our final submitted model.

N	Albedo†			KITTI (%)		Timing (ms)	
	all	noc	occ	F1-bg	F1-all	Sintel	KITTI
2	0.867	0.369	7.186	4.81	5.12	138	353
3	0.817	0.357	6.658	4.80	5.11	281	685
<u>4</u>	0.802	0.359	6.418	4.71	5.07	422	1022
5	**0.770**	**0.342**	**6.198**	**4.67**	**5.01**	557	1358

Table 3: Impact of the numer of frames N in our model.

For the attention mechanism of our proposed model, we



(a) Frame1 & Frame2 (b) Attention Map1 (c) Attention Map2 (d) Refined Flow

Figure 6: **Attention map visualizations on the Sintel (test) final pass.** Dot A and dot B in Frame 1 move to dot A′ and dot B′ in Frame 2. We show the attention map and final refined optical flow for the last three columns. The top row visualizes RAFT [35], and our MFCFlow generates the bottom row. The green square indicates the original position of the query point, and the blue square denotes the most relevant point to the query point. The brighter colors mean higher attention weights (more similar). Arrows in the last column highlight significant improvements.

Type	Sintel (train)		Albedo† (test)			KITTI (train)	
	Clean	Final	all	noc	occ	EPE	F1
	all	all					
p	0.681	**1.125**	0.822	**0.343**	6.896	**0.561**	**1.184**
c	**0.668**	1.140	**0.817**	0.357	**6.658**	0.573	1.214
+p	0.692	1.130	0.851	0.375	6.891	0.595	1.333

Table 4: Ablation results about attention mechanism.

TopK	Sintel (train)		Albedo† (test)			KITTI (train)	
	Clean	Final	all	noc	occ	EPE	F1
	all	all					
60%	0.671	1.138	0.824	0.362	6.693	0.574	1.215
70%	0.673	1.140	0.806	0.357	**6.498**	0.573	1.211
80%	0.671	**1.137**	**0.802**	**0.352**	6.509	**0.572**	**1.210**
90%	0.670	1.140	0.835	0.367	6.773	0.573	1.211
100%	**0.668**	1.140	0.817	0.357	6.658	0.573	1.214

Table 5: Impact of the percentage of reference motion features in our model.

compare the performance of different inputs of the attention function: (1) positional attention replaces content attention, denoted as (p), (2) only content attention, denoted as (c), (3) positional attention adds to the content attention, denoted as (+p). As illustrated in Table 4, the addition of positional encoding benefits our model with minor improvements. Based on comprehensive considerations such as model complexity, we choose the model with only content attention in our final model. In Figure 6, we show the attention map visualizations on the Sintel (test) final pass. RAFT and MFCFlow inevitably make many noisy correlations. However, with temporal coherence, MFCFlow filters out some noisy correlations, so further refined flow is more precise, demonstrating the benefits of the attention mechanism.

We also ablate the percentage of reference motion features, filtering the most relevant features for following reweighting. Results are shown in Table 5, and we determine the model corresponding to 80%.

5. Conclusions

We explore the ambiguity caused by occlusions, for which we propose a novel multi-frame recurrent optical flow estimation framework (MFCFlow) that aggregates motion features in temporal sequences to diminish regional ambiguity. We introduce an attention-based Motion-guided Feature Compensation cell (MFC cell), where previous motions will be utilized to recover current ambiguous features, effectively fusing temporal coherence with respect to the time scale. Furthermore, a TopK attention selector is employed to filter out irrelevant references, significantly reducing noisy correlations in subsequent feature matching. Extensive experiments on different optical flow benchmarks show that our MFCFlow significantly improves predictions in occluded regions. Notably, we find that motion features in detail regions, such as the edge of the contour, may not achieve reasonable compensation. In future work, we will explore high-level feature consistency for better coherence modeling. For example, exploring the potential constraint from geometry motion and contour-based features may be conducive to modeling temporal coherence.

Acknowledgement. This research was supported by National Science and Technology Major Project from Minister of Science and Technology, China(2018AAA0103100), National Natural Science Foundation of China(61873255, 62103399), Shanghai Municipal Science and Technology Major Project (ZHANGJIANG LAB) under Grant 2018SHZDZX01, Youth Innovation Promotion Association, Chinese Academy of Sciences(2021233) and Shanghai Academic Research Leader(22XD1424500).

References

[1] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *European conference on computer vision*, pages 25–36. Springer, 2004.

[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] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. *arXiv preprint arXiv:2104.13371*, 2021.

[4] 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.

[5] Pierre Godet, Alexandre Boulch, Aurélien Plyer, and Guy Le Besnerais. Starflow: A spatiotemporal recurrent cell for lightweight multi-frame optical flow estimation. *arXiv preprint arXiv:2007.05481*, 2020.

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

[7] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8981–8989, 2018.

[8] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. A lightweight optical flow cnn—revisiting data fidelity and regularization. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2555–2569, 2020.

[9] Junhwa Hur and Stefan Roth. Mirrorflow: Exploiting symmetries in joint optical flow and occlusion estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 312–321, 2017.

[10] Junhwa Hur and Stefan Roth. Iterative residual refinement for joint optical flow and occlusion estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5754–5763, 2019.

[11] 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.

[12] Eddy Ilg, Tonmoy Saikia, Margret Keuper, and Thomas Brox. Occlusions, motion and depth boundaries with a generic network for disparity, optical flow or scene flow estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 614–630, 2018.

[13] Joel Janai, Fatma Guney, Anurag Ranjan, Michael Black, and Andreas Geiger. Unsupervised learning of multi-frame optical flow with occlusions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 690–706, 2018.

[14] 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.

[15] 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.

[16] Yang Jiao, Guangming Shi, and Trac D Tran. Optical flow estimation via motion feature recovery. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2558–2562. IEEE, 2021.

[17] Armand Joulin, Moustapha Cissé, David Grangier, Hervé Jégou, et al. Efficient softmax approximation for gpus. In *International conference on machine learning*, pages 1302–1310. PMLR, 2017.

[18] Daniel Kondermann, Rahul Nair, Stephan Meister, Wolfgang Mischler, Burkhard Güssefeld, Katrin Honauer, Sabine Hofmann, Claus Brenner, and Bernd Jähne. Stereo ground truth with error bars. In *Asian conference on computer vision*, pages 595–610. Springer, 2014.

[19] Pengpeng Liu, Michael Lyu, Irwin King, and Jia Xu. Selflow: Self-supervised learning of optical flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4571–4580, 2019.

[20] Xiaozhang Liu, Hui Liu, and Yuxiu Lin. Video frame interpolation via optical flow estimation with image inpainting. *International Journal of Intelligent Systems*, 35(12):2087–2102, 2020.

[21] 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.

[22] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

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

[24] Michal Neoral, Jan Šochman, and Jiří Matas. Continual occlusion and optical flow estimation. In *Asian Conference on Computer Vision*, pages 159–174. Springer, 2018.

[25] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[26] AJ Piergiovanni and Michael S Ryoo. Representation flow for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9945–9953, 2019.

[27] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170, 2017.

[28] 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.

[29] Mehdi S. M. Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[30] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

[31] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, page 1100612. International Society for Optics and Photonics, 2019.

[32] 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.

[33] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Models matter, so does training: An empirical study of cnns for optical flow estimation. *IEEE transactions on pattern analysis and machine intelligence*, 42(6):1408–1423, 2019.

[34] Shuyang Sun, Zhanghui Kuang, Lu Sheng, Wanli Ouyang, and Wei Zhang. Optical flow guided feature: A fast and robust motion representation for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1390–1399, 2018.

[35] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020.

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

[37] Longguang Wang, Yulan Guo, Li Liu, Zaiping Lin, Xinpu Deng, and Wei An. Deep video super-resolution using hr optical flow estimation. *IEEE Transactions on Image Processing*, 29:4323–4336, 2020.

[38] Pichao Wang, Xue Wang, Fan Wang, Ming Lin, Shuning Chang, Wen Xie, Hao Li, and Rong Jin. Kvt: k-nn attention for boosting vision transformers. *arXiv preprint arXiv:2106.00515*, 2021.

[39] Manuel Werlberger, Thomas Pock, Markus Unger, and Horst Bischof. Optical flow guided tv-l 1 video interpolation and restoration. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 273–286. Springer, 2011.

[40] Li Xu, Jiaya Jia, and Yasuyuki Matsushita. Motion detail preserving optical flow estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1744–1757, 2011.

[41] Shengyu Zhao, Yilun Sheng, Yue Dong, Eric I Chang, Yan Xu, et al. Maskflownet: Asymmetric feature matching with learnable occlusion mask. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6278–6287, 2020.