AccFlow: Backward Accumulation for Long-Range Optical Flow

Guangyang Wu[1,2*] Xiaohong Liu[1†] Kunming Luo[5‡] Xi Liu[2‡] Qingqing Zheng[3]
Shuaicheng Liu[2] Xinyang Jiang[4] Guangtao Zhai[1] Wenyi Wang[2†]
[1]Shanghai Jiao Tong University [2]University of Electronic Science and Technology of China
[3]Shenzhen Institute of Advanced Technology [4]Microsoft Research Asia
[5]Hong Kong University of Science and Technology

Abstract

Recent deep learning-based optical flow estimators have exhibited impressive performance in generating local flows between consecutive frames. However, the estimation of long-range flows between distant frames, particularly under complex object deformation and large motion occlusion, remains a challenging task. One promising solution is to accumulate local flows explicitly or implicitly to obtain the desired long-range flow. Nevertheless, the accumulation errors and flow misalignment can hinder the effectiveness of this approach. This paper proposes a novel recurrent framework called AccFlow, which recursively backward accumulates local flows using a deformable module called as AccPlus. In addition, an adaptive blending module is designed along with AccPlus to alleviate the occlusion effect by backward accumulation and rectify the accumulation error. Notably, we demonstrate the superiority of backward accumulation over conventional forward accumulation, which to the best of our knowledge has not been explicitly established before. To train and evaluate the proposed AccFlow, we have constructed a large-scale high-quality dataset named CVO, which provides ground-truth optical flow labels between adjacent and distant frames. Extensive experiments validate the effectiveness of AccFlow in handling long-range optical flow estimation. Codes are available at https://github.com/mulns/AccFlow.

1. Introduction

Optical flow is ideally a dense field of motion vectors that depicts the pixel-wise correspondence of two video frames. Since a variety of downstream applications (*e.g.*, video editing [3, 14, 54], action recognition [42], and object tracking [1]) significantly benefit from the accuracy of flow estimation, optical flow estimation turns out to be a long-standing fundamental task in computer vision [41, 39, 40, 27, 26, 55, 25, 13].

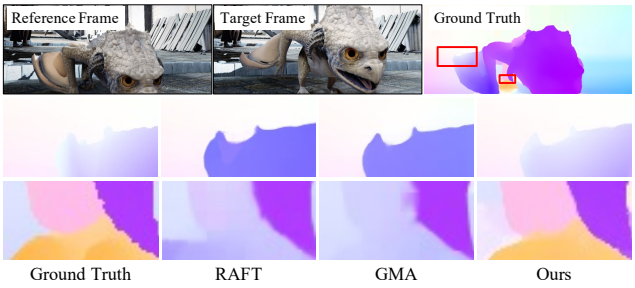


Figure 1: Comparisons of our method with RAFT [48] and GMA [19] on HS-Sintel dataset [18]. Zoom-in regions are annotated in red boxes. Our method outperforms other methods especially for occluded area.

Recent advances [6, 47, 48] resort to deep learning to estimate optical flow and achieve promising accuracy. Although remarkable performance has been achieved in the *local* flow estimation between two adjacent frames, it is non-trivial to estimate the *long-range* flow that records the pixel correspondence between two distant frames.

The long-range optical flow is a grounded research topic that has plenty of practical applications. For instance, in video completion [8], long-range optical flow is beneficial to the detail compensation between distant frames; in video key-point propagation [11], since the long-range optical flow performs holistic pixel-tracking in nature, it frees the quantity limitation of tracked pixels; in video super-resolution [31], it enables better inter-frame alignment in one sliding window; and in segmentation mask propagation [51], it provides an explicit approach to propagate masks to distant frames, improving the interpretability compared to the implicit matching. The above examples take a glance at the wide applications of long-range optical flow. More significantly, the success of this task has the potential to break through the performance bottleneck of relevant tasks.

† Corresponding authors. ‡ Equal contribution. * Work was partially finished at the University of Electronic Science and Technology of China.

Surprisingly, even though the long-range optical flow is significant and can benefit many related tasks, few works put effort on this research line. One possible reason is the lack of public datasets that provide ground-truth bidirectional cross-frame optical flows for training and validation. In literature, the early attempt to address this long-range optical flow task is Lim *et al*. [22], which proposed a method based on the *forward* flow accumulation, in which the flows of adjacent frames are added successively along the motion trajectories. The recent work [18] follows this idea and reasons the occlusion regions from high-frame-rate frames. Apart from these, one can simply estimate the long-range flow by employing methods specified for local flow [19, 48]. As shown in Figure 1, since the influence of occlusion is positively related to the time interval between two frames, the accuracy of flow estimation from these methods would be deteriorated severely or even be unacceptable when the time interval is beyond a threshold. In addition, one can also traverse all pixels in a frame and employ the pixel-tracking methods [38, 11] to produce the long-range dense flow, which has huge computational overheads and cannot be used in applications requiring dense flow. To sum up, a considerate long-range optical flow method should address the following challenging issues:

1) **Occlusion** As the time interval increases, the flow estimation of two distant frames suffers significant degradation owing to the inter-frame occlusion. Therefore, without a specific design, the common methods that aim at dealing with local flows perform poorly. Janai *et al*. [18] formulate it as an energy minimization problem and found it highly non-convex, so they exploit the linearity of small motions and reasons about occlusions from multiple frames. However, this strategy is based on high-frame-rate videos ($\geq$ 240 FPS) and not applicable on regular videos.

2) **Accumulation error** Although flow accumulation is a promising solution to tackle long-range flow estimation, it also brings the accumulation error, resulting in inaccurate estimation in non-occluded regions. Therefore, the effectiveness of accumulation error compensation is critical. Lim *et al*. [22] and Janai *et al*. [18] constrained the photo consistency of warped frames to shrink accumulated error. However, the photo consistency loss is not comprehensive for flow estimation as revealed in [17, 24].

3) **Efficiency** The computational complexity of long-range optical flow should be controlled at an appropriate level to support the downstream tasks in practice. Therefore, the pixel-tracking methods [38, 11], which iterative estimate the per-pixel long-range displacement, do not satisfy this requirement.

To address the above issues, we propose a novel framework, named AccFlow, to estimate long-range optical flow by progressively backward accumulating local flows with effective corrections. More specifically, to alleviate the occlusion effect, we propose the *backward accumulation*, a new accumulation strategy distinct from the *forward accumulation* pipeline, and elaborate a corresponding deep module, named AccPlus. More details about the difference between backward and forward accumulation can be found in Section 3.1 and 3.2. The AccFlow framework consists of three components: an arbitrary optical flow estimator, the AccPlus module, and an adaptive blending module. The arbitrary optical flow estimator is used to estimate local flows and long-range initial flow. The AccPlus performs the backward accumulation in feature domain. The adaptive blending module rectifies the accumulated error. Furthermore, to train and validate our AccFlow, we elaborately build a large-scale synthetic dataset, named CVO (cross-frame video optical flows). Different from other synthetic flow datasets [5, 6], the CVO includes *comprehensive* cross-frame bidirectional flow annotations. The CVO also includes more challenging cases that have large pixel displacement and severe occlusion.

The contributions of this paper can be summarized as follows:

- We propose a novel **backward accumulation** strategy to alleviate the long-range occlusion.
- We build the CVO, a new large-scale synthetic dataset with comprehensive **cross-frame** optical flow annotations.
- We propose the **AccFlow framework** which is simple yet effective to predict the long-range optical flow and achieves the state-of-the-art results on several benchmarks.

2. Related Works

2.1. Adjacent Frame Optical Flow Estimation

Optical flow methods can be categorized into two-frame and multi-frame methods according to the number of input frames. For two-frame methods, traditional algorithms [4, 45, 36] obtain optical flow by minimizing well-designed energy functions based on the brightness constancy assumption. By training a convolutional network on a synthetic dataset, FlowNet [6] first established a deep learning approach for optical flow estimation. After that, the performance of optical flow estimation is gradually improved by various works, such as FlowNet2 [16], PWC-Net [47], and IRR-PWC [15]. Recently, RAFT [48] proposed a new paradigm to estimate optical flow by introducing 4D correlation volume and recurrent network. Following RAFT, graph reasoning [30], global motion aggregation [19], kernel patch attention [29], and cross-attention transformer [44] are further proposed to improve the accuracy and efficiency.

The purpose of multi-frame optical flow estimation is to estimate the optical flow of adjacent frames by utilizing the temporal information of multiple video frames. Traditional methods achieve multi-frame optical flow estimation by phase-based representations of local image structure [7, 12], spatial-temporal regularization term [18, 52, 43], constant velocity prior [18, 49, 37, 46, 50], constant acceleration assumption [2, 20], and directional prior [32]. Recently, deep-based multi-frame methods are proposed to fuse flow prediction [35] or feature [34, 9] from previous frame pair into the current estimation process.

Although these optical flow methods have achieved remarkable performance, they mainly focus on estimating optical flow of two adjacent frames, leaving the long-range optical flow of non-adjacent frames rarely being explored.

2.2. Non-adjacent Frame Optical Flow Estimation

Lim *et al*. [23] proposed the early work to obtain the cross-frame optical flow, where the Lucas-Kanade method [28] is used to produce optical flow at a high frame rate and the accumulation strategy is designed to generate optical flow at a standard frame rate. After that, this accumulation method is improved by accumulation error modeling and correction [21, 22, 18]. Janai *et al*. [18] cast this task as an energy minimization problem, and opt for a data-driven hypothesis generation strategy for optimization. Recently, Harley *et al*. [11] proposed a deep CNN network, PIPs, to estimate cross-frame sparse optical flow from the perspective of per-pixel tracking over the video sequence. Although PIPs has achieved state-of-the-art performance for video pixel tracking, it is difficult to obtain long-range dense optical flow due to the lack of spatial coherence information. In this paper, we deeply analyze the drawbacks of existing accumulation strategies and propose a new accumulation framework for obtaining long-range dense optical flow.

3. Methods

Let $\mathcal{I} = \{\mathbf{I}_1, \dots, \mathbf{I}_N\}$ denote a video sequence with N image frames $\mathbf{I}_t \in \mathbb{R}^{w\times h\times 3}$ of size $w \times h$ and 3 color channels. Let $\mathbf{F}_{i,j} \in \mathbb{R}^{w\times h\times 2}$ denote the optical flow field from the reference image $\mathbf{I}_i$ to the target image $\mathbf{I}_j$. Specifically, for each pixel $\mathbf{x} \in \Omega_i = \{1, \dots, w\} \times \{1, \dots, h\}$ in reference image $\mathbf{I}_i$, $\mathbf{F}_{i,j}(\mathbf{x}) \in \mathbb{R}^2$ describes the apparent motion from frame I_i to I_j.

Our goal is to estimate the long-range optical flow field $\mathbf{F}_{1,N}$ by accumulating all intermediate local flow fields $\{\mathbf{F}_{1,2}, \dots, \mathbf{F}_{N-1,N}\}$. To achieve this, Lim *et al*. [22] and Janai *et al*. [18] formulate it as a dense pixel tracking task and obtain the long-range flow by tracking through pixel trajectories. In this paper, we refer to these approaches as the *forward accumulation*. In Section 3.1, we revisit the forward accumulation process and provide a formalization of it. The essential problem inherent in this process is analyzed, and a solution referred to as *backward accumulation* is proposed in Section 3.2. Subsequently, we introduce in Section 3.3 the proposed AccFlow framework that accomplishes the aforementioned backward accumulation to mitigate the occlusion effect and rectify the accumulated error. Additionally, we introduce the proposed CVO dataset which provides synthesized video with ground-truth long-range optical flow between distant frames in Section 3.4.

3.1. Revisiting the Forward Accumulation

Generally, the accumulation process is a recursive procedure to fuse all intermediate local flows together. For brevity, we define the fusion of two adjacent optical flows $\mathbf{F}_{i,k}$ and $\mathbf{F}_{k,j}$ as $\oplus$, and we present the fused flow $\mathbf{F}_{i,j}$ as:

$$\mathbf{F}_{i,j} = \mathbf{F}_{i,k} \oplus \mathbf{F}_{k,j} \tag{1}$$

where $i, k, j \in [1, N]$ denote three time stamps satisfying $i < k < j$. Since the adjacent flows $\mathbf{F}_{i,k}$ and $\mathbf{F}_{k,j}$ start at different frames (*i.e*., frame $\mathbf{I}_i$ and $\mathbf{I}_k$), in order to obtain the target flow $\mathbf{F}_{i,j}$ which starts at frame $\mathbf{I}_i$, we need to warp the start point of each motion vector in $\mathbf{F}_{k,j}$ to align them with $\mathbf{F}_{i,k}$, and then add the two flows pixel-wise. Let $\widetilde{\mathbf{F}}^i_{k,j}$ denote the warped $\mathbf{F}_{k,j}$ starting at frame $\mathbf{I}_i$, we have:

$$\widetilde{\mathbf{F}}^i_{k,j}(\mathbf{x}) = \mathbf{F}_{k,j}(\mathbf{x} + \mathbf{F}_{i,k}(\mathbf{x})) \tag{2}$$

for each pixel $\mathbf{x}$ in reference image $\mathbf{I}_i$. Then we obtain the target flow $\mathbf{F}_{i,j}$ by:

$$\mathbf{F}_{i,j}(\mathbf{x}) = \mathbf{F}_{i,k}(\mathbf{x}) + \widetilde{\mathbf{F}}^i_{k,j}(\mathbf{x}). \tag{3}$$

However, as Janai *et al*. [18] revealed, the reference pixel $\mathbf{x} \in \Omega_i$ can be forward occluded in frame $\mathbf{I}_k$, which leads to wrong warping results in Equation (1)-(3). Therefore, researchers usually speculate on the occlusion mask and solve the occluded regions by estimation. For brevity, we define the binary occlusion mask $\mathbf{O}_{i,k}$, where $\mathbf{O}_{i,k}(\mathbf{x}) \in \{0, 1\}$ specifies whether pixel $\mathbf{x} \in \Omega_i$ is forward occluded from frame $\mathbf{I}_i$ to $\mathbf{I}_k$. Equation (1)-(3) valid only when pixel $\mathbf{x} \in \Omega_i$ is not occluded in frame I_k (*i.e*., $\mathbf{O}_{i,k}(\mathbf{x}) = 0$). As for occluded pixels (*i.e*., $\mathbf{O}_{i,k}(\mathbf{x}) = 1$), its optical flow has to be estimated by some carefully designed occlusion solvers. For easy notation, function $solveOcc$ denotes occlusion solvers in general , and $\mathbf{P}_{i,j} \in \mathbb{R}^{w\times h\times 2}$ denote the estimated flows in occluded region, where

$$\mathbf{P}_{i,j} = solveOcc(\mathbf{F}_{i,k}, \mathbf{F}_{k,j}, \mathbf{O}_{i,k}). \tag{4}$$

Therefore, Equation (3) can be re-formulated as:

$$\mathbf{F}_{i,j}(\mathbf{x}) = \begin{cases} \mathbf{F}_{i,k}(\mathbf{x}) + \widetilde{\mathbf{F}}^i_{k,j}(\mathbf{x}) & \text{if } \mathbf{O}_{i,k}(\mathbf{x}) = 0, \\ \mathbf{P}_{i,j}(\mathbf{x}) & \text{if } \mathbf{O}_{i,k}(\mathbf{x}) = 1. \end{cases} \tag{5}$$

Algorithm 1: The Forward Accumulation

Input: $\{\mathbf{F}_{t,t+1} \mid t \in [1, N-1]\}$
Output: $\mathbf{F}_{1,N}$
for $t \leftarrow 2$ **to** $N-1$ **:**
  $\mathbf{O}_{1,t} \leftarrow getOcc(\mathbf{F}_{1,t}, \mathbf{F}_{t,t+1})$
  $\mathbf{P}_{1,t+1} \leftarrow solveOcc(\mathbf{F}_{1,t}, \mathbf{F}_{t,t+1}, \mathbf{O}_{1,t})$
  for $\mathbf{x} \in \Omega_1$ **:**
    $\widetilde{\mathbf{F}}^1_{t,t+1}(\mathbf{x}) \leftarrow \mathbf{F}_{t,t+1}(\mathbf{x} + \mathbf{F}_{1,t}(\mathbf{x}))$
    if $\mathbf{O}_{1,t}(\mathbf{x}) = 0$ **:**
      $\mathbf{F}_{1,t+1}(\mathbf{x}) \leftarrow \mathbf{F}_{1,t}(\mathbf{x}) + \widetilde{\mathbf{F}}^1_{t,t+1}(\mathbf{x})$
    elif $\mathbf{O}_{1,t}(\mathbf{x}) = 1$ **:**
      $\mathbf{F}_{1,t+1}(\mathbf{x}) \leftarrow \mathbf{P}_{1,t+1}(\mathbf{x})$

Algorithm 2: The Backward Accumulation

Input: $\{\mathbf{F}_{t,t+1} \mid t \in [1, N-1]\}$
Output: $\mathbf{F}_{1,N}$
for $t \leftarrow N-1$ **to** 2 **:**
  $\mathbf{O}_{t-1,t} \leftarrow getOcc(\mathbf{F}_{t-1,t}, \mathbf{F}_{t,N})$
  $\mathbf{P}_{t-1,N} \leftarrow solveOcc(\mathbf{F}_{t-1,t}, \mathbf{F}_{t,N}, \mathbf{O}_{t-1,t})$
  for $\mathbf{x} \in \Omega_{t-1}$ **:**
    $\widetilde{\mathbf{F}}^{t-1}_{t,N}(\mathbf{x}) \leftarrow \mathbf{F}_{t,N}(\mathbf{x} + \mathbf{F}_{t-1,t}(\mathbf{x}))$
    if $\mathbf{O}_{t-1,t}(\mathbf{x}) = 0$ **:**
      $\mathbf{F}_{t-1,N}(\mathbf{x}) \leftarrow \mathbf{F}_{t-1,t}(\mathbf{x}) + \widetilde{\mathbf{F}}^{t-1}_{t,N}(\mathbf{x})$
    elif $\mathbf{O}_{1,t}(\mathbf{x}) = 1$ **:**
      $\mathbf{F}_{t-1,N}(\mathbf{x}) \leftarrow \mathbf{P}_{t-1,N}(\mathbf{x})$

The forward accumulation process recursively performs the above operations. Specifically, with the time index t increases from 2 to $N-1$, we recursively produce $\mathbf{F}_{1,t+1}$ by fusing the pre-obtained flow $\mathbf{F}_{1,t}$ and the local flow $\mathbf{F}_{t,t+1}$ as follows:

$$\mathbf{F}_{1,t+1} = \mathbf{F}_{1,t} \oplus \mathbf{F}_{t,t+1}, \tag{6}$$

where for each pixel $\mathbf{x} \in \Omega_1$ in reference image $\mathbf{I}_1$, we have

$$\mathbf{F}_{1,t+1}(\mathbf{x}) = \begin{cases} \mathbf{F}_{1,t}(\mathbf{x}) + \widetilde{\mathbf{F}}^1_{t,t+1}(\mathbf{x}) & \text{if } \mathbf{O}_{1,t}(\mathbf{x}) = 0, \\ \mathbf{P}_{1,t+1}(\mathbf{x}) & \text{if } \mathbf{O}_{1,t}(\mathbf{x}) = 1, \end{cases} \tag{7}$$

where the occlusion mask $\mathbf{O}_{1,t}$ is usually estimated as well. We denote the occlusion reasoning methods as $getOcc$ in general:

$$\mathbf{O}_{1,t} = getOcc(\mathbf{F}_{1,t}, \mathbf{F}_{t,t+1}). \tag{8}$$

For clarity, we present the pseudocode of the forward accumulation process in Algorithm 1.

3.2. Backward Accumulation

Previous research [18] has shown that the forward accumulation can generate high quality motion hypotheses for visible regions, but the occluded regions limit its performance. In this subsection, we first analyze the occlusion area in the forward accumulation process, then propose a new solution to alleviate the occlusion effect.

Let $\Delta = |k - i| \geq 1$ denote the time interval, we define the proportion of occluded area of $\mathbf{O}_{i,k}$ as:

$$\alpha^i_\Delta = \frac{\sum_{\mathbf{x} \in \Omega_i} \mathbf{O}_{i,k}(\mathbf{x})}{h \times w}, \tag{9}$$

where $\alpha^i_\Delta \in [0, 1]$. We begin by analyzing the case of a one-dimensional object moving with constant velocity, assuming that the object is of length δw pixels, the canvas length is $M \gg \delta w$, the velocity of the object is v pixels per frame, and the background is fixed. From time $t = 1$ to $t = k$, the proportion of forward occluded area is calculated as:

$$\alpha^1_{|k-1|} = \frac{\min\{v \times |k-1|, \delta w\}}{M}, \tag{10}$$

which is positively correlated with the time interval $|k - 1|$. Similar conclusions can be extended to two-dimensional cases. Thus, the inequality

$$\alpha^i_{\Delta+1} \geq \alpha^i_\Delta, \tag{11}$$

holds for linear motion.

While the assumption of linear motion may not always hold in practical scenarios, our experiments show that Equation (11) remains valid when a significant number of samples are tested. The statistical results over 5000 samples are provided in terms of box-plot in Figure 2, which demonstrates that the α^i_Δ is positively correlated with Δ as Equation (10) indicates. This conclusion is important for the following analysis.

Algorithm 1 shows that the occlusion proportion α^1_{t-1} of $\mathbf{O}_{1,t}$ increases progressively with t increases, which significantly burdens the occlusion solver. Although existing techniques [19, 48] can powerfully solve occlusion with deep neural networks (DNN), the constant increment of the occlusion proportion is still a challenge that might consume substantial computational resources.

To address the above critical issue, we propose a simple solution, named the *backward accumulation*, where we reverse the accumulation order without extra computational complexity involved. As analyzed in Equation (3)-(4), the alignment operation introduces errors in the forward occluded regions, and as revealed in Equation (11), they are proportionally correlated with the time interval. In each step of accumulation process, we can simplify the problem as the alignment of two optical flows, one of which has a larger magnitude (pre-obtained from the last step) and another one has a smaller magnitude (the local flow). The forward accumulation chooses to align two flows along the larger one, which essentially leads to a larger occlusion area. Therefore, we propose to align the two flows along the smaller

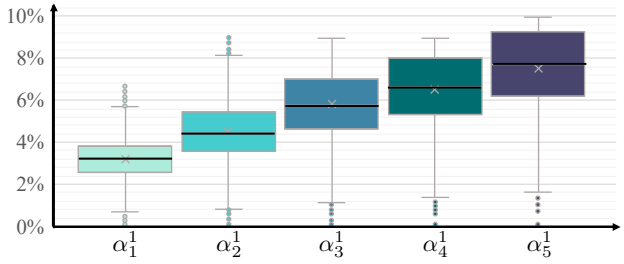


Figure 2: Box-plot of the occlusion proportion α_Δ^1 over 5000 samples, the occlusion proportion (Y-axis) increases with the time interval Δ (X-axis) increases.

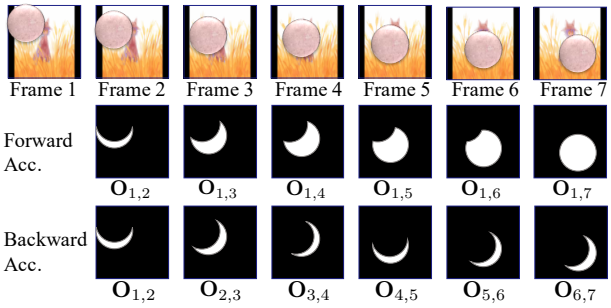


Figure 3: Visualization of occlusion masks during accumulation. White regions denote occluded area.

one. Specifically, with time variable t decreases from $N-1$ to 2, we recursively produce the long-range flow $\mathbf{F}_{t-1,N}$ by fusing the pre-obtained flow $\mathbf{F}_{t,N}$ and the local flow $\mathbf{F}_{t-1,t}$ as follows:

$$\mathbf{F}_{t-1,N} = \mathbf{F}_{t-1,t} \oplus \mathbf{F}_{t,N}, \tag{12}$$

where for each pixel $\mathbf{x} \in \Omega_{t-1}$ in reference image $\mathbf{I}_{t-1}$, we have

$$\mathbf{F}_{t-1,N}(\mathbf{x}) = \begin{cases} \mathbf{F}_{t-1,t}(\mathbf{x}) + \widetilde{\mathbf{F}}_{t,N}^{t-1}(\mathbf{x}) & \text{if } \mathbf{O}_{t-1,t}(\mathbf{x}) = 0, \\ \mathbf{P}_{t-1,N}(\mathbf{x}) & \text{if } \mathbf{O}_{t-1,t}(\mathbf{x}) = 1, \end{cases} \tag{13}$$

and the occlusion mask is obtained by:

$$\mathbf{O}_{t-1,t} = getOcc(\mathbf{F}_{t-1,t}, \mathbf{F}_{t,N}). \tag{14}$$

By doing this, we form the backward accumulation process presented in Algorithm 2.

As evident from the recursive process, the occluded regions are pixels with $\mathbf{O}_{t-1,t}(\mathbf{x}) = 1, \mathbf{x} \in \Omega_{t-1}$, at each step. The occlusion proportion defined in Equation (9) is α_1^{t-1} here. During the backward accumulation, although the reference image undergoes changes, the occluded region remains at a minimum level, particularly when compared to the forward accumulation method where the occluded region progressively increases. We visualize this observation in Figure 3. The reduced occluded area enables the occlusion solver to handle the occlusion more efficiently.

3.3. AccFlow Framework

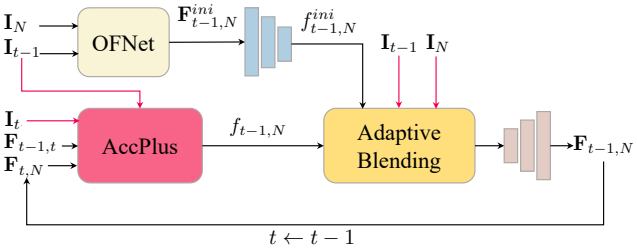
In this section, we present AccFlow, a deep framework that employs the backward accumulation to estimate accurate long-range optical flow. The framework consists of three components, an arbitrary optical flow estimator OFNet (*e.g.*, RAFT, GMA, *etc.*), the AccPlus module, and the adaptive fusion module. Initially, local flows $\{\mathbf{F}_{t,t+1} \mid t \in [1, N-1]\}$ are obtained from the pretrained OFNet as inputs of AccFlow. The AccFlow recursively produces the long-range flow $\mathbf{F}_{t-1,N}$ with time t decreases from $N-1$ to 2 and the recurrent structure is shown in Figure 4a.

The AccPlus Module. Following the Algorithm 2, we implement the backward accumulation in the AccPlus module to perform flow fusion in feature domain as shown in Figure 4b. At each stage, given the local flow $\mathbf{F}_{t-1,t}$ and pre-obtained flow $\mathbf{F}_{t,N}$, we encode them into motion features $f_{t-1,t}$ and $f_{t,N}$ with a motion encoder. The motion encoder spatially downscales features by $1/4$ times. The occlusion mask $\mathbf{O}_{t-1,t}$ is determined by $getOcc$ which is a simple warping operation in this paper. More details about the encoder and $getOcc$ are provided in appendix. Afterwards, we warp the motion features $f_{t,N}$ to align them with $f_{t-1,t}$ by deformable convolution and produce $\widetilde{f}_{t,N}$. In the AccPlus, we implement $solveOcc$ in Algorithm 2 by a set of convolutional layers. Specifically, we concatenate $\widetilde{f}_{t,N}$ and $f_{t-1,t}$ along the channel dimensional, where $f_{t-1,t}$ provides the spatial coherence information for handling occlusion. The concatenated feature is then processed by multiple convolutional layers. The resulting output features, denoted as $p_{t-1,N}$, are then merged with $\widetilde{f}_{t,N}$ and $f_{t-1,t}$ to produce the final target motion feature $f_{t-1,N}$.

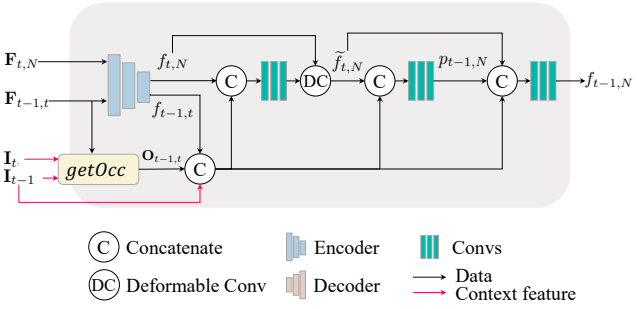
The adaptive blending module. Directly decoding the output features $f_{t-1,N}$ of AccPlus and passing them to next stage may result in the accumulation error. To mitigate this issue, an adaptive blending module is added to suppress the accumulation error by using the directly estimated long-range flow as prior information. Specifically, we first establish an initial long-range optical flow $\mathbf{F}_{t-1,N}^{ini}$ with the pretrained OFNet, and then encode it into a motion feature $f_{t-1,N}^{ini}$ with the motion encoder (share parameters with the one in AccPlus). Subsequently, the adaptive blending module takes the two motion features (*i.e.*, $f_{t-1,N}^{ini}$ and $f_{t-1,N}$) and corresponding video frames as inputs to calculate an adaptive confidence mask. The confidence mask is then used to fuse them with attention mechanism, and the output motion features are decoded into the optical flow $\mathbf{F}_{t-1,N}$ with a motion decoder. Details of the motion decoder are provided in appendix.

3.4. CVO Dataset

Existing optical flow datasets only provide the local optical flow annotations. In order to provide the ground-truth



(a) AccFlow Framework.



(b) AccPlus Module.

Figure 4: Illustration of the network structure. (a) The AccFlow framework. Time t decreases from $N-1$ to 2 to obtain long-range flow $\mathbf{F}_{1,N}$. OFNet is an arbitrary flow estimator. (b) The AccPlus module, an efficient module that implements the backward accumulation in feature domain. The red arrows signify the encoding of images into context features by a context encoder, which adheres to the structure outlined in [48].

(GT) long-range optical flows, we construct a cross-frame video optical flow dataset (CVO), consisting of 12K synthetic video sequences and GT optical flow labels across different frame intervals. This dataset is essential for the research on long-range optical flow estimation and other related tasks.

Dataset Collection We generate the CVO dataset using Kubric [10], which is a data generation pipeline for creating semi-realistic synthetic multi-object videos. We first simulate the movement of multiple objects, and then render frames along with optical flow annotations. For each video sequence, we render 7 frames of size 512×512 at 60 FPS (frame per second) in conjunction with the bidirectional optical flow of adjacent frames. In addition, we provide cross-frame bidirectional optical flows across different frame intervals. All the cross-frame flows take the first frame as reference. We further render the RGB video frames with and without random motion blur, which is denoted as *Clean* and *Final* sets. We partition all video sequences into two subsets, 11K sequences and 500 sequences, which serve as the training and validation splits, respectively.

Comparisons with Existing Datasets The CVO dataset contains richer annotations compared with existing optical flow datasets [5, 6] since it provides cross-frame bidirectional optical flow annotations. Moreover, the CVO contains more challenging samples with large motion and complex occlusion. We compare the flow magnitudes among different datasets by plotting the statistical histograms in Figure 5. Even though the FlyingThings3D [6] has similar flow magnitude distribution compared to CVO, the CVO contains more extreme large motions (flow magnitude ≥ 125 pixels). According to our experiments, the proposed CVO is sufficient to support researches on long-range optical flow estimation and other related tasks.

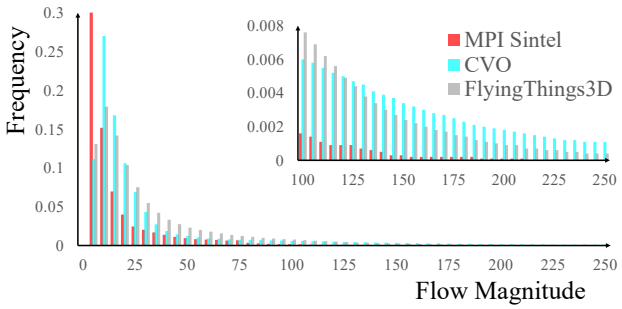


Figure 5: The histogram comparisons of the flow magnitude between the training set of CVO and public datasets, such as MPI Sintel [5] and FlyingThings3D [33].

4. Experiments

4.1. Validation Benchmarks

CVO: We adopt the CVO testing set, which consists of *Clean* and *Final* splits, as one of our validation benchmarks. Each split contains 500 sequences for evaluation. In each sequence, there are 7 frames of size 512×512, and the default GT optical flow $\mathbf{F}^{gt}_{1,7}$. If experiments on other frame intervals are desired, we provide the corresponding GT flow $\mathbf{F}^{gt}_{1,i}, i \in [2, 6]$ (denoted as CVO-i).

HS-Sintel: MPI Sintel [5] is a commonly used optical flow benchmark generated from the realistic animated film. However, it only provides GT flows at 24 FPS. Therefore, we use the High-Speed Sintel videos [18], namely HS-Sintel, as an alternative. Specifically, Janai *et al.* [18] selected a subset of 19 sequences from the MPI Sintel training set (clean pass) and re-rendered them 24 FPS to 1008 FPS with $4\times$ resolution. Unfortunately, the GT flows at other frame rates of HS-Sintel are not publicly available. Therefore, we use the GT flows at 24 FPS of MPI Sintel as labels to evaluate the estimates from video sequences at 1008 FPS of HS-Sintel.

4.2. Implementation Details

Loss function: During the recurrent process to obtain the target flow $\mathbf{F}_{1,N}$, the AccFlow also produces intermediate

Method	HS-Sintel			CVO (*Clean*)			CVO (*Final*)			Inference
	ALL	NOC	OCC	ALL	NOC	OCC	ALL	NOC	OCC	time (s)
RAFT	2.141	1.124	7.169	5.687	2.798	13.233	6.653	3.812	13.891	0.129
RAFT-*Lim*	3.868	1.845	12.63	11.96	6.573	31.10	12.34	6.938	31.45	0.956
RAFT-*w*	1.921	1.004	6.623	5.259	2.274	12.59	5.508	2.493	12.90	0.525
Acc+RAFT (ours)	1.709	1.163	5.639	3.170	1.623	8.113	3.283	1.714	8.261	0.813
GMA	2.291	1.330	7.139	5.757	2.775	13.58	6.265	3.530	13.71	0.234
GMA-*Lim*	3.871	1.764	12.79	12.22	6.708	31.40	12.42	7.038	31.61	2.159
GMA-*w*	1.924	1.043	6.458	5.136	2.137	12.49	5.515	2.502	12.81	1.167
Acc+GMA (ours)	1.568	1.091	5.003	3.583	1.807	8.868	3.752	1.979	9.030	1.499
RAFT*	2.567	1.426	7.717	4.445	1.948	11.73	4.537	2.003	11.70	0.129
RAFT*-*Lim*	3.657	1.611	12.36	23.34	6.543	32.90	13.02	7.033	33.82	0.956
RAFT*-*w*	2.139	1.059	6.963	3.738	1.052	10.41	3.808	1.162	10.14	0.525
Acc+RAFT* (ours)	1.383	0.930	4.546	2.634	1.155	7.302	2.707	1.249	7.295	0.813
GMA*	2.520	1.469	7.600	4.638	2.342	11.33	4.633	2.114	11.36	0.234
GMA*-*Lim*	3.306	1.381	11.70	11.39	5.833	31.28	11.68	6.130	31.35	2.159
GMA*-*w*	1.888	0.946	6.516	3.832	1.082	10.38	3.807	1.159	10.10	1.167
Acc+GMA* (ours)	1.434	0.950	4.770	2.732	1.181	7.438	2.808	1.261	7.495	1.499
SlowFlow	2.58†	0.87†	9.45†	-	-	-	-	-	-	$\geq$ 500
PIPs	-	-	-	8.568	6.351	21.55	8.954	6.718	22.06	$\geq$ 500
GMFlow	2.055	1.024	7.132	5.801	2.680	13.521	6.506	3.402	14.21	0.341

Table 1: Comparisons of AccPlus framework with other methods on two benchmarks in terms of EPE $\downarrow$ on all regions (ALL) and occluded regions (OCC). The best and the second-best results are marked in red and blue, respectively. '-*Lim*' denotes the flow accumulation method in [22]. '-*w*' denotes the warm-start method (details in Section 4.3). For the SlowFlow [18], we refer to data in their paper (denoted with †). We report the inference time of 7 frames of size 512×512 per sample on an NVIDIA GTX3090 GPU.

flows $\mathbf{F}_{t,N}, t \in [1, N-2]$. Therefore, we train the network by supervising all the flow outputs with L1 loss:

$$\mathcal{L} = \frac{1}{N-2} \sum_{i=1}^{N-2} \|\mathbf{F}_{i,N} - \mathbf{F}_{i,N}^{gt}\|_1. \tag{15}$$

Training details: We train the AccFlow with the mixture of 'clean' and 'final' pass of CVO training set. We augment the training data by randomly cropping the input frames into patches of size 256×256. Other training hyperparameters (*e.g.*, learning rate and batch size) follow the default settings from [48]. By replacing the OFNet with different existing optical flow estimators, we train four models for comparison. Specifically, we embed the officially pretrained RAFT [48] and GMA [19] in AccFlow framework, respectively. On the one hand, we fix the parameter of OFNet and train other parameters from scratch, and produce Acc+RAFT and Acc+GMA, respectively. On the other hand, we fine-tune the parameter of OFNet and produce Acc+RAFT* and Acc+GMA*, respectively.

4.3. Alternative Approaches

Previously, several works [22, 18] have been focused on optical flow accumulation. Therefore, for more comprehensive comparisons, we consider some other alternative approaches to estimate long-range optical flow.

Direct estimation. One of the naive methods is to directly estimate long-range flow with two distant reference images. Other than RAFT and GMA, we also compare the GMFlow [53] which formulates the optical flow as a global matching problem to solve large motion. For fair comparisons, we also fine-tune the RAFT and GMA with training set of CVO, denoted as RAFT* and GMA*, respectively.

Pixel tracking. Another intuitive way is to use pixel tracking method to iteratively estimate the per-pixel long-range displacement. We use the SOTA pixel tracking method PIPs [11] to achieve this. Such process is time-consuming so we only test this method on CVO testing set.

Warm start. Zachary *et al*. [48] propose to estimate optical flow with warm start. This method can also be applied in flow accumulation, that is, we use the pre-obtained $\mathbf{F}_{1,t}$ as an initialized flow input to estimate $\mathbf{F}_{1,t+1}$. This procedure is essentially an implicit forward accumulation process, thus we include it into comparisons.

4.4. Comparisons with Existing Methods

We compare the existing methods in terms of the average End-Point-Error (EPE) applied to all pixels (ALL) and occlusion regions (OCC). In Table 1, we compare our AccFlow with previous methods on two benchmarks, and our AccFlow outperforms all the previous methods by a large margin especially for occluded regions. Specifically, we notice that it is challenge for direct methods (the 1,5,9,13,18-*th* rows in Table 1) to produce long-range optical flow due to the extreme large motion and occlusion problems. For forward accumulation, the explicit methods (*i.e.*, [22] and [18]) fail to handle the constantly increased occlusion which result in inferior performance. PIPs can accurately estimate sparse motion but suffers from the lack of spatial coherence information for dense flow estimation. Moreover, the im-

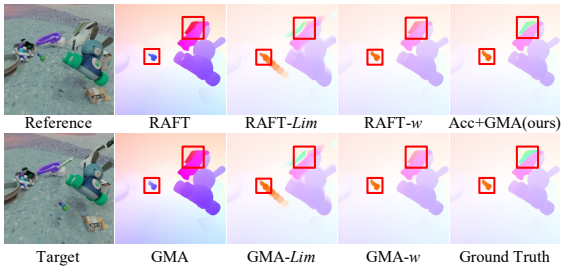


Figure 6: Visual quality comparisons on CVO dataset. Two small objects with large motions are emphasized with red boxes. More results can be found in the supplementary.

Acc+RAFT F.	Acc+RAFT B.	AB	HS-Sintel ALL	HS-Sintel NOC	HS-Sintel OCC	CVO (Final) ALL	CVO (Final) NOC	CVO (Final) OCC
✓			2.238	5.758	5.758	3.328	1.914	7.716
	✓		1.740	1.303	4.711	2.709	1.252	7.299
✓		✓	1.716	0.936	5.895	3.229	0.873	8.823
	✓	✓	1.383	0.930	4.546	2.707	1.249	7.295

Table 2: Ablation study of AccFlow framework (reported in EPE ↓). 'F.' denotes a modified AccPlus that accumulates the local flows in forward manner, 'B.' is the proposed AccPlus with backward accumulation, and 'AB' denotes the adaptive blending module.

plicit forward accumulation method (*i.e.*, warm start) is not specially designed for this task and fall short in tackling occlusion problem, but still brings certain performance gain compared with direct methods. Compared to all these methods, the AccFlow framework can decrease the average EPE error by large margin, which justifies the effectiveness of our framework for occlusion correction and non-occlusion correspondence enhancement.

Moreover, the qualitative comparisons are shown in Figure 6, where two small objects with large motion are annotated in red boxes. It can be seen that our AccFlow can produce accurate optical flows while the compared methods suffer from significant errors especially for occluded area.

4.5. Ablation Study

Backward VS. Forward accumulation: In Section 3.2, we demonstrate that the backward accumulation is less susceptible to occlusion effect than the forward one. In order to fairly compare the two methods, we design a modified AccPlus module which implements the forward accumulation in Table 2 (denoted as 'F.'). It is worth noting that the modification only change the inputs of network and no additional computational complexity is introduced. Detailed structure of the forward version of AccPlus is provided in appendix. In Table 2, we compare the backward accumulation with the forward one in terms of EPE under the same experimental settings. We can find that the backward version can deal with the occluded area more effectively than the forward version by large margin. This is because the backward accumulation has stable and minimum occlusion proportion at each step of iterations.

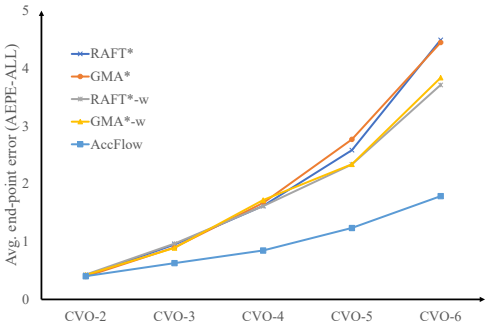


Figure 7: Average EPE ↓ (ALL) of long-range flows from the compared methods in different estimation ranges.

Adaptive blending module: In Section 3.3, we design the AccFlow framework not only to address occlusion problem but also suppress the accumulation error. Specifically, the adaptive blending module takes a directly estimated long-range flow as prior to rectify the cumulated flow. To evaluate this, we train networks w/. and w/o. the adaptive blending module (denoted as 'AB') in Table 2. The EPE is reduced by large margin especially for non-occluded area (NOC), which demonstrates the necessity of adaptive blending module for mitigating accumulated error.

Accumulation for different frame ranges In Figure 7, we show the results of long-range optical flow estimation in different estimation ranges. When the range increases, the EPE of the flows from our proposed AccFlow (Acc+GMA*) increases slower than that from direct estimation and the warm start methods. This observation shows the robustness of our proposed framework in different estimation ranges.

5. Conclusion

We propose the backward accumulation strategy for improved long-range optical flow estimation, surpassing prior methods. AccFlow employs feature domain backward accumulation and DNN-based error correction. Experimental results effectively address occlusion and accumulation errors. Ablation studies confirm superiority and adaptive blending's necessity. AccFlow notably reduces EPE on several benchmarks. In conclusion, AccFlow offers a simple, potent solution for flow accumulation, with scalability.

6. Acknowledgment

The work was supported in part by the Shanghai Pujiang Program under Grant 22PJ1406800, in part by the Guang-Dong Basic and Applied Basic Research Foundation under Project (No.2023A1515010644), and in part by Sichuan Provincial Key Laboratory of Intelligent Terminals under Grant SCITLAB-20016.

References

[1] Aseem Behl, Omid Hosseini Jafari, Siva Karthik Mustikovela, Hassan Abu Alhaija, Carsten Rother, and Andreas Geiger. Bounding boxes, segmentations and object coordinates: How important is recognition for 3d scene flow estimation in autonomous driving scenarios? In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 2574–2583, 2017. 1

[2] Michael J Black and Padmanabhan Anandan. Robust dynamic motion estimation over time. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 296–203, 1991. 3

[3] Nicolas Bonneel, James Tompkin, Kalyan Sunkavalli, Deqing Sun, Sylvain Paris, and Hanspeter Pfister. Blind video temporal consistency. *ACM Trans. Graph.*, 34:196:1–196:9, 2015. 1

[4] Thomas Brox, Andres Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow etimation based on a theory for warping. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pages 25–36, 2004. 2

[5] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pages 611–625, 2012. 2, 6

[6] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 2758–2766, 2015. 1, 2, 6

[7] David J Fleet and Allan D Jepson. Computation of component image velocity from local phase information. *Int. J. Comput. Vis.*, 5:77–104, 1990. 3

[8] Chen Gao, Ayush Saraf, Jia-Bin Huang, and Johannes Kopf. Flow-edge guided video completion. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pages 713–729, 2020. 1

[9] Pierre Godet, Alexandre Boulch, Aurélien Plyer, and Guy Le Besnerais. Starflow: A spatiotemporal recurrent cell for lightweight multi-frame optical flow estimation. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2462–2469, 2021. 3

[10] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 3749–3761, 2022. 6

[11] Adam W. Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022. 1, 2, 3, 7

[12] David J Heeger. Optical flow using spatiotemporal filters. *Int. J. Comput. Vis.*, 1:279–302, 1988. 3

[13] Shan Huang, Xiaohong Liu, Tao Tan, Menghan Hu, Xiaoer Wei, Tingli Chen, and Bin Sheng. Transmrsr: Transformer-based self-distilled generative prior for brain MRI super-resolution. *Arxiv preprint:2306.06669*, 2023. 1

[14] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pages 624–642, 2022. 1

[15] Junhwa Hur and Stefan Roth. Iterative residual refinement for joint optical flow and occlusion estimation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 5754–5763, 2019. 2

[16] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2462–2470, 2017. 2

[17] Joel Janai, Fatma Guney, Anurag Ranjan, Michael Black, and Andreas Geiger. Unsupervised learning of multi-frame optical flow with occlusions. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pages 690–706, 2018. 2

[18] Joel Janai, Fatma Guney, Jonas Wulff, Michael J Black, and Andreas Geiger. Slow flow: Exploiting high-speed cameras for accurate and diverse optical flow reference data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3597–3607, 2017. 1, 2, 3, 4, 6, 7

[19] Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. Learning to estimate hidden motions with global motion aggregation. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 9772–9781, 2021. 1, 2, 4, 7

[20] Ryan Kennedy and Camillo J Taylor. Optical flow with geometric occlusion estimation and fusion of multiple frames. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 364–377, 2015. 3

[21] SukHwan Lim, John Apostolopoulos, and AE Gamal. Benefits of temporal oversampling in optical flow estimation. In *Proc. IEEE Int. Conf. Image Process. (ICIP)*, pages 2567–2570, 2004. 3

[22] SukHwan Lim, John G. Apostolopoulos, and Abbas El Gamal. Optical flow estimation using temporally oversampled video. *IEEE Trans. Image Process.*, 14:1074–1087, 2005. 2, 3, 7

[23] SukHwan Lim and Abbas El Gamal. Optical flow estimation using high frame rate sequences. In *Proc. IEEE Int. Conf. Image Process. (ICIP)*, pages 925–928, 2001. 3

[24] Pengpeng Liu, Michael Lyu, Irwin King, and Jia Xu. Selflow:self-supervised learning of optical flow. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 4571–4580, 2019. 2

[25] Xiaohong Liu, Lei Chen, Wenyi Wang, and Jiying Zhao. Robust multi-frame super-resolution based on spatially weighted half-quadratic estimation and adaptive BTV regularization. *IEEE Trans. Image Process.*, 27(10):4971–4986, 2018. 1

[26] Xiaohong Liu, Lingshi Kong, Yang Zhou, Jiying Zhao, and Jun Chen. End-to-end trainable video super-resolution based on a new mechanism for implicit motion estimation and compensation. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*, pages 2405–2414. IEEE, 2020. 1

[27] Xiaohong Liu, Kangdi Shi, Zhe Wang, and Jun Chen. Exploit camera raw data for video super- resolution via hidden markov model inference. *IEEE Trans. Image Process.*, 30:2127–2140, 2021. 1

[28] Bruce D Lucas, Takeo Kanade, et al. *An iterative image registration technique with an application to stereo vision*, volume 81. Vancouver, 1981. 3

[29] Ao Luo, Fan Yang, Xin Li, and Shuaicheng Liu. Learning optical flow with kernel patch attention. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 8906–8915, 2022. 2

[30] Ao Luo, Fan Yang, Kunming Luo, Xin Li, Haoqiang Fan, and Shuaicheng Liu. Learning optical flow with adaptive graph reasoning. In *AAAI*, pages 1890–1898, 2022. 2

[31] Ziwei Luo, Youwei Li, Shen Cheng, Lei Yu, Qi Wu, Zhihong Wen, Haoqiang Fan, Jian Sun, and Shuaicheng Liu. BSRT: improving burst super-resolution with swin transformer and flow-guided deformable alignment. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh. (CVPRW)*, pages 997–1007, 2022. 1

[32] Daniel Maurer, Michael Stoll, and Andrés Bruhn. Directional priors for multi-frame optical flow. In *Proc. Brit. Mach. Vis. Conf. (BMVC)*, page 106, 2018. 3

[33] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 4040–4048, 2016. 6

[34] Michal Neoral, Jan Šochman, and Jiří Matas. Continual occlusion and optical flow estimation. In *ACCV*, pages 159–174, 2018. 3

[35] 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, 2019. 3

[36] Jerome Revaud, Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Epicflow: Edge-preserving interpolation of correspondences for optical flow. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1164–1172, 2015. 2

[37] Agustín Salgado and Javier Sánchez. Temporal constraints in large optical flow estimation. In *International Conference on Computer Aided Systems Theory*, pages 709–716, 2007. 3

[38] Peter Sand and Seth J. Teller. Particle video: Long-range motion estimation using point trajectories. *Int. J. Comput. Vis.*, 80:72–91, 2008. 2

[39] Zhihao Shi, Xiaohong Liu, Chengqi Li, Linhui Dai, Jun Chen, Timothy N. Davidson, and Jiying Zhao. Learning for unconstrained space-time video super-resolution. *IEEE Trans. Broadcast.*, 68(2):345–358, 2022. 1

[40] Zhihao Shi, Xiaohong Liu, Kangdi Shi, Linhui Dai, and Jun Chen. Video frame interpolation via generalized deformable convolution. *IEEE Trans. Multim.*, 24:426–439, 2022. 1

[41] Zhihao Shi, Xiangyu Xu, Xiaohong Liu, Jun Chen, and Ming-Hsuan Yang. Video frame interpolation transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 17461–17470. IEEE, 2022. 1

[42] K Simonyan and A Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proc. Adv. Neural Inform. Process. Syst. (NeurIPS)*, pages 568–576, 2014. 1

[43] Michael Stoll, Sebastian Volz, and Andrés Bruhn. Joint trilateral filtering for multiframe optical flow. In *Proc. IEEE Int. Conf. Image Process. (ICIP)*, pages 3845–3849, 2013. 3

[44] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 17602–17611, 2022. 2

[45] Deqing Sun, Stefan Roth, and Michael J Black. Secrets of optical flow estimation and their principles. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2432–2439, 2010. 2

[46] Deqing Sun, Erik Sudderth, and Michael Black. Layered image motion with explicit occlusions, temporal consistency, and depth ordering. *Proc. Adv. Neural Inform. Process. Syst. (NeurIPS)*, 23:2226–2234, 2010. 3

[47] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 8934–8943, 2018. 1, 2

[48] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pages 402–419, 2020. 1, 2, 4, 6, 7

[49] Sebastian Volz, Andres Bruhn, Levi Valgaerts, and Henning Zimmer. Modeling temporal coherence for optical flow. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 1116–1123, 2011. 3

[50] Chia-Ming Wang, Kuo-Chin Fan, Cheng-Tzu Wang, and Tong-Yee Lee. Estimating optical flow by integrating multi-frame information. *Journal of Information Science & Engineering*, 24:1719–1731, 2008. 3

[51] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2566–2576, 2019. 1

[52] Joachim Weickert and Christoph Schnörr. Variational optic flow computation with a spatio-temporal smoothness constraint. *Journal of mathematical imaging and vision*, 14:245–255, 2001. 3

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

[54] Yichao Yan, Bingbing Ni, Wendong Zhang, Jun Tang, and Xiaokang Yang. Cross-modality motion parameterization for fine-grained video prediction. *Computer Vision and Image Understanding*, 183:11–19, 2019. 1

[55] Guanghao Yin, Zefan Qu, Xinyang Jiang, Shan Jiang, Zhenhua Han, Ningxin Zheng, Xiaohong Liu, Huan Yang, Yuqing Yang, Dongsheng Li, and Lili Qiu. Online video streaming super-resolution with adaptive look-up table fusion. *arXiv preprint:2303.00334*. 1