# A Simple and Effective Fusion Approach for Multi-frame Optical Flow Estimation

Zhile Ren[1], Orazio Gallo[2], Deqing Sun[2], Ming-Hsuan Yang[3], Erik B. Sudderth[4], and Jan Kautz[2(✉)]

[1] Brown University, Providence, USA
[2] NVIDIA, Santa Clara, USA
jkautz@nvidia.com
[3] UC Merced, Merced, USA
[4] UC Irvine, Irvine, USA

**Abstract.** To date, top-performing optical flow estimation methods only take pairs of consecutive frames into account. While elegant and appealing, the idea of using more than two frames has not yet produced state-of-the-art results. We present a simple, yet effective fusion approach for multi-frame optical flow that benefits from longer-term temporal cues. Our method first warps the optical flow from previous frames to the current, thereby yielding multiple plausible estimates. It then fuses the complementary information carried by these estimates into a new optical flow field. At the time of writing, our method ranks first among published results in the MPI Sintel and KITTI 2015 benchmarks.

**Keywords:** Multi-frame optical flow · Temporal optical flow fusion

## 1 Introduction

Despite recent advances in optical flow estimation, it is still challenging to account for complicated motion patterns. At video rates, even such complicated motion patterns are smooth for longer than just two consecutive frames. This suggests that information from frames that are adjacent in time could be used to improve optical flow estimates. Indeed, numerous methods have been developed [2,3,9,10]. However, none of the top three optical flow algorithms on the major benchmark datasets uses more than two frames [4,6].

We observe that, for some types of motion and in certain regions, past frames may carry more valuable information than recent ones, even if the optical flow changes abruptly—as is the case of occlusion regions and out-of-boundary pixels. Kennedy and Taylor [8] also leverage this observation, and select which *one* of multiple flow estimates from adjacent frames is the best for a given pixel. We propose a method to *fuse* the available information. Specifically, we first estimate per-frame optical flow using a two-frame network module, and then warp multiple optical flow estimates from the past to the current frame, which we can fuse with a second neural network module.


L. Leal-Taixé and S. Roth (Eds.): ECCV 2018 Workshops, LNCS 11134, pp. 706–710, 2019.
https://doi.org/10.1007/978-3-030-11024-6_53

Our approach offers several advantages. First, it allows to fully capitalize on motion information from past frames. Second, our fusion network is agnostic to the algorithm that generates the two-frame optical flow estimates; any standard method can be used as an input, making our framework flexible and straightforward to upgrade when improved two-frame algorithms become available. Finally, if the underlying optical flow algorithm is differentiable, our approach can be trained end-to-end. Extensive experiments show that the proposed algorithm outperforms published state-of-the-art, two-frame optical flow methods by significant margins on the KITTI [6] and Sintel [4] benchmarks. To further validate our results, we present alternative baseline approaches incorporating recurrent neural networks with the state-of-the-art deep-learning optical flow estimation methods, and show that the fusion approach achieves significant performance gains.

## 2 Proposed Model: Temporal FlowFusion

For clarity reasons, we focus on three-frame optical flow estimation. Given three input frames $\mathbf{I}_{t-1}$, $\mathbf{I}_t$, and $\mathbf{I}_{t+1}$, our aim is to estimate the optical flow from frame $t$ to frame $t+1$, $\mathbf{w}^f_{t\to t+1}$. The superscript '$f$' indicates that it fuses information from all the frames. We use two-frame methods, such as PWC-Net [11], to estimate three motion fields, $\mathbf{w}_{t\to t+1}$, $\mathbf{w}_{t-1\to t}$, and $\mathbf{w}_{t\to t-1}$. We backward warp $\mathbf{w}_{t-1\to t}$ using $\mathbf{w}_{t\to t-1}$: $\widehat{\mathbf{w}}_{t\to t+1} = \mathcal{W}(\mathbf{w}_{t-1\to t}; \mathbf{w}_{t\to t-1})$, where $\mathcal{W}(\mathbf{x}; \mathbf{w})$ denotes warping the input $\mathbf{x}$ using the flow $\mathbf{w}$.

Now we have two candidates for the same frame: $\widehat{\mathbf{w}}_{t\to t+1}$ and $\mathbf{w}_{t\to t+1}$, we take inspiration from the work of Ilg *et al.* who perform optical flow fusion in the spatial domain for two-frame flow estimation [7]. We extend this approach to the temporal domain. Our fusion network takes two flow estimates $\widehat{\mathbf{w}}_{t\to t+1}$ and $\mathbf{w}_{t\to t+1}$, the corresponding brightness constancy errors $E_{\widehat{\mathbf{w}}} = |\mathbf{I}_t - \mathcal{W}(\mathbf{I}_{t+1}; \widehat{\mathbf{w}}_{t\to t+1})|$ and $E_{\mathbf{w}} = |\mathbf{I}_t - \mathcal{W}(\mathbf{I}_{t+1}; \mathbf{w}_{t\to t+1})|$ as well as the current frame $\mathbf{I}_t$. A visualization of the network structure is shown at Fig. 1. The dotted lines indicate that two sub-networks share the same weights, while the double vertical lines denote the feature concatenation.

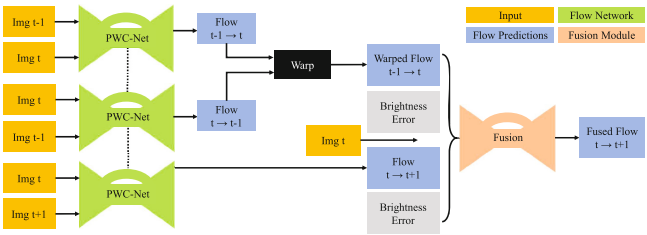


**Fig. 1.** Architecture of the proposed fusion approach.

We also propose two deep-learning baseline methods, shown at Fig. 2. **FlowNetS++**: FlowNetS [5] is a standard U-Net structure. We copy the encoded features from the previous pair of images to the current frame. **FlowNetS + GRU**: We use GRU-RCN [1] to extract abstract representations from videos and propagate encoded features in previous frames through time in a GRU-RCN unit and introduce a network structure, which we dub FlowNetS + GRU. We preserve the overall U-Net structure and apply GRU-RCN units at different levels of the encoder with different spatial resolutions. Encoded features at the sixth level are the smallest in resolution.

## 3 Experimental Results

We test two architectures as building blocks: FlowNetS [5] for its wide adoption, and PWC-Net [11] for its efficiency and performance on standard benchmarks. We follow Sun *et al.* [11] to design our training procedure and loss function. For consistency among different multi-frame algorithms, we use three frames as inputs.

For fusion networks, the network structure is similar to FlowNet2 [7] except for the first convolution layer, because our input to the fusion network has different channels. For the single optical flow prediction output by our fusion network, we set $\alpha = 0.005$ in the loss function [11] and use learning rate 0.0001 for fine-tuning.

We perform an ablation study of the two-frame and multi-frame methods using the virtual KITTI and Monkaa datasets, as summarized in Table 1. The Fusion approach consistently outperforms all other methods, including those using the GRU units. On the MPI Sintel [4] and KITTI benchmark [6], PWC-Fusion outperforms all two-frame optical flow methods including the state-of-the-art PWC-Net. This is also the first time a multi-frame optical flow algorithm consistently outperforms two-frame approaches across different datasets. We provide some visual results in Fig. 3 (Tables 2 and 3).

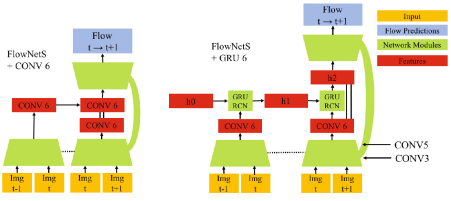


**Fig. 2.** Baseline network structures.

**Table 1.** Ablation study on the virtual KITTI dataset.

| | FlowNetS | FlowNetS++ | GRU 3 | GRU 4 | GRU 5 | GRU 6 | Fusion |
|---|---|---|---|---|---|---|---|
| EPE All | 6.12 | 5.90 | 5.26 | 5.40 | 5.15 | 5.32 | 5.00 |
| EPE Inside | 4.03 | 3.87 | 3.61 | 3.64 | 3.58 | 3.59 | 3.14 |
| EPE Outside | 28.97 | 27.57 | 23.26 | 24.60 | 22.28 | 24.25 | 25.15 |
| EPE Occlusion | 7.44 | 7.11 | 5.93 | 6.27 | 5.82 | 6.18 | 6.14 |

| | PWC-Net | GRU 3 | GRU 4 | GRU 5 | GRU 6 | Fusion |
|---|---|---|---|---|---|---|
| EPE All | 2.34 | 2.17 | 2.13 | 2.12 | 2.16 | 2.07 |
| EPE Inside | 1.60 | 1.44 | 1.41 | 1.40 | 1.42 | 1.37 |
| EPE Outside | 10.43 | 10.01 | 9.94 | 10.02 | 9.86 | 9.71 |
| EPE Occlusion | 2.41 | 2.29 | 2.24 | 2.24 | 2.26 | 2.27 |

**Table 2.** Results of the MPI Sintel [4].

| | EPE | Match | Unmatch | d0-10 | d10-60 | d60-140 | s0-10 | s10-40 | s40+ |
|---|---|---|---|---|---|---|---|---|---|
| PWC-Fusion | **4.566** | **2.216** | 23.732 | **4.664** | **2.017** | **1.222** | 0.893 | 2.902 | 26.810 |
| PWC-Net | 4.596 | 2.254 | **23.696** | 4.781 | 2.045 | 1.234 | 0.945 | 2.978 | **26.620** |
| ProFlow | 5.015 | 2.659 | 24.192 | 4.985 | 2.185 | 1.771 | 0.964 | 2.989 | 29.987 |
| DCFlow | 5.119 | 2.283 | 28.228 | 4.665 | 2.108 | 1.440 | 1.052 | 3.434 | 29.351 |
| FlowFieldsCNN | 5.363 | 2.303 | 30.313 | 4.718 | 2.020 | 1.399 | 1.032 | 3.065 | 32.422 |
| MR-Flow | 5.376 | 2.818 | 26.235 | 5.109 | 2.395 | 1.755 | 0.908 | 3.443 | 32.221 |
| LiteFlowNet | 5.381 | 2.419 | 29.535 | 4.090 | 2.097 | 1.729 | **0.754** | **2.747** | 34.722 |
| S2F-IF | 5.417 | 2.549 | 28.795 | 4.745 | 2.198 | 1.712 | 1.157 | 3.468 | 31.262 |

**Table 3.** Results of the KITTI [6].

| | Fl-all-Occ | Fl-fg-Occ | Fl-bg-Occ | Fl-all-Ncc | Fl-fg-Ncc | Fl-bg-Ncc |
|---|---|---|---|---|---|---|
| PWC-Fusion | **7.17** | **7.25** | **7.15** | **4.47** | **4.25** | **4.52** |
| PWC-Net | 7.90 | 8.03 | 7.87 | 5.07 | 5.04 | 5.08 |
| LiteFlowNet | 9.38 | 7.99 | 9.66 | 5.49 | 5.09 | 5.58 |
| MirrorFlow | 10.29 | 17.07 | 8.93 | 7.46 | 12.95 | 6.24 |
| SDF | 11.01 | 23.01 | 8.61 | 8.04 | 18.38 | 5.75 |
| UnFlow | 11.11 | 15.93 | 10.15 | 7.46 | 12.36 | 6.38 |
| MRFlow | 12.19 | 22.51 | 10.13 | 8.86 | 17.91 | 6.86 |
| ProFlow | 15.04 | 20.91 | 13.86 | 10.15 | 17.9 | 8.44 |

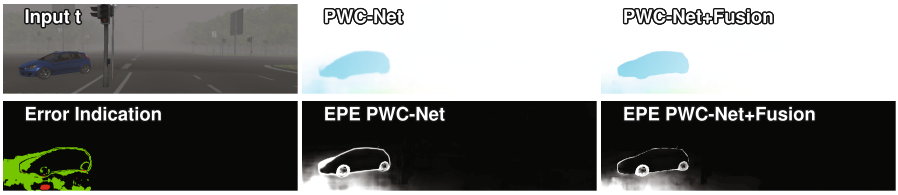


**Fig. 3.** Visual results of our fusion method. Green in the indication map means that PWC-Net+Fusion is more accurate than PWC-Net, and red means the opposite. (Color figure online)

## 4 Conclusions

We have presented a simple and effective fusion approach for multi-frame optical flow estimation. Multiple frames provide new information beyond what is available from two adjacent frames, in particular for occluded and out-of-boundary pixels. Thus we propose fusing the warped previous flow with the current flow estimate. Extensive experiments demonstrate the benefit of our approach: it outperforms both two-frame baselines and sensible multi-frame baselines based on GRUs. Moreover, it is top-ranked among all published flow methods on the MPI Sintel and KITTI 2015 benchmark.

**Acknowledgement.** We thank Fitsum Reda and Jinwei Gu for help with implementations, Xiaodong Yang for helpful discussions about RNN models, and Simon Baker for insightful discussions about multi-frame flow estimation.

## References

1. Ballas, N., Yao, L., Pal, C., Courville, A.: Delving deeper into convolutional networks for learning video representations. arXiv preprint arXiv:1511.06432 (2015)
2. Black, M.J., Anandan, P.: Robust dynamic motion estimation over time. In: CVPR (1991)
3. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping. In: Pajdla, T., Matas, J. (eds.) ECCV 2004. LNCS, vol. 3024, pp. 25–36. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24673-2_3

4. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7577, pp. 611–625. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33783-3_44
5. Dosovitskiy, A., et al.: Flownet: learning optical flow with convolutional networks. In: ICCV (2015)
6. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: CVPR (2012)
7. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: evolution of optical flow estimation with deep networks. In: CVPR (2017)
8. Kennedy, R., Taylor, C.J.: Optical flow with geometric occlusion estimation and fusion of multiple frames. In: Tai, X.-C., Bae, E., Chan, T.F., Lysaker, M. (eds.) EMMCVPR 2015. LNCS, vol. 8932, pp. 364–377. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-14612-6_27
9. Maurer, D., Bruhn, A.: Proflow: learning to predict optical flow. In: BMVC (2018)
10. Sand, P., Teller, S.: Particle video: long-range motion estimation using point trajectories. IJCV **80**(1), 72–91 (2008)
11. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In: CVPR (2018)