SelFlow: Self-Supervised Learning of Optical Flow

Pengpeng Liu[†]*, Michael Lyu[†], Irwin King[†], Jia Xu[§]
[†] Chinese University of Hong Kong, [§] Tencent AI Lab

Abstract

We present a self-supervised learning approach for optical flow. Our method distills reliable flow estimations from non-occluded pixels, and uses these predictions as ground truth to learn optical flow for hallucinated occlusions. We further design a simple CNN to utilize temporal information from multiple frames for better flow estimation. These two principles lead to an approach that yields the best performance for unsupervised optical flow learning on the challenging benchmarks including MPI Sintel, KITTI 2012 and 2015. More notably, our self-supervised pre-trained model provides an excellent initialization for supervised fine-tuning. Our fine-tuned models achieve state-of-the-art results on all three datasets. At the time of writing, we achieve EPE=4.26 on the Sintel benchmark, outperforming all submitted methods.

1. Introduction

Optical flow estimation is a core building block for a variety of computer vision systems. Despite decades of development, accurate flow estimation remains an open problem due to one key challenge: occlusion. Recent studies learn to estimate optical flow end-to-end from images using convolutional neural networks (CNNs) [2, 14, 5, 4, 18]. However, training fully supervised CNNs requires a large amount of labeled training data, which is extremely difficult to obtain for optical flow. In the absence of large-scale real-world annotations, existing methods turn to pre-train on synthetic labeled datasets [2, 10] and then fine-tune on small annotated datasets [5, 4, 18]. However, there usually exists a large gap between the distribution of synthetic data and natural scenes. To train a stable model, we have to carefully follow specific learning schedules across different datasets [5, 4, 18].

One promising direction is to develop unsupervised optical flow learning methods. The basic idea is to warp target image towards reference image according to the estimated optical flow, then minimize the difference between reference image and warped target image using a photometric loss [7, 16]. Such idea works well for non-occluded pixels but turns to provide misleading information for occluded pixels. Recent methods propose to exclude occluded pixels when computing the photometric loss or employ additional spatial and temporal smoothness terms to regularize flow estimation [11, 20, 6]. Most recently, DDFlow [8] proposes a data distillation approach, which employs random cropping to create occlusions for self-supervision. Unfortunately, these methods fails to generalize well for all natural occlusions. As a result, there is still a large performance gap comparing with state-of-the-art fully supervised methods.

Is it possible to effectively learn optical flow with occlusions? In this paper, we show that a self-supervised approach can learn to estimate optical flow with any form of occlusions from unlabeled data. Our work is based on distilling reliable flow estimations from non-occluded pixels, and using these predictions to guide the optical flow learning for hallucinated occlusions. Figure 1 illustrates our idea to create synthetic occlusions by perturbing superpixels. We further utilize temporal information from multiple frames to improve flow prediction accuracy within a simple CNN architecture. The resulted learning approach yields the highest accuracy among all unsupervised optical flow learning methods on Sintel and KITTI benchmarks.

Surprisingly, our self-supervised pre-trained model provides an excellent initialization for supervised fine-tuning. At the time of writing, our fine-tuned model achieves the highest reported accuracy (EPE=4.26) on the Sintel benchmark. Our approach also significantly outperforms all published optical flow methods on the KITTI 2012 benchmark, and achieves highly competitive results on the KITTI 2015 benchmark. To the best of our knowledge, it is the first time that a supervised learning method achieves such remarkable accuracies without using any external labeled data.

2. Method

In this section, we present our self-supervised approach to learning optical flow from unlabeled data. To this end, we train two CNNs (NOC-Model and OCC-Model) with the same network architecture. We distill reliable non-occluded flow estimations from NOC-Model to guide the learning of OCC-Model for those occluded pixels. Only OCC-Model is needed at testing. We build our network based on PWC-Net [18] and further extend it to multi-frame optical flow

*To appear at 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition.

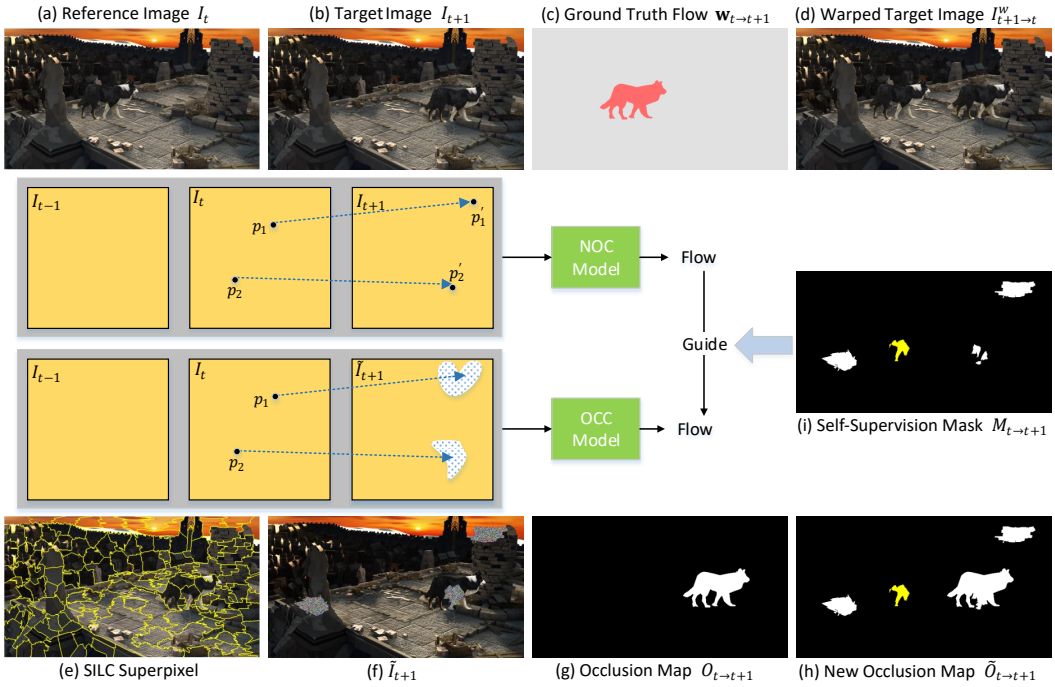

Figure 1. A toy example to illustrate our self-supervised learning idea. We first train our NOC-model with the classical photometric loss (measuring the difference between the reference image (a) and the warped target image(d)), guided by the occlusion map (g). Then we perturbate randomly selected superpixels in the target image (b) to hallucinate occlusions. Finally, we use reliable flow estimations from our NOC-Model to guide the learning of our OCC-Model for those newly occluded pixels (denoted by self-supervision mask (i), where value 1 means the pixel is non-occluded in (g) but occluded in (h)). Note the yellow region is part of the moving dog. Our self-supervised approach learns optical flow for both moving objects and static scenes.

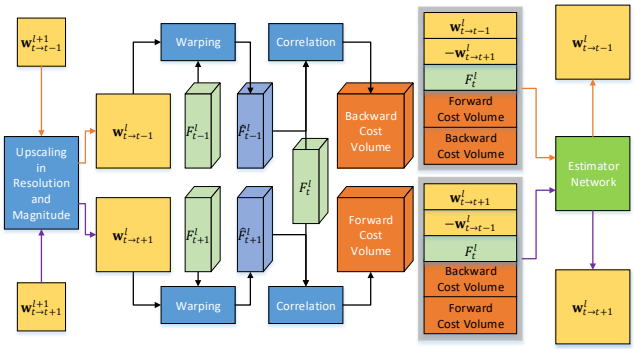

Figure 2. Our network architecture at each level (similar to PWC-Net [18]). $\dot{\mathbf{w}}^l$ denotes the initial coarse flow of level l and $\hat{F}^l$ denotes the warped feature representation. At each level, we swap the initial flow and cost volume as input to estimate both forward and backward flow concurrently. Then these estimations are passed to layer $l-1$ to estimate higher-resolution flow.

estimation (Figure 2). For supervised fine-tuning, We initialize the model with the pre-trained OCC-Model on each dataset.

Notation. Given three consecutive RGB images I_{t-1}, I_t, I_{t+1}, our goal is to estimate the forward flow from I_t to I_{t+1}. Let $\mathbf{w}_{i\to j}$ denote the flow from I_i to I_j, e.g., $\mathbf{w}_{t\to t+1}$ denotes the forward flow from I_t to I_{t+1}. After obtaining optical flow, we can backward warp target image to reconstruct reference image using Spatial Transformer Network.

Here, we use $I^w_{j\to i}$ to denote warping I_j to I_i with flow $\mathbf{w}_{i\to j}$. Similarly, we use $O_{i\to j}$ to denote the occlusion map from I_i to I_j, where value 1 means the pixel in I_i is not visible in I_j. In our self-supervised setting, we create new target image $\widetilde{I}_{t+1}$ by injecting random noise on superpixels for occlusion generation. We can inject noise to any of three consecutive frames as shown in Figure 1. For brevity, here we choose I_{t+1} as an example. If we let I_{t-1}, I_t and $\widetilde{I}_{t+1}$ as input, then $\widetilde{\mathbf{w}}, \widetilde{O}, \widetilde{I}^w$ represent the generated flow, occlusion map and warped image.

Occlusion Hallucination. For occlusion estimation, we employ forward-backward consistency check as [19, 11, 8]. During self-supervised training, we hallucinate occlusions by perturbing local regions with random noise. In a newly generated target image, the pixels corresponding to noise regions automatically become occluded. For occlusion hallucination, we first generate superpixels, then randomly select several superpixels and fill them with noise. Figure 1 shows a simple example, where only the dog is moving. Initially, occlusion map between I_t and I_{t+1} is (g). After randomly selecting several superpixels from (e) to inject noise, occlusion map change to (h).

NOC-to-OCC as Self-Supervision. Our self-training idea is built on top of the classical photometric loss [11, 20, 6], which is highly effective for non-occluded pixels. Figure 1

	Method	Sintel Clean		Sintel Final		KITTI 2012			KITTI 2015	
		train	test	train	test	train	test	test(Fl)	train	test(Fl)
Unsupervised	BackToBasic+ft [7]	–	–	–	–	11.3	9.9	–	–	–
	DSTFlow+ft [16]	(6.16)	10.41	(6.81)	11.27	10.43	12.4	–	16.79	39%
	UnFlow-CSS [11]	–	–	(7.91)	10.22	3.29	–	–	8.10	23.30%
	OccAwareFlow+ft [20]	(4.03)	7.95	(5.95)	9.15	3.55	4.2	–	8.88	31.2%
	MultiFrameOccFlow-None+ft [6]	(6.05)	–	(7.09)	–	–	–	–	6.65	–
	MultiFrameOccFlow-Soft+ft [6]	(3.89)	7.23	(5.52)	8.81	–	–	–	6.59	22.94%
	DDFlow+ft [8]	(2.92)	**6.18**	3.98	7.40	2.35	3.0	8.86%	5.72	14.29%
	Ours	**(2.88)**	6.56	**(3.87)**	**6.57**	**1.69**	**2.2**	**7.68%**	**4.84**	**14.19%**
Supervised	FlowNetS+ft [2]	(3.66)	6.96	(4.44)	7.76	7.52	9.1	44.49%	–	–
	FlowNetC+ft [2]	(3.78)	6.85	(5.28)	8.51	8.79	–	–	–	–
	SpyNet+ft [14]	(3.17)	6.64	(4.32)	8.36	8.25	10.1	20.97%	–	35.07%
	FlowNet2+ft [5]	(1.45)	4.16	(2.01)	5.74	(1.28)	1.8	8.8%	(2.3)	11.48%
	UnFlow-CSS+ft [11]	–	–	–	–	(1.14)	1.7	8.42%	(1.86)	11.11%
	LiteFlowNet+ft [4]	**(1.35)**	4.54	(1.78)	5.38	(1.05)	1.6	7.27%	(1.62)	9.38%
	PWC-Net+ft-CVPR [18]	(2.02)	4.39	(2.08)	5.04	(1.45)	1.7	8.10%	(2.16)	9.60%
	PWC-Net+ft-axXiv [17]	(1.71)	3.45	(2.34)	4.60	(1.08)	**1.5**	6.82%	(1.45)	7.90%
	ProFlow+ft [9]	(1.78)	**2.82**	–	5.02	(1.89)	2.1	7.88%	(5.22)	15.04%
	ContinualFlow+ft [13]	–	3.34	–	4.52	–	–	–	–	10.03%
	MFF+ft [15]	–	3.42	–	4.57	–	1.7	7.87%	–	**7.17%**
	Ours+ft	(1.68)	3.74	**(1.77)**	**4.26**	**(0.76)**	**1.5**	**6.19%**	(**1.18**)	8.42%

Table 1. Comparison with state-of-the-art learning based optical flow estimation methods. All numbers are EPE except for the last column of KITTI 2012 and KITTI 2015 testing sets, where we report percentage of erroneous pixels over all pixels (Fl-all). Parentheses mean that the training and testing are performed on the same dataset.

illustrates our main idea. Suppose pixel p_1 in image I_t is not occluded in I_{t+1}, and pixel p'_1 is its corresponding pixel. If we inject noise to I_{t+1} and let I_{t-1}, I_t, $\widetilde{I}_{t+1}$ as input, p_1 then becomes occluded. Good news is we can use the flow estimation of NOC-Model as annotation to guide OCC-Model to learn the flow of p_1 from I_t to $\widetilde{I}_{t+1}$. In the example of Figure 1, self-supervision is only employed to (i), which represents those pixels non-occluded from I_t to I_{t+1} but become occluded from I_t to $\widetilde{I}_{t+1}$.

Loss Functions. We first apply photometric loss L_p to non-occluded pixels, which is defined as follows:

$$L_p = \sum_{i,j} \frac{\sum \psi(I_i - I^w_{j\to i}) \odot (1 - O_i)}{\sum (1 - O_i)} \tag{1}$$

where $\psi(x) = (|x|+\epsilon)^q$ is a robust loss function, $\odot$ denotes element-wise multiplication. We set $\epsilon = 0.01$, $q = 0.4$ for all experiments. Only L_p is necessary to train NOC-Model.

To train our OCC-Model to estimate optical flow of occluded pixels, we define a self-supervision loss L_o for those synthetic occluded pixels (Figure 1(i)). First, we compute a self-supervision mask M to represent these pixels, where $M_{i\to j} = \text{clip}(\widetilde{O}_{i\to j} - O_{i\to j}, 0, 1)$.

Then, we define our self-supervision loss L_o as,

$$L_o = \sum_{i,j} \frac{\sum \psi(\mathbf{w}_{i\to j} - \widetilde{\mathbf{w}}_{i\to j}) \odot M_{i\to j}}{\sum M_{i\to j}} \tag{2}$$

For our OCC-Model, we train with a simple combination of $L_p + L_o$ for both non-occluded pixels and occluded pixels.

3. Experiments

We evaluate and compare our method with state-of-the-art unsupervised and supervised learning methods on public optical flow benchmarks including MPI Sintel [1], KITTI 2012 [3] and KITTI 2015 [12]. As shown in Table 1, we achieve state-of-the-art results for both unsupervised and supervised optical flow learning on all datasets under all evaluation metrics.

Unsupervised Learning. On Sintel final benchmark, we reduce the previous best EPE from 7.40 [8] to 6.57, with 11.2% relative improvements. On KITTI datasets, the improvement is more significant. For the training dataset, we achieve EPE=1.69 with 28.1% improvement on KITTI 2012 and EPE=4.84 with 15.3% relative improvement on KITTI 2015 compared with previous best unsupervised method DDFlow. On KITTI 2012 testing set, we achieve Fl-all=7.68%, which is better than state-of-the-art supervised methods including FlowNet2 [5], PWC-Net [18], ProFlow [9], and MFF [15]. On KITTI 2015 testing benchmark, we achieve Fl-all 14.19%, better than all unsupervised methods. Our unsupervised results also outperform some fully supervised methods such as ProFlow [9].

Supervised Fine-tuning. We achieve state-of-the-art results on all three datasets, with Fl-all=6.19% on KITTI 2012

Occlusion Handling	Multiple Frame	Self-Supervision Rectangle	Self-Supervision Superpixel	Sintel Clean			Sintel Final			KITTI 2012			KITTI 2015		
				ALL	NOC	OCC	ALL	NOC	OCC	ALL	NOC	OCC	ALL	NOC	OCC
✗	✗	✗	✗	(3.85)	(1.53)	(33.48)	(5.28)	(2.81)	(36.83)	7.05	1.31	45.03	13.51	3.71	75.51
✗	✓	✗	✗	(3.67)	(1.54)	(30.80)	(4.98)	(2.68)	(34.42)	6.52	1.11	42.44	12.13	3.47	66.91
✓	✗	✗	✗	(3.35)	(1.37)	(28.70)	(4.50)	(2.37)	(31.81)	4.96	0.99	31.29	8.99	3.20	45.68
✓	✓	✗	✗	(3.20)	(1.35)	(26.63)	(4.33)	(2.32)	(29.80)	3.32	0.94	19.11	7.66	2.47	40.99
✓	✗	✗	✓	(2.96)	(1.33)	(23.78)	(4.06)	(2.25)	(27.19)	1.97	0.92	8.96	5.85	2.96	24.17
✓	✓	✓	✗	(2.91)	(1.37)	(22.58)	(3.99)	(2.27)	(26.01)	1.78	0.96	7.47	5.01	2.55	21.86
✓	✓	✗	✓	**(2.88)**	**(1.30)**	**(22.06)**	**(3.87)**	**(2.24)**	**(25.42)**	**1.69**	**0.91**	**6.95**	**4.84**	**2.40**	**19.68**

Table 2. Ablation study. We report EPE of our unsupervised results under different settings over all pixels (ALL), non-occluded pixels (NOC) and occluded pixels (OCC). Note that we employ Census Transform when computing photometric loss by default.

and Fl-all=8.42% on KITTI 2015. Most importantly, our method yields EPE=4.26 on the Sintel final dataset, achieving the highest accuracy on the Sintel benchmark among all submitted methods. All these show that our method reduces the reliance of pre-training with synthetic datasets.

Ablation Study. To demonstrate the usefulness of individual technical steps, we conduct a rigorous ablation study (Table 2). Multi-frame formulation, occlusion handling and self-supervision can constantly improve performance. For self-supervision, we employ two strategies for occlusion hallucination: rectangle and superpixel. Both strategies improve performance significantly, especially for occluded pixels. Comparing superpixel noise injection with rectangle noise injection, superpixel setting has several advantages. First, the shape of superpixel is random and edges are more correlated to motion boundaries. Second, pixels in the same superpixel usually have similar motion patterns. As a result, superpixel setting achieves slightly better performance.

4. Conclusion

We present a self-supervised approach to learning accurate optical flow. Our method injects noise into superpixels to create occlusions, and let one model guide the another to learn optical flow for occluded pixels. Extensive experiments show our method significantly outperforms all existing unsupervised optical flow learning methods. After finetuning with our unsupervised model, we achieve state-of-the-art flow estimation accuracy on all leading benchmarks.

References

[1] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012.

[2] 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 *ICCV*, 2015.

[3] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012.

[4] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *CVPR*, 2018.

[5] 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 *CVPR*, 2017.

[6] Joel Janai, Fatma Güney, Anurag Ranjan, Michael J. Black, and Andreas Geiger. Unsupervised learning of multi-frame optical flow with occlusions. In *ECCV*, 2018.

[7] J Yu Jason, Adam W Harley, and Konstantinos G Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *ECCV*, 2016.

[8] Pengpeng Liu, Irwin King, Michael R. Lyu, and Jia Xu. Ddflow: Learning optical flow with unlabeled data distillation. In *AAAI*, 2019.

[9] D. Maurer and A. Bruhn. Proflow: Learning to predict optical flow. In *BMVC*, 2018.

[10] 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 *CVPR*, 2016.

[11] Simon Meister, Junhwa Hur, and Stefan Roth. UnFlow: Unsupervised learning of optical flow with a bidirectional census loss. In *AAAI*, New Orleans, Louisiana, 2018.

[12] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *CVPR*, 2015.

[13] Michal Neoral, Jan ochman, and Ji Matas. Continual occlusions and optical flow estimation. In *ACCV*, 2018.

[14] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *CVPR*, 2017.

[15] 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 *IEEE Winter Conference on Applications of Computer Vision*, 2019.

[16] Zhe Ren, Junchi Yan, Bingbing Ni, Bin Liu, Xiaokang Yang, and Hongyuan Zha. Unsupervised deep learning for optical flow estimation. In *AAAI*, 2017.

[17] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Models matter, so does training: An empirical study of cnns for optical flow estimation. *arXiv preprint arXiv:1809.05571*, 2018.

[18] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018.

[19] Narayanan Sundaram, Thomas Brox, and Kurt Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *ECCV*, 2010.

[20] Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, and Wei Xu. Occlusion aware unsupervised learning of optical flow. In *CVPR*, 2018.