Guided Optical Flow Learning

Yi Zhu[1] Zhenzhong Lan[2] Shawn Newsam[1] Alexander G. Hauptmann[2]
[1]University of California, Merced [2]Carnegie Mellon University
{yzhu25,snewsam}@ucmerced.edu {lanzhzh,alex}@cs.cmu.edu

Abstract

We study the unsupervised learning of CNNs for optical flow estimation using proxy ground truth data. Supervised CNNs, due to their immense learning capacity, have shown superior performance on a range of computer vision problems including optical flow prediction. They however require the ground truth flow which is usually not accessible except on limited synthetic data. Without the guidance of ground truth optical flow, unsupervised CNNs often perform worse as they are naturally ill-conditioned. We therefore propose a novel framework in which proxy ground truth data generated from classical approaches is used to guide the CNN learning. The models are further refined in an unsupervised fashion using an image reconstruction loss. Our guided learning approach is competitive with or superior to state-of-the-art approaches on three standard benchmarks yet is completely unsupervised and can run in real time.

1. Introduction

Optical flow contains valuable information for general image sequence analysis due to its capability to represent motion. It is widely used in vision tasks such as human action recognition [18], semantic segmentation [8], video frame prediction [15], video object tracking etc.

Classical approaches for estimating optical flow are often based on a variational model and solved as an energy minimization process [11, 4, 5]. They remain top performers on a number of evaluation benchmarks; however, most of them are too slow to be used in real time applications. Due to the great success of Convolutional Neural Network (CNN), several works [7, 16] have proposed using CNNs to estimate the motion between image pairs and have achieved promising results. Although they are much more efficient than classical approaches, these methods require supervision and cannot apply to real world data where the ground truth is not easily accessible. Thus, some recent works [1, 20] have investigated unsupervised learning through novel loss functions but they often perform worse than supervised ones.

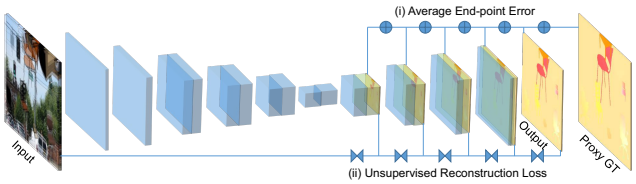

Figure 1. An overview of our proposed guided learning framework. $\oplus$ denotes computing the per-pixel endpoint error with respect to the proxy ground truth flow. $\bowtie$ represents the inverse warping and unsupervised reconstruction loss with respect to the input image pairs.

To improve the accuracy of unsupervised CNNs for optical flow estimation, we propose to use the results of classical methods as guidance for our unsupervised learning process. We refer to this as novel *guided optical flow learning* as shown in Fig. 1. Specifically, there are two stages. (i) We generate proxy ground truth flow using classical approaches, and then train a supervised CNN with them. (ii) We fine tune the learned models by minimizing an image reconstruction loss. By training the CNNs using proxy ground truth, we hope to provide a good initialization point for subsequent network learning. By fine tuning the models on target datasets, we hope to overcome the risk that CNN might have learned the failure cases of the classical approaches. The entire learning framework is thus unsupervised.

Our contributions are two-fold. First, we demonstrate that supervised CNNs can learn to estimate optical flow well even when only guided using noisy proxy ground truth data generated from classical methods. Second, we show that fine tuning the learned models for target datasets by minimizing a reconstruction loss further improves performance. Our proposed guided learning is completely unsupervised and achieves competitive or superior performance to state-of-the-art, real time approaches on standard benchmarks.

2. Method

Given an adjacent frame pair I_1 and I_2, our goal is to learn a model that can estimate the per-pixel motion field (U, V) between the two images accurately and efficiently.

U and V are the horizontal and vertical displacements, respectively. We describe our proxy ground truth guided framework in Section 2.1, and the unsupervised fine tuning strategy in Section 2.2.

2.1. Proxy Ground Truth Guidance

Current approaches to the supervised training of CNNs for estimating optical flow use synthetic ground truth datasets. These synthetic motions/scenes are quite different from real ones which limits the generalizability of the learned models. And, even constructing synthetic dataset requires a lot of manual effort [6]. The current largest synthetic datasets with dense ground truth optical flow, Flying Chairs [7] and FlyingThings3D [16], consist of only 22k image pairs which is not ideal for deep learning especially for such an ill-conditioned problem as motion estimation. In order for CNN-based optical flow estimation to reach its full potential, a learning framework is needed that can scale the size of the training data. Unsupervised learning is one ideal way to achieve this scaling because it does not require ground truth flow.

Classical approaches to optical flow estimation are unsupervised in that there is no learning process involved [11, 4, 5, 2, 12]. They only require the image pairs as input, with some extra assumptions (like image brightness constancy, gradient constancy, smoothness) and information (like motion boundaries, dense image matching). These non-CNN based classical methods currently achieve the best performance on standard benchmarks and are thus considered the state-of-the-art. Inspired by their good performance, we conjecture that *these approaches can be used to generate proxy ground truth data for training CNN-based optical flow estimators*.

In this work, we choose FlowFields [2] as our classical optical flow estimator. To our knowledge, it is one of the most accurate flow estimators among the published work. We hope that by using FlowFields to generate proxy ground truth, we can learn to estimate motion between image pairs as effectively as using the true ground truth.

For fair comparison, we use the "FlowNet Simple" network as descried in [7] as our supervised CNN architecture. This allows us to compare our guided learning approach to using the true ground truth, particularly with respect to how well the learned models generalize to other datasets. We use endpoint error (EPE) as our guided loss since it is the standard error measure for optical flow evaluation

$$L_{\text{epe}} = \frac{1}{N} \sum \sqrt{(U - U')^2 + (V - V')^2}, \qquad (1)$$

where N denotes the total number of pixels in I_1. U and V are the proxy ground truth flow fields while U' and V' are the flow estimates from the CNN.

2.2. Unsupervised Fine Tuning

As stated in Section 1, a potential drawback to using classical approaches to create training data is that the quality of this data will necessarily be limited by the accuracy of the estimator. If a classical approach fails to detect certain motion patterns, a network trained on the proxy ground truth is also likely to miss these patterns. This leads us to ask if there is other unsupervised guidance that can improve the network training?

The unsupervised approach of [20] treats optical flow estimation as an image reconstruction problem based on the intuition that if the estimated flow and the next frame can be used to reconstruct the current frame then the network has learned useful representations of the underlying motions. During training, the loss is computed as the photometric error between the true current frame I_1 and the inverse-warped next frame I_1'

$$L_{\text{reconst}} = \frac{1}{N} \sum_{i,j}^{N} \rho(I_1(i,j) - I_1'(i,j)), \qquad (2)$$

where $I_1'(i,j) = I_2(i + U_{i,j}, j + V_{i,j})$. The inverse warp is performed using a spatial transformer module [13] inside the CNN. We use a robust convex error function, the generalized Charbonnier penalty $\rho(x) = (x^2 + \epsilon^2)^\alpha$, to reduce the influence of outliers. This reconstruction loss is similar to the brightness constancy objective in classical variational formulations but is quite different from the EPE loss in the proxy ground truth guided learning. We thus propose fine tuning our model using this reconstruction loss as an additional unsupervised guide.

During fine tuning, the total energy we aim to minimize is a simple weighted sum of the EPE loss and the image reconstruction loss

$$L(U, V; I_1, I_2) = L_{\text{epe}} + \lambda \cdot L_{\text{reconst}}, \qquad (3)$$

where λ controls the level of reconstruction guidance. Note that we could add additional unsupervised guides like a gradient constancy assumption or an edge-aware weighted smoothness loss [10] to further fine tune our models.

An overview of our guided learning framework with both the proxy ground truth guidance and the unsupervised fine tuning is illustrated in Fig. 1.

3. Experiments

3.1. Datasets

Flying Chairs [7] is a synthetic dataset designed specifically for training CNNs to estimate optical flow. It is created by applying affine transformations to real images and synthetically rendered chairs. The dataset contains 22,872 image pairs: 22,232 training and 640 test samples according to the standard evaluation split.

MPI Sintel [6] is also a synthetic dataset derived from a short open source animated 3D movie. There are 1,628 frames, 1,064 for training and 564 for testing. It is the most widely adopted benchmark to compare optical flow estimators. In this work, we only report performance on its final pass because it contains sufficiently realistic scenes including natural image degradations.

KITTI Optical Flow 2012 [9] is a real world dataset collected from a driving platform. It consists of 194 training image pairs and 195 test pairs with sparse ground truth flow. We report the average EPE in total for the test set.

We consider guided learning with and without fine tuning. In the no fine tuning regime, the model is trained using the proxy ground truth produced using a classical estimator. In the fine tuning regime, the model is first trained using the proxy ground truth and then fine tuned using both the proxy ground truth and the reconstruction guide. The Sintel and KITTI datasets are too small to produce enough proxy ground truth to train our model from scratch so the models evaluated on these datasets are first pretrained on the Chairs dataset. These models are then either applied to the Sintel and KITTI datasets without fine tuning or are fine tuned using the target dataset (proxy ground truth).

3.2. Implementation

As shown in Fig. 1, our architecture consists of contractive and expanding parts. In the no fine tuning learning regime, we calculate the per-pixel EPE loss for each expansion. There are 5 expansions resulting in 5 losses. We use the same loss weights as in [7]. The models are trained using Adam optimization with the default parameter values $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is set to 10^{-4} and divided by half every 100k iterations after the first 300k. We end our training at 600k iterations.

In the fine tuning learning regime, we calculate both the EPE and reconstruction loss for each expansion. Thus there are a total of 10 losses. The generalized Charbonnier parameter α is set to 0.25 in the reconstruction loss. λ is 0.1. We use the default Adam optimization with a fixed learning rate of 10^{-6} and training is stopped at 10k iterations.

We apply the same intensive data augmentation as in [7] to prevent over-fitting in both learning regimes. The proxy ground truth is computed using the FlowFields binary kindly provided by authors in [2].

3.3. Results and Discussion

We have three observations given the results in Table 1.

Observation 1: *We can use proxy ground truth generated by state-of-the-art classical flow estimators to train CNNs for optical flow prediction*. A model trained using the FlowFields proxy ground truth achieves an average EPE of 3.34 on Chairs which is comparable to the 2.71 achieved by the model trained using the true ground truth. Note that the proxy ground truth is still quite noisy with an average EPE of 2.45 away from the true ground truth.

Method	Chairs	Sintel	KITTI
FlowFields [2]	2.45	5.81	3.5
FlowNetS (Ground Truth) [7]	2.71	8.43	9.1
UnsupFlowNet [20]	5.30	11.19	11.3
FlowNetS (FlowFields)	3.34	8.05	9.7
FlowNetS (FlowFields) + Unsup	3.01	7.96	9.5

Table 1. Results reported using average EPE, lower is better. Bottom section shows our guided learning results, the models are trained using the FlowFields proxy ground truth. The last row includes fine tuning.

The model trained using the FlowFields proxy ground truth (EPE 3.34) performs worse than the FlowFields estimator (EPE 2.45), which is expected. This is because FlowFields adopts a hierarchical approach which is non-local in the image space. It also uses dense correspondence to capture image details. Thus, FlowFields itself can output crisp motion boundaries and accurate flow. However, unlike the CNN model, it cannot run in real time.

Observation 2: *Sometime, training using proxy ground truth can generalize better than training using the true ground truth.* The model trained using the Chairs proxy ground truth (computed with FlowFields) performs better (EPE 8.05) on Sintel than the model trained using the Chairs true ground truth (EPE 8.43). We make similar observations for KITTI[1]. This improved generalization might result from over-fitting when training with the true ground truth since the three datasets are quite different with respect to object and motion types. The proxy is noisier which could serve as a form of data augmentation for unseen motion types.

In addition, we experiment on directly training a Sintel model from scratch without using the pretrained Chairs model. We use the same implementation details. The performance is about one and half pixel worse in terms of EPE than using the pretrained model. Therefore, pretraining CNNs on a large dataset (with either true or proxy ground truth data) is important for optical flow estimation.

Observation 3: *Our proposed fine tuning regime improves performance on all three datasets.* Fine tuning results in an average EPE decrease from 3.34 to 3.01 for Chairs, 8.05 to 7.96 for Sintel, and 9.7 to 9.5 for KITTI. Note that an average EPE of 3.01 for Chairs is very close to performance of the supervised model FlowNetS (EPE 2.71). This demonstrates that image reconstruction loss is effective as an additional unsupervised guide for motion learning. It can act like fine tuning without requiring ground truth flow of the target dataset.

We also investigate training a network from scratch using a joint training regime. That is, using both L_{epe} and L_{reconst}, not only using L_{reconst} in the fine tuning stage. The per-

[1]Note that FlowNetS's performance on KITTI (EPE 9.1) is fine tuned.

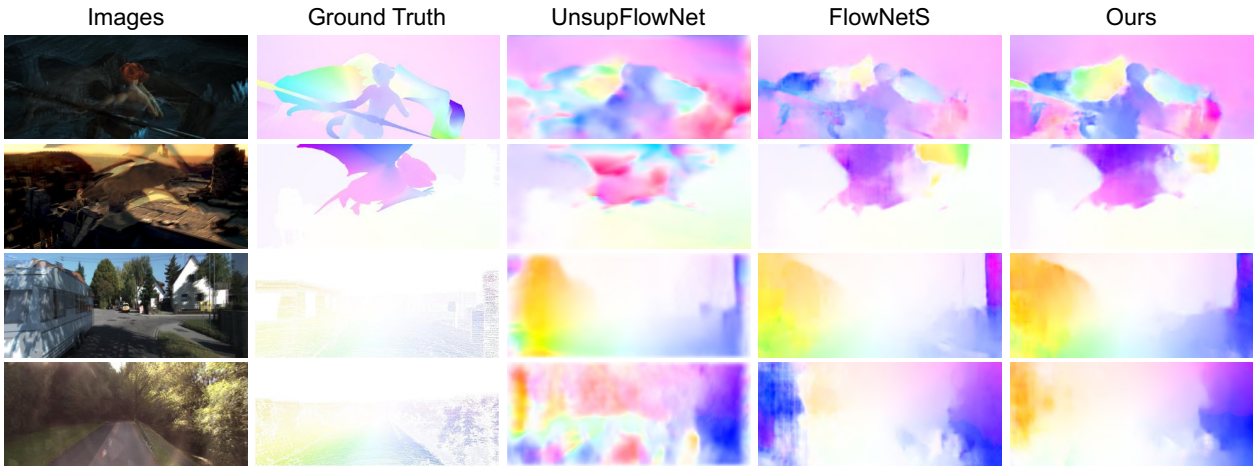

Figure 2. Visual examples of predicted optical flow from different methods. Top two are from Sintel, and bottom two from KITTI.

formance was worse on all three benchmarks. The reason might be that pretraining using just the proxy ground truth prevents the model from becoming trapped in local minima. It thus can provide a good initialization for further network learning. A joint training regime using both losses may hurt the network's convergence in the beginning.

However, we expect unsupervised learning to bring more complementarity. Image reconstruction loss may not be the most appropriate guidance for learning optical flow prediction. We will explore how to best incorporate additional unsupervised objectives in future work.

3.4. Comparison to State-of-the-Art

We compare our proposed method to recent state-of-the-art approaches. We only consider approaches that are fast because optical flow is often used in time sensitive applications. We evaluated all CNN-based approaches on a workstation with Intel Core I7 with 4.00GHz and a Nvidia Titan X GPU. For classical approaches, we just use their reported runtime. As shown in Table 2, our method performs the best for Sintel even though it does not require the true ground truth for training. For Chairs, we achieve on par performance with [7]. For KITTI, we perform inferior to [19]. This is likely because the flow in KITTI is caused purely by the motion of the car so the segmentation into layers performed in [19] helps in capturing motion boundaries. Our approach outperforms the state-of-the-art unsupervised approaches of [1, 20] by a large margin, thus demonstrating the effectiveness of our proposed guided learning using proxy ground truth and image reconstruction. Visual comparison of Sintel and KITTI results are shown in Fig. 2. We can see that UnsupFlowNet [20] is able to produce reasonable flow fields estimation, but quite noisy. And it doesn't perform well in highly saturated and very dark regions. Our results are much more detailed and smoothed due to the proxy guidance and unsupervised fine tuning.

Method	Chairs	Sintel	KITTI	Runtime
EPPM [3]	–	8.38	9.2	0.25
PCA-Flow [19]	–	8.65	**6.2**	0.19*
DIS-Fast [14]	–	10.13	14.4	0.02*
FlowNetS [7]	**2.71**	8.43	9.1	0.06
UnsupFlowNet [20]	5.30	11.19	11.3	0.06
USCNN [1]	–	8.88	–	–
Ours	3.01	**7.96**	9.5	0.06

Table 2. State-of-the-art comparison, runtime is reported in seconds per frame. Top: Classical approaches. Middle: CNN-based approaches. Bottom: Ours. * indicates the algorithm is evaluated using CPU, while the rest are on GPU.

4. Conclusion

We propose a guided optical flow learning framework which is unsupervised and results in an estimator that can run in real time. We show that proxy ground truth data produced using state-of-the-art classical estimators can be used to train CNNs. This allows the training sets to scale which is important for deep learning. We also show that training using proxy ground truth can result in better generalization than training using the true ground truth. And, finally, we also show that an unsupervised image reconstruction loss can provide further learning guidance.

More broadly, we introduce a paradigm which can be integrated into future state-of-the-art motion estimation networks [17] to improve performance. In future work, we plan to experiment with large-scale video corpora to learn non-rigid real world motion patterns rather than just learning limited motions found in synthetic datasets.

Acknowledgements This work was funded in part by a National Science Foundation CAREER grant, #IIS-1150115. We gratefully acknowledge NVIDIA Corporation through the donation of the Titan X GPU used in this work.

References

[1] A. Ahmadi and I. Patras. Unsupervised Convolutional Neural Networks for Motion Estimation. In *ICIP*, 2016.

[2] C. Bailer, B. Taetz, and D. Stricker. Flow Fields: Dense Correspondence Fields for Highly Accurate Large Displacement Optical Flow Estimation. In *ICCV*, 2015.

[3] L. Bao, Q. Yang, and H. Jin. Fast Edge-Preserving PatchMatch for Large Displacement Optical Flow. In *CVPR*, 2014.

[4] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High Accuracy Optical Flow Estimation Based on a Theory for Warping. In *ECCV*, 2004.

[5] T. Brox and J. Malik. Large Displacement Optical Flow: Descriptor Matching in Variational Motion Estimation. *PAMI*, 33:500–513, March 2011.

[6] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A Naturalistic Open Source Movie for Optical Flow Evaluation. In *ECCV*, 2012.

[7] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. v.d. Smagt, D. Cremers, and T. Brox. FlowNet: Learning Optical Flow with Convolutional Networks. In *ICCV*, 2015.

[8] K. Fragkiadaki, P. Arbelez, P. Felsen, and J. Malik. Learning to Segment Moving Objects in Videos. In *CVPR*, 2015.

[9] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *CVPR*, 2012.

[10] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised Monocular Depth Estimation with Left-Right Consistency. *arXiv preprint arXiv:1609.03677*, 2016.

[11] B. K. Horn and B. G. Schunck. Determining Optical Flow. *Artificial Intelligence*, 17:185–203, 1981.

[12] Y. Hu, R. Song, and Y. Li. Efficient Coarse-to-Fine PatchMatch for Large Displacement Optical Flow. In *CVPR*, 2016.

[13] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial Transformer Network. In *NIPS*, 2015.

[14] T. Kroeger, R. Timofte, D. Dai, and L. V. Gool. Fast Optical Flow using Dense Inverse Searchn. In *ECCV*, 2016.

[15] M. Mathieu, C. Couprie, and Y. LeCun. Deep Multi-Scale Video Prediction beyond Mean Square Error. In *ICLR*, 2016.

[16] N. Mayer, E. Ilg, P. Husser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In *CVPR*, 2016.

[17] A. Ranjan and M. J. Black. Optical Flow Estimation using a Spatial Pyramid Network. *arXiv preprint arXiv:1611.00850*, 2016.

[18] K. Simonyan and A. Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. In *NIPS*, 2014.

[19] J. Wulff and M. Black. Efficient Sparse-to-Dense Optical Flow Estimation using a Learned Basis and Layers. In *CVPR*, 2015.

[20] J. J. Yu, A. W. Harley, and K. G. Derpanis. Back to Basics: Unsupervised Learning of Optical Flow via Brightness Constancy and Motion Smoothness. In *ECCVW*, 2016.