Back to Basics: Unsupervised Learning of Optical Flow via Brightness Constancy and Motion Smoothness

Jason J. Yu(✉), Adam W. Harley, and Konstantinos G. Derpanis

Department of Computer Science, Ryerson University, Toronto, Canada
{jjyu,aharley,kosta}@scs.ryerson.ca

Abstract. Recently, convolutional networks (convnets) have proven useful for predicting optical flow. Much of this success is predicated on the availability of large datasets that require expensive and involved data acquisition and laborious labeling. To bypass these challenges, we propose an unsupervised approach (i.e., without leveraging groundtruth flow) to train a convnet end-to-end for predicting optical flow between two images. We use a loss function that combines a data term that measures photometric constancy over time with a spatial term that models the expected variation of flow across the image. Together these losses form a proxy measure for losses based on the groundtruth flow. Empirically, we show that a strong convnet baseline trained with the proposed unsupervised approach outperforms the same network trained with supervision on the KITTI dataset.

1 Introduction

Visual motion estimation is a core research area of computer vision. Most prominent has been the recovery of the apparent motion of image brightness patterns, i.e., optical flow. Much of this work has centred on extracting the pixelwise velocities between two temporal images within a variational framework [7,13].

Recently, convolutional networks (convnets) have proven useful for a variety of per-pixel prediction tasks, including optical flow [3]. Convnets are high-capacity models that approximate the complex, non-linear transformation between input imagery and the output. Success with convnets has relied almost exclusively on fully-supervised schemes, where the target value (i.e., the label) is provided during training. This is problematic for learning optical flow because directly obtaining the motion field groundtruth from real scenes — the quantity that optical flow attempts to approximate — is not possible.

In this paper, we propose an end-to-end unsupervised approach to train a convnet for predicting optical flow between two images based on a standard variational loss. Rather than rely on imagery as well as the corresponding groundtruth flow for training, we use the images alone. In particular, we use a loss function that combines a data term that measures photometric constancy over time with a spatial term that models the expected variation of flow across the image. The photometric

G. Hua and H. Jégou (Eds.): ECCV 2016 Workshops, Part III, LNCS 9915, pp. 3–10, 2016.
DOI: 10.1007/978-3-319-49409-8_1

loss measures the difference between the first input image and the (inverse) warped subsequent image based on the predicted optical flow by the network. The smoothness loss measures the difference between spatially neighbouring flow predictions. Together, these two losses form a proxy for losses based on the groundtruth flow.

Recovering optical flow between two frames is a well studied problem, with much previous work founded on variational formulations [2,7,12,13]. Our loss is similar to the objective functions proposed for two-frame motion estimation; however, rather than optimize the velocity map between input frames, we use it to optimize the convnet weights over the training set of imagery.

Several recent works [3,10] have proposed convnets that learn the mapping between input image frames and the corresponding flow. Each of these approaches is presented in a supervised setting, where images and their corresponding groundtruth flows are provided. This setting assumes the availability of a large, annotated dataset. Existing flow datasets (e.g., KITTI [6]) are too small to support training accurate networks. Computer generated scenes and their corresponding flow [3,4,10] provide a means to address this issue. Although some recent efforts have attempted to semi-automate the data creation process [4], creating large, diverse imagery remains laborious. Another possibility is using the output of an existing optical flow estimator to provide the groundtruth [14]. This training approach may result in learning both correct flow prediction and the failure aspects of the flow estimator used for training. In this work, we avoid these drawbacks by learning flow in an unsupervised manner, using only the input imagery.

Concurrent work has proposed unsupervised methods to circumvent the need of vasts amounts of labeled data for training. A spatiotemporal video autoencoder [11] was introduced that incorporates a long short-term memory (LSTM) architecture for unsupervised flow and image frame prediction. Here, we present a simpler feedforward convnet model targeting flow prediction alone. Most closely related to the current paper is recent work that proposed a convnet for depth estimation trained in an unsupervised manner [5]. In a similar fashion to the proposed approach, a photometric loss warps one image to another, and a smoothness loss term is used to bias the predictor towards smooth depth estimates. Unlike the current work, the manner in which the photometric loss is handled (via a linear Taylor series approximation) precludes end-to-end learning.

Contributions. In the light of previous research, we make the following contributions. First, we present an unsupervised approach to training a convnet in an end-to-end manner for predicting optical flow between two images. The limited but valuable groundtruth flow is reserved for fine-tuning the network and cross-validating its parameters. Second, we demonstrate empirically that a strong convnet baseline trained with our unsupervised approach outperforms the same network trained with supervision on KITTI, where insufficient groundtruth flow is available for training.

2 Technical Approach

Given an RGB image pair as input, $\mathbf{X} \in \mathbb{R}^{H \times W \times 6}$, our objective is to learn a non-linear mapping (approximated by a convnet) to the corresponding optical flow, $\mathbf{Y} \in \mathbb{R}^{H \times W \times 2}$, where H and W denote the image height and width, respectively. In Sect. 2.1, we outline our unsupervised loss. Section 2.2 provides details on how the unsupervised loss is integrated with a reference convnet architecture.

2.1 Unsupervised Loss

The training set is comprised of pairs of temporally consecutive images, $\{I(x, y, t), I(x, y, t+1)\}$. Unlike prior work, we do not assume access to the corresponding velocity pixel labels, cf. [3]. Instead, we return to traditional means for scoring a given solution, via a loss that combines a photometric loss between the first image and the warped second image, and a loss related to the smoothness of the velocity field prediction [7]:

$$\begin{aligned} &\mathcal{L}(\mathbf{u}, \mathbf{v}; I(x, y, t), I(x, y, t+1)) = \\ &\quad \ell_{\text{photometric}}(\mathbf{u}, \mathbf{v}; I(x, y, t), I(x, y, t+1)) + \\ &\quad \lambda \ell_{\text{smoothness}}(\mathbf{u}, \mathbf{v}), \end{aligned} \tag{1}$$

where $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{H \times W}$ are the horizontal and vertical components of the predicted flow field, respectively, and λ is a regularization parameter that weighs the relative importance of smoothness of the predicted flow. Note, the photometric loss can be replaced or augmented with other measures, such as the image gradient constancy [2].

Given the predicted flow, the photometric loss is computed as the difference between the first image and the backward/inverse warped second image:

$$\begin{aligned} &\ell_{\text{photometric}}(\mathbf{u}, \mathbf{v}; I(x, y, t), I(x, y, t+1)) = \\ &\sum_{i,j} \rho_D(I(i, j, t) - I(i + u_{i,j}, j + v_{i,j}, t+1)), \end{aligned} \tag{2}$$

where ρ_D is the data penalty function. We consider the robust generalized Charbonnier penalty function $\rho(x) = (x^2 + \epsilon^2)^\alpha$ to mitigate the effects of outliers [13].

To compute the non-rigid backward warp, we use the recently proposed spatial transformer module [8]. This allows the learning to be performed with standard backpropagation in an end-to-end fashion. In brief, the spatial transformer can be described as two parts that work in sequence: (i) a sampling grid generator and (ii) a differentiable image sampler. (The spatial transformer localization step is not needed here as flow prediction, (u, v), provides the necessary parameters for the mapping between image points across frames.) The sampling grid is generated by the following pointwise transformation:

$$\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = W_{(u,v)} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} x_1 + u \\ y_1 + v \end{pmatrix}, \tag{3}$$

where (x_1, y_1) are the coordinates in the first image and (x_2, y_2) are the sampling coordinates in the second image. The bilinear sampling step can be written in the following (sub-)differentiable form:

$$I_{\text{warp}}(x_1, y_1, t+1) = \sum_{j}^{H} \sum_{i}^{W} I(i, j, t+1) M(1 - |x_2 - i|) M(1 - |y_2 - j|), \tag{4}$$

where $M(\cdot) = \max(0, \cdot)$. For details about backpropagating through this module, see [8].

Regions with insufficient image structure support multiple equally scoring velocities, e.g., the aperture problem. To address this ambiguity, we introduce a standard robust (piecewise) smoothness loss:

$$\begin{aligned} \ell_{\text{smoothness}}(\mathbf{u}, \mathbf{v}) = & \\ \sum_{j}^{H} \sum_{i}^{W} & [\rho_S(u_{i,j} - u_{i+1,j}) + \rho_S(u_{i,j} - u_{i,j+1}) \\ & + \rho_S(v_{i,j} - v_{i+1,j}) + \rho_S(v_{i,j} - v_{i,j+1})], \end{aligned} \tag{5}$$

where $\rho_S(\cdot)$ is the (spatial) smoothness penalty function realized by the generalized Charbonnier function.

A summary of our proposed unsupervised approach for flow prediction is provided in Fig. 1.

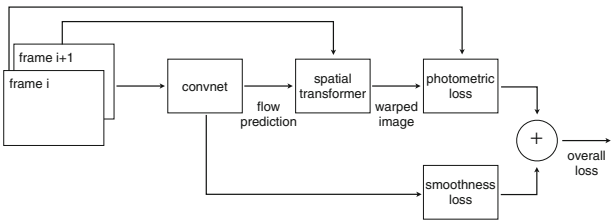

Fig. 1. Overview of our unsupervised approach.

2.2 Network Architecture

We use "FlowNet Simple" [3] as a reference network. This architecture consists of a contractive part followed by an expanding part. The contractive part takes as input two RGB images stacked together, and processes them with a cascade of strided convolution layers. The expanding part implements a "skip-layer" architecture that combines information from various levels of the contractive part with "upconvolving" layers to iteratively refine the coarse flow predictions. The FlowNet Simple architecture is illustrated in Fig. 2.

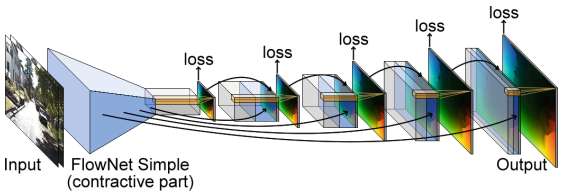

Fig. 2. "FlowNet Simple" architecture. Two images are taken as input, and an optical flow prediction is generated using a multi-stage refinement process. Feature maps from the contractive part, as well as intermediate flow predictions, are used in the "upconvolutional" part.

In this work, we use a loss comprised of a final loss and several intermediate losses placed at various stages of the expansionary part. The intermediate losses are meant to guide earlier layers more directly towards the final objective [9]. In FlowNet, the endpoint error (EPE), a standard error measure for optical flow, is used as the supervised training loss. As a proxy to per-pixel groundtruth flow, we replace the EPE with the proposed unsupervised loss, (1).

3 Empirical Evaluation

3.1 Datasets

Flying Chairs. This synthetic dataset is realized by applying affine mappings to publicly available colour images and a rendered set of 3D chair models. The dataset contains 22,232 training and 640 test image pairs with groundtruth flow. To cross-validate the hyper-parameters and monitor for overfitting in learning, we set aside 2,000 image pairs from the training set. We use both photometric and geometric augmentations to avoid overfitting. The photometric augmentations are comprised of additive Gaussian noise applied to each image, contrast, multiplicative colour changes to the RGB channels, gamma and additive brightness. The geometric transformations are comprised of 2D translations, left-right flipping, rotations and scalings.

KITTI 2012. This dataset consists of images collected on a driving platform. There are 194 and 195 training and testing image pairs, respectively, with sparse groundtruth flow. The training set is used for cross-validation and to monitor the learning progress. For training, we use the raw KITTI data from the city, residential and road classes, where groundtruth flow is unavailable. To avoid training on related testing imagery, we remove all raw images that are visually similar with the testing ones, including their temporal neighbours ± 20 frames. The (curated) raw data is comprised of 82,958 image pairs. We include both photometric and geometric augmentations. We use the same type of photometric augmentations as applied to Flying Chairs. The geometric transformations consist of left-right flipping and scalings. We also use a small relative translation.

3.2 Training Details

We use the "FlowNet Simple" architecture provided in the publicly available FlowNet Caffe code [1]. For the photometric loss, the generalized Charbonnier parameter, α, is set to 0.25 and 0.38 for Flying Chairs and KITTI, respectively. For the smoothness loss, α is set to 0.37 and 0.21 for Flying Chairs and KITTI, respectively. The smoothness weight, λ, is set to 1 for Flying Chairs and 0.53 for KITTI. We use Adam as the optimization method, where its parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is set to 1.6e–5 for Flying Chairs and 1.0e–5 for KITTI and we divide it by half every 100,000 iterations.

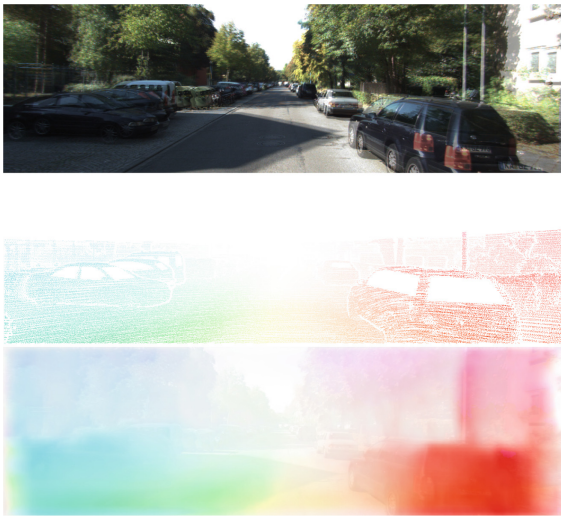

Fig. 3. KITTI example. (top-to-bottom) Input image frames overlaid, groundtruth flow, and predicted flow from unsupervised network overlaid on the first input image.

Table 1. Average endpoint error (EPE) flow results. The reported EPEs for supervised FlowNet are the best results of the FlowNet Simple architecture [3] without variational smoothing post-processing. "Avg. All" and "Avg. NOC" refer to the EPE taken over all the labeled pixels and all non-occluded labeled pixels, respectively.

Approach	Chairs	KITTI			
		Avg. All		Avg. NOC	
	Test	Train	Test	Train	Test
EpicFlow [12]	2.9	3.5	3.8	1.8	1.5
DeepFlow [15]	3.5	4.6	5.8	2.0	1.5
LDOF [2]	3.5	13.7	12.4	5.0	5.6
FlowNet [3]	2.7	7.5	9.1	5.3	5.0
FlowNet (ours)	5.3	11.3	9.9	4.3	4.6

The batch size is set to four image pairs. In total, we train using 600,000 iterations for Flying Chairs and 400,000 for KITTI. In initial tests, we noticed that the unsupervised approach had difficulties in regions that were highly saturated or very dark. Adding photometric augmentation compounds this issue by making these regions even less discriminable. To address this issue, we pass the geometrically augmented images directly to the photometric loss prior to photometric augmentation. Further, we apply a local 9×9 response normalization to the geometrically augmented images to ameliorate multiplicative lighting factors.

3.3 Results

Table 1 provides a summary of results. As expected, FlowNet trained with the groundtruth flow on Flying Chairs outperforms the unsupervised one. Note, however, the scenario where sufficient dense groundtruth is available is generally unrealistic with real imagery. Conversely, KITTI exemplifies an automative scenario, where abundant dense groundtruth flow with real images is unavailable. To sidestep this issue, previous work [3] used synthetic data as a proxy for supervised training. On the non-occluded (NOC) metric, the unsupervised approach improves upon the supervised one [3] on the KITTI training set. This improvement persists on the official test set. Considering all pixels (i.e., occluded and non-occluded) the proposed approach remains competitive to the supervised one. Figure 3 shows an example flow prediction result on KITTI. While the performance of the unsupervised approach lags behind the state-of-the-art, it operates in realtime with a testing runtime of 0.03 s on an NVIDIA GTX 1080 GPU.

4 Discussion and Summary

We presented an end-to-end unsupervised approach to training convnets for optical flow prediction. We showed that the proposed unsupervised training approach yields competitive and even superior performance to a supervised one. This opens up avenues for further improvement by leveraging the vast amounts of video that can easily be captured with commodity cameras taken in the domain of interest, such as automotive applications. Furthermore, this is a general learning framework that can be extended in a variety of ways via more sophisticated losses to enhance convnet-based mappings between temporal input imagery and flow.

References

1. FlowNet Caffe code (v1.0). http://lmb.informatik.uni-freiburg.de/resources/software.php
2. Brox, T., Malik, J.: Large displacement optical flow: descriptor matching in variational motion estimation. PAMI **33**(3), 500–513 (2011)
3. Dosovitskiy, A., Fischer, P., Ilg, E., Häusser, P., Hazirbas, C., Golkov, V., van der Smagt, P., Cremers, D., Brox, T.: FlowNet: learning optical flow with convolutional networks. In ICCV, pp. 2758–2766 (2015)

4. Gaidon, A., Wang, Q., Cabon, Y., Vig, E.: Virtual worlds as proxy for multi-object tracking analysis. In: CVPR (2016)
5. Garg, R., Vijay Kumar, B.G., Carneiro, G., Reid, I.: Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 740–756. Springer, Heidelberg (2016). doi:10.1007/978-3-319-46484-8_45
6. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: the KITTI dataset. IJRR **32**, 1231–1237 (2013)
7. Horn, B.K.P., Schunck, B.G.: Determining optical flow. AI **17**(1–3), 185–203 (1981)
8. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: NIPS (2015)
9. Lee, C., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: AISTATS (2015)
10. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: CVPR (2016)
11. Patraucean, V., Handa, A., Cipolla, R.: Spatio-temporal video autoencoder with differentiable memory. CoRR, abs/1511.06309 (2015)
12. Revaud, J., Weinzaepfel, P., Harchaoui, Z., Schmid, C.: EpicFlow: edge-preserving interpolation of correspondences for optical flow. In: CVPR (2015)
13. Sun, D.Q., Roth, S., Black, M.J.: A quantitative analysis of current practices in optical flow estimation and the principles behind them. IJCV **106**(2), 115–137 (2014)
14. Tran, D., Bourdev, L.D., Fergus, R., Torresani, L., Paluri, M.: Deep end2end voxel2voxel prediction. In: Workshop on DeepVision (2016)
15. Weinzaepfel, P., Revaud, J., Harchaoui, Z., Schmid, C.: DeepFlow: large displacement optical flow with deep matching. In: ICCV (2013)