Your Objective Is Wrong: Rethink Unsupervised learning of Optical Flow

Anonymous CVPR submission

Paper ID ****

Abstract

The ABSTRACT is to be in fully-justified italicized text, at the top of the left-hand column, below the author and affiliation information. Use the word "Abstract" as the title, in 12-point Times, boldface type, centered relative to the column, initially capitalized. The abstract is to be in 10-point, single-spaced type. Leave two blank lines after the Abstract, then begin the main text. Look at previous CVPR abstracts to get a feel for style and length.

1. Introduction

1.1. Related Work

1.2. Novel Contribution

We extend FlowNet [4] in this work, in summary our contributions are three folds. First, we proposed to combine traditional layered approach for optical flow estimation with deep learning. The proposed approach does not require pre-segmentation of images, instead, the separation of layers is automatically done during training the network. Second, a soft-masks module is proposed. This soft-masks module implements a channel-wise maxout operation among masks. As a result, the estimated optical flow will be separated to layers. each of which will contain optical flow that is estimated using a quadratic function. Third, we extend the FlowNet by adding the proposed soft-mask module in output layers, the resulting network is trained to compare with both supervised and unsupervised optical flow estimation approaches using neural networks. The empirical results show that the proposed network structure achieves comparable or lower error in each experimental group.

2. Methodology

The proposed approach and corresponding analysis will be introduced in this section.

2.1. Annotation

Given a pair of images $I_a, I_b \in \mathbb{R}^{H\times W\times C}$ as input, where H, W and C are height, width and channels of the input images. The proposed approach is going to estimate an optical flow field $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{H\times W}$, where $\mathbf{u}$ and $\mathbf{v}$ are the horizontal and vertical components of the optical flow field to be estimated that transform image from I_a to I_b. The original formulation of optical flow estimation is proposed by Horn and Schunck in [8]. In classical formulation, an objective function is composed with a combination of a data term which makes a local constancy assumption of some image property and a spatial term that models how the flow is expected to vary across images. We write the classical optical flow objective function as:

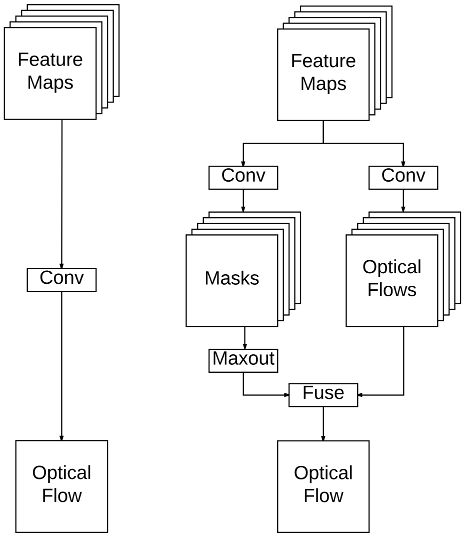
$$E(\mathbf{u}, \mathbf{v}) = \sum_{i}^{H}\sum_{j}^{W}(I_1(i+u_{ij}, j+v_{ij}) - I_0(i,j))^2 + \varphi(\mathbf{u}, \mathbf{v}) \quad (1)$$

where $\varphi(\mathbf{u}, \mathbf{v}))$ is a regularization term that constrains smoothness of optical flow.

Nowadays, the above objective is still being used by many optical flow estimation using deep neural network based on unsupervised training framework [1][14][18]. We also use above objective when training our network and comparing with results of unsupervised methods. Experiments results are presented in Section 3.

2.2. Soft-masks module

FlowNet [4] is the first work that uses deep convolutional neural network for flow estimation. The network architecture used by FlowNet is very similar to classical structure of auto-encoder, where optical flows are generated using deconvolution at each scale level of the image pyramid. To refine estimation of the flows, shortcuts are built to connect layers of corresponding level in encoder and decoder. Let's take a look at a single computation of convolution, and for simplicity, let's assume f represents both horizontal and vertical components of an output flow. Given $X \in \mathbb{R}^{s\times s\times c}$, representing a volume feature vector in the input feature volume, inputted to output layer, where s is kernel size and



a. Normal output b. Soft-masks module

Figure 1. An illustration of the structure of the proposed soft-masks module compared with normal linear optical flow output.

c is number of channels. FlowNet employs a linear activation to compute optical flow:

$$f = X^T W + b \tag{2}$$

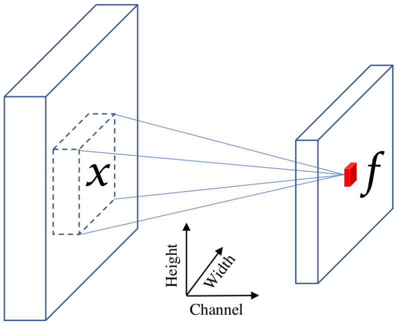


Figure 2. An illustration of annotation used in convolution.

Given actual optical flow field a nonlinear and piece-wise smooth representation of motions contained in images, using linear function to fit the flow field is less accurate. We introduce a soft-masks module in this paper. The proposed module can be used to replace the linear output of optical flow estimation. We will show that by using this module, we are able to separate optical flow field to multiple layers. Flow estimation in each layer is smooth and can be estimated using a quadratic function, which results in a more accurate and flexible optical flow estimation.

An illustration of soft-masks module is shown in Figure 1. The essential part of the soft-masks module is its dual-branch structure which contains mask branch and optical flow branch. The same input feature maps represented as a set of volume feature vectors, $X \in \mathbb{R}^{s \times s \times c}$ thus are imported to both branches. The most significant contribution of this work is to separate one optical flow field to multiple layers. For a separation of k layers. k masks will be generated in mask branch as illustrated in Figure 1, which requires k convolutional filters labeled as $\{W_n^m, b_n^m\}_{n=1}^k$ being used in mask branch. Correspondingly, the same number of filters are used in optical flow branch which are labeled as $\{W_n^f, b_n^f\}_{n=1}^k$. Then the mask and intermediate optical flow could be computed as following:

$$\begin{aligned} m_n &= X^T W_n^m + b_n^m && \text{for } n = 1 \ldots k \\ f_n &= X^T W_n^f + b_n^f && \text{for } n = 1 \ldots k \end{aligned} \tag{3}$$

Thus, give k filters used in both branches, we will obtain k corresponding pairs of mask and intermediate optical flow. Basically, by using k filters in optical flow branch and generating k intermediate optical flow, we assume each filter will work independently and being active only to a single type or a few types of object motions. Correspondingly, filters in mask branch are expected to have some behaviors that each of the generated masks which indeed are active maps should be high active for certain types of motions in some region and low active for other types of motions in other regions. This inspires us to use a maxout operation to extract mask entries with maximal activation along channel axis. So, after maxout operation, for each mask $m_n, n = 1 \ldots k$, all entries will be zero-out except entries whose activation value are maximal in the some region among all masks. We denote masks after maxout as $\{m'_n\}_{n=1}^k$. Thus, there is no intersection among masks in $\{m'_n\}_{n=1}^k$ and the union of all $m'_n, n = 1 \ldots k$ has activation in full region. The maxout can be represented as following:

$$m'_n = \begin{cases} m_n, & \text{if } m_n = \max\limits_{p=1 \ldots k}(m_p) \\ 0, & \text{otherwise} \end{cases} \quad \text{for } n = 1 \ldots k \tag{4}$$

It could be seen from Equation 5, the masks after maxout are not converted to binary masks. This is the reason why the module is called soft-masks module.

Masks after maxout operation are applied to corresponding intermediate optical flows by element-wise multiplication which is shown in below:

$$f'_n = \begin{cases} m'_n \times f_n, & \text{if } m'_n \neq 0 \\ 0, & \text{otherwise} \end{cases} \quad \text{for } n = 1 \ldots k \tag{5}$$

Results of the above computation is a set of disjoint optical flow layers, each of which represent a certain type of

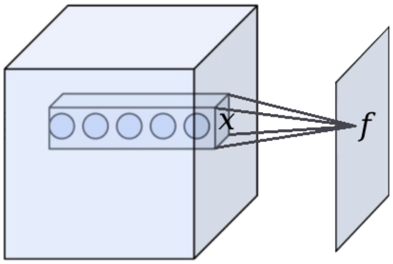


Figure 3. An illustration of how maxout is working for soft-masks module.

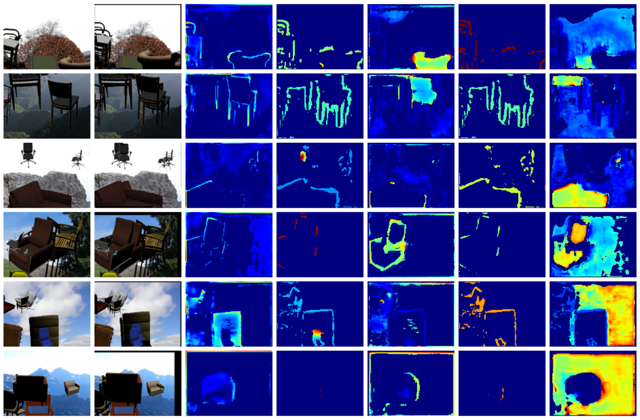


Figure 4. Examples of generated masks using the proposed soft-mask module. In this example, five masks are generated for each input image pair.Mask values are rearranged to ones between 0 and 1 for rendering.

motion. An illustration of how soft-masks module works is show in Figure 4 and results of generated masks are show in Figure

2.3. Quadratic fitting of optical flows

Objects move in different ways in images, which results in different types of motion. The underlying optical flows thus are non-linear and locally piece-wise smooth.

There are two consequences of using soft-masks module that could make estimation of the optical flow easier. The first advantage of using soft-masks module originate from usage of maxout in generating masks. By keeping only the maximal value among all masks, the optical flows are forced to be separated to multiple disjoint layers. Theoretically proofing maxout will result in a precise cut along motion boundary is challenging and is still under our investigation. However, qualitative results shown in Figure 4 demonstrate that soft-masks module enable resulting masks to separate flow field to pieces according to detected motion types. In addition, the masks detects boundary of each piece as well, which enables estimation of optical flows on boundary more accurate. Secondly, The estimation of optical flows using the proposed soft-masks module is quadratic in terms of feature maps X inputed to the module.

To show this, given the computation of masks and intermediate optical flows shown in Equation 3, the computation of non-zero f'_n could be written as:

$$\begin{aligned} f'_n &= m'_n \times f_n \\ &= (X^T W_n^m + b_n^m) \times (X^T W_n^f + b_n^f) \\ &= W_n^m X X^T W_n^f + X^T(b_n^f W_n^m + b_n^m W_n^f) + b_n^m b_n^f \end{aligned} \tag{6}$$

As shown in above equation, the representation of f'_n is quadratic in terms of variable X.

To better illustrate the difference of using soft-masks module with respect to linear output, a 1D example is show in Figure 5. In the example, function values are smooth in three separate ranges. The improvement of fitting data using piecewise quadratic function over linear function could be seen from this example.

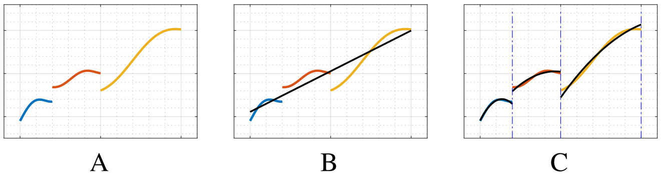


Figure 5. A: Given data. B: Fitting using linear function. C: Fitting using piece-wise quadratic function.

2.4. Regularization for Unsupervised Training

Training an unsupervised neural network for optical flow estimation is available by using network similar to FlowNet as a base optical flow inferring network followed by a spatial transform network (STN) [10]. Existing works such as: DSTFlow [14], USCNN [1], and back-to-basic unsupervised FlowNet (bb-FlowNet) [18] all follow the same framework and train their train neural networks unsupervisely to minimize an objective defined in Equation 1. To show the proposed soft-masks module can improve flow estimation under the same framework, we add the soft-masks module to FlowNet and use it as a base optical flow inferring network in unsupervised training framework.

Smoothness restriction which is used by all above mentioned unsupervised approaches plays a significant role in regularizing the local consistancy of optical flows. We follow the traditional regularization term used to constraint deformation field [16][2] and define the regularization term used in Equation 1 as following:

$$\begin{aligned} \varphi((u)_\xi, (v)_\xi) = &\sum\left(\left(\frac{\partial^2 \mathbf{u}}{\partial \mathbf{x}^2}\right)^2 + \left(\frac{\partial^2 \mathbf{u}}{\partial \mathbf{y}^2}\right)^2 + 2\left(\frac{\partial^2 \mathbf{u}}{\partial \mathbf{x} \partial \mathbf{y}}\right)^2\right) + \\ &\sum\left(\left(\frac{\partial^2 \mathbf{v}}{\partial \mathbf{x}^2}\right)^2 + \left(\frac{\partial^2 \mathbf{v}}{\partial \mathbf{y}^2}\right)^2 + 2\left(\frac{\partial^2 \mathbf{v}}{\partial \mathbf{x} \partial \mathbf{y}}\right)^2\right) \end{aligned}$$

3. Empirical Evaluation

3.1. Benchmark

We evaluate our performance on three standard optical flow benchmarks: Flying Chairs [4], Sintel [3], and KITTI [7]. We compare the performance of the proposed approach to both supervised methods such as: FlowNet(S/C) [4], FlowNet 2.0 [9], SPyNet [13] and DeepFlow [17], EpicFlow [15] and unsupervised methods including: DSTFlow [14], USCNN [1], and back-to-basic unsupervised FlowNet (bb-FlowNet) [18].

3.2. Proposed Network Structure

The goal of this paper is not to show a brand new design of a network and superior performance could be obtained using the network, instead, we are going to show performance improvement could be obtained by replacing normal optical flow output layer with soft-masks module. Based on this intention, we choose FlowNetS and FlowNetC as our base networks and replace their optical flow output layers with soft-masks module. Since we highlight the layered flow estimation (LOFE) in this paper, we term the resulting networks modified from FlowNet as FlowNetS+LOFE and FlowNetC+LOFE.

One thing worth noticing is that although we chosen FlowNet to work on in our work, but since soft-masks module basically works as an add-on to a network, it can be used on wider selections of networks and vision tasks such as image segmentation [11][12], depth estimation in range images [6][5], etc..

3.3. Evaluation of Soft-masks Module

Since we replaced simple linear output layer in FlowNet(S/C) with a relatively more complicated soft-masks module, one obvious question about our result is whether we obtained them by using a model with more coefficients rather than by the way of how soft-masks module works. Therefore, to better investigate the mechanism of soft-masks module, we compared the FlowNetC+LOFE with another two networks in which we slightly changed the structures of the soft-masks module.

In the first network, given the proposed structure as FlowNetC+LOFE, we removed maxout operation from soft-masks module and keep all other configuration the same. We term the resulting network as FlowNetC+LOFE-v1. In this case FlowNetC+LOFE-v1 will have the exact same number of coefficients as FlowNetC+LOFE. For the second network, we further totally removed the mask branch from soft-masks module and leave the intermediate optical flow branch only. The second network is denoted as FlowNetC+LOFE-v2.

3.4. Training Details

3.5. Results

4. Discussion and Summary

References

[1] A. Ahmadi and I. Patras. Unsupervised convolutional neural networks for motion estimation. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 1629–1633. IEEE, 2016. 1, 3, 4

[2] J. Ashburner, K. J. Friston, et al. Nonlinear spatial normalization using basis functions. *Human brain mapping*, 7(4):254–266, 1999. 3

[3] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag, Oct. 2012. 4

[4] A. Dosovitskiy, P. Fischery, E. Ilg, P. Husser, C. Hazirbas, V. Golkov, P. v. d. Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2758–2766, Dec 2015. 1, 4

[5] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2658, 2015. 4

[6] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014. 4

[7] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3354–3361. IEEE, 2012. 4

[8] B. K. Horn and B. G. Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981. 1

[9] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 4

[10] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, 2015. 3

[11] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 4

[12] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1520–1528, 2015. 4

[13] A. Ranjan and M. J. Black. Optical flow estimation using a spatial pyramid network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 4

[14] Z. Ren, J. Yan, B. Ni, B. Liu, X. Yang, and H. Zha. Unsupervised deep learning for optical flow estimation. In *Artificial Intelligence (AAAI-17), Proceedings of the Thirty-First AAAI Conference on*, pages 1495–1501, 2017. 1, 3, 4

[15] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid. Epicflow: Edge-preserving interpolation of correspondences for optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1164–1172, 2015. 4

[16] T. Rohlfing, C. R. Maurer, D. A. Bluemke, and M. A. Jacobs. Volume-preserving nonrigid registration of mr breast images using free-form deformation with an incompressibility constraint. *IEEE transactions on medical imaging*, 22(6):730–741, 2003. 3

[17] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. Deepflow: Large displacement optical flow with deep matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1385–1392, 2013. 4

[18] J. J. Yu, A. W. Harley, and K. G. Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. *CoRR*, abs/1608.05842, 2016. 1, 3, 4