Layered Optical Flow Estimation using a soft-mask Module in Deep Neural Networks

Anonymous CVPR submission

Paper ID 4263

Abstract

Using a layered representation for motion estimation has the advantage of being able to cope with discontinuities and occlusions. In this paper, we learn to estimate optical flow by combining a layered motion representation with deep learning. Instead of pre-segmenting the image to layers, the proposed approach automatically generates a layered representation of optical flow using the proposed soft-mask module. The essential components of the soft-mask module are maxout and fuse operations, which enable a disjoint layered representation of optical flow and more accurate flow estimation. We show that by using masks the motion estimate results in a quadratic function of input features in the output layer. The proposed soft-mask module can be added to any existing optical flow estimation networks by replacing their flow output layer. In this work we use FlowNet as the base network to which we add the soft-mask module. The resulting network is tested on three well known benchmarks with both supervised and unsupervised flow estimation tasks. Evaluation results show that the proposed network achieve better results compared with respect to the original FlowNet.

1. Introduction

Optical flow estimation is a crucial and challenging problem which numerous applications in computer vision. Traditional differential methods for estimating optical flow include patch based methods, such as affine flow and Lucas-Kanade [23] and variational methods such as Horn and Schunck in [15]. which includes a regularization term and provide a global solution of optical flow. Various of successful improvement of these initial formulations have been proposed over many years.

Layered models of optical flow offer an easy performance boost for optical flow estimation [36][19][8]. Disjointly splitting the optical flow into layers enables an easier modeling of optical flow in each layer. Such representation is especially helpful for small object motion estimation, as many optical flow estimation techniques are biased towards motion in large areas. Layered representation also improves flow computation on flow field boundaries by handling the smoothness constraint separately in each layer.

FlowNet proposed by Dosovitskiy [9] was the first work for use a deep neural network to end-to-end optical flow estimation. The concept of training FlowNet is fundamentally different from established differential approaches. As traditional differential optical flow estimation techniques perform well and are well established, several deep learning based approaches tried to bridge the gap between traditional approaches and deep learning based approaches by using the best on both sides. For example, Ranjan and Blacks [26] use a pyramid representation of flow and residual flows to address large flow displacement estimation. Several approaches [27][1][40] investigated the basic principles of flow estimation and proposed unsupervised training of networks.

Our work combines the idea of using layered optical flow representation with a deep neural network structure. Unlike previous approaches [39][4][20], where the layered representation has was generated separately, the layered representation in the proposed approach is inferred internally and automatically when training the neural network. We achieve this by designing a soft-mask module. The soft-mask module is a network structure which splits optical flow to layers using disjoint real valued masks. As the masks are not binary we use the term 'soft' to refer to them. The soft-mask module offers a more accurate flow estimation due to two unique characteristics. The first is its ability to represent estimated flow using disjoint layers, which results in a more focused and simpler flow estimation for each layer. Second, compared with the linear flow output in FlowNet, the flow estimated using the soft-mask module is a quadratic in terms of input features, which allows the soft-mask module a better abilitys to fit more complicated optical flow patterns. The idea of using the soft-mask module is similar to maxout networks proposed by Goodfellow [13], where the output of a neuron is the max of a set of inputs. The

proposed soft-mask module extends the maxout operation to 2D. In addition, instead of keeping max value only, we zero-out non-max values and use these values as mask-out region when masks are fused with layered optical flows.

In this work, the soft-mask module is added to FlowNet by replacing the output layer of the network with the soft-mask module. More generally, the soft-mask module may be used in other per-pixel prediction tasks such as semantic segmentation [22] and single image depth estimation [11]. We that by using the soft-mask module we boost the performance of FlowNet when tested on several public datasets. We further show that, both supervised and unsupervised flow estimation methods benefit from using the soft-mask module.

1.1. Related Work

Our work effectively combines ideas from using layered representation in classical optical flow approaches with recent deep learning approaches. Our literature review thus focuses on the work most relevant to this.

Layered approaches. Using layered approaches in motion estimation have been demonstrated as an approach for motion estimation by using layers to overcome discontinuities and occlusions. A layered approach has been proposed by Darrell and Pentland [7][8] where they incorporate a Bayesian model for segmentation and robust statistics. Wang and Adelson [35][36] use affine layers to represent the motion field. Similarly, recent work by Sun *et al*. [32][33] use affine motion to regularize the flow in each layer, while Jepson and Black [19] formalize the problem using probabilistic mixture models. Yang *et al*. [39] fit a piecewise adaptive flow field using piecewise parametric models while maintaining a global inter-piece flow continuity constraint. Exploiting recent advances in semantic scene segmentation, [31] use different flow types for segmented object in different layers. Hur and Roth [16] treat semantic segmentation and flow estimation as a joint problem. Additional methods for joint motion and segmentation estimation includes [3][6][24][30][34][38][42].

Deep learning approaches. Deep neural networks have been shown to be successful in many computer vision tasks including object recognition [14] and dense prediction problems [41][22]. FlowNet attempts to solve optical flow estimation using a deep neural network. FlowNet provides an end-to-end optical flow learning framework which serves as a base model for many later work [17][1][27][40]

1.2. Novel Contribution

In this work, we extend FlowNet [9] to improve its performance in several ways. First, we propose to combine a 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 when training the network. Second, a soft-mask module is proposed. This soft-mask 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 the output layers. The resulting network is trained and compared with both supervised and unsupervised optical flow estimation approaches using neural networks. Experimental results show that the proposed network structure achieves comparable or lower error in each experimental group.

2. Methodology

2.1. Objective

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 estimates 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 that transform image I_a to image I_b. Following the optical flow objective by Horn and Schunck [15], which combines a data term enforcing the optical flow constraint equation and a regularization term enforcing smoothness. The classical optical flow objective thus is defined as:

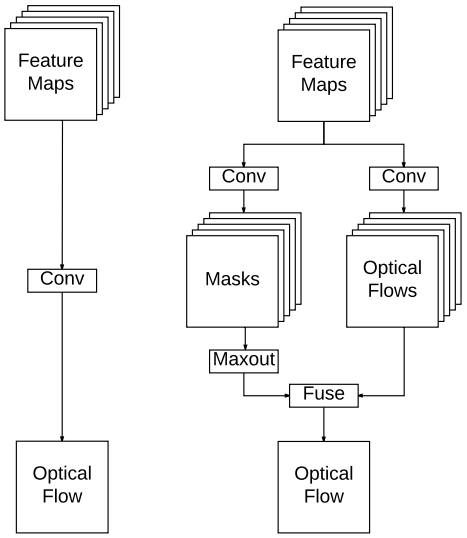
$$E(\mathbf{u},\mathbf{v}) = \sum_{i}^{H}\sum_{j}^{W}(I_0(i+u_{ij}, j+v_{ij}) - I_1(i,j))^2 + \lambda\cdot\varphi(\mathbf{u},\mathbf{v}) \tag{1}$$

In this equation $\varphi(\mathbf{u},\mathbf{v}))$ is a regularization term that constrains the smoothness of optical flow and λ is a weight coefficient.

The above objective is used by several deep neural network based for optical flow estimation based on an unsupervised training framework [1][27][40].

2.2. soft-mask module

FlowNet [9] was the first work to use deep convolutional neural network for optical flow estimation. The network architecture used by FlowNet is very similar to the structure of a classical auto-encoder, where optical flows are generated using deconvolution at each scale level of the image pyramid. To refine flow estimations, shortcuts are built to connect layers of corresponding levels in encoder and decoder layers. Consider a single computation of convolution, and for simplicity, assume that f represents both horizontal and vertical components of an output flow. Given $X \in \mathbb{R}^{s\times s\times c}$, representing an input feature volume vector, where s is kernel size and c is number of channels, FlowNet employs a



a. Normal output b. soft-mask module

Figure 1. Illustration of the structure of the proposed soft-mask module compared with traditional linear optical flow network.

linear activation to compute optical flow:

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

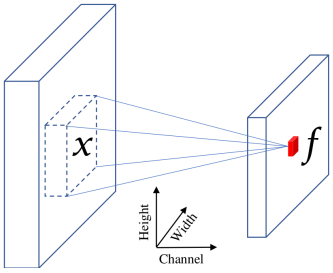


Figure 2. An illustration of annotation used in convolution.

Given that actual optical flow fields are nonlinear and piecewise smooth, using linear function to fit the flow field shifts non-linearity to the convolutional layers making the learning there more difficult. Using the soft-mask module proposed in this paper to replace the linear output of optical flow estimation, we are able to separate optical flow field to multiple layers. The flow estimation in each layer is smooth and is easier to estimate using a linear function. This results in a more accurate and flexible optical flow estimation.

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

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

Thus, given k filters, we will obtain k corresponding pairs of mask and intermediate optical flow. By using k filters in the optical flow branch and generating k intermediate optical flow fields, we assume that each filter will work independently and model a single type or a few types of object motions. Correspondingly, filters in the mask branch are expected to mask out parts with consistent motions by being high in certain regions and low in others. This leads us to use a maxout operation to extract mask entries with maximal activation along the channel axis. After the maxout operation, for each mask $m_n (n = 1 \dots k)$, all entries will be zero-out except for entries whose activation values are maximal in the some region among all masks. We denote the masks after maxout using $\{m'_n\}_{n=1}^k$. Following the maxout operation, there is no intersection among masks and the union of all $m'_n, n = 1 \dots k$ has activation in the full region. The maxout is given by:

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

Note that, the masks after maxout are not converted to binary values. thus resulting in soft-masks. By using soft masks we are able to not only mask out irrelevant parts, but also prioritize values in each layer.

In the proposed approach, masks after the maxout operation are applied to corresponding intermediate optical flow field by element-wise multiplication as shown below:

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

The results of the above computation is a set of disjoint optical flow layers, each of which represents a certain type of motion. An illustration of how the soft-mask module works is show in Figure 3 and results of generated masks are shown in Figure 4.

2.3. Quadratic fitting of optical flow

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

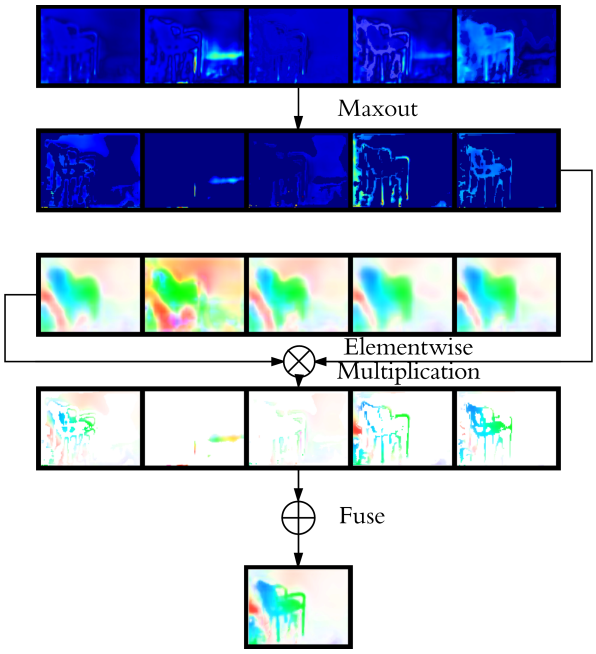


Figure 3. Pipeline of soft-mask module.

There are two consequences of using the proposed soft-mask module that could make the estimation of the optical flow easier. The first advantage of using the soft-mask module originates from using maxout in mask generation. By keeping only the maximal value among all masks, the optical flow is separated to multiple disjoint layers. Theoretically proving that maxout result in a precise cut along motion boundaries is challenging and is still under investigation. However, qualitative results as shown in Figure 4 demonstrate that the soft-mask module enable resulting masks to separate the flow field into pieces according to detected motion types. In addition, the masks detect the boundary of each piece, which allows for the estimation of optical flows on boundaries to be more accurate. The second advantage of using the proposed soft-mask module is that the output is quadratic in terms of feature maps X fed to the module. To see this, consider the computation of masks and intermediate optical flows shown in Equation 3. The computation of non-zero f'_n could be written as:

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

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

To better illustrate the advantage in using the soft-mask module with respect to linear output. Consider the 1D example shown in Figure 5. In this example, function values are smooth in three separate domains. The improvement of fitting data using a piecewise quadratic function is shown by comparing B and C.

2.4. Regularization for Unsupervised Training

Training an unsupervised neural network for optical flow estimation is possible by using a network similar to FlowNet for base optical flow inference followed by a spatial transform network (STN) [18] to generate warped iamges for comparison. Existing work such as: DSTFlow [27], USCNN [1], and back-to-basic unsupervised FlowNet (bb-FlowNet) [40] all follow the same framework and train their networks without supervision to minimize the objective defined in Equation 1. To show that proposed soft-mask module can improve flow estimation using the same framework, we add the soft-mask module to FlowNet and use it as a base optical flow inference network in an unsupervised training framework.

The smoothness term which is used by all above unsupervised approaches plays a significant role in regularizing the local consistancy of optical flow. We use the bending energy regularization [29][2] defined by:

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

3. Empirical Evaluation

3.1. Benchmark

We evaluate our performance on three standard optical flow benchmarks: Flying Chairs [9], Sintel [5], and KITTI [12]. We compare the performance of the proposed approach to both supervised methods such as: FlowNet(S/C) [9], SPyNet [26], DeepFlow [37], and EpicFlow [28], and unsupervised methods including: DSTFlow [27], USCNN [1], and back-to-basic unsupervised FlowNet (bb-FlowNet) [40].

Recently, FlowNet 2.0, a follow-up work of FlowNet, achieved state of the art results on most datasets. The architecture of FlowNet 2.0 [17] uses several FlowNets and contains cascade training of the FlowNets in different phases. Since the focus of this paper is on using the soft-mask module to boost performance of a single network, we do not include FlowNet 2.0 in our evaluation. Note that the proposed soft-mask module can be incorporated into FlowNet 2.0.

3.2. Network Structures

The goal of this paper is to show how the performance of existing optical flow networks can be improved by replacing the normal optical flow output layer with the proposed

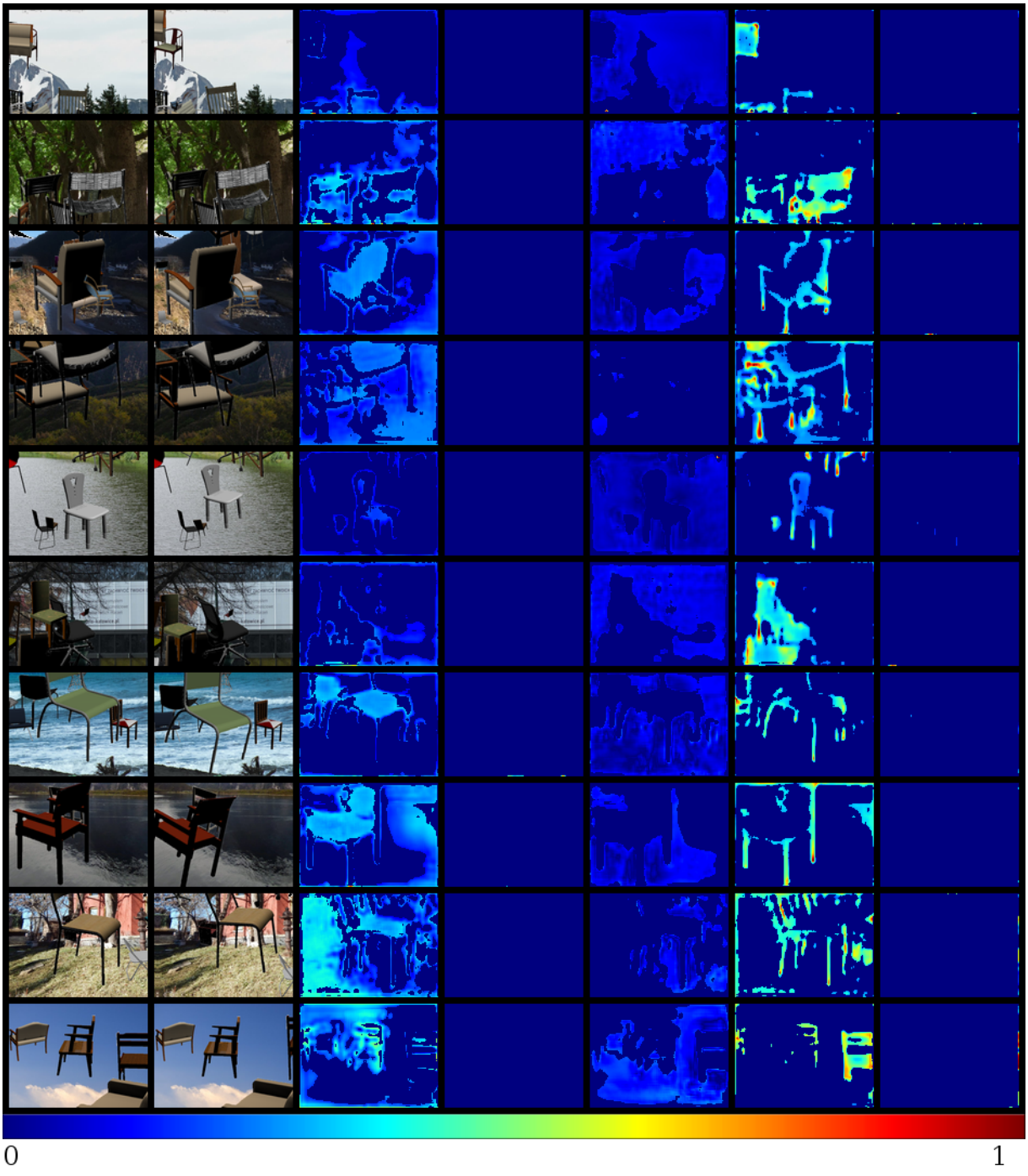


Figure 4. Examples of generated masks using the proposed soft-mask module. In this example, five masks are generated for each input image pair (left). Note that column 4 and 7 are not empty. Some small structures such as flows on the boundaries are captured by masks in these two columns. The rendering of the figure is best viewed electronically.

soft-mask module. We choose FlowNetS and FlowNetC as the base networks and replace their optical flow output layers with a soft-mask module. Using the layered optical flow estimation (LOFE) proposed in this paper, we term the resulting modified networks: FlowNetS+LOFE and FlowNetC+LOFE, respectively.

While we evaluate FlowNet in this work, the soft-mask module can be added to other networks and used on various vision tasks such as image segmentation [22][25] and depth estimation in range images [11][10].

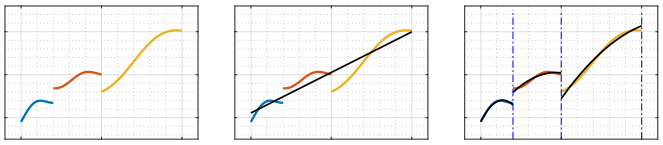


Figure 5. A: Given data. B: Fitting using linear function. C: Fitting using piecewise quadratic function.

3.3. Training Details

Both supervised and unsupervised networks are trained using Adam [21] optimization with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We use a small batch size of 8 across all dataset with 3000 iterations per epoch. The initial learning rate is set to be $(1e-4)$ and decrease to half every 60 epochs until the network converge. All experiments were finished on a single Nvidia 1080 GPU.

Various types of data augmentation are used during training. We apply rotation at random within $[-17°, 17°]$. A random translation within $[-50, 50]$ pixels is applied to both horizontal and vertical directions. In addition, following [26] we include additive white Gaussian noise sampled uniformly from $\mathcal{N}(0, 0.1)$. We also apply color jitter with additive brightness, contrast and saturation samlped from a Gaussian, $\mathcal{N}(0, 0.4)$. All data augmentation are done using GPU during training.

3.4. Results

Evaluations were done with compared methods in two groups according to whether the training of the methods is unsupervised or supervised. Table 1 shows the endpoint error (EPE) of the proposed network and several well known methods. The endpoint error measures the distance in pixels between known optical flow vectors and estimated ones. Except for EpicFlow and DeepFlow, all the other methods were trained in a supervised manner. We compare results of unsupervised methods in Table 2.Note that FlowNet [17] and SPyNet [26], report additional fine-tuned network results which are not used here since we are interested in evaluating the impact of using the soft-mask module, and since fine tuning may vary between methods and thus mask differences.

Supervised methods. The proposed FlowNetS+LOFE and FlowNetC+LOFE tested in this group use $k = 10$ for the number of layers. As can be seen from Table 1, FlowNetS+LOFE and FlowNetC+LOFE achieve the best performance on all tests except for the training set of the Sintel Clean dataset. We observe that, the performance of FlowNetS and FlowNetC are both boosted by replacing the optical flow output layer with the soft-mask module. Considering the computation time we observe a small time increment when using the soft-mask module, and which is in an acceptable range.

Unsupervised methods. Training optical flow estimation networks without supervision is straight forward by using the objective shown in Equation 1. The difficult part is to decide the weight coefficient λ in the equation. To choose an appropriate λ, we did a grid search for the best value in a range of $[0, 10]$. We stopped the search when the improvement is small. The results are shown in Table2. As can be observed, the proposed networks achieve the best performance except for KITTI dataset.

3.5. Evaluation of the soft-mask Module

Since we replaced the simple linear output layer in FlowNet(S/C) with a more complex soft-mask module, we would like to verify whether the improved results are obtained due to the way then soft-mask module works and not simply due to having a model with more coefficients. To better investigate the operation of the soft-mask module, we compared the FlowNetC+LOFE with two other networks in which we slightly changed the structures of the soft-mask module.

In the first network, given the proposed structure as FlowNetC+LOFE, we removed the maxout operation from the soft-mask module and kept the remaining configuration the same. We denote the resulting work FlowNetC+LOFE/no-maxout. In this case FlowNetC+LOFE/no-maxout will have the exact same number of coefficients as FlowNetC+LOFE. For the second network, we removed the mask branch from soft-mask module leaving only the intermediate optical flow only. The second network is denoted as FlowNetC+LOFE/no-masks. For all three networks, we use $k = 10$ in the soft-mask module. We use original FlowNetC as a baseline in this comparison. In order to obtain an unbiased comparison result, we trained and tested each of these networks on the Flying Chairs dataset five times. The average EPE and standard deviation is reported in Table 3.

As can be seen from Table 3, the proposed FlowNet+LOFE performed better than its two variants. This comparison leads to two conclusions. First, the better performance obtained by adding the soft-mask module to FlowNet is not because a larger model being used. This is since the no-maxout version of the proposed network has identical complexity to the proposed network. Thus we conclude that the maxout operation make optical flow estimation an easier task by separating optical flows to different layers. Second, it can be seen from the table that the performance of the no-maxout version is slightly better than that of the no-masks version. While a possible explanation is that the model of no-maxout is bigger than the model of no-masks, note that FlowNet, the smallest model in this comparison, achieved a better performance

Figure 6. Examples of predicted flows compared with results from FlowNet. A/B: input image pairs. C: FlowNetS, D: FlowNetC, E: FlowNetS+LOFE, F: FlowNetC+LOFE.

Table 1. Average end point errors (EPE) of the proposed networks compared to several existing methods. EpicFlow and DeepFlow are traditional methods which do not use neural networks. All other methods in the table are trained with supervised data. Bold font indicates the most accurate results among the network based methods.

Method	Flying Chairs	Sintel Clean		Sintel Final		KITTI		Time (s)
	Test	Train	Test	Train	Test	Train	Test	
EpicFlow	2.94	2.40	4.12	3.7	6.29	3.47	3.8	16
DeepFlow	3.53	3.31	5.38	4.56	7.21	4.58	5.8	17
FlowNetS	2.71	4.50	7.42	5.45	8.43	8.26	-	0.12
FlowNetC	2.19	4.31	7.28	5.87	8.81	9.35	-	0.23
SPyNet	2.63	**4.23**	**6.82**	5.67	8.49	9.12	-	0.11
FlowNetS+LOFE	2.53	4.35	7.11	**5.32**	**8.25**	**8.03**	-	0.17
FlowNetC+LOFE	**2.08**	**4.23**	7.01	5.51	8.51	9.14	-	0.30

than the no-masks model. We hypothesize that the reason leading to this result is a fact that the no-maxout version of model is using a quadratic function to fit optical flow instead of the linear function used by the no-masks version.

We investigate the relationship between k the number of masks and flow layers used in the soft-mask module, and network performance in terms of EPE. Experiments were done using the Flying Chairs dataset. We start with a soft-mask module with $k = 2$, then set $k = 5$x, where $x = 1, \ldots, 8$. As can be observed in Figure 7, there is an immediate benefit to using the soft-mask module with respect to FlowNetC, where $k = 2$ will efficiently boost performance. We see in Figure 7 convergence after $k = 10$ and a slightly increase when $k > 25$. This may be due to slight overfitting when separating the optical flow to too many layers.

4. Conclusion

We describe a new approach for optical flow estimation by combining traditional layered flow representation with deep learning method. rather than pre-segmenting images to layers, the proposed approach automatically learns a layered representation of optical flow using the proposed soft-

Table 2. EPE errors of methods that are trained without supervision. The results of compared methods are taken directly from the corresponding paper.

Method	Flying Chairs	Sintel Clean		Sintel Final		KITTI	
		Train	Test	Train	Test	Train	Test
DSTFlow	5.11	6.93	10.40	7.82	11.11	**10.43**	-
USCNN	-	-	-	8.88	-	-	-
BB-FlowNet	5.36	-	-	-	-	11.32	**9.93**
FlowNetS+LOFE	**4.81**	**6.56**	10.10	**7.62**	10.98	10.78	10.82
FlowNetC+LOFE	4.92	6.78	**9.98**	7.77	**10.19**	11.01	11.25

Table 3. Comparison of the proposed FlowNet+LOFE and its two variants.

	Flying Chairs
FlowNetC	2.19 ± 0.021
FlowNetC+LOFE	2.08 ± 0.018
FlowNetC+LOFE/no-maxout	2.16 ± 0.021
FlowNetC+LOFE/no-masks	2.22 ± 0.021

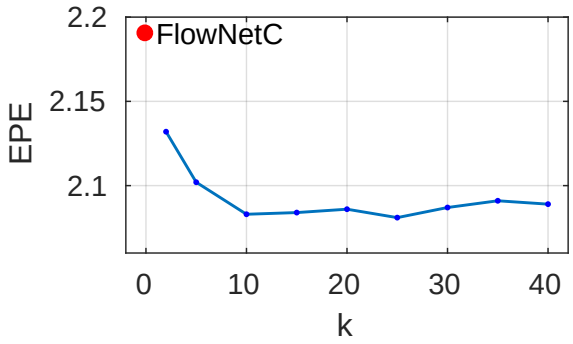


Figure 7. EPE as a function of k, the number of layers used in soft-mask module.

mask module. The soft-mask module has the advantage of splitting flow to layers in which the computation of the flow is quadratic in terms of input features. For evaluation, we use FlowNet as our base net to add the soft-mask module. The resulting networks are tested on three well known benchmark with both supervised and unsupervised flow estimation tasks. Experimental results show that the proposed network achieve better results with respect to the original FlowNet.

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, 2, 4

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

[3] S. Birchfield and C. Tomasi. Multiway cut for stereo and motion with slanted surfaces. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 1, pages 489–495. IEEE, 1999. 2

[4] M. J. Black and A. D. Jepson. Estimating optical flow in segmented images using variable-order parametric models with local deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(10):972–986, 1996. 1

[5] 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

[6] D. Cremers and S. Soatto. Motion competition: A variational approach to piecewise parametric motion segmentation. *International Journal of Computer Vision*, 62(3):249–265, 2005. 2

[7] T. Darrell and A. Pentland. Robust estimation of a multi-layered motion representation. In *Visual Motion, 1991., Proceedings of the IEEE Workshop on*, pages 173–178. IEEE, 1991. 2

[8] T. Darrell and A. P. Pentland. Cooperative robust estimation using layers of support. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5):474–487, 1995. 1, 2

[9] 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, 2, 4

[10] 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. 5

[11] 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. 2, 5

[12] 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

[13] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio. Maxout networks. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pages III–1319–III–1327. JMLR.org, 2013. 1

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[15] B. K. Horn and B. G. Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981. 1, 2

[16] J. Hur and S. Roth. *Joint Optical Flow and Temporally Consistent Semantic Segmentation*, pages 163–177. Springer International Publishing, Cham, 2016. 2

[17] 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. 2, 4, 6

[18] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, 2015. 4

[19] A. Jepson and M. J. Black. Mixture models for optical flow computation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 760–761, Jun 1993. 1, 2

[20] S. X. Ju, M. J. Black, and A. D. Jepson. Skin and bones: Multi-layer, locally affine, optical flow and regularization with transparency. In *Computer Vision and Pattern Recognition, 1996. Proceedings CVPR'96, 1996 IEEE Computer Society Conference on*, pages 307–314. IEEE, 1996. 1

[21] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[22] 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. 2, 5

[23] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'81, pages 674–679, San Francisco, CA, USA, 1981. Morgan Kaufmann Publishers Inc. 1

[24] E. Mémin and P. Pérez. Hierarchical estimation and segmentation of dense motion fields. *International Journal of Computer Vision*, 46(2):129–155, 2002. 2

[25] 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. 5

[26] 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. 1, 4, 6

[27] 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, 2, 4

[28] 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

[29] 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. 4

[30] A. Roussos, C. Russell, R. Garg, and L. Agapito. Dense multibody motion estimation and reconstruction from a handheld camera. In *Mixed and Augmented Reality (ISMAR), 2012 IEEE International Symposium on*, pages 31–40. IEEE, 2012. 2

[31] L. Sevilla-Lara, D. Sun, V. Jampani, and M. J. Black. Optical flow with semantic segmentation and localized layers. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[32] D. Sun, E. B. Sudderth, and M. J. Black. Layered image motion with explicit occlusions, temporal consistency, and depth ordering. In *Advances in Neural Information Processing Systems*, pages 2226–2234, 2010. 2

[33] D. Sun, E. B. Sudderth, and M. J. Black. Layered segmentation and optical flow estimation over time. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1768–1775. IEEE, 2012. 2

[34] M. Unger, M. Werlberger, T. Pock, and H. Bischof. Joint motion estimation and segmentation of complex scenes with label costs and occlusion modeling. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1878–1885. IEEE, 2012. 2

[35] J. Y. Wang and E. H. Adelson. Layered representation for motion analysis. In *Computer Vision and Pattern Recognition, 1993. Proceedings CVPR'93., 1993 IEEE Computer Society Conference on*, pages 361–366. IEEE, 1993. 2

[36] J. Y. Wang and E. H. Adelson. Representing moving images with layers. *IEEE Transactions on Image Processing*, 3(5):625–638, 1994. 1, 2

[37] 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

[38] K. Yamaguchi, D. McAllester, and R. Urtasun. Robust monocular epipolar flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1862–1869, 2013. 2

[39] J. Yang and H. Li. Dense, accurate optical flow estimation with piecewise parametric model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1019–1027, 2015. 1, 2

[40] 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, 2, 4

[41] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1537, 2015. 2

[42] C. Zitnick, N. Jojic, and S. B. Kang. Consistent segmentation for optical flow estimation. In *Computer Vision, 2005. ICCV*

2005. Tenth IEEE International Conference on, volume 2, pages 1308–1315. IEEE, 2005. 2