AD-NET: ATTENTION GUIDED NETWORK FOR OPTICAL FLOW ESTIMATION USING DILATED CONVOLUTION

Mingliang Zhai[⋆]*, Xuezhi Xiang*[⋆]*, Rongfang Zhang*[⋆]*, Ning Lv*[⋆] *and Abdulmotaleb El Saddik*[†]

[⋆] Harbin Engineering University, School of Information and Communication Engineering
Harbin 150001, China
[†]University of Ottawa, School of Electrical Engineering and Computer Science
Ottawa ON K1N 6N5, Canada

ABSTRACT

Variational models for optical flow estimation usually define an energy function that contains prior assumptions to explore rudimentary statistics of images. However, such methods cannot learn motion knowledge from the pre-prepared data and have many parameters that need to be set manually. Nowadays, convolutional neural networks (CNNs) have been used in optical flow estimation successfully, which can learn weights from the training dataset and can predict optical flow end-to-end. In this paper, we propose an attention guided network for learning optical flow, named AD-Net, which contains several attention units for modelling the relativities between the channels. Further, we introduce dilated convolution into supervised network for reducing the loss of motion details. In addition, some prior auxiliary constraints are embedded in the supervised network as auxiliary loss terms. Our proposed approach is tested on MPI-Sintel and KITTI2012 datasets and can preserve motion edges and details effectively.

Index Terms— Optical flow estimation, deep learning, attention mechanism, dilated convolution

1. INTRODUCTION

Inferring optical flow is one of the key challenges in fields such as autonomous driving and action recognition. The estimation of optical flow has been developed for many years. Traditional approaches [1, 2, 3] always define an energy function based on prior knowledge, which capture rudimentary statistics of images. We regard these methods as knowledge-driven methods. However, knowledge-driven approaches usually need pre-define the prior constraints and are too slow to be used in real-world application.

This work was supported in part by the National Natural Science Foundation of China under Grant 61401113, in part by the Natural Science Foundation of Heilongjiang Province of China under Grant LC201426, and in part by the Fundamental Research Funds for the Central Universities of China under Grant HEUCF190801.

Nowadays, CNNs are widely used for optical flow estimation [4, 5, 6, 7, 8], which can automatically learn knowledge from training data. From that, we regard these learning approaches as data-driven methods. The first research of learning optical flow is proposed by Dosovitskiy *et al.* [4], which introduces two end-to-end trained networks FlowNetS and FlowNetC based on encoder-decoder architecture. However, [4] still cannot compare to most of knowledge-driven methods [9, 10, 11]. To further improve accuracy, FlowNet2.0 [5] cascades several sub-networks (FlowNet) to form a larger network and trains these small networks one-by-one. Although FlowNet2.0 achieves good performance, the amount of parameters is large due to stacking several separate networks. Moreover, the training process of FlowNet2.0 is complicated and the training process is time consuming. The FlowNet1.0 and FlowNet2.0 are both based on encoder-decoder networks, which are easy to loss the details of motion due to strided convolution and cannot lay emphasis on the important motion details. In addition, [4] and [5] are trained on synthetic data with supervised manner by using endpoint loss, which only rely on data and ignore the advantages of many prior constraints used in knowledge-driven methods. [6, 7, 8] are unsupervised methods which do not need to train on labelled data. Although unsupervised methods do not require large amounts of labelled data for training, the accuracy of these methods is slightly lower than the supervised methods.

Recently, attention mechanism is widely used in many computer vision areas such as image classification [12], object detection [13] and image super-resolution [14]. The mainly purpose of attention mechanism is to capture important feature information for guiding feature learning, which can be viewed as a unit to bias the allocation of available processing resources towards the most informative parts of an input image [12, 13, 14]. Inspired by the methods mentioned above, we introduce attention mechanism into optical flow learning network for paying much attention to the important motion details. Dilated convolution is first proposed in semantic segmentation task [15] for improving the perceptive field of convolutional kernel. In optical flow field,

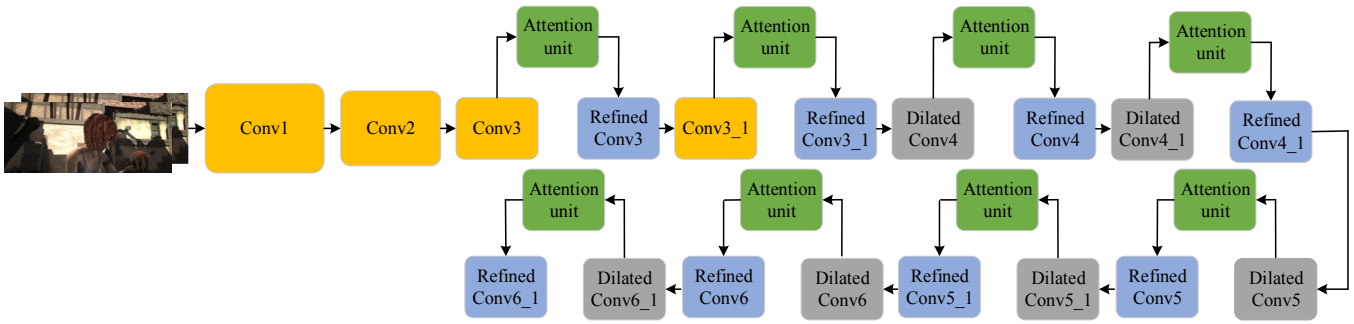

Fig. 1. An overview of our proposed network based on attention unit and dilated convolution (contracting part). Starting with conv3, we introduce attention units for refining feature maps. The last six layers are implemented by dilated convolution.

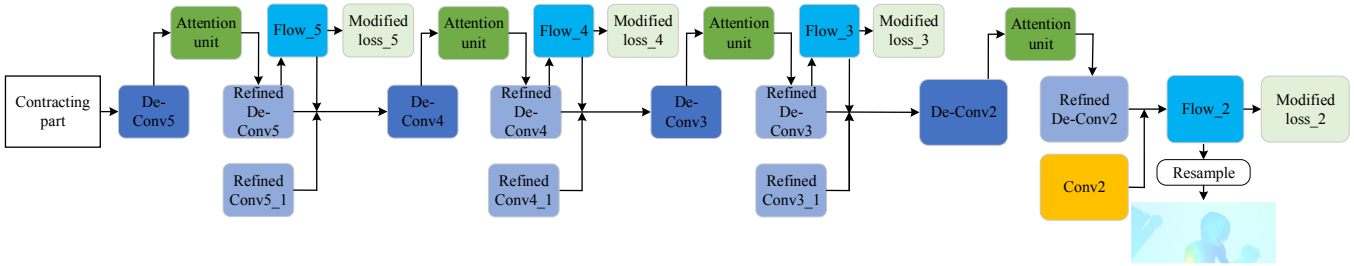

Fig. 2. An overview of our proposed network based on attention unit and prior auxiliary constraints (expanding prat). The expanding part contains a series of deconvolution layers, and we embed attention unit after each deconvolution operation. The modified loss is combined with prior auxiliary assumptions.

Zhu and Newsam [16] design an unsupervised network using dilated convolution. In contrast to [16], our method introduces dilated convolution into the supervised network, which improves the size of feature map without large computational burden and preserves the details of motion. Xiang *et al.* [17] combine prior assumptions with supervised loss term, which can not only use prior knowledge but also use large amounts of data during training. In addition, we also employ the prior auxiliary constraints in our training loss. In summary, we propose a novel network for learning optical flow, called AD-Net, which is combined with the attention mechanism, dilated convolution and prior auxiliary constraints.

We summarize our contributions as follows:

1) We introduce attention mechanism into learning optical flow network.

2) We learn optical flow in supervised manner incorporating dilated convolution operation.

3) In addition, our supervised network is combined with prior auxiliary constraints which are widely used in knowledge-driven methods.

2. METHOD

In Section 2.1, we first explain how to integrate attention mechanism into learning optical flow. Then, we mainly introduce how we incorporate dilated convolution for optical flow estimation in Section 2.2. The training loss is detailed in Section 2.3. Our network is based on encoder-decoder architecture. The encoder part is shown in Fig. 1, which is combined with attention unit and dilated convolution. The decoder part is shown in Fig. 2, which is combined with attention unit and the prior auxiliary constraints. The detail of attention unit is shown in Fig. 3.

2.1. Attention Guided Networks

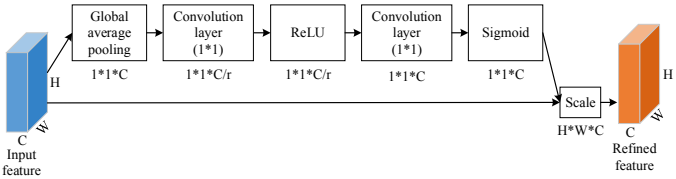

Fig. 3. The attention unit proposed in [12] contains five parts: global average pooling, convolution layer with 1*1 kernel, ReLU, sigmoid and scale layer. H, W and C are height, width and channel number of feature map.

Existing deep convolutional neural network based methods for optical flow estimation such as [4, 5, 6] have mostly focused on accuracy, but do not pay much attention to the important motion details. Inspired by [12], we introduce at-

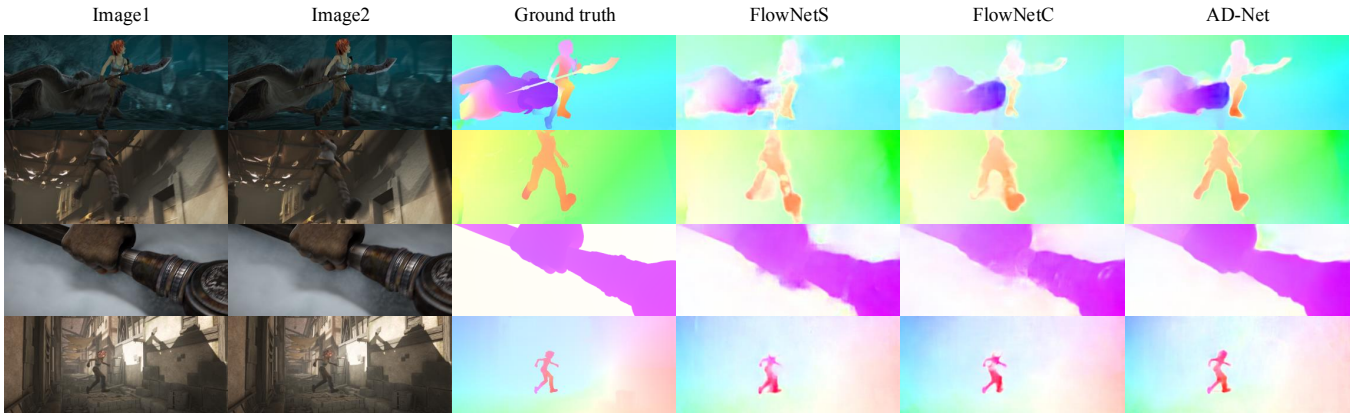

Fig. 4. Visual examples of predicted optical flow from different methods on MPI-Sintel dataset (final version). The results of FlowNetS, FlowNetC and AD-Net are shown from left to right.

tention mechanism into flow estimation network. Fig. 1 and Fig. 2 show the modified encoder and decoder parts based on FlowNetS, which are similar to FlowNetS but incorporate the attention units and dilated convolution (described in Section 2.2). The numbers of feature maps for the layers from Conv1 to De-Conv2 are 64, 128, 256, 256, 512, 512, 512, 512, 1024, 1024, 512, 256, 128, 64 respectively. As shown in Fig. 1 and Fig. 2, from Conv3 to De-Conv2, we implement the attention units that are designed to change the weights of different channels according to the global information. The detail of attention unit is shown in Fig. 3. H and W are the height and width of image. First, the input feature ($H * W * C$) are fed into a global average pooling layer followed by two $1 * 1$ convolution layers, which produces a $1*1*C$ tensor where C is the number of channels of input. The value in this $1 * 1 * C$ tensor represents the weight of different channels. Then, the $1 * 1 * C$ tensor is reshaped as $H * W * C$. r is the reduction ratio. In our network, r is set to 16. Finally, the input feature is weighted by element-wise multiplication operation.

2.2. Dilated Convolution

The previous encoder-decoder networks for optical flow estimation, such as FlowNetS and FlowNetC, usually loss motion details due to convolutional stride and deconvolution operation, which may impede dense prediction tasks such as optical flow estimation. Detailed spatial information is desired in optical flow estimation. To solve this problem, one strategy is to keep the resolution of the feature maps unchanged directly. However, this approach increases the number of parameters and computational burden of the model.

Dilated convolution, is first proposed in semantic segmentation [15], which can increase the receptive field of the convolutional kernel without reducing the resolution of feature map. Using dilated convolution in network can extract more comprehensive feature information while keeping the spatial resolution of feature maps unchanged, which is significant for dense optical flow estimation. So, we adopt dilated convolution in our network. From Fig. 1, we can see that six dilated convolution layers are embedded in contracting part. In FlowNetS, the original minimum resolution of feature map is $1/64$ of input. After using dilated convolution, the minimum resolution of feature map is kept at $1/8$ of input. From Conv4 to Conv6_1, the rates of dilation are set as 2, 2, 4, 4, 8, 8 respectively.

2.3. Training loss

Most supervised methods [4, 5] only use endpoint error (EPE) as a loss term to guide the training of network. The goal of training is to minimize the EPE loss. However, these methods overemphasize the factor of deep learning and ignore advantages of prior constraints used in knowledge-driven methods. The previous work [17] proposes a loss function for training optical flow, which combines the prior assumptions used in knowledge-driven methods with the EPE loss term. The whole loss function contain four terms: brightness constancy loss, gradient constancy loss, smoothness loss and EPE loss. In our network, we also use the same loss function during training. From Fig. 2, we can find that the modified loss is calculated at different stages.

3. EXPERIMENTS

3.1. Training Details

Our network was trained on FlyingChairs and FlyingThings3D datasets. We used Caffe [18] as deep learning framework. The network were trained on a NVIDIA 1080Ti GPU. We used Adam optimizer and first trained our network on FlyingChairs dataset with 1200k iterations. The learning rate was set to 0.0001 which was later divided by 2 every 200k iterations after the first 400k. The batch size on FlyingChairs dataset was set to 8. Second, we fine-tuned our network on

Table 1. Performance comparison on public benchmarks

Method	Sintel clean		Sintel final		KITTI 2012
	Train	Test	Train	Test	Train
DenseFlow [7]	-	-	-	10.07	-
FlowNet2-S [5]	3.79	-	4.93	-	-
UnsupFlow [6]	-	-	-	-	11.3
Occ-Aware [8]	5.23	8.02	6.34	9.08	12.95
FlowNetS [4]	4.50	7.42	5.45	8.43	8.26
FlowNetC [4]	4.31	7.28	5.87	8.81	9.35
AD-Net	**3.05**	**6.46**	**4.71**	**7.51**	**6.01**

FlyingThings3D dataset with 500k iterations. The learning rate was set to 0.00001 and divided it by 2 every 100k after the first 200k iterations. The batch size on FlyingThings3D dataset was set to 4. The whole training process is same as the process of training sub-network proposed in [5]. We further tested our model on MPI-Sintel and KITTI2012 datasets. The weights of the modified loss function were set according to [17].

3.2. Results

In Table 1, we compared our network with recent data-based approaches [4, 5, 6, 7, 8] on MPI-Sintel and KITTI2012 benchmarks. Among them, [4, 5] are supervised approaches, and [6, 7, 8] are unsupervised approaches. We use the average endpoint error (AEE) as the criterion for evaluation. Table 2 reports the performance comparison on public benchmarks. Our method achieves better results than FlowNetS and FlowNetC on Sintel and KITTI datasets. We also performs better than unsupervised methods [6, 7, 8]. Comparing to FlowNet2-S [5], our method can performs better results on Sintel dataset. [5] also proposes other network that stacks several sub-networks. The stacked network can obtain better results than our method. However, each sub-network needs to be trained one by one and the total number of iterations is more than our iterations. The running time of our method is 0.4s per frame on KITTI2012 dataset. The experimental results show that using attention mechanism and dilated convolution is beneficial for optical flow estimation. We further show some example estimations on Sintel final dataset in Fig. 4. We can find that our method can preserve more motion details than FlowNetS and FlowNetC and can obtain clear motion edges.

3.3. Ablation Study

In this section, we conduct several experiments to evaluate the effectiveness of our proposed method. To facilitate comparison, we trained the follows network only on FlyingChairs dataset with short training schedule proposed in [4], and compared to original FlowNetS. On MPI-Sintel final dataset (training), the AEE of original FlowNetS is 5.45. The AEE of our models are shown in Table 2.

Table 2. Ablation study

Attention mechanism	Dilated convolution	Auxiliary constraints	Sintel (AEE)
No	No	No	5.45
Yes	No	No	5.16
No	Yes	No	4.91
No	No	Yes	5.02
Yes	Yes	Yes	4.71

Attention mechanism. We added attention units into FlowNetS network. From Table 2, we can find that the AEE drops about 5.3% compared with FlowNetS.

Dilated convolution. We further only introduced dilated convolution into FlowNetS. The experimental results show that the dilated convolution has about 9.9% improvement on the results.

Prior auxiliary constraints. From Table 2, we can find that using prior auxiliary constraints can reduce the average endpoint error about 7.9%, which proves the effectiveness of our method.

4. CONCLUSION

In this paper, we introduce attention mechanism, dilated convolution and prior auxiliary constraints into the supervised network for learning optical flow. The attention mechanism performs feature refinement in network, which can correct the weight of the feature map according to the global information. Moreover, unlike previous works that typically reduce spatial resolution of the feature maps, we employ dilated convolution which preserves motion details without large parameters. In addition, we further combine the prior constraints with CNNs, which can constrain the relationship between the image and optical flow during training. We evaluate the proposed method both numerically and qualitatively on the benchmark datasets, such as MPI-Sintel and KITTI2012. The experimental results indicate that flow fields estimated by our network are more sharper and have rich motion details.

5. REFERENCES

[1] B. Horn and B. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, no. 1, pp. 185 – 203, 1981.

[2] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2010, pp. 2432–2439.

[3] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 500–513, March 2011.

[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)*, Dec 2015, pp. 2758–2766.

[5] 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 *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 1647–1655.

[6] Jason J. Yu, Adam W. Harley, and Konstantinos G. Derpanis, "Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness," in *Computer Vision – ECCV 2016 Workshops*, 2016, pp. 3–10.

[7] Y. Zhu and S. Newsam, "Densenet for dense flow," in *2017 IEEE International Conference on Image Processing (ICIP)*, Sept 2017, pp. 790–794.

[8] Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, Peng Wang, and Wei Xu, "Occlusion aware unsupervised learning of optical flow," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[9] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "Deepflow: Large displacement optical flow with deep matching," in *2013 IEEE International Conference on Computer Vision (ICCV)*, Dec 2013, pp. 1385–1392.

[10] L. Bao, Q. Yang, and H. Jin, "Fast edge-preserving patchmatch for large displacement optical flow," *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 4996–5006, Dec 2014.

[11] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "Epicflow: Edge-preserving interpolation of correspondences for optical flow," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1164–1172.

[12] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[13] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu, "Reverse attention for salient object detection," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[14] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu, "Image super-resolution using very deep residual channel attention networks," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[15] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, April 2018.

[16] Y. Zhu and S. Newsam, "Learning optical flow via dilated networks and occlusion reasoning," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, Oct 2018, pp. 3333–3337.

[17] X. Xiang, M. Zhai, R. Zhang, Y. Qiao, and A. El Saddik, "Deep optical flow supervised learning with prior assumptions," *IEEE Access*, vol. 6, pp. 43222–43232, 2018.

[18] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, 2014, pp. 675–678.