MA-NET: MULTI-SCALE ATTENTION-AWARE NETWORK FOR OPTICAL FLOW ESTIMATION

Mu Li* *Baojiang Zhong** *Kai-Kuang Ma*†

* School of Computer Science and Technology, Soochow University, Suzhou, China
†School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

ABSTRACT

Existing methods for optical flow estimation can perform well on images with small offsets of large objects. However, they could fail when there are a number of small and fast-moving objects. In particular, current deep learning methods often impose a down-sampling of the input data for a reduction of computational complexity. This practice inevitably incurs a loss of image details and causes small objects be ignored. To solve the problem, a *multi-scale attention-aware network* (MA-Net) is proposed, in which coarse-scale and fine-scale features are extracted in parallel and then attention is paid to them for producing the optical flow estimation. Extensive experiments conducted on the Sintel and KITTI2015 datasets show that the proposed MA-Net can capture fast-moving small objects with high accuracy and thus deliver superior performance over a number of state-of-the-art methods.

Index Terms— Optical flow estimation, deep learning, multi-scale attention, local feature, global feature

1. INTRODUCTION

With the rapid progress and great success of deep learning, a number of deep learning-based methods have been proposed for optical flow estimation in recent years. A pioneering work is the FlowNet proposed by Dosovitskiy et al [1] in 2015, which is an end-to-end approach constructed by adopting the U-Net architecture [2]. Ilg et al [3] proposed the ‘FlowNet 2.0’ by stacking several FlowNet networks into a large model. Sun et al [4] developed the PWC-Net, which uses a feature warping layer and iterative optimization. Teed and Deng [5] proposed the RAFT, in which down-sampling and iterative optimization are combined without using U-Net architecture.

In the above-mentioned networks, data down-sampling is generally performed during the process of feature extraction. Unfortunately, this practice could incur a loss of small objects motion details, leading to inferior estimation performance. Several attempts have been made by using attention modules to solve the problem. The attention mechanism in computer vision is originated from our human visual system—we tend to selectively focus on part of information and ignore other in a real-world scene. Attention-aware networks have been widely used in deep learning [6, 7, 8]. For the task of optical flow estimation, Zhai et al [9] integrated the *squeeze-and-excitation* (SE) module [6] into the FlowNet [1], in which channel attention is exploited but spatial attention is ignored. Xiang et al [10] introduced a spatial attention module and then integrated channel and spatial attention in their network.

To fully explore the potential of attention mechanism, a *multi-scale attention-aware network* (MA-Net) is proposed in this paper for optical flow estimation. Its novelty lies in that the attention is paid to coarse-scale and fine-scale image features in parallel via a multi-scale network architecture. In our MA-Net, channel attention and spatial attention are both exploited. For the former, inspired by the work of Woo et al [11] in image classification, a lightweight channel attention module is established and used. For the latter, the criss-cross attention module developed by Huang et al [8] for semantic segmentation is employed. Note that both of the two modules are of low computational complexity; thus the MA-Net also has rather low computational burden. On the other hand, it can deliver superior performance over existing methods.

The rest of the paper is organized as follows. Details of our proposed MA-Net for conducting optical flow estimation are described in Section 2. Extensive experimental results are documented and discussed in Section 3. Finally, a conclusion is drawn in Section 4.

2. THE PROPOSED METHOD

In this section, the channel and spatial attention modules are presented first, and then our MA-Net is described in detail, which is developed by incorporating these two modules into a multi-scale network architecture.

This work was supported in part by the Natural Science Foundation of the Jiangsu Higher Education Institutions of China under Grant 21KJA520007, in part by the National Natural Science Foundation of China under Grant 61572341, in part by the Priority Academic Program Development of Jiangsu Higher Education Institutions, in part by Collaborative Innovation Center of Novel Software Technology and Industrialization, and in part by the NTU-WASP Joint Project under Grant M4082184 and AcRF Tier 1 2018-T1-002-153.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

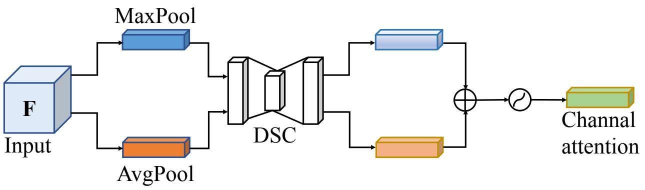

Fig. 1: Channel attention module, where ‘DSC’ denotes the multi-layer depthwise separable convolution.

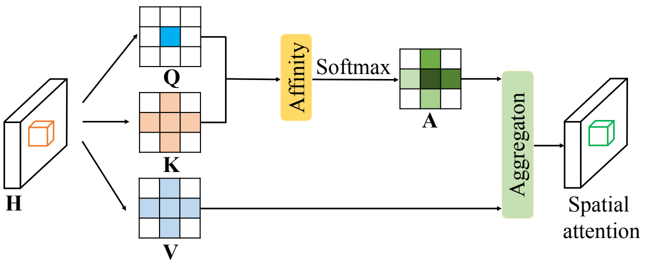

Fig. 2: The spatial attention module.

2.1. The channel attention module

Given an intermediate feature map $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ as input, the channel attention module will produce a channel attention map $\mathbf{M_c} \in \mathbb{R}^{C\times 1\times 1}$ as its output. Fig. 1 demonstrates the flowchart of this module. First, the spatial information of $\mathbf{F}$ is aggregated with average-pooling and max-pooling in parallel, yielding two spatial context descriptors, denoted as $\mathbf{F^c_{avg}}$ and $\mathbf{F^c_{max}}$, respectively. Both descriptors are then forwarded to a shared network block. In our work, for a computational complexity reduction, this block is composed of multi-layer deepwise separable convolutions (DSC) with a hidden layer, and the hidden activation size is set to $\mathbb{R}^{C/r\times 1\times 1}$, where r is the reduction ratio. Finally, the output feature vectors from the shared block are merged to produce the channel attention map $\mathbf{M_c}$ by using element-wise summation:

$$\begin{aligned}\mathbf{M_c} &= \sigma(\mathrm{DSC}(\mathrm{AvgPool}(\mathbf{F})) + \mathrm{DSC}(\mathrm{MaxPool}(\mathbf{F}))) \\ &= \sigma(\mathbf{W_1}(\mathbf{W_0}(\mathbf{F^c_{avg}})) + \mathbf{W_1}(\mathbf{W_0}(\mathbf{F^c_{max}})))\end{aligned} \tag{1}$$

where σ denotes the sigmoid function, $\mathbf{W_0} \in \mathbb{R}^{C/r\times C}$ and $\mathbf{W_1} \in \mathbb{R}^{C\times C/r}$ are the DSC weights.

2.2. The spatial attention module

Fig. 2 shows the structure of the spatial attention module. By taking a feature map $\mathbf{H} \in \mathbb{R}^{C\times H\times W}$ as input, the module firstly applies two convolutional layers with 1×1 filters on $\mathbf{H}$ to generate two higher-level feature maps $\mathbf{Q}$ and $\mathbf{K}$, which are both of the size $C' \times H \times W$, where C' is the number of channels and usually taken to be less than C for dimension reduction. Then, $\mathbf{Q}$ and $\mathbf{K}$ are used to produce an attention map $\mathbf{A} \in \mathbb{R}^{(H+W-1)\times H\times W}$ via the *affinity* operation [8].

At the same time, another convolutional layer with 1×1 filters is applied on $\mathbf{H}$ to generate a higher-level feature map $\mathbf{V} \in \mathbb{R}^{C\times H\times W}$ for feature adaption. At each position u in the spatial dimension of $\mathbf{V}$, a vector $\mathbf{V_u} \in \mathbb{R}^{C}$ and a set $\mathbf{\Phi_u} \in \mathbb{R}^{(H+W-1)\times C}$ can be obtained. The set $\mathbf{\Phi_u}$ is a collection of feature vectors in $\mathbf{V}$ that are in the same row or column as u. The contextual information is collected by the *aggregation* operation as follows:

$$\mathbf{M_s} = \sum_{i\in|\mathbf{\Phi_u}|} \mathbf{A_{i,u}}\mathbf{\Phi_{i,u}} \tag{2}$$

where $\mathbf{A_{i,u}}$ is a scalar value at channel i and position u in $\mathbf{A}$, and $\mathbf{M_s} \in \mathbb{R}^{C\times H\times W}$ is the yielded spatial attention map.

2.3. Multi-scale attention guided network

The RAFT [5] is followed to develop our MA-Net, with the intention to preserve image features at fine and coarse scales in the optical flow estimation process. Fig. 3 presents the architecture of the proposed MA-Net. First, four convolution layers are exploited to extract image features at increasingly high levels, which are with kernel sizes of 7×7, 5×5, 3×3 and 3×3, respectively. After that, a *multi-scale channel attention* block and a *multi-scale spatial attention* block are sequentially used, yielding attention-aware features as output. Then, the similarity of the feature matrix is extracted based on two sets of yielded features, with each set corresponding to one of the two input images.

At the same time, “context encoder” takes the first frame in the sequence as its input and produces a set of feature maps as output, which records the initial location of each object presented in this frame. Based on the set of feature maps, the location offsets of the follow-up frames are measured to estimate the optical flow of moving objects. Results of above-mentioned similarity calculation and context information are finally forwarded to an update module [5] to produce the optical flow estimation result through iterative optimization. In what follows, we will focus on describing the two multi-scale attention blocks, which are proposed and exploited in this work to improve the performance of optical flow estimation.

In the multi-scale channel attention block, we exploit two convolutional layers to extract features in parallel. These two layers are with kernel size of 5×5 and 3×3 for simulating a coarse and fine receptive field, respectively. In consequence, they are dedicated to extract coarse-scale and fine-scale image features, respectively. By setting the stride size to 1 and the padding to half of the size of the convolutional kernels, the output feature maps are of the same size as the input ones. Then, the channel attention module specified in Section 2.1 is imposed on each of the yielded feature maps. Finally, the obtained attention maps are combined with the feature maps at the coarse and fine scales, respectively:

$$\mathbf{F}' = \mathbf{M_c} \otimes \mathbf{F} \tag{3}$$

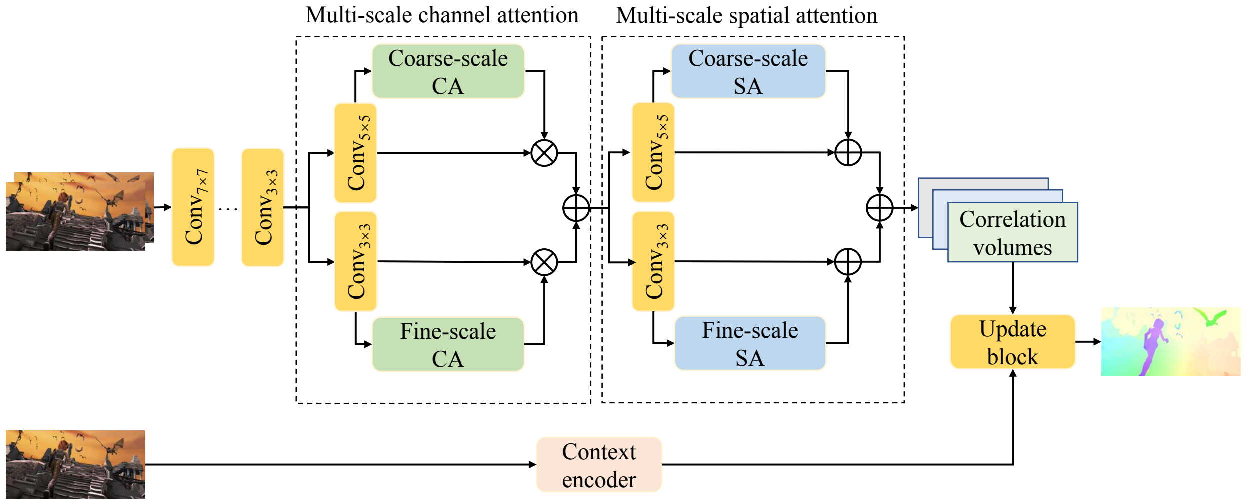

Fig. 3: Architecture of the MA-Net, where channel attention and spatial attention are denoted as 'CA' and 'SA', respectively.

Table 1: Performance comparison of different optical flow estimation methods on Sintel and KITTI2015.

Training Data	Method	Sintel (train)		Sintel (test)		KITTI15 (train)	
		Clean	Final	Clean	Final	F1-epe	F1-all
C+T	HD3 [12]	3.84	8.77	-	-	13.17	24
	LiteFlowNet [13]	2.48	4.04	-	-	10.39	28.5
	PWC-Net [4]	2.55	3.93	-	-	10.35	33.7
	VCN [14]	2.21	3.68	-	-	8.36	25.1
	FlowNet2 [3]	2.02	3.54	3.96	6.02	10.08	30
	RAFT [5]	1.43	2.71	-	-	5.04	17.4
	MA-Net (Ours)	**1.39**	**2.70**	-	-	**4.93**	**17.3**
C+T+S/K	FlowNet2 [3]	1.45	2.01	4.16	5.74	2.30	6.8
	HD3 [12]	1.87	1.17	4.79	4.67	1.31	4.1
	IRR-PWC [15]	1.92	2.51	3.84	4.58	1.63	5.3
	VCN [14]	1.88	3.20	2.93	4.57	1.76	4.43
	MaskFlowNet [16]	1.76	3.14	2.68	4.26	-	-
	RAFT [5]	0.77	1.20	2.08	3.41	**0.64**	1.5
	MA-Net (Ours)	**0.71**	**1.15**	**1.92**	**3.33**	0.65	**1.5**

where $\otimes$ denotes the element-wise multiplication, $\mathbf{M_c}$ and $\mathbf{F}$ are the input attention and feature maps, respectively, and $\mathbf{F}'$ is the yielded attention-aware feature map. It should be noted that during the operation of element-wise multiplication, the attention values are broadcasted along the spatial dimension. As a result, channel attention has been paid to coarse-scale and fine-scale features in parallel.

The multi-scale spatial attention block is developed with the same structure as the former block. After coarse-scale and fine-scale features are extracted by using two convolutional layers with different kernel sizes, the spatial attention module specified in Section 2.2 is imposed on them in parallel to produce attention-aware feature maps as follows:

$$\mathbf{H}' = \mathbf{M_s} + \mathbf{H} \tag{4}$$

where $\mathbf{H}$ and $\mathbf{M_s}$ are the input feature and spatial attention maps, respectively, and $\mathbf{H}'$ denotes the yielded attention-aware feature map. As a result, for both the coarse-scale and fine-scale branches, spatial attention has been added to enhance the extracted features and augment the pixel-wise representation.

It is worth mentioning that the two developed multi-scale blocks serve as the last two layers of the network, where the data are of the lowest resolution due to down-sampling. Therefore, they are of low computational complexity.

3. EXPERIMENTS

3.1. Training Details

The used datasets include the FlyingChairs, FlyingThings, Sintel and KITTI2015. In fact, the datasets we chosen contain

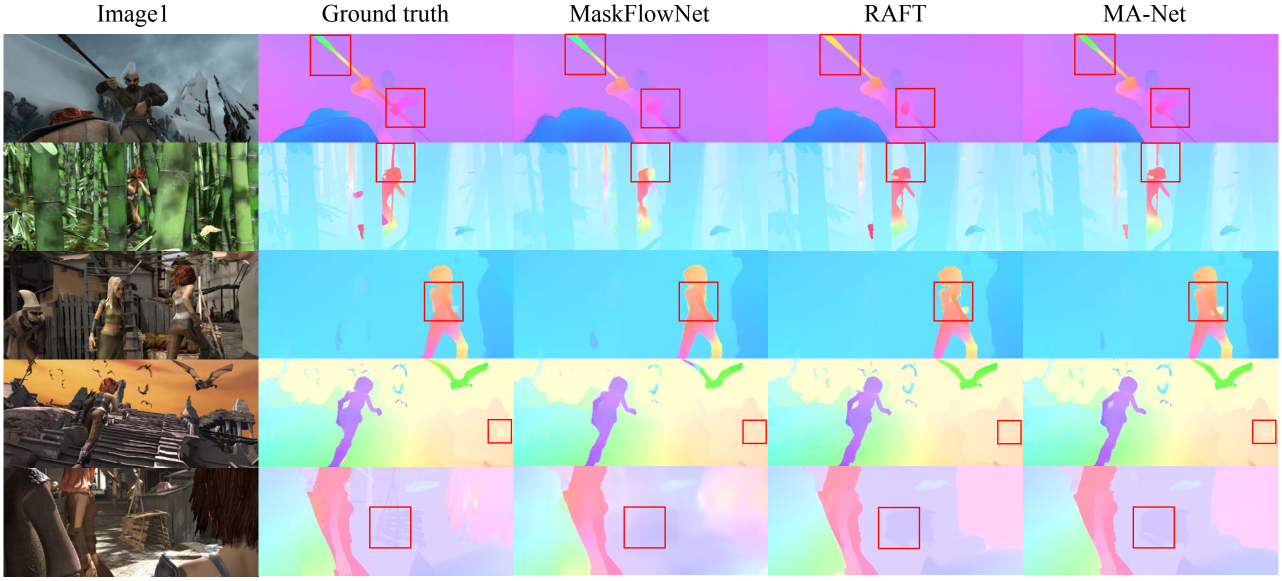

Fig. 4: Visual examples of predicted optical flow by using different methods on a set of images taken from the Sintel dataset.

a wealth of fast-moving small objects. Our proposed MA-Net is trained with the FlyingChairs and FlyingThings first, and then finetuned with the Sintel and KITTI2015. For network training on the FlyingChairs and FlyingThings, the adam optimizer is used, the initial learning rate is set to 0.000125, and the batch size is set to 6. Two 2080Ti GPUs were used in this network training stage.

3.2. Results

We compared our MA-Net with a number of state-of-the-art networks on the Sintel and KITTI2015 datasets. The End Point Error (EPE) is used as a measure of performance. The results are documented in Table. 1, where 'C+T' denotes the results trained with FlyingChairs (C) and FlyingThings (T) and tested on the training sets of Sintel and KITTI2015, and 'C+T+S/K' denotes the results trained with Sintel and KITTI2015 training sets after training on FlyingChairs (C) and FlyingThings (T), and tested with the Sintel and KITTI2015 test sets.

From Table. 1, it is seen that our MA-Net can produce the lowest EPE values (therefore, the best performance) among the compared methods in most of the comparison cases. Fig. 4 demonstrates a subjective comparison of different networks with a set of test images taken from the Sintel dataset. One can see that our method can capture tiny objects with higher accuracy when compared with others.

3.3. Ablation Study

To verify the contribution of each component of our MA-Net, an ablation study is conducted on the Sintel final (training) dataset, as shown in Table. 2. In the first to third study cases, the spatial attention module, the channel attention module and the DSC block in the channel attention module are in turns removed, respectively. In the fourth case, the multi-scale network architecture is not exploited; that is, attention is paid to fine-scale features only, as practiced in existing networks. In the fifth case, all components are used, and we arrive at the MA-Net. The performance of MA-Net is a little bit lower than the third case, where the DSC block is not used. On the other hand, however, a clear reduction of computational complexity is achieved in the MA-Net.

Table 2: Ablation study

Channel attention	Spatial attention	DSC block	Used scales (coarse, fine)	Sintel (EPE)
Yes	No	Yes	Coarse and fine	1.156
No	Yes	Yes	Coarse and fine	1.166
Yes	Yes	No	Coarse and fine	1.149
Yes	Yes	Yes	Fine scale only	1.172
Yes	Yes	Yes	Coarse and fine	1.150

4. CONCLUSION

A multi-scale attention-aware network has been proposed for optical flow estimation, in which attention-aware fine-scale and coarse-scale image features are extracted in parallel and jointly used to produce the optical flow estimation results. Experimental results obtained on benchmark datasets have clearly shown that the proposed MA-Net is able to capture fast-moving small objects with high accuracy and thus deliver superior performance over a number of existing methods.

5. REFERENCES

[1] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox, "FlowNet: Learning optical flow with convolutional networks," in *International Conference on Computer Vision (ICCV)*, 2015, pp. 2758–2766.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241.

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

[4] D. Sun, X. Yang, M. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8934–8943.

[5] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 402–419.

[6] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[7] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3146–3154.

[8] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-Cross attention for semantic segmentation," in *International Conference on Computer Vision (ICCV)*, 2019, pp. 603–612.

[9] M. Zhai, X. Xiang, R. Zhang, N. Lv, and A. El-Saddik, "AD-Net: Attention guided network for optical flow estimation using dilated convolution," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019, pp. 2207–2211.

[10] X. Xiang, M. Zhai, R. Zhang, N. Lv, and A. El-Saddik, "Optical flow estimation using spatial-channel combinational attention-based pyramid networks," in *IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 1272–1276.

[11] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[12] Z. Yin, T. Darrell, and F. Yu, "Hierarchical discrete distribution decomposition for match density estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6044–6053.

[13] T. Hui, X. Tang, and C. C. Loy, "LiteFlowNet: A lightweight convolutional neural network for optical flow estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8981–8989.

[14] G. Yang and D. Ramanan, "Volumetric correspondence networks for optical flow," in *Advances in Neural Information Processing Systems*, 2019, pp. 793–803.

[15] J. Hur and S. Roth, "Iterative residual refinement for joint optical flow and occlusion estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5754–5763.

[16] S. Zhao, Y. Sheng, Y. Dong, E. I. Chang, and Y. Xu, "MaskFlowNet: Asymmetric feature matching with learnable occlusion mask," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6277–6286.