MSCFNet: A Lightweight Network With Multi-Scale Context Fusion for Real-Time Semantic Segmentation

Guangwei Gao, *Member, IEEE*, Guoan Xu, Yi Yu, *Member, IEEE*, Jin Xie, *Member, IEEE*, Jian Yang, *Member, IEEE*, and Dong Yue, *Fellow, IEEE*

***Abstract*— In recent years, how to strike a good trade-off between accuracy, inference speed, and model size has become the core issue for real-time semantic segmentation applications, which plays a vital role in real-world scenarios such as autonomous driving systems and drones. In this study, we devise a novel lightweight network using a multi-scale context fusion (MSCFNet) scheme, which explores an asymmetric encoder-decoder architecture to alleviate these problems. More specifically, the encoder adopts some developed efficient asymmetric residual (EAR) modules, which are composed of factorization depth-wise convolution and dilation convolution. Meanwhile, instead of complicated computation, simple deconvolution is applied in the decoder to further reduce the amount of parameters while still maintaining the high segmentation accuracy. Also, MSCFNet has branches with efficient attention modules from different stages of the network to well capture multi-scale contextual information. Then we combine them before the final classification to enhance the expression of the features and improve the segmentation efficiency. Comprehensive experiments on challenging datasets have demonstrated that the proposed MSCFNet, which contains only 1.15M parameters, achieves 71.9% Mean IoU on the Cityscapes testing dataset and can run at over 50 FPS on a single Titan XP GPU configuration.**

***Index Terms*— Real-time semantic segmentation, lightweight network, encoder–decoder architecture, context fusion.**

I. Introduction

Autonomous driving technology has been widely studied to enhance the driving experience and relieve traffic pressure [1], [2]. With the help of cameras, it becomes easier to perceive and understand the surrounding environment [3]–[5]. Semantic segmentation, which aims at assigning a category to each pixel for the given image, is a challenging research topic in the field of computer vision. Recently, deep convolutional neural networks (DCNNs) [6]–[8] have shown their impressive capabilities on image classification with high resolution. Especially the fully convolutional network (FCN) [9], which is a pioneer CNN for semantic segmentation task. The encoder-decoder network has also become a popular structure for solving the segmentation problem. Although achieving remarkable results, most of the previous networks [10]–[14] ignored the segmentation efficiency, namely, their calculation and storage requirements are so high that it is difficult to meet the demands of real-world applications where information needs to interact quickly with the environment. Meanwhile, electric equipments, such as robotics, cellphones, and telemedicine, etc., having small memory capacity and limited computational cost, cannot support the enormous complex algorithms.

Therefore, it is a primary trend to design lightweight and efficient networks to overcome the above problems. The smaller-scale network means a faster inference speed and less redundancy [15]. However, most of the existing real-time research works [16]–[19] mainly focus on shallowing the networks and reducing parameters to shorten the time-consuming at the expense of model accuracy.

In this work, we devise a novel lightweight and efficient network, called multi-scale context fusion network (MSCFNet), to get a better balance between the accuracy and efficiency for real-time semantic segmentation task. Like most of the previous works, our proposed model also explores an asymmetric encoder-decoder structure. As presented in Fig. 1, our MSCFNet has multiple branches with efficient attention

Manuscript received 25 December 2020; revised 4 March 2021, 28 April 2021, and 16 June 2021; accepted 15 July 2021. Date of publication 2 August 2021; date of current version 5 December 2022. This work was supported in part by the National Key Research and Development Program of China under Project 2018AAA0100102 and Project 2018AAA0100100; in part by the National Natural Science Foundation of China under Grant 61972212, Grant 61772568, and Grant 61833011; in part the Natural Science Foundation of Jiangsu Province under Grant BK20190089; in part by the Six Talent Peaks Project, Jiangsu Province, under Grant RJFW-011; and in part by the Open Fund Project of Provincial Key Laboratory for Computer Information Processing Technology (Soochow University) under Grant KJS1840. The Associate Editor for this article was H. Lu. *(Corresponding authors: Guangwei Gao; Guoan Xu.)*

Guangwei Gao is with the Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, Nanjing 210023, China, also with the Digital Content and Media Sciences Research Division, National Institute of Informatics, Tokyo 101-8430, Japan, and also with the Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou 215006, China (e-mail: csggao@gmail.com).

Guoan Xu is with the College of Automation and College of Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing 210023, China (e-mail: xga_njupt@163.com).

Yi Yu is with the Digital Content and Media Sciences Research Division, National Institute of Informatics, Tokyo 101-8430, Japan (e-mail: yiyu@nii.ac.jp).

Jin Xie and Jian Yang are with the School of Computer Science and Technology, Nanjing University of Science and Technology, Suzhou 210094, China (e-mail: csjxie@njust.edu.cn; csjyang@njust.edu.cn).

Dong Yue is with the College of Automation and College of Artificial Intelligence, Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, Nanjing 210023, China (e-mail: medongy@vip.163.com).

Digital Object Identifier 10.1109/TITS.2021.3098355

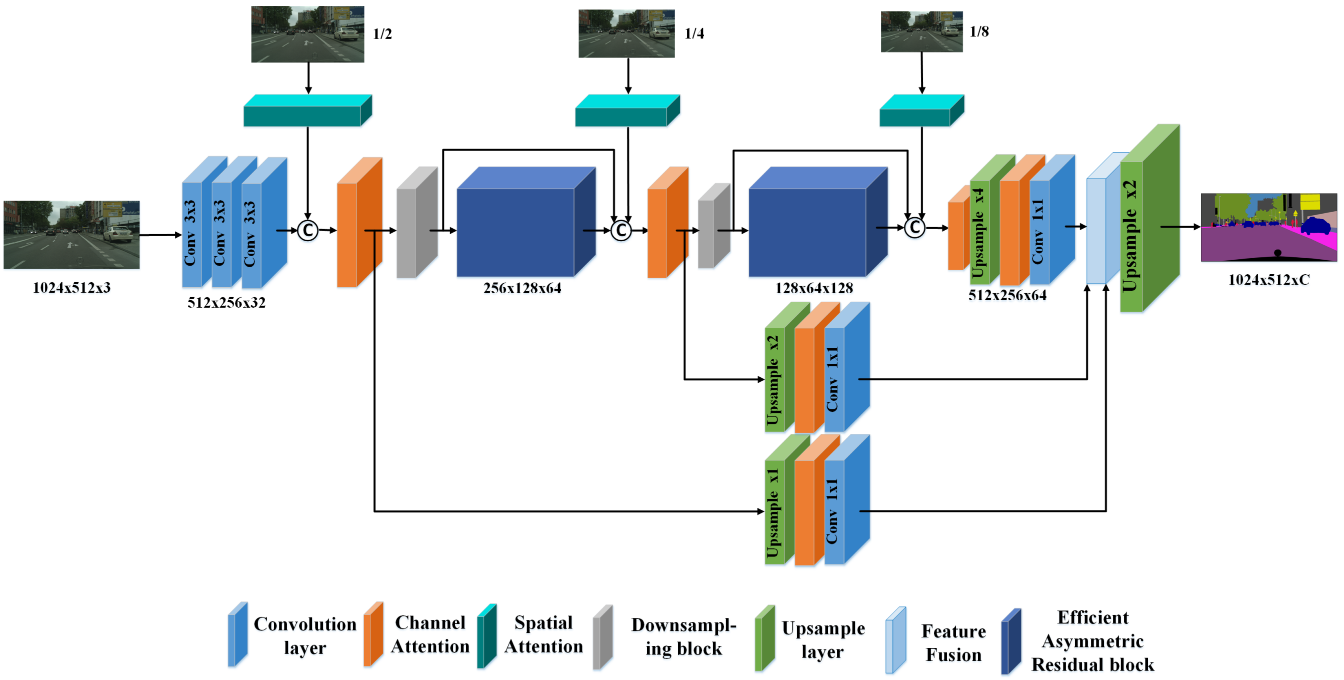

Fig. 1. The procedure of our proposed MSCFNet. The sizes (width, height, and channel) of the intermediate features are given in the process of network. "C" denotes the concatenation, "×1, ×2 and ×4" mean the upsampling factor, "1/2, 1/4 and 1/8" indicate the ratio of the original image scale. (Best viewed in color).

mechanisms from different stages of the network, containing multi-scale contextual information for segmentation purpose. Although a small number of parameters and calculations have been added, the general execution improves a lot. The core unit of our MSCFNet is an efficient asymmetric residual (EAR) module with dilated factorized depth-wise separable convolution, which allows us to extract attentive and cooperative feature information on a large receptive field efficiently and quickly.

Our main contributions can be listed as three-fold:

- We devise an efficient asymmetric residual (EAR) module and construct a lightweight semantic segmentation network with a multi-scale context fusion scheme, which fuses the attentive features adaptively, contributing to the efficiency and effectiveness of the segmentation task.
- Short-range and long-range connections with efficient spatial and channel attention presented in our method facilitate the local and contextual information interaction greatly, contributing to the improvement of the performance.
- Our network achieves prominent performance on both Cityscapes and CamVid datasets without any other data augment skills. It has 1.15M model size, while achieves a mean intersection over union mIoU) of 71.9% and 69.3% on Cityscapes and CamVid datasets, respectively.

II. Related Work

A. Factorization Convolution

Factorization convolution is often used to improve the efficiency where a traditional two-dimensional convolution is replaced by two one-dimensional convolutions. Xception [20] and MobileNet [21] applied depth-wise separable convolution, where each input channel and each filter kernel is divided into a group, which operates individually. ERFNet [18], DABNet [22], and LEDNet [23] decomposed a 3 × 3 convolution into a 3 × 1 and a 1 × 3 convolution. They are all beneficial from the factorization convolution, which can reduce the amount of computational burdens.

B. Dilation Convolution

Dilation convolution is used to insert zeros between two adjacent kernel values of the standard convolution to achieve the purpose of enlarging the receptive field without adding the parameters. For example, DeepLab series [11], [24], [25] suggested a spatial pyramid pooling module that adopts various dilation rates arranged as a pyramid. LEDNet [23] designed a split-shuffle-non-bottleneck (SS-nbt) module using the dilation convolution to construct an asymmetric encoder-decoder architecture. Dilation8 [26] proposed a multi-scale context aggregation network by dilated convolutions. EDANet [27] incorporated dilated convolution and dense connection to attain high efficiency.

C. Lightweight Segmentation Networks

Lightweight segmentation networks are eagerly required to attain the desired balance between the prediction accuracy and the related inference efficiency [28]–[35]. ENet [16] was the first lightweight architecture used in real-time applications, which trimmed the amount of the convolution filters to decrease the calculation. ESPNet [36] proposed an effective spatial pyramid module, which can collect multi-scale contextual information. ICNet [37] used a strategy called image

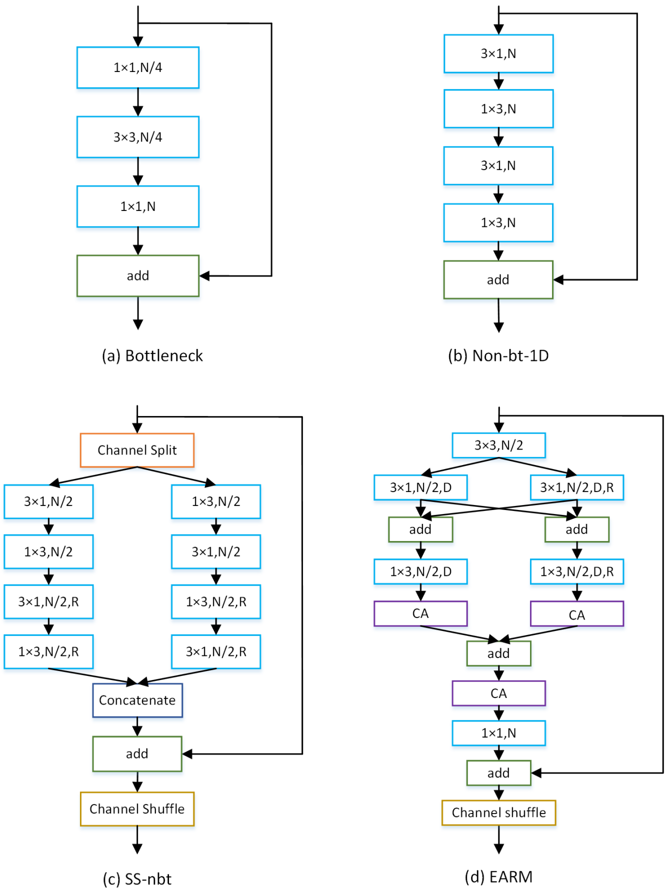

Fig. 2. Comparisons of various types of residual modules. "N" is the number of the output channels, "R" represents the dilation rate of kernel and "D" indicates a depth-wise convolution. (a) Bottleneck [16]. (b) Non-bt-1D of ERFNet [18]. (c) SS-nbt of LEDNet [23]. (d) Our EAR module.

cascade to improve the segmentation efficiency. BiseNet [38] introduced two branches, one is to retain shallow spatial information and the other is to extract deep contextual information. LEDNet [23] showed the benefit brought by the channel split and channel shuffle operations. Although these networks have made a relatively satisfactory trade-off between performance and speed, there is still adequate room for further promotion.

D. Attention Mechanism

Attention mechanism has been broadly adopted in the field of pattern recognition and computer vision. Its essence is to imitate the human visual mechanism to learn a weight distribution of the image features and apply these weights to the original features. CCNet [39] devised an efficient criss-cross attention module to capture the image dependencies. GCNet [40] and ANN [41] further observed the non-local attention mechanism and achieved promising performance for the semantic segmentation task. DANet [13] used the channel and spatial attention tricks simultaneously to model the semantic inter-dependencies. Some of the above works performed sophisticated matrix multiplication on the pixel level, which is not suitable for lightweight applications.

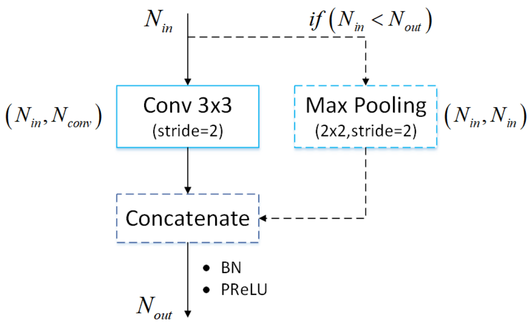

Fig. 3. Downsampling block structure. N_{in}: input channel, N_{out}: output channel, N_{conv}: output channel after convolution, BN: Batch Normalization.

SENet [42], which is a lightweight threshold mechanism, has been widely used to model the correlation of all channels. It first employs a global average pooling to squeeze the global spatial information into channel descriptors and then uses two fully connected layers to capture cross-channel interaction. GENet [43] introduced a pair of operators, consisting of gathering feature responses from a large scale and exciting this information to local features. CBAM [44] sequentially inferred attention maps along spatial dimension and channel dimension separately, and then the input feature maps are multiplied to the attention maps for adaptive feature refinement. GSoP-Net [45] introduced higher-order representation across from lower to higher layers to effectively explore those statistical information. By dissecting the mechanism in SENet, ECANet [46] proposed an effective yet efficient cross-channel interaction scheme avoiding channel dimensionality reduction. It performs well on the tasks of object detection and image classification in terms of parameters and computations.

In contrast to the above approaches, in our method, we design a lightweight semantic segmentation network with a contextual fusion structure to speed up the network training, reduce the model size, and meanwhile ensure the effectiveness of the final inference results. Regarding the popular attention mechanism, we inject spatial and channel attention based on the different conditions of the network at different stages. Simultaneous consideration of these schemes boosts the final segmentation accuracy.

III. Methodology

In this section, we first illustrate our efficient asymmetric residual (EAR) module and then introduce the efficient attention and context fusion modules. Finally, we elaborate the whole network architecture which consists of an initial block, three input injection modules, two downsampling blocks, two EAR blocks, and two context branches. The entire structure of the proposed MSCFNet is given in Fig. 1.

A. EAR Module

Lightweight networks have witnessed a lot of residual designs (See Fig. 2). Inspired by these designs, we devise the efficient asymmetric residual (EAR) module with their common advantages to achieve a better result under the circumstance of limited computational capacity. Our EAR module is shown in Fig. 2 (d). Firstly, the number of input channels is reduced to half by a 3 × 3 convolution at the bottleneck. The reason why we use a 3×3 convolution instead of a 1×1 convolution which has fewer parameters is that when using 1 × 1 convolution, the residual block must construct deeper for a larger receptive field capturing more contextual information, the computational cost and memory requirements must increase. The following is a two-branch structure. One branch applies factorization convolution to depth-wise convolution so that it can collect local and short-range feature information. Specifically, a standard 3 × 3 depth-wise convolution is divided into a 3 × 1 convolution and a 1 × 3 convolution. They would have the same size of the receptive field, while the latter has a fewer number of parameters. Another branch adopts dilation convolution enlarging receptive field to the factorization depth-wise convolution to capture complex and long-range feature information. To avoid the gridding artifacts, we use different dilation rates in different EAR modules, which are not integer powers of 2.

For the sake of sharing information for different branches, we put the feature interaction operations between 3 × 1 and a 1 × 3 convolution in the two branches. In such a way, the contextual information extracted by the two branches can complement each other. The feature maps from each branch are then sent to the channel attention module for better extracting discriminative features. And then, the two low-dimension branches are fused and fed into the channel attention module for the same purpose. Following is a 1 × 1 point-wise convolution to recover the related channels of the feature maps. Finally, For the sake of evading the drawback of information independence between channels caused by depth-wise convolution, we explore a channel shuffle followed the combination of the output of 1 × 1 point-wise convolution and the input to facilitate the channel information exchanging and sharing. The above operations can be expressed as follow:

$$x_b = C_{3\times3}\left(\rho\left(x_{EARin}\right)\right), \tag{1}$$

$$y_1 = CA\left(C_{1\times3}\left(C_{3\times1}\left(x_b\right) + C_{3\times1,d}\left(x_b\right)\right)\right), \tag{2}$$

$$y_2 = CA\left(C_{1\times3,d}\left(C_{3\times1,d}\left(x_b\right) + C_{3\times1}\left(x_b\right)\right)\right), \tag{3}$$

$$y_{EARout} = S\left(C_{1\times1}\left(CA\left(\rho\left(y_1 + y_2\right)\right)\right) + x_{EARin}\right), \tag{4}$$

where x_{EARin} and y_{EARout} represent the input and output of the EAR module, x_b is the output of 3×3 convolution, y_1 and y_2 are the outputs of two branches in EAR module, $C_{m\times n}$ denotes convolution operation with the kernel size $m\times n$, d is the dilation rate, ρ is the PReLU nonlinear activation function, S means channel shuffle operation.

B. Efficient Attention

Generally, owing to the small number of network layers, a lightweight network can hardly extract deep enough features thoroughly like a large network. Consequently, producing representative features and combining them may be an essential manner to enhance the segmentation performance. To this end, we borrow the idea of the attention mechanism in our model. Attention is beneficial to both information integration and object feature emphasizing. Our efficient attention mechanisms include both spatial attention and channel attention modules.

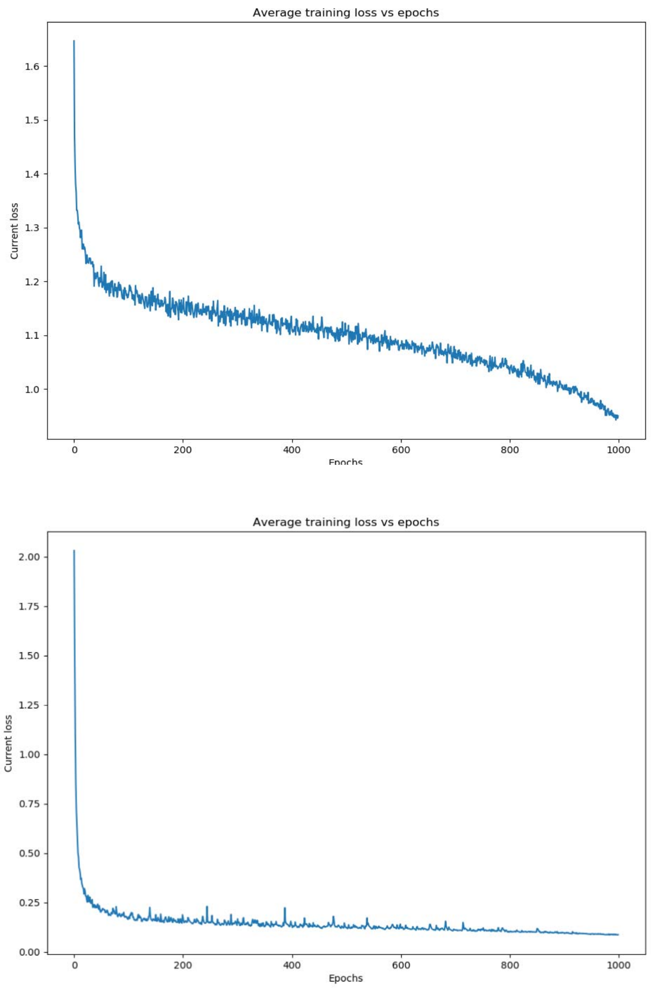

Fig. 4. The convergence curve of our MSCFNet on Cityscapes (top) and CamVid (bottom) datasets. (Best viewed in color).

1) Spatial Attention: The spatial attention we used is motivated by CBAM [44], exploring the inter-spatial relationship of the input features to generate attention maps that depict where to highlight or suppress. The average pooling and max pooling operations are first adopted, and then concatenating them to formulate a feature descriptor followed by a convolution layer to produce the desired spatial weight maps. Finally, we multiply the attention maps and the input features of the module to obtain the final generated features. This procedure can be formulated as follows:

$$SA\left(F\right) = \sigma\left(f^{7\times7}\left(Concat\left[AvgP\left(F\right), MaxP\left(F\right)\right]\right)\right) \times F, \tag{5}$$

where $F \in R^{C\times H\times W}$ denotes the input features, σ is the sigmoid activation function, $f^{7\times7}$ denotes a standard convolution with the kernel size 7×7, $Concat$ means the concatenate operation, $AvgP$ and $MaxP$ represent the average pooling and max pooling operation, respectively.

The spatial attentions are placed between the three injection modules and the main branch (see Fig. 1). The outputs of spatial attention are calculated in the following:

$$F_{sa}^{n} = SA\left(\frac{1}{\beta} Input\right), \quad \beta = 2^n, \quad n = 1, 2, 3, \tag{6}$$

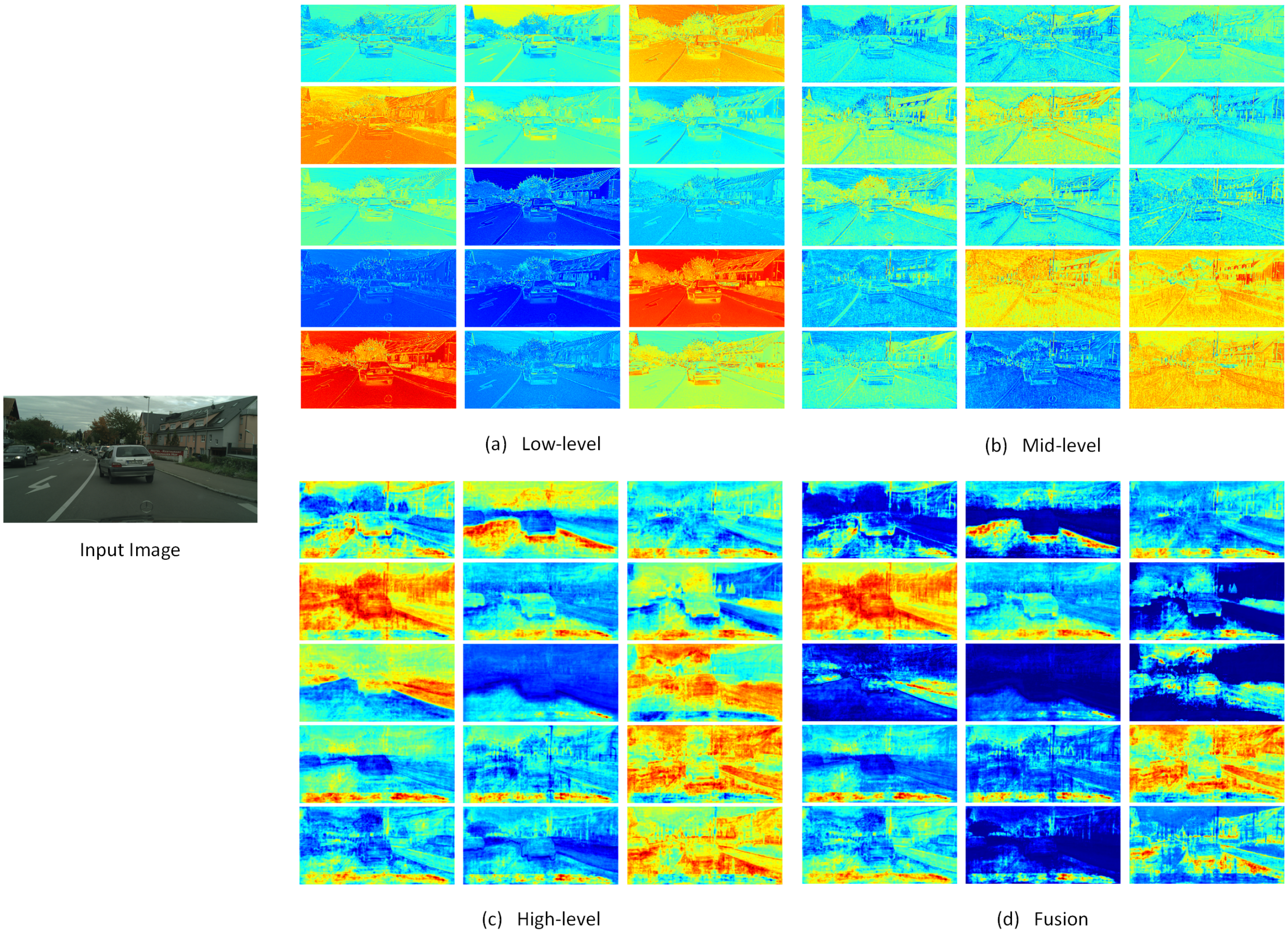

Fig. 5. Feature maps with different levels on the Cityscapes validation set.

where $Input$ is the observed image, β denotes the factor of the reduction and SA means the spatial attention operation.

2) Channel Attention: The channel attention we adapted is derived from ECANet [46], which just occupies a little computational resource but improves the performance greatly by comparison. We settle three channel attention modules followed by each feature combination operation as well as one before context fusion in the main branch, and two in the middle of the context branches. They can be simply separated into two categories, one is the attention operation in the encoder process, and the other is the attention operation in the decoder process (see Fig. 1). The procedure of the above channel attention can be formulated as follows:

$$CA\left(F\right) = \sigma\left(f^{k\times k}\left(T\left(AvgP\left(F\right)\right)\right)\right) \times F, \tag{7}$$

where T represents the compression, transposition and extension operations of the tensor dimensions, $f^{k\times k}$ denotes a standard convolution with adaptive selection of kernel size k.

The attentive features of the encoding process can be calculated as follow:

$$F^n_{ca} = CA\left(Concat\left(O_n, F^n_{sa}\right)\right), \quad n = 1, 2, 3, \tag{8}$$

where F^n_{sa} and F^n_{ca} represent the outputs of spatial attention and channel attention respectively, O_1 is the output of the initial block, O_2 and O_3 are the outputs of the two EAR blocks, CA denotes the channel attention operation.

And the attentive features in the decoding process can be described as follow:

$$F_{Bri} = CA\left(Up\left(I_i, \alpha\right)\right), \quad \alpha = 2^{i-1}, \quad i = 1, 2, 3, \tag{9}$$

where F_{Br1}, F_{Br2} and F_{Br3} denote the output features of the main branch and the other two projection branches, Up is the bilinear interpolate operation in the corresponding branches, I_1, I_2, I_3 indicate the low-level, intermediate-level, high-level contextual information respectively, and α is the magnification factor.

In summary, spatial attention focuses on indicating "where" to highlight while channel attention focuses on indicating "what" a given feature is. By considering these two attention mechanisms simultaneously at different stages in the network, our method can adaptively promote the representational power of the extracted features and facilitate the local and contextual information interaction greatly, which have been validated to be effective in the experimental section.

TABLE I

ABLATION STUDY RESULTS ON CAMVID TESTING SET. CF: CONTEXT FUSION, FC: FEATURE CONCATENATION, FA: FEATURE ADDING, CA: CHANNEL ATTENTION, SA: SPATIAL ATTENTION; R: DILATION RATE. SUPERSCRIPT'†' DENOTES THE FINAL VERSION

	CF		CA				
Models	FC	FA	SE	ECA	SA	mIoU (%)	Param
(a) Context Fusion							
MSCFNet	✗	✗	✗	✗	✗	67.82	1.1429M
MSCFNet	✓	✗	✗	✗	✗	68.42	1.1533M
MSCFNet	✗	✓	✗	✗	✗	68.87	1.1483M
(b) Attention Module							
MSCFNet	✗	✓	✓	✗	✗	68.96	1.2696M
MSCFNet	✗	✓	✗	✓	✗	69.16	1.1483M
MSCFNet	✗	✓	✗	✓	✓	69.30	1.1486M
(c) Dilation Rate							
MSCFNet (R=1,1,2,1,2,5,7,9,17)	✗	✓	✗	✓	✓	68.27	0.7684M
MSCFNet (R=1,1,2,2,5,1,1,2,2,4,4,8,8,16,16)	✗	✓	✗	✓	✓	69.04	1.1486M
MSCFNet (R=1,1,2,2,5,1,2,5,7,9,2,5,7,9,17)†	✗	✓	✗	✓	✓	69.30	1.1486M

TABLE II

ABLATION STUDY RESULTS BY GRADUALLY ADDING INTERMEDIATE CONTEXTUAL FEATURES ON THE CAMVID TESTING SET

Models	mIoU (%)	Param
MSCFNet-High_level	67.91	1.1432M
MSCFNet-High_level + Mid_level	68.49	1.1474M
MSCFNet-High_level + Mid_level + Low_level	69.30	1.1486M

TABLE III

EVALUATION RESULTS ON THE CITYSCAPES TESTING SET

Methods	Pretrained	Input Size	mIoU (%)↑	Speed (FPS)↑	Param↓
SegNet [17]	ImageNet	640 × 360	56.1	17	29.50M
ENet [16]	No	512 × 1024	58.3	77	**0.36M**
FSSNet [33]	-	512 × 1024	58.8	51	-
ESPNet [36]	No	512 × 1024	60.3	113	**0.36M**
CGNet [47]	No	1024 × 2048	64.8	50	0.50M
NDNet [30]	No	1024 × 2048	65.3	40	0.50M
ContextNet [48]	No	1024 × 2048	66.1	18	0.85M
EDANet [27]	No	512 × 1024	67.3	81	0.68M
ERFNet [18]	No	512 × 1024	68.0	42	2.10M
Fast-SCNN [49]	-	1024 × 2048	68.0	124	1.11M
BiseNet [38]	ImageNet	768 × 1536	68.4	106	5.80M
ICNet [37]	ImageNet	1024 × 2048	69.5	30	26.50M
DABNet [22]	No	1024 × 2048	70.1	28	0.80M
FarSeeNet [28]	-	512 × 1024	70.2	69	-
DFANet [35]	ImageNet	512 × 1024	70.3	**160**	7.80M
LEDNet [23]	No	512 × 1024	70.6	71	0.90M
EdgeNet [32]	-	512 × 1024	71.0	31	-
MSCFNet (ours)	No	512 × 1024	**71.9**	50	1.15M

C. Context Fusion

Previous methods usually learnt finer-scale predictions in a stage-by-stage manner. That means, the net in each stage is trained by the initialization of the previous stage net, which may result in the contextual cues cannot complement each other. Meanwhile, some remarkable networks have shown that good performance usually stems from a fusion of hierarchical information. Hence, we adopt this idea in our method before the final classification to integrate the multi-scale contextual information. The features F_{cf} from the context fusion operation can be formulated as follow:

$$F_{cf} = \rho\left(F_{Br1} + F_{Br2} + F_{Br3}\right). \tag{10}$$

Therefore, it can be seen that the context fusion module combining attentive contextual information from different stages of the network, which can alleviate the limitations caused by the spatial statistics of pixels loss.

D. Network Architecture

In this subsection, we introduce the entire lightweight network as presented in Fig. 1. MSCFNet has an asymmetric structure with an encoder structure and the related decoder structure, finally followed by a widely used classification layer.

1) Encoder: In the main flow, the initial feature extraction module includes three 3×3 convolutions, in which the first one uses stride 2 to extract feature information and reduce the size simultaneously. Then the downsampling block we employed has two alternative outputs of a 3×3 convolution with stride 2 and a 2 × 2 max-pooling with stride 2. If the amount of the input channels is larger than or equal to that of the output channels, the block is just the single 3 × 3 convolution. Otherwise, the max-pooling operation is added, the concatenation of these two branches forms the final downsampling outputs. Please see Fig. 3 for more details [27].

The downsampling operation amplifies the receptive field for collecting more contextual information. Nevertheless, the reduction in terms of the resolution of the input feature maps often leads to spatial and boundary resolution loss. By taking these into account, we just perform downsampling three times progressively and obtain 1/8 resolution of the original feature map to gather deeper context but maintain more image details. Followed each downsampling block is the EAR block, which includes different numbers of consecutive EAR modules. The first block has 5 EAR modules, while the second consists of 10 EAR modules for dense feature extraction. To better promote feature propagation and contextual information relationship, we apply inter-block concatenation to combine high-level and low-level features. Also, we employ dilation convolution in the blocks as depicted

TABLE IV

PER-CLASS IOU(%) PERFORMANCE ON THE CITYSCAPES TESTING SET. LIST OF CATEGORIES: ROAD, SKY, CAR, VEGETATION, BUILDING, SIDE-WALK, PEDESTRIAN, BUS, TRAFFIC SIGN, BICYCLE, TERRAIN, TRAFFIC LIGHT, RIDER, POLE, TRAIN, MOTORCYCLE, WALL, FENCE AND TRUCK. 'CLA': 19 CLASSES

Methods	Roa	Sky	Car	Veg	Bui	Sid	Ped	Bus	TSi	Bic	Ter	TLi	Rid	Pol	Tra	Mot	Wal	Fen	Tru	Cla
SegNet [17]	96.4	91.8	89.3	87.0	84.0	73.2	62.8	43.1	45.1	51.9	63.8	39.8	42.8	35.7	44.1	35.8	28.4	29.0	38.1	57.0
ENet [16]	96.3	90.6	90.6	88.6	75.0	74.2	65.5	50.5	44.0	55.4	61.4	34.1	38.4	43.4	48.1	38.8	32.2	33.2	36.9	58.3
ESPNet [36]	97.0	92.6	92.3	90.8	76.2	77.5	67.0	52.5	46.3	57.2	63.2	35.6	40.9	45.0	50.1	41.8	35.0	36.1	38.1	60.3
CGNet [47]	95.5	92.9	90.2	89.6	88.1	78.7	74.9	59.5	63.9	60.2	67.6	59.8	54.9	54.1	25.2	47.3	40.0	43.0	44.1	64.8
ERFNet [18]	97.7	94.2	92.8	91.4	89.8	81.0	76.8	60.1	65.3	61.7	68.2	59.8	57.1	56.3	51.8	47.3	42.5	48.0	50.8	68.0
ICNet [37]	97.1	93.5	92.6	91.5	89.7	79.2	74.6	**72.7**	63.4	70.5	68.3	60.4	56.1	61.5	51.3	53.6	43.2	48.9	51.3	69.5
DABNet [22]	97.9	92.8	93.7	91.8	90.6	82.0	78.1	63.7	67.7	66.8	70.1	63.5	57.8	59.3	56.0	51.3	45.5	50.1	52.8	70.1
LEDNet [23]	**98.1**	**94.9**	90.9	**92.6**	**91.6**	79.5	76.2	64.0	**72.8**	**71.6**	61.2	61.3	53.7	**62.8**	**52.7**	44.4	47.7	49.9	**64.4**	70.6
EdgeNet [32]	**98.1**	**94.9**	**94.3**	92.4	**91.6**	**83.1**	80.4	60.9	71.4	67.7	69.7	**67.2**	61.1	62.6	52.5	55.3	45.4	50.6	50.0	71.0
MSCFNet (ours)	97.7	94.3	94.1	92.3	91.0	82.8	**82.7**	66.1	71.4	70.2	**70.2**	67.1	**62.7**	61.2	51.9	**57.6**	**49.0**	**52.5**	50.9	**71.9**

TABLE V

COMPARISON RESULTS ON THE CAMVID TESTING SET

Methods	Pretrained	mIoU (%)↑	Param↓
ENet [16]	No	51.3	**0.36M**
SegNet [17]	ImageNet	55.6	29.50M
FCN-8s [50]	ImageNet	57.0	134.50M
NDNet [30]	No	57.2	0.50M
DeepLabLFOV [51]	ImageNet	61.6	37.30M
DFANet [35]	ImageNet	64.7	7.80M
Dilation8 [26]	ImageNet	65.3	140.80M
CGNet [47]	No	65.6	0.50M
BiseNet [38]	ImageNet	65.6	5.80M
DABNet [22]	No	66.4	0.76M
ICNet [37]	ImageNet	67.1	26.50M
MSCFNet (ours)	No	**69.3**	1.15M

in Section II-B. The dilation rates in the first block are $\{1, 1, 2, 2, 5\}$, and the second are $\{1, 2, 5, 7, 9, 2, 5, 7, 9, 17\}$, respectively. We choose this scheme to mitigate the gridding artifacts and enlarge the receptive field for more context of larger scope.

For better feature reuse, we insert efficient spatial attention in the shortcut connections between the input features, which is handled by three input injection modules with $1/2$, $1/4$, $1/8$ ratios, and two downsampling blocks, as well as the last convolution layer respectively. Moreover, to integrate contextual semantic information and allocate channel information resources of the feature maps, we use effective channel attention modules followed the above three identical places of the feature concatenations. What's more, we apply long-range shortcut connections in our network. The two context branches are composed of efficient channel attentions covering contextual information from different scales. Also, this operation alleviates the contradiction that shallow features lack semantic information and deep features lack boundary and detailed information.

2) Decoder: Our Decoder is asymmetrical relative to the Encoder. It entirely uses a ×4 upsampling and a final ×2 deconvolution operations to restore the original input image size and output the final segmentation prediction simultaneously. Compared to some existing networks directly upsampling 8 times or others decoding step by step, our two-step decoder structure combines the superiority of the two, which not only ensures the simplicity of the calculation, but also guarantees the maximum recovery of the decoded information.

IV. EXPERIMENTS

A. Implementation Protocol

1) Datasets: We use two popular benchmarks of urban street scenes – Cityscapes [52] and CamVid to assess the effectiveness of our proposed network. The Cityscapes dataset contains 19 semantic categories including 5000 fine-annotated samples with the size of 2048×1024 that are separated into three subsets: 2975 samples used for training, 500 samples used for validation, and the other 1525 samples used for testing. The CamVid is another smaller dataset for self-driving scenarios. It has 11 semantic categories with 367 training, 101 validation, and 233 testing samples, of which the image size is 960×720. For Cityscapes and CamVid, the original input is resized to 1024×512 and 480×360 for network training, respectively.

2) Parameter Settings: For the Cityscapes dataset, we train our network end-to-end by exploiting the stochastic gradient descent (SGD) [53] method with a batch size of 4 to leverage the hardware memory. The momentum is set as 0.9 and also the related weight decay is set as 1×10^{-4}. The "poly" policy for learning rate is also adopted, where the learning rate is adaptively adjusted according to the following equation after every iteration:

$$lr = lr_{base} \times \left(1 - \frac{iteration}{max\_iteration}\right)^{power}, \tag{11}$$

where lr_{base} is the initial learning rate, lr is the learning rate after each iteration, $iteration$ is the index of the current iteration, $max\_iteration$ is the maximum number of iterations in each epoch. We configure the initial learning rate as 4.5×10^{-2} and the power is 0.9.

When performing training on the CamVid dataset, we adjust the optimization method to Adam with a batch size of 8. The momentum is set as 0.9 and the weight decay is set as 2×10^{-4}. Also, we employ the "poly" learning rate policy with the initial learning rate of 1×10^{-3}. Fig. 4 plots the curves of the loss function vs. the number of iterations on the Cityscapes

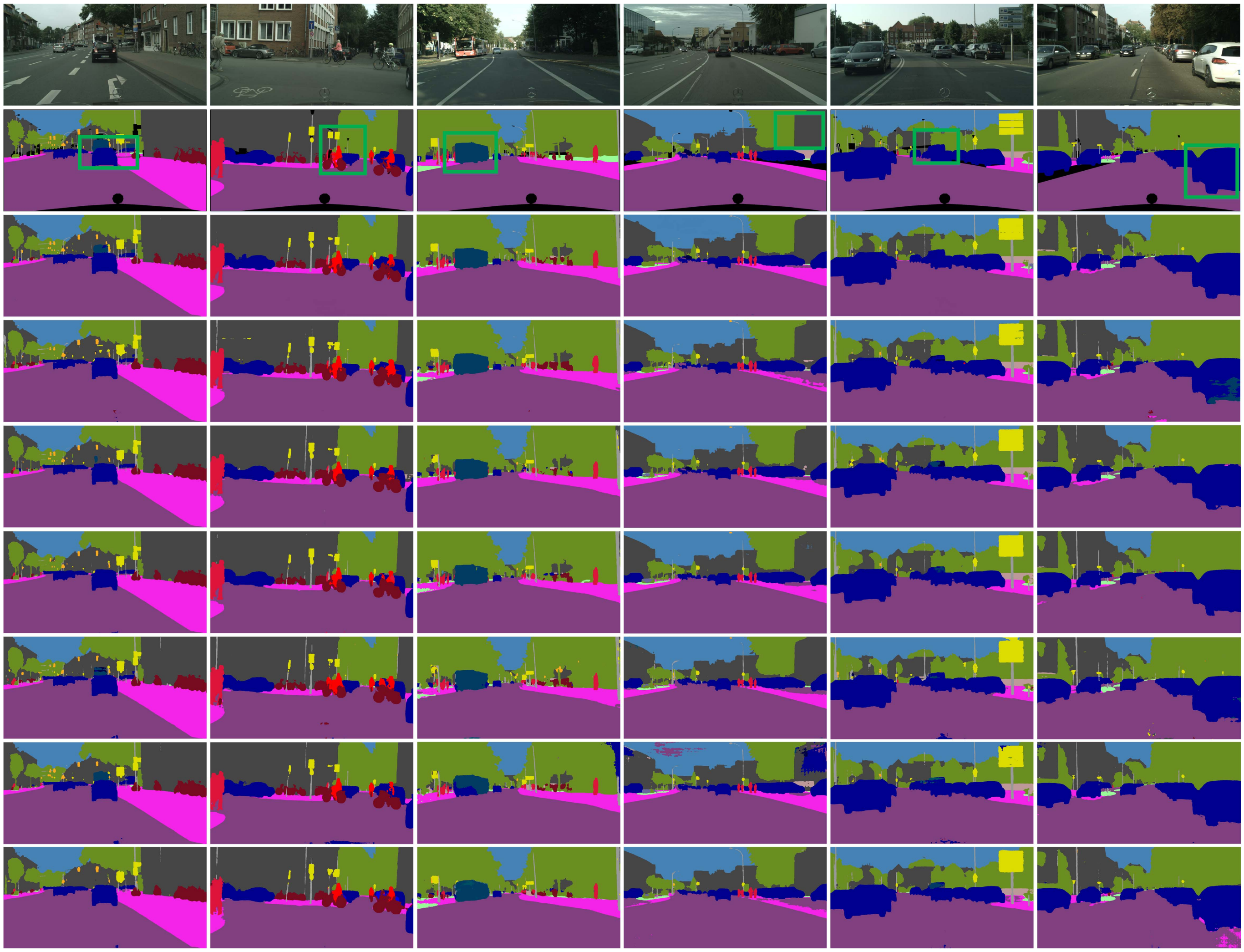

Fig. 6. The comparative results on the Cityscapes val dataset. From top to the bottom are successively the original observed images, ground truths, segmentation results from our MSCFNet, LEDNet [23], DABNet [22], ERFNet [18], EDANet [27], CGNet [47] and ESPNet [36]. (Best viewed in color).

and CamVid datasets. The two curves drop smoothly and converge eventually, indicating that our MSCFNet can be well trained.

B. Ablation Study

In this part, we conduct comparative studies to demonstrate the feasibility and effectiveness of our presented method. All the ablation studies are conducted on the CamVid training set, validation set, and evaluated on its testing set.

1) Context Fusion: To study how the contextual features affect the segmentation accuracy, we have performed experiments without any attention mechanism in this part. From Table I (a) we could observe that the segmentation results without the context fusion mechanism are more than 1% lower than those used, which has confirmed that multi-scale contextual features play a critical role in dense pixel classification tasks. As for the way of context fusion, feature adding has better performance than that of feature concatenation. In Table II, we also study how the different levels of features affect the segmentation results. Fig. 5 depicts the heat maps from the low-level, mid-level, and high-level contextual features.

2) Attention Module: We use the attention mechanism for better extracting spatial features and promoting channel information interaction. Table I (b) demonstrates the effectiveness brought by channel and spatial attention. Adding channel attention SE [42] can bring 0.1% slightly better accuracy. However, ECA [46] works better with fewer parameters. When we add spatial attention to the injection branches, the segmentation accuracy increases from 69.16% to 69.30%. From the above studies, we conclude that by using the channel and spatial attention mechanisms simultaneously, we can gain better segmentation performance at the cost of negligible parameters increasing.

3) Dilation Rate: As depicted in Table I (c), we design three experiments to study the effect of the number of EAR modules and the dilation rate. First, we reduce the EAR modules to 3 in the first block and to 6 in the second block. We find that the segmentation accuracy is 1.03% lower than that of our final version, indicating that more EAR modules can improve the performance. Then setting the same number of modules, we adopt the idea of DABNet [22] with the dilation rates are the power of 2 in the second EAR block.

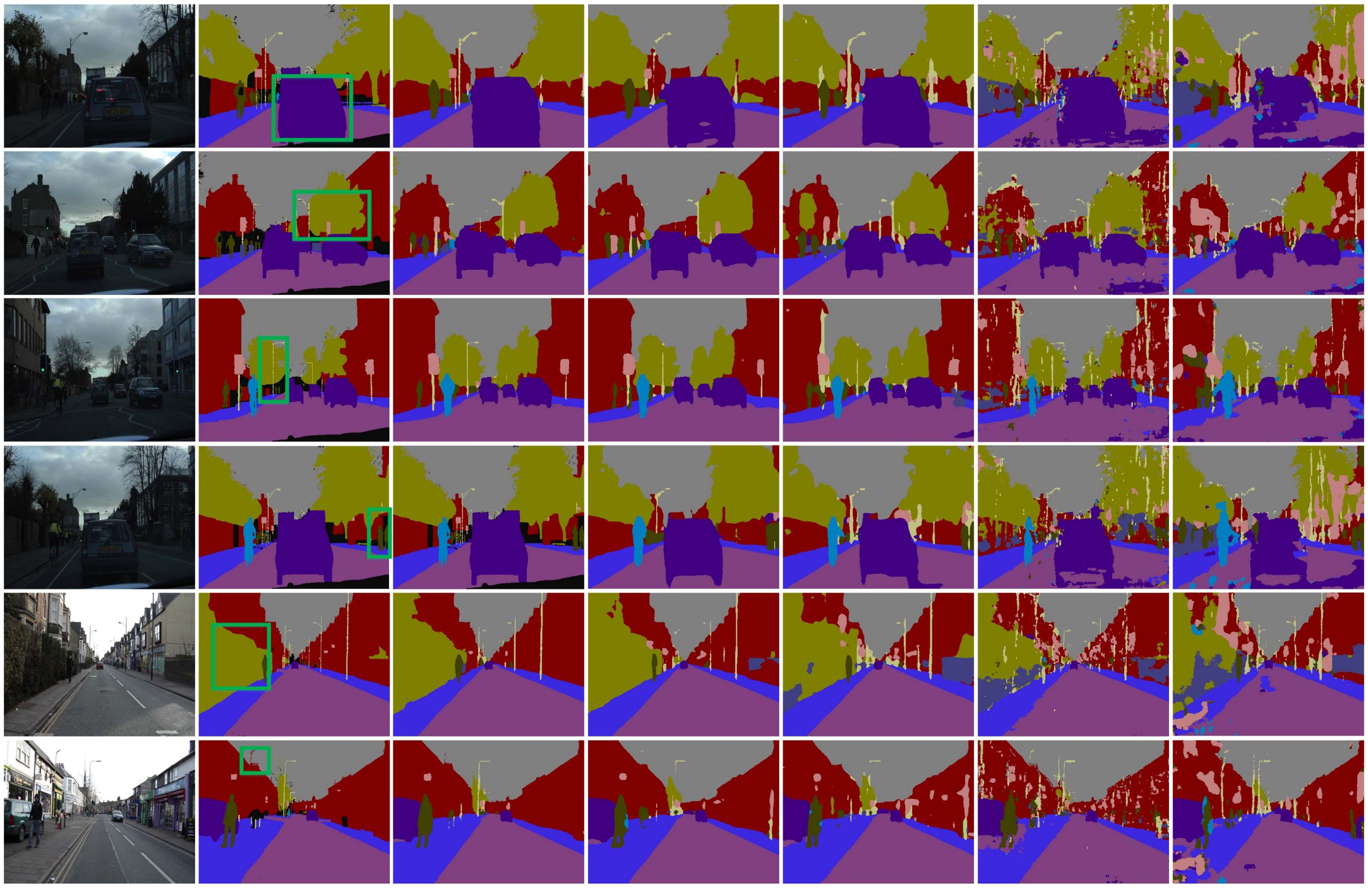

Fig. 7. The comparative results on the CamVid testing set. From left to right are original observed images, ground truths, segmentation outputs from our MSCFNet, DABNet [22], CGNet [47], SegNet [17] and ENet [16]. (Best viewed in color).

The results also validate that our settings can achieve better performance.

C. Evaluation Results

The performance of our MSCFNet is evaluated with several state-of-the-art ones in this part on the above mentioned Cityscapes and CamVid datasets: FCN-8s [50], DeepLabLFOV [51], Dilation8 [26], ENet [16], FSSNet [33], SegNet [17], ERFNet [18], Fast-SCNN [49], ESPNet [36], CGNet [47], NDNet [30], ContextNet [48], ICNet [37], BiseNet [38], EDANet [27], DABNet [22], FarSeeNet [28], LEDNet [23], DFANet [35], and EdgeNet [32].

As can be observed from Table III to Table V, the comparison results verify that our MSCFNet attains a better balance between segmentation accuracy and efficiency. For the Cityscapes dataset, our MSCFNet only has a 1.15M model size but yields 71.9% class mIoU and 88.4% category mIoU, respectively, even 74.2% class mIoU on the validation set. Considering the efficiency, MSCFNet only occupies 15% of the number of parameters of the ICNet but achieves a faster speed and a better result. Although DABNet is almost 0.4M smaller than our network, it delivers poor accuracy with 1.8% lower than our MSCFNet. For the CamVid dataset, we can see that our model also gets remarkable performance with a smaller capacity, and achieves 69.3% class mIoU on the CamVid testing set, which is superior to most of the existing competitive methods. Although DFANet has a faster speed, it occupies almost 7× number of parameters than our MSCFNet. As shown in these tables, our network makes a good trade-off among model size, inference speed, and segmentation accuracy. The qualitative comparisons with some respective methods are also shown in Fig. 6 and Fig. 7, which qualitatively verify the effectiveness of our MSCFNet.

V. Conclusion

In summary, we have proposed a multi-scale context fusion network (MSCFNet), which improves both segmentation accuracy and inference speed for lightweight semantic segmentation task in this work. We designed an efficient asymmetric residual (EAR) module, which adopts factorization depth-wise convolution and dilation convolution to capture object features with a lower number of parameters and computational budgets using different receptive fields. Moreover, MSCFNet had branches with efficient attention modules from different stages extracting multi-scale contextual information. Then, the features from these connections were combined to enhance the expression of the features and facilitate the local and contextual information interaction greatly. Extensive experiments on the CamVid and Cityscapes datasets have validated that our architecture can attain a better balance between efficiency and accuracy than several comparative approaches.

References

[1] H. Lu, Q. Liu, D. Tian, Y. Li, H. Kim, and S. Serikawa, "The cognitive Internet of vehicles for autonomous driving," *IEEE Netw.*, vol. 33, no. 3, pp. 65–73, May 2019.

[2] G. Gao, Y. Yu, M. Yang, H. Chang, P. Huang, and D. Yue, "Cross-resolution face recognition with pose variations via multilayer locality-constrained structural orthogonal procrustes regression," *Inf. Sci.*, vol. 506, pp. 19–36, Jan. 2020.

[3] W. Zhao, H. Lu, and D. Wang, "Multisensor image fusion and enhancement in spectral total variation domain," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 866–879, Apr. 2018.

[4] G. Gao, J. Yang, X.-Y. Jing, F. Shen, W. Yang, and D. Yue, "Learning robust and discriminative low-rank representations for face recognition with occlusion," *Pattern Recognit.*, vol. 66, pp. 129–143, Jun. 2017.

[5] G. Gao, Y. Yu, J. Xie, J. Yang, M. Yang, and J. Zhang, "Constructing multilayer locality-constrained matrix regression framework for noise robust face super-resolution," *Pattern Recognit.*, vol. 110, Feb. 2021, Art. no. 107539.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[7] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural Comput.*, vol. 29, no. 9, pp. 2352–2449, Sep. 2017.

[8] G. Gao, Y. Yu, J. Yang, G.-J. Qi, and M. Yang, "Hierarchical deep CNN feature set-based representation learning for robust cross-resolution face recognition," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Dec. 3, 2020, doi: 10.1109/TCSVT.2020.3042178.

[9] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[10] H. Lu *et al.*, "Wound intensity correction and segmentation with convolutional neural networks," *Concurrency Comput. Pract. Exper.*, vol. 29, no. 6, p. e3927, Mar. 2017.

[11] L.-C. 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 Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[12] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 3684–3692.

[13] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[14] K. Sun *et al.*, "High-resolution representations for labeling pixels and regions," 2019, *arXiv:1904.04514*. [Online]. Available: http://arxiv.org/abs/1904.04514

[15] R. Lan, L. Sun, Z. Liu, H. Lu, C. Pang, and X. Luo, "MADNet: A fast and lightweight network for single-image super resolution," *IEEE Trans. Cybern.*, vol. 51, no. 3, pp. 1443–1453, Mar. 2021.

[16] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," 2016, *arXiv:1606.02147*. [Online]. Available: http://arxiv.org/abs/1606.02147

[17] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[18] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "ERFNet: Efficient residual factorized ConvNet for real-time semantic segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 263–272, Jan. 2018.

[19] S. Mehta, M. Rastegari, L. Shapiro, and H. Hajishirzi, "ESPNetv2: A light-weight, power efficient, and general purpose convolutional neural network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9190–9200.

[20] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.

[21] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: http://arxiv.org/abs/1704.04861

[22] G. Li, I. Yun, J. Kim, and J. Kim, "DABNet: Depth-wise asymmetric bottleneck for real-time semantic segmentation," 2019, *arXiv:1907.11357*. [Online]. Available: http://arxiv.org/abs/1907.11357

[23] Y. Wang *et al.*, "Lednet: A lightweight encoder-decoder network for real-time semantic segmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1860–1864.

[24] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 2018, pp. 801–818.

[25] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587

[26] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: http://arxiv.org/abs/1511.07122

[27] S.-Y. Lo, H.-M. Hang, S.-W. Chan, and J.-J. Lin, "Efficient dense modules of asymmetric convolution for real-time semantic segmentation," in *Proc. ACM Multimedia Asia (MMAsia)*, Dec. 2019, pp. 1–6.

[28] Z. Zhang and K. Zhang, "FarSee-net: Real-time semantic segmentation by efficient multi-scale context aggregation and feature space super-resolution," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 8411–8417.

[29] Z. Yang *et al.*, "Small object augmentation of urban scenes for real-time semantic segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 5175–5190, 2020.

[30] Z. Yang *et al.*, "NDNet: Narrow while deep network for real-time semantic segmentation," *IEEE Trans. Intell. Transp. Syst.*, early access, Apr. 27, 2020, doi: 10.1109/TITS.2020.2987816.

[31] V.-C. Miclea and S. Nedevschi, "Real-time semantic segmentation-based stereo reconstruction," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 4, pp. 1514–1524, Apr. 2020.

[32] H.-Y. Han, Y.-C. Chen, P.-Y. Hsiao, and L.-C. Fu, "Using channel-wise attention for deep CNN based real-time semantic segmentation with class-aware edge information," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 2, pp. 1041–1051, Feb. 2021.

[33] X. Zhang, Z. Chen, Q. M. J. Wu, L. Cai, D. Lu, and X. Li, "Fast semantic segmentation for scene perception," *IEEE Trans. Ind. Informat.*, vol. 15, no. 2, pp. 1183–1192, Feb. 2019.

[34] S. Zhou, D. Nie, E. Adeli, J. Yin, J. Lian, and D. Shen, "High-resolution encoder–decoder networks for low-contrast medical image segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 461–475, 2020.

[35] H. Li, P. Xiong, H. Fan, and J. Sun, "DFANet: Deep feature aggregation for real-time semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9522–9531.

[36] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, "ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 552–568.

[37] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "ICNet for real-time semantic segmentation on high-resolution images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 405–420.

[38] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 325–341.

[39] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.

[40] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1971–1980.

[41] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, "Asymmetric non-local neural networks for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 593–602.

[42] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7132–7141.

[43] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi, "Gather-excite: Exploiting feature context in convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 9401–9411.

[44] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[45] Z. Gao, J. Xie, Q. Wang, and P. Li, "Global second-order pooling convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3024–3033.

[46] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11534–11542.

[47] T. Wu, S. Tang, R. Zhang, J. Cao, and Y. Zhang, "CGNet: A light-weight context guided network for semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 1169–1179, 2021.

[48] R. P. K. Poudel, U. Bonde, S. Liwicki, and C. Zach, "ContextNet: Exploring context and detail for semantic segmentation in real-time," 2018, *arXiv:1805.04554*. [Online]. Available: http://arxiv.org/abs/1805.04554

[49] R. P. K. Poudel, S. Liwicki, and R. Cipolla, "Fast-SCNN: Fast semantic segmentation network," 2019, *arXiv:1902.04502*. [Online]. Available: http://arxiv.org/abs/1902.04502

[50] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[51] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*. [Online]. Available: http://arxiv.org/abs/1412.7062

[52] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[53] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. Int. Conf. Comput. Statist. (COMPSTAT)*, Sep. 2010, pp. 177–186.

Guangwei Gao (Member, IEEE) received the Ph.D. degree in pattern recognition and intelligence systems from Nanjing University of Science and Technology, Nanjing, in 2014. He was an Exchange Student of the Department of Computing, The Hong Kong Polytechnic University, in 2011 and 2013. He was a Project Researcher with the National Institute of Informatics, Tokyo, Japan, in 2019. He is currently an Associate Professor with the Institute of Advanced Technology, Nanjing University of Posts and Telecommunications. His research interests include pattern recognition and computer vision. He has served as a Reviewer for IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CYBERNETICS, *Pattern Recognition*, *Neurocomputing*, *Pattern Recognition Letters*, *AAAI*, *ICPR*, and *ICIP*.

Guoan Xu received the B.S. degree in measurement control technology and instrumentation from Changshu Institute of Technology, Jiangsu, China, in 2019. He is currently pursuing the M.S. degree with the College of Automation and College of Artificial Intelligence, Nanjing University of Posts and Telecommunications. His research interest includes image semantic segmentation.

Yi Yu (Member, IEEE) received the Ph.D. degree in information and computer science from Nara Women's University, Japan. She is currently an Assistant Professor with the National Institute of Informatics (NII), Japan. Before joining NII, she was a Senior Research Fellow with the School of Computing, National University of Singapore. Her research covers large-scale multimedia data mining and pattern analysis, location-based mobile media service, and social media analysis. She and her team received the Best Paper Award from the IEEE ISM 2012, the 2nd prize in Yahoo Flickr Grand Challenge 2015, were in the top winners (out of 29 teams) from ACM SIGSPATIAL GIS Cup 2013, and the Best Paper Runner-Up in APWeb-WAIM 2017, recognized as finalist of the World's First 10K Best Paper Award in ICME 2017.

Jin Xie (Member, IEEE) received the Ph.D. degree from the Department of Computing, Hong Kong Polytechnic University in 2012. From 2013 to 2017, he was a Research Scientist with New York University Abu Dhabi, Abu Dhabi, United Arab Emirates. He is currently a Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology. His current research interests include image forensics, computer vision, machine learning, 3-D computer vision with the convex optimization, and deep learning methods.

Jian Yang (Member, IEEE) received the Ph.D. degree in pattern recognition and intelligence systems from Nanjing University of Science and Technology (NUST) in 2002. In 2003, he was a Post-Doctoral Researcher at the University of Zaragoza. From 2004 to 2006, he was a Post-Doctoral Fellow at the Biometrics Centre, The Hong Kong Polytechnic University. From 2006 to 2007, he was a Post-Doctoral Fellow at the Department of Computer Science, New Jersey Institute of Technology. He is currently a Chang-Jiang Professor with the School of Computer Science and Engineering, NUST. He is the author of more than 100 scientific papers in pattern recognition and computer vision. His papers have been cited more than 4000 times in the Web of Science, and 9000 times in the Scholar Google. His research interests include pattern recognition, computer vision, and machine learning. Currently, he is/was an Associate Editor of *Pattern Recognition Letters*, IEEE TRANSACTIONS NEURAL NETWORKS AND LEARNING SYSTEMS, and *Neurocomputing*. He is a fellow of IAPR.

Dong Yue (Fellow, IEEE) received the Ph.D. degree in engineering from South China University of Technology, Guangzhou, China, in 1995. He is currently a Professor and the Dean of the Institute of Advanced Technology and the College of Automation and AI at Nanjing University of Posts and Telecommunications. He has authored or coauthored more than 250 articles in international journals and two books. He holds more than 50 patents. His current research interests include analysis and synthesis of networked control systems, multiagent systems, optimal control of power systems, and the Internet of Things. He served as the Associate Editor for *IEEE Industrial Electronics Magazine*, IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, and IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS. He was the Associate Editor of the *Journal of the Franklin Institute*and *International Journal of Systems Sciences*, and the Guest Editor of Special Issue on New Trends in Energy Internet: Artificial Intelligence-Based Control, Network Security, and Management.