E2CANet: An Efficient and Effective Convolutional Attention Network for Semantic Segmentation

Yuerong Mu and Qiang Guo (✉)

School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan, China
guoqiang@sdufe.edu.cn

Abstract. Many semantic segmentation methods employ various attention mechanisms to improve segmentation accuracy. However, as the accuracy of the model increases, the computational cost is relatively expensive, which is not favorable for some practical applications. To solve this problem, this paper presents an efficient and effective convolutional attention network (E2CANet), which is designed to achieve a good trade-off between segmentation accuracy and computational efficiency. E2CANet adopts an encoder-decoder architecture with skip connections to preserve details and semantic information. For the encoder, we use cheap convolutional operations to introduce two different attentions, i.e. global attention and multi-scale attention, which can significantly reduce the computational cost while highlighting important features and suppressing unnecessary ones. A lightweight All-MLP decoder, which only consists of six linear layers, is used to aggregate features from the encoder. The simple design of this decoder is also the key to reduce computational complexity. Extensive experiments are performed on ADE20K, Cityscapes, and COCO-stuff datasets. The proposed E2CANet delivers very competitive results on all datasets. Especially, E2CANet-Tiny (a lightweight version of E2CANet) achieves 41.92% mIoU on ADE20K dataset with less than 4.4M parameters, which demonstrates the efficiency and effectiveness of our method. Code is available at https://github.com/muyuerong/E2CANet.

Keywords: Semantic segmentation · Multi-scale attention · Global attention · Feature extraction · Lightweight All-MLP decoder

1 Introduction

Semantic segmentation is an important task of computer vision, aiming at labelling each pixel in an image as a corresponding semantic category. Different from image classification that identifies the categories of the whole image, semantic segmentation classifies each pixel in the image. It has a wide range of applications such as medical image segmentation [29], autonomous driving [37], saliency detection [19], human-machine interaction [25], and many other fields.

In recent years, attention mechanisms are introduced into semantic segmentation to focus on regions or features of interest. The attention mechanisms assign different weights to different input data for concentrating on the most relevant parts. Using attention mechanisms can help the segmentation model to better understand the input image and thus improve segmentation accuracy. In semantic segmentation, common attention mechanisms include spatial attention and channel attention. Channel attention pays attention to 'what' is meaningful given an input image. Different from channel attention, spatial attention focuses on 'where' is an informative part. Based on the above two attentions, some attention modules are designed to further improve segmentation accuracy. Large kernel attention (LKA) was proposed in VAN [14] to build channel attention and spatial attention. It combines the long-range dependence of self-attention and the advantage of large kernel convolution to make full use of contextual information. In RepLKNet [10], it can be found that using large kernel convolutions can significantly improve the effective receptive fields compared to increase the number of layers with smaller kernel size, and can even leverage more shape information. However, traditional large kernel convolutions are computationally expensive. To tackle this issue, Guo et al. [13] designed a multi-scale convolutional attention module, which introduces large kernels by using multi-branch lightweight strip convolutions. Besides, Woo et al. [36] proposed a convolutional block attention module, which sequentially applies channel attention and spatial attention. The module learns what and where to emphasize or suppress and refines intermediate features effectively.

Inspired by the impact of attention mechanisms on segmentation accuracy improvement, we propose an efficient and effective convolutional attention network (E2CANet) to achieve high segmentation accuracy and low computational cost. Our network adopts an encoder-decoder architecture with skip connections. The encoder uses a common hierarchical structure with four stages. Each stage contains a novel convolutional attention block, which introduces global attention and multi-scale attention by using cheap convolution operations. The former employs global average pooling and max pooling to obtain global context information, and the latter uses strip convolutions at different scales to capture multi-scale features. E2CANet sequentially utilizes these two attentions to focus on useful features and suppress useless ones. To further improve the efficiency of E2CANet, a lightweight All-MLP decoder is introduced, which aggregates the features from different stages and obtains the final segmentation results. Such an encoder-decoder design can allow our network with a good trade-off between segmentation accuracy and computational cost. Extensive experimental results on three public datasets demonstrate the advantages of E2CANet in terms of segmentation performance and the number of parameters.

In summary, our contributions are as follows:

- To increase the segmentation accuracy, we propose a convolutional attention block (CA block) that is composed of global attention and multi-scale attention. Based on the CA block, E2CANet is designed for tackling the task of semantic segmentation.

- To obtain a powerful representation and make E2CANet more efficient, we use a lightweight decoder without computationally complex modules to aggregate features of four stages.
- We conduct extensive experiments on the ADE20K, Cityscapes, and COCO-stuff datasets to validate the high segmentation accuracy and low computational complexity of the proposed E2CANet.

2 Related Work

In this section, we briefly review some common methods for extracting global and multi-scale features in semantic segmentation, as well as some representative attention mechanisms.

2.1 Global Feature Extraction

Global features can help semantic segmentation models better understand the overall context. FCN [24] is a significant advancement in the field of semantic segmentation, which defines a skip connection to combine deep semantic information with shallow detail information. However, it ignores the global information of an image. As a variant of FCN, ParseNet [23] obtains global features by using global average pooling. Besides, a context encoding module is introduced in EncNet [43] to capture global context information, which can significantly improve segmentation results with only a small additional computational cost compared to FCN. Unlike the above three networks, SENet [18] applies the global context to recalibrate the weights of different channels, but the global context information is not fully utilized. To obtain the global features, Wang et al. [34] proposed NLNet to model the long-range dependencies by using self-attention mechanism. It first computes the pairwise relations between the query and all positions to form the attention map, and then aggregates the features of all positions by weighted sum. However, global context information modeled by NLNet are almost the same for different positions within an image, which results in waste of calculations. Inspired by the global context modeling capabilities of NLNet, Cao et al. [4] designed a lightweight global context block in GCNet, which is able to efficiently models global context information.

2.2 Multi-scale Feature Extraction

Multi-scale features play an important role in the field of computer vision and image processing. To capture multi-scale contextual information, PSPNet [44] uses the pyramid pooling module to gather multi-scale information. Different from PSPNet, DeepLabV3 [6] designs the multi-scale dilated convolution to extract features at different scales. Both PSPNet and DeepLabV3 apply $n \times n$ square convolutions to extract features at different scales. However, using square convolutions, especially the kernel size being larger than seven, introduces a significantly increase in computational complexity. To solve this issue, Szegedy et

al. [31] used $1 \times n$ and $n \times 1$ strip convolutions in parallel instead of $n \times n$ convolutions to reduce the number of parameters. Besides, HRNet [33] extracts multi-scale features by connecting high-to-low resolution convolutions in parallel. The convolutional multi-scale fusion module is utilized to integrate multi-scale feature hierarchies in HRFormer [42]. To capture different scale semantic dependencies, Jiao et al. [22] proposed a multi-scale dilated attention, which uses different dilation rates for different heads.

2.3 Attention Mechanism

In recent years, various attention mechanisms are used to adaptively select important features, resulting in the improvement of accuracy and efficiency [32] [15] [16]. Channel attention and spatial attention are employed in semantic segmentation to focus on the important information in the channel dimension and spatial dimension, respectively. A representative model of the channel attention is SENet, which designs a "Squeeze-and-Excitation" block (SE block) to adaptively recalibrate channel-wise feature responses by explicitly modeling interdependencies between channels. However, SE block adopts global average pooling to capture global information, which limits the modeling capability. To address this issue, a global second-order pooling block [12] was proposed to gather global features. Moreover, spatial attention was first presented in STN [21] to pay attention to the most relevant regions. A learnable spatial transformer module is introduced in STN, which can be inserted into existing convolutional architectures, performing spatial transformation on the input images. Woo et al. [36] proposed a convolutional block attention module (CBAM), which performs global average pooling and max pooling to adaptively learn channel and spatial attention weights. The experimental results in CBAM showed that combining channel and spatial attentions outperforms using only one attention independently.

3 Method

This section introduces the efficient and effective convolutional attention network for semantic segmentation. As illustrated in Fig. 1, E2CANet mainly contains two parts: (i) a convolutional attention encoder that is designed to better extract global features and multi-scale features; (ii) a lightweight All-MLP decoder that aims to aggregate features from the encoder and obtain the semantic segmentation mask. In the rest of this section, we detail the proposed encoder and decoder designs.

3.1 Convolutional Attention Encoder

The encoder of E2CANet adopts a common hierarchical structure to extract feature maps with high-level semantic information. A simple stem block is applied for capturing low-level features, followed by four stages to further extract global and multi-scale features as shown in Fig. 2. The first stage is composed of a

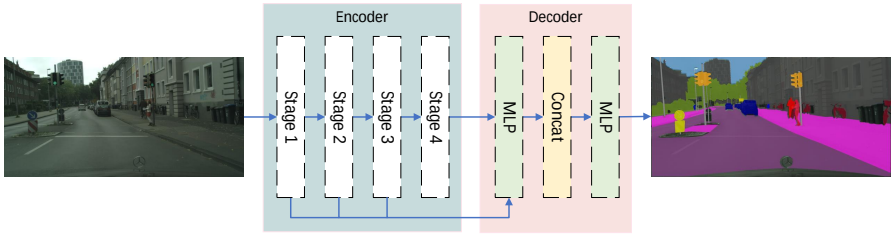

Fig. 1: The overall architecture of the proposed E2CANet

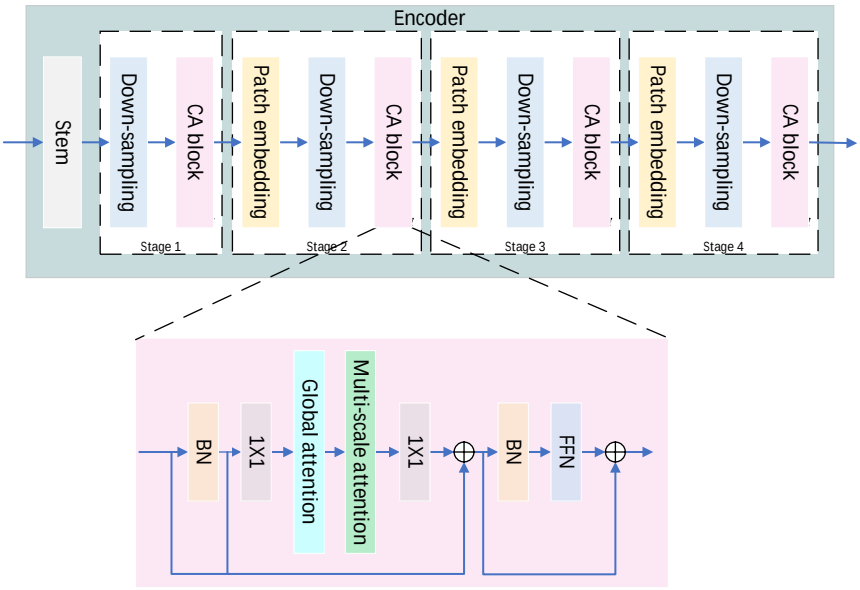

Fig. 2: The architecture of the proposed encoder. It is composed of a stem block, patch embedding blocks, down-sampling blocks, and convolutional attention blocks.

down-sampling block and a stack of convolutional attention blocks. The other three stages have a patch embedding block before each down-sampling block. The stem block stakes two 3×3 convolution layers. Each layer has a batch normalization, and a GELU activation is applied only for the first layer. The patch embedding block, containing a 3×3 convolution and a BN layer, is adopted to down-sample the spatial dimension by $4\times$ and increase the channel dimension by $2\times$. As for the down-sampling block, we utilize a convolution with stride 2 and kernel size 3×3, followed by a batch normalization layer. To meet the needs of devices with less memory and less computational resources, we design a lightweight version named E2CANet-Tiny. Compared to E2CANet, it reduces

Table 1: Detailed settings of different versions of the proposed E2CANet. C and L represent the numbers of channels and building blocks, respectively.

stage	E2CANet	E2CANet-Tiny
1	C=64, L=2	C=32, L=3
2	C=128, L=2	C=64, L=3
3	C=320, L=4	C=160, L=5
4	C=512, L=2	C=256, L=2

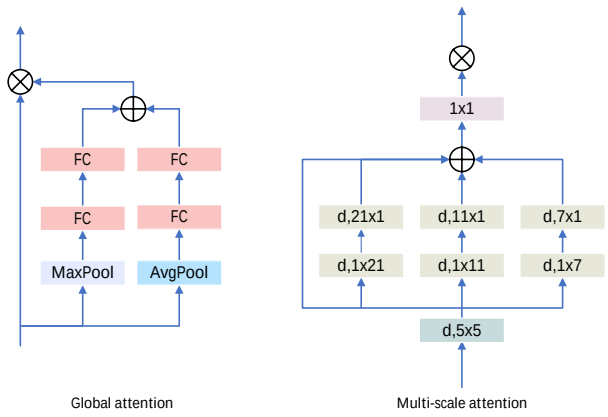

Fig. 3: The structures of global attention and multi-scale attention

the number of parameters by lowing the channel numbers and changing the number of building blocks. Detailed settings of E2CANet and E2CANet-Tiny are listed in Table 1.

Convolutional Attention Block. The structural design of CA block is inspired by the transformer encoder in ViT. It is mainly composed of global attention, multi-scale attention, skip connections, and feed-forward network (FFN), which is shown in the bottom of Fig. 2. Skip connections preserve the details and semantic information of the input data, which can avoid detail information loss and improve the training speed. FFN is used to further extract features based on global attention and multi-scale attention, and enhance the ability of representation of our model. It consists of two 1×1 convolutions and a 3×3 convolution. Global attention and multi-scale attention are the main components of the CA block. In the following, we will detail these two attentions.

In global attention, global average pooling and max pooling are commonly adopted to aggregate global information. Global average pooling computes the average of all pixel values in the feature maps, compressing the original two-

dimensional feature maps into a one-dimensional feature vector. It can significantly reduce the number of parameters in the model while retaining the global information in feature maps. Different from global average pooling, global max pooling extracts the maximum value of all pixels, highlighting the most important features. Empirically, exploiting both types of pooling, which can extract richer high-level features, is more effective than using each independently. Therefore, we use global average pooling and max pooling simultaneously to obtain average-pooled and max-pooled features.

Fig. 3 (left) shows the structure of the global attention, which first performs the global average pooling and global max pooling on the input feature map to generate average-pooled and max-pooled features. Then features are forwarded to two fully connected layers to further extract the global context information. Finally, we utilize the element-wise summation to merge the attention map that is used as weights to reweigh the input. The global attention can be formulated as:

$$\begin{aligned} M_g(F) &= MLP(GAP(F)) + MLP(MAX(F)), \\ F_g &= M_g(F) \otimes F, \end{aligned}$$

where F is the input feature map, GAP and MAX denote the global average pooling and max pooling, respectively. M_g is the global attention map, and F_g is the output of global attention. $\otimes$ represents the element-wise multiplication.

In multi-scale attention, a multi-scale convolutional architecture is designed to extract features at different scales, which is depicted in Fig. 3 (right). To aggregate local information, a 5×5 depth-wise convolution is used, followed by four branches. One of these branches is an identity connection. The other three branches employ large kernel convolutions with different sizes to capture features at different scales. Using large kernel convolutions can improve the segmentation accuracy due to the reason that it can enlarge the receptive fields. However, traditional $n \times n$ large kernel convolutions are costly, especially n is larger than seven. To tackle this issue, we utilize a depth-wise $1 \times n$ convolution and a $n \times 1$ depth-wise convolution to approximate the $n \times n$ depth-wise convolutions. Specifically, the kernel sizes of other three branches are set to 7, 11, and 21, respectively. Then, we obtain the attention weights by using the 1×1 convolution. The multi-scale attention weights are computed as:

$$\begin{aligned} F' &= Conv_{5\times5}(F), \\ Conv_{n_i}(F) &= Conv_{i\times1}\big(Conv_{1\times i}(F)\big), \\ M_m(F') &= Conv_{1\times1}\big(F' + Conv_{n_7}(F') + Conv_{n_{11}}(F') + Conv_{n_{21}}(F')\big), \\ F_m &= M_m(F') \otimes F, \end{aligned}$$

where $Conv_{1\times1}$ and $Conv_{5\times5}$ represent depth-wise convolutions with kernel sizes of 1 and 5, $Conv_{n_i}$ and i denote strip convolutions and kernel sizes of strip convolutions, respectively. M_m is the multi-scale attention map, and F_m is the final output.

3.2 Lightweight All-MLP Decoder

The decoder is responsible for mapping the features extracted by the encoder to generate the segmentation mask. As shown in Fig. 1, we adopt a lightweight All-MLP decoder, which only consists of six linear layers. It first upsamples the features of different stages of the encoder to the same size as Stage 1. Then, a MLP layer is adopted to fuse the features and another MLP layer takes the fused feature to predict the segmentation masks. This lightweight decoder allows E2CANet to achieve a good trade-off between segmentation accuracy and computational cost.

4 Experiments

In this section, we evaluate the proposed E2CANet on ADE20K, Cityscapes, and COCO-stuff. We first summarize the datasets and the implementation details. Then, the contributions of each component are investigated in ablation studies on ADE20K. Finally, to verify the efficiency and effectiveness of E2CANet, we compare it with some state-of-the-art segmentation methods.

Datasets. ADE20K [45] is a large-scale dataset for scene parsing with more than 20, 000 images and 150 semantic tags of different categories, which covers a wide range of different scenes from indoor to outdoor, nature to urban. Each image has been labeled in detail to semantically classify each pixel in the image.

Cityscapes [9] is a semantic understanding image dataset on urban street scenes. It mainly contains street scenes from 50 different cities, with 5, 000 high-quality pixel-level annotated images of driving scenes in urban environments (2, 975 for train, 500 for validation, 1, 525 for test, with 19 categories) and 20, 000 roughly annotated images.

COCO-stuff [2], a large-scale dataset for scene understanding, is an extension of the COCO (Common Object in Context) dataset, which annotates new categories from images in the COCO dataset in order to provide a more comprehensive understanding of the scenes in the image. This dataset comprises over 200, 000 images and is labeled with 80 different object categories and 91 different pixel-level scene categories in each image. These pixel-level scene categories are marked as segmentation masks that can be used to train and evaluate semantic segmentation models.

We implement E2CANet based on MMSegmentation [8], which is an open source semantic segmentation toolbox. All the experiments are performed on the Pytorch [28] platform with a V100 GPU. The batch size is set to 8 for Cityscapes and 16 for other datasets. The iteration number is set to 160K for ADE20K and Cityscapes, and 80K for COCO-stuff. For all experiments, we use mean Intersection over Union (mIoU) and the number of parameters to serve as the evaluation metrics. For optimization, a learning rate of 0.01, a momentum of 0.9, and a weight decay of 0.5×10^{-3} are used in training.

Table 2: Comparisons of different pooling methods used in global attention.

Methods	param (M)	mIoU (%)
AvgPool	4.4	40.7
MaxPool	4.4	40.6
MaxPool&AvgPool	4.4	41.9

4.1 Ablation Study

In this subsection, we conduct extensive ablation studies of our E2CANet on ADE20K to perform detail analysis of our proposed method in three aspects. We evaluate the benefits of using both global average pooling and max pooling, and examine the influence of different scale branches. Then, we verify the effectiveness of using global attention and multi-scale attention sequentially.

Selections of Pooling Methods in Global Attention. Table 2 shows the results of using three variants of the pooling methods in global attention, which contains global average pooling, global max pooling, and joint use of both poolings. From Table 2, it can be observed that using global average pooling performs slightly better than max pooling. We can also find using both types of pooling can improve mIoU by 1.2% compared to only using average pooling with no increase in the number of parameters. Therefore, we use both global average pooling and max pooling in E2CANet.

Ablations on Multi-scale Attention Design. We perform ablation studies to verify the importance of three different scale branches and 1×1 convolution. $n \times n$ branch is composed of a depth-wise $1 \times n$ convolution and a $n \times 1$ depth wise convolution. In Table 3, it can be found that using three different scale branches is more efficient than using any two of them. The segmentation results are further improved by adding 1×1 convolution for channel mixing. The results demonstrate that branches at different scales and the 1×1 convolution are important for the final performance.

Usages of the Global and Multi-scale Attentions. In this ablation study, we compare four different ways of using the multi-scale (m.s.) attention and global (glo.) attention, i.e. m.s. attention only, glo. attention only, m.s.-glo. attention, and glo.-m.s. attention. The main difference between m.s.-glo. attention and glo.-m.s. attention is the attention order. The results are shown in Table 4. It can be observed that only using m.s. attention gains 1.2% mIoU over glo. attention. We can also find that using both attentions sequentially performs better than using them individually. This proves that both attentions play an important role in E2CANet. As each attention plays different roles, the order may affect the segmentation accuracy. In Table 4, the last two rows show that

Table 3: Ablation studies on multi-scale attention design. Br: branch.

7×7 Br	11×11 Br	21×21 Br	1×1 Conv	mIoU (%)
✘	✓	✓	✓	40.2
✓	✘	✓	✓	40.4
✓	✓	✘	✓	40.7
✓	✓	✓	✘	23.3
✓	✓	✓	✓	41.9

Table 4: Usages of the global and multi-scale attention.

Methods	mIoU (%)
m.s. attention	40.3
glo. attention	39.1
m.s.-glo. attention	40.9
glo.-m.s. attention	41.9

glo.-m.s. attention performs better than m.s.-glo. attention. Therefore, glo.-m.s. attention is adopted in our E2CANet.

4.2 Comparisons with State-of-the-Art Methods

In this subsection, to verify the efficiency and effectiveness of the proposed E2CANet, we compare it with several recent transformer-based segmentation methods include HRFormer [42], EfficientViT [3], Mask2Former [7], and SegFormer [39] and CNN-based state-of-the-art segmentation methods include ENet [27], SegNet [1], CGNet [38], MoblieNetV2 [30], BiSeNetV2 [41], PIDNet [40], DDRNet [26], RegSeg [11], and PSPNet [44] on ADE20K, Cityscapes, and COCO-Stuff datasets.

Comparisons with Transformer-based Methods. We compare E2CANet-Tiny and E2CANet with the state-of-the-art transformer-based methods, and the results are shown in Table 5. SegFormer-B0 and EfficientViT-B0 are the lightweight models of SegFormer and EfficientViT for fast inference, and the number of parameters of E2CANet-Tiny are similar to them. From Table 5, we can find that E2CANet-Tiny gains 4.5% mIoU over SegFormer-B0 with the similar GFLOPs (floating point operations) on ADE20K and 3.7% mIoU over EfficientViT-B0 on Cityscapes. HRFormer-S and SegFormer-B1 are both the small scaled versions of HRFormer and SegFormer, which have the similar number of parameters to E2CANet. It can be also seen that E2CANet yields 2.7% mIoU improvement compared to HRFormer-S, and the GFLOPs

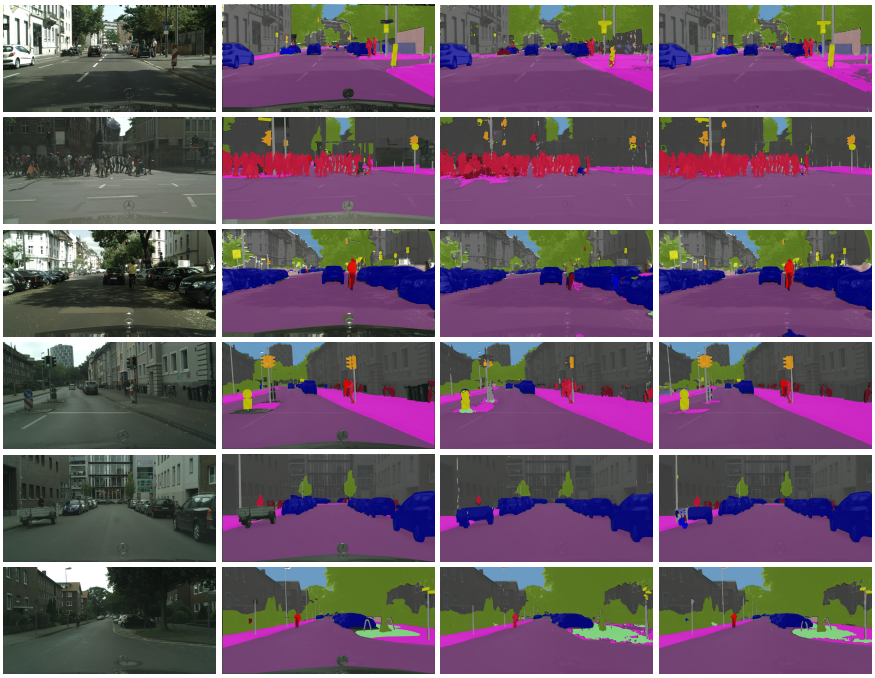

Fig. 4: Visual segmentation results of SegFormer-B0 and E2CANet-Tiny on the Cityscapes dataset

of E2CANet are seven times smaller than the latter on COCO-Stuff. Besides, our proposed method surpasses SegFormer-B1 by 2.3% mIoU and EfficientViT-B1 by 0.5% with less computational cost on Cityscapes. Although the mIoU of Mask2Former is 2.8% higher than E2CANet on ADE20K, the number of parameters and GFLOPs of our method are far less. We visualize some segmentation results on Cityscapes dateset in Figs. 4 and 5. We observe that E2CANet-Tiny obtains finer segmentation results than SegFormer-B0 in Fig. 4. As shown in Fig. 5, E2CANet achieves improved results compared to SegFormer-B1. Moreover, it yields competitive performance in comparion to HRFormer-S, and slightly better than the latter in some details. Compared to E2CANet-Tiny, E2CANet achieves higher segmentation performance, which demonstrate that as the number of parameters increase, the segmentation accuracy increases.

Comparisons with CNN-based Methods. We further report the comparison results of the E2CANet-Tiny against to other CNN-based methods on Cityscapes. Table 6 shows that ENet and SegNet obtain the worst segmentation results. The mIoU of E2CANet-Tiny is about 9.2% higher than that of MobileNetV2 with a little increase of the number of the parameters. It can be found that our network yields 1% mIoU improvement compared to PSPNet

Table 5: Comparisons with state-of-the-art transformer-based methods on ADE20K, COCO-stuff, and Cityscapes datasets.

Methods	params	ADE20K		COCO-stuff		Cityscapes	
	(M)	GFLOPs	mIoU (%)	GFLOPs	mIoU (%)	GFLOPs	mIoU (%)
SegFormer-B0 [39]	3.8	8.4	37.4	8.4	35.6	125.5	76.2
EfficientViT-B0 [3]	7	-	-	-	-	-	75.7
E2CANet-Tiny	4.4	7.9	41.9	7.9	36.5	75.1	79.4
HRFormer-S [42]	13.5	109.5	44.0	109.5	37.9	835.7	80.0
SegFormer-B1 [39]	13.7	15.9	42.2	15.9	40.2	243.7	78.5
EfficientViT-B1 [3]	48	-	-	-	-	-	80.3
Mask2Former [7]	44	70.1	47.2	-	-	52.3	79.4
E2CANet	14.1	15.2	44.4	15.2	40.6	152.0	80.8

Table 6: Comparisons with state-of-the-art CNN-based methods on Cityscapes.

Methods	Params (M)	GFLOPs	mIoU (%)
ENet [27]	0.4	45.36	58.3
SegNet [1]	29.5	3.8	58.3
CGNet [38]	0.5	6	64.8
MoblieNetV2 [30]	2.1	9.1	70.2
BiSeNetV2 [41]	-	21.1	72.6
PSPNet [44]	65.6	286	78.4
PIDNet-S [40]	73.6	47.6	78.6
DDRNet-23-Slim [26]	5.7	36.3	77.4
RegSeg [11]	3.34	39.1	78.3
PEM-STDC1 [5]	17	92	78.3
E2CANet-Tiny	4.4	75.1	79.4

with a dramatically decrease of the number of parameters. We also compare our method to three recent real-time semantic segmentation methods, e.g., PIDNet, DDRNet, and RegSeg. It can be find that our method achieves very competitive performance against these three methods. The results in Table 6 verify the efficiency and effectiveness of the E2CANet-Tiny.

5 Conclusion

In this paper, we present an efficient and effective convolutional attention network for semantic segmentation. The model introduces global attention and multi-scale attention by using cheap convolutional operations to highlight useful features and suppress unnecessary ones. Moreover, a lightweight All-MLP de-

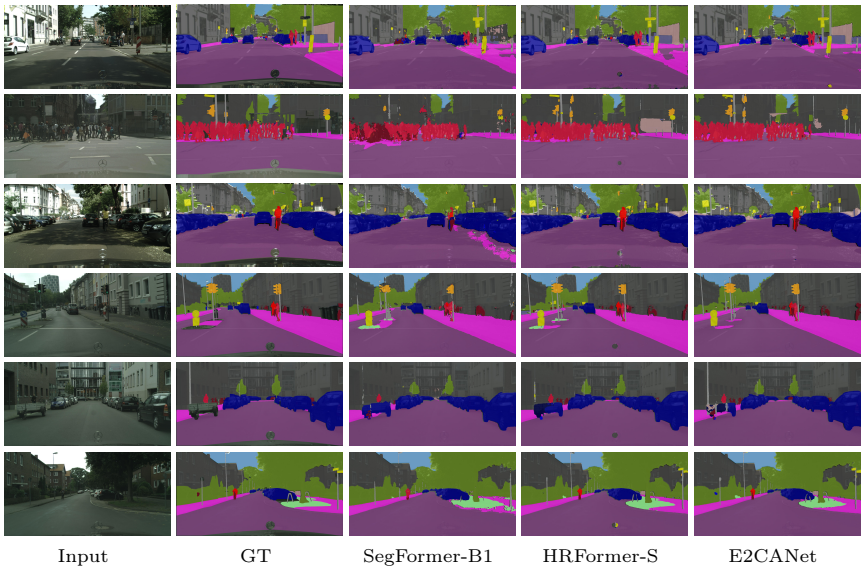

Fig. 5: Visual segmentation results of SegFormer-B1, HRFormer-S, and E2CANet on the Cityscapes dataset

coder is used to further reduce the computational complexity, which allows for a good trade-off between segmentation accuracy and computational cost. Extensive experiments on three commonly used datasets show that the proposed model achieves very competitive performance compares to some state-of-the-art methods.

In the future, we will explore the low rank approximation strategy as used in [35] [17] [20] to further improve the inference speed and reduce the number of parameters.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China (61873145), in part by the Natural Science Foundation of Shandong Province for Excellent Young Scholars (ZR2017JL029), and in part by the Science and Technology Innovation Program for Distinguished Young Scholars of Shandong Province Higher Education Institutions (2019KIN045).

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(12), 2481–2495 (2017) 10, 12
2. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1209–1218 (2018) 8

3. Cai, H., Li, J., Hu, M., Gan, C., Han, S.: Efficientvit: Lightweight multi-scale attention for high-resolution dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17302–17313 (2023) 10, 12
4. Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H.: Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019) 3
5. Cavagnero, N., Rosi, G., Cuttano, C., Pistilli, F., Ciccone, M., Averta, G., Cermelli, F.: Pem: Prototype-based efficient maskformer for image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15804–15813 (2024) 12
6. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2017) 3
7. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1290–1299 (2022) 10, 12
8. Contributors, M.: MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation (2020) 8
9. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3213–3223 (2016) 8
10. Ding, X., Zhang, X., Han, J., Ding, G.: Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11963–11975 (2022) 2
11. Gao, R.: Rethinking dilated convolution for real-time semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4675–4684 (2023) 10, 12
12. Gao, Z., Xie, J., Wang, Q., Li, P.: Global second-order pooling convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3024–3033 (2019) 4
13. Guo, M.H., Lu, C.Z., Hou, Q., Liu, Z., Cheng, M.M., Hu, S.M.: Segnext: Rethinking convolutional attention design for semantic segmentation. Advances in Neural Information Processing Systems **35**, 1140–1156 (2022) 2
14. Guo, M.H., Lu, C.Z., Liu, Z.N., Cheng, M.M., Hu, S.M.: Visual attention network. Computational Visual Media **9**(4), 733–752 (2023) 2
15. Guo, M.H., Xu, T.X., Liu, J.J., Liu, Z.N., Jiang, P.T., Mu, T.J., Zhang, S.H., Martin, R.R., Cheng, M.M., Hu, S.M.: Attention mechanisms in computer vision: A survey. Computational Visual Media **8**(3), 331–368 (2022) 4
16. Guo, Q., Fang, L., Wang, R., Zhang, C.: Multivariate time series forecasting using multiscale recurrent networks with scale attention and cross-scale guidance. IEEE Transactions on Neural Networks and Learning Systems pp. 1–15 (2023) 4
17. Guo, Q., Zhang, Y., Qiu, S., Zhang, C.: Accelerating patch-based low-rank image restoration using kd-forest and lanczos approximation. Information Sciences **556**, 177–193 (2021) 13
18. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7132–7141 (2018) 3

19. Hui, S., Guo, Q., Geng, X., Zhang, C.: Multi-guidance cnns for salient object detection. ACM Transactions on Multimedia Computing, Communications and Applications **19**(3), 1–19 (2023) 1
20. Jaderberg, M., Vedaldi, A., Zisserman, A.: Speeding up convolutional neural networks with low rank expansions. In: Proceedings of the British Machine Vision Conference. pp. 1–15 (2014) 13
21. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. Advances in Neural Information Processing Systems **28** (2015) 4
22. Jiao, J., Tang, Y.M., Lin, K.Y., Gao, Y., Ma, A.J., Wang, Y., Zheng, W.S.: Dilateformer: Multi-scale dilated transformer for visual recognition. IEEE Transactions on Multimedia **25**, 8906–8919 (2023) 4
23. Liu, W., Rabinovich, A., Berg, A.C.: Parsenet: Looking wider to see better. International Conference on Learning Representations. (2016) 3
24. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440 (2015) 3
25. Oberweger, M., Wohlhart, P., Lepetit, V.: Hands deep in deep learning for hand pose estimation. In: Proceedings of the 20th Computer Vision Winter Workshop. pp. 21–30 (2015) 1
26. Pan, H., Hong, Y., Sun, W., Jia, Y.: Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes. IEEE Transactions on Intelligent Transportation Systems **24**(3), 3448–3460 (2022) 10, 12
27. Paszke, A., Chaurasia, A., Kim, S., Culurciello, E.: Enet: A deep neural network architecture for real-time semantic segmentation. arXiv preprint arXiv:1606.02147 (2016) 10, 12
28. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems **32** (2019) 8
29. Qiao, Q., Wang, W., Qu, M., Su, K., Jiang, B., Guo, Q.: Medical image segmentation via single-source domain generalization with random amplitude spectrum synthesis. In: M. G. Linguraru et al. (Eds.) MICCAI 2024, LNCS 15009, pp. 435–445, 2024. pp. 435–445. Springer (2024) 1
30. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4510–4520 (2018) 10, 12
31. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9 (2015) 4
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017) 4
33. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. IEEE transactions on Pattern Analysis and Machine Intelligence **43**(10), 3349–3364 (2020) 4
34. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7794–7803 (2018) 3

35. Wen, W., Wu, C., Wang, Y., Chen, Y., Li, H.: Learning structured sparsity in deep neural networks. Advances in Neural Information Processing Systems **29** (2016) 13
36. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision. pp. 3–19 (2018) 2, 4
37. Wu, J., Jiao, J., Yang, Q., Zha, Z.J., Chen, X.: Ground-aware point cloud semantic segmentation for autonomous driving. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 971–979 (2019) 1
38. Wu, T., Tang, S., Zhang, R., Cao, J., Zhang, Y.: Cgnet: A light-weight context guided network for semantic segmentation. IEEE Transactions on Image Processing **30**, 1169–1179 (2020) 10, 12
39. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems **34**, 12077–12090 (2021) 10, 12
40. Xu, J., Xiong, Z., Bhattacharyya, S.P.: Pidnet: A real-time semantic segmentation network inspired by pid controllers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19529–19539 (2023) 10, 12
41. Yu, C., Gao, C., Wang, J., Yu, G., Shen, C., Sang, N.: Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. International Journal of Computer Vision **129**, 3051–3068 (2021) 10, 12
42. Yuan, Y., Fu, R., Huang, L., Lin, W., Zhang, C., Chen, X., Wang, J.: Hrformer: High-resolution vision transformer for dense predict. Advances in Neural Information Processing Systems **34**, 7281–7293 (2021) 4, 10, 12
43. Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., Agrawal, A.: Context encoding for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7151–7160 (2018) 3
44. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2881–2890 (2017) 3, 10, 12
45. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 633–641 (2017) 8