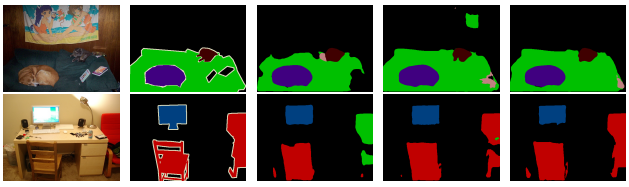
Supplement Material
SCARF: A Semantic Constrained Attention Refinement Network for Semantic Segmentation

Xiaofeng Ding[1], Chaomin Shen[2], Zhengping Che[3], Tieyong Zeng[4], Yaxin Peng[1]✉
[1]Department of Mathematics, School of Science, Shanghai University
[2]School of Computer Science *and* Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University [3]Didi Chuxing [4]The Chinese University of Hong Kong
{dxfeng, yaxin.peng}@shu.edu.cn, cmshen@cs.ecnu.edu.cn
chezhengping@didiglobal.com, zeng@math.cuhk.edu.hk

A. Supplement Material

We report some additional experimental results in the supplement material due to the page limit in the main submission. This supplement material consists of experiments on Pascal VOC 2012 dataset.

Visualization of SCARF model. We visualize the segmentation results in Fig. 1. The CA block improves the classification accuracy of the foreground categories since it eliminates the local noise by aggregating the category-prior contextual information. In addition, the BCA block further improves the segmentation results of background category, implying that the BCA block overcomes the contextual information confusion problem and enhances the differences between foreground and background categories.



(a) image (b) GT (c) baseline (d) CA (e) BCA

Figure 1. Visualization of segmentation results on PASCAL VOC 2012 validation set: (a) raw image, (b) ground truth, (c) baseline model, (d) SCARF with CA, (e) SCARF with BCA.

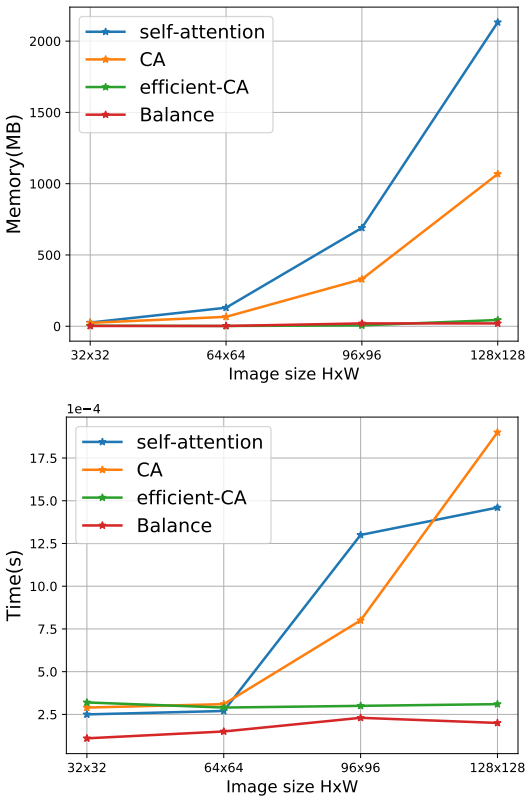


Figure 2. Comparison of computational costs (memory and time) for self-attention, CA, efficient CA and BCA methods.

Ablation studies for computation cost. In the manuscript, we have conducted the ablation studies for computation cost (memory and time) of our proposed model (full network). In this section, we further design one simple yet effective experiment to directly evaluate the computation cost of different contextual information aggregation methods (only module). Suppose the input feature channel number D is 128, and the category number C is 21. The image size HW varies from 32×32 to 128×128. Fig. 2 illustrates the computational cost of different methods, including self-

Method	mIoU	Memory	Time	Parameters
PSPNet	82.6	2300M	0.101s	513,965kb
DANet	82.6	2111M	0.107s	520,785kb
EncNet	82.9	1976M	0.102s	473,140kb
Baseline	82.6	1880M	0.092s	459,636kb
SCARF	**85.0**	1904M	0.099s	470,534kb

Table 1. Cost comparisons on the PASCAL VOC 2012 test set.

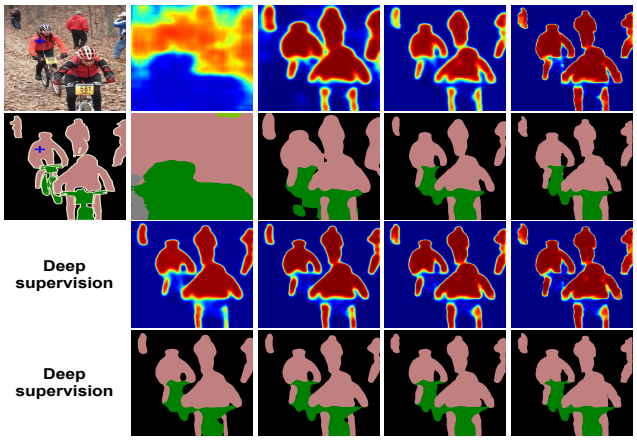
attention, CA, efficient CA and balance methods. The computation cost of self-attention and CA methods highly increases with the image size due to the generation of the attention A with computational complexity $O(HW \times HW)$. Our proposed efficient CA method highly reduces both the memory and time costs of the CA method with computational complexity $O(HW)$. In addition, our methods (efficient CA and balance) achieve extremely low computational costs over self-attention method for feature fusion of contextual information.

Furthermore, we provide the comparisons of computation cost with some recent methods, including PSPNet, DANet, and EncNet. All methods are tested with the same experiment environment, preprocessing, and encoder architecture. We report the segmentation accuracy on the PASCAL VOC test set, the computation cost (memory and time) of single scale inference, and the model parameters. Table 1 shows that SCARF achieves the superior performance over other methods with low computation and parameter costs. Specifically, our model do not lead to high computation and parameter costs since we adopt small feature channel $C = 128$ for all decoder layers. Most of the computation and parameter costs are from the encoder architecture.

Methods	DS	Non-dilated	Dilated
SCARF		80.6	81.0
SCARF	$\surd$	**80.7**	**81.6**

Table 2. Ablation results for deep supervision to SCARF network on PASCAL VOC 2012 validation set (mIoU (%))

Ablation studies of deep supervision to SCARF network. As shown in Table 2, deep supervision improves the performance of SCARF network from 80.6%/81.0% to 80.7%/81.6%, indicating the significant effect of deep supervision to the SCARF network. Furthermore, we show the segmentation of all layers in SCARF network with/without deep supervision on the PASCAL VOC 2012 validation set. As shown in Fig. 3, SCARF network has a rough trend to learn the segmentation even without deep supervision, implying the consistency of the SCARF network and deep supervision. SCARF network without deep supervision, however, leads to segmentation confusion shown in the first (second) row second column of Fig. 3, decreasing both the effect of attention process, which to some extent explains the influence of deep supervision to SCARF network.



(a) image (b) L^4 (c) L^3 (d) L^2 (e) L^1

Figure 3. Visualization of segmentation results and attention maps without (with) deep supervision for different layers on PASCAL VOC 2012 validation set. SCARF network without deep supervision (first and second rows) leads to the segmentation confusion of high level layer L^4. In addition, SCARF network tends to learn the segmentation iteratively from high level L^4 to low level L^1 both with and without deep supervision.

.