SCARF: A Semantic Constrained Attention Refinement Network for Semantic Segmentation

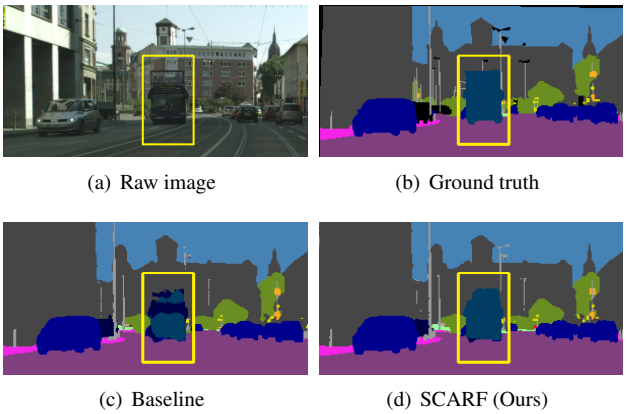
Xiaofeng Ding[1], Chaomin Shen[2], Zhengping Che[3], Tieyong Zeng[4], Yaxin Peng[1]✉
[1]Department of Mathematics, School of Science, Shanghai University
[2]School of Computer Science *and* Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University [3]Didi Chuxing [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

Abstract

Semantic segmentation has achieved great progress by exploiting the contextual dependencies. In this paper, we propose an end-to-end Semantic Constrained Attention Re-Finement (SCARF) network, based on semantic constrained contextual dependencies, to fully utilize the semantic information across different layers. Our novelties lie in the following aspects: Firstly, we present a general framework for capturing the non-local contextual dependencies. Secondly, within the framework, we introduce an efficient Category Attention (CA) block to capture semantic-related context by using the category constraint from coarse segmentation, which reduces the computational complexity from $O(n^2)$ to $O(n)$ for image with n pixels. Thirdly, we overcome the contextual information confusion problem by balancing the non-local contextual dependencies and the local consistency adaptively using a category-wise learning weight. Finally, we fully utilize the multi-scale semantic-related contextual information by refining the segmentation iteratively across layers with semantic constraint. Extensive evaluations demonstrate that our SCARF network significantly improves the segmentation results and achieves superior performance 85.0% mIoU on PASCAL VOC 2012, 55.0% mIoU on PASCAL Context, and 82.1% mIoU on Cityscapes.

1. Introduction

Semantic segmentation, aiming to assign pixel-wise category labels for a given image, has been widely applied to various real-world applications such as autonomous driving [8, 13] and medical diagnosing [21]. Being a fundamental task in computer vision, it achieves great success. Nevertheless, challenges still remain. For instance, Fig. 1(c) illustrates one problem of inaccurate segmentation led by objects of different categories (bus and truck) have similar feature representations. This paper attempts to provide a solution to this issue by aggregating features of same category to enhance the intra-class feature representation, as the result shown in Fig. 1(d).



(a) Raw image (b) Ground truth (c) Baseline (d) SCARF (Ours)

Figure 1. Example of a segmentation result from the Cityscapes validation set. Compared with the baseline model, our network called SCARF enhances the consistency of the intra-class features.

Pioneer work for semantic segmentation includes the Fully Convolutional Network (FCN) [26]. Nevertheless, due to the short-range receptive field caused by its fixed structure, FCN inherently does not make full use of the contextual information, leaving room for improvement. Recently, the contextual information has shown its capability for semantic segmentation. Specifically, it is aggregated with multi-scale dilation convolutions via the Atrous Spatial Pyramid Pooling (ASPP) module in DeepLab methods [6, 5], and extracted with pyramid pooling module in PSPNet [46]. The contextual information, however, is aggregated non-adaptively in these methods, not satisfying the

different contextual dependencies to different pixels.

To aggregate the contextual information adaptively, the attention-based methods are extensively studied. Point-wise attention based methods [47, 38] connect all pixels in the feature maps through the predicted attention map. Meanwhile, self-attention [32], a non-local contextual aggregation method, is employed by segmentation [36, 12, 45, 19] to aggregate the contextual information for each position. These methods, however, have two problems. First, they lead to the high computational complexity $O(HW \times HW)$, with HW the image size. Second, the attention mechanism is not clear with decreased pair-wise relationship from structure reasoning, since the attention map is completely learned by back-propagation without prior information.

To address the above two problems, we propose a novel network, called Semantic Constrained Attention ReFinement (SCARF), based on the semantic constrained contextual dependencies. We first present a general framework for capturing the non-local contextual information. Within this framework, we introduce a Category Attention (CA) block to capture the semantic context by using the category constraint from coarse segmentation. More specifically, we utilize the covariance matrix of coarse segmentation to generate the attention map, since the inner product of category probability vectors inherently represents the pairwise relationship of different pixels, enhancing the structure reasoning of the attention block. With the help of the associative law, we provide an efficeint version of CA, called the efficient CA, reducing the computational complexity from $O(HW \times HW)$ to $O(HW)$. In addition, given different contextual dependencies to different categories, we adaptively balance the non-local contextual dependencies and the local consistency by introducing a category-wise learning weight, overcoming the contextual information confusion problem. Finally, our model refines the segmentation iteratively across layers, fully utilizing the semantic information. In summary, our main contributions are as follows:

- We introduce an efficient CA block to capture semantic-related context by using the category constraint from coarse segmentation, reducing the computational complexity from $O(HW \times HW)$ to $O(HW)$.
- Our model adaptively balances the non-local contextual dependencies and the local consistency by introducing a category-wise learning weight, overcoming the contextual information confusion problem.
- Our model refines the segmentation iteratively across layers, fully using coarse segmentation information.

2. Related Work

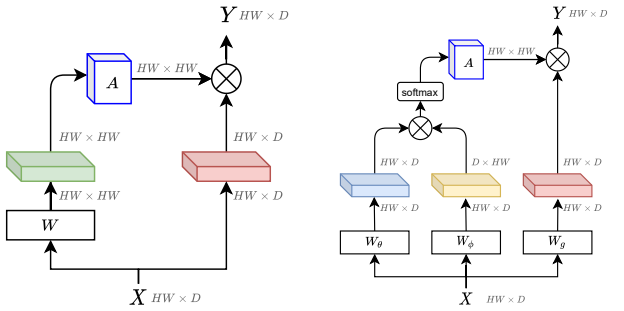
Semantic segmentation. As a pioneer, FCN [26] first introduces the pretrained network for semantic segmentation. Following FCN, various nice works have made remarkable progress by exploiting the feature context, including single-branch and multi-branch methods. For single-branch methods, SegNet [2], U-Net [29], RefineNet [25] adopt encoder-decoder structure to fuse multi-level features; Deeplab [6, 5] and PSPNet [46] collect the multi-scale context by designing the pyramid modules; CCNet [20], CFNet [45] and DANet [12] employ the self-attention method [36] to aggregate long-range spatial information; Conditional random field methods [22, 48, 33, 4] are also employed to harvest the feature context. For multi-branch methods, DFN [40], UPerNet [37] and GSCNN [31] improve the segmentation branch by introducing other visual concepts, such as boundary, textual and shape branches; FuseNet [14], RTFNet [30], DFM-RTFNet [35] and SNE-RoadSeg+ [34] learn informative features with data-fusion, which improves segmentation by using addition visual modalities, such as depth and thermal images.

Attention mechanism. A rich body of literature investigates approaches for attaching attention mechanism to the Deep Neural Networks (DNNs). Inspired by SENet [18], which attaches attention on the channel to select desired feature maps, BiSeNet [39], DFN [40] and EncNet [44] utilize channel-wise attention for semantic segmentation. Meanwhile, [36] models pair-wise relationships by calculating the correlation matrix of feature maps as the attention map, and it matches the self-attention mechanism [32]. Following this, DANet [12] takes both channel-wise attention and self-attention into account, and CCNet [20] replaces the self-attention block with criss-cross attention to reduce computational complexity. Similarly, PSANet [47] and CPNet [38] obtain the pair-wise attention map adaptively via a single convolution layer. In addition, AA [10] integrates different types of attention modules and proposes an attention aggregation framework.

Coarse-to-fine methods. Coarse-to-fine is a hierarchical context mining mechanism widely used in applications such as face detection [11], shape detection [1], and optical flow [3]. Recently, a lot of deep network based segmentation methods [25, 43, 50] adopt coarse-to-fine strategy. RefineNet [25] proposes a multi-path refinement network to enable high-resolution prediction using long-range residual connections. ACFNet [43] utilizes the coarse segmentation to calculate the class center of each category on fine stage. CiSS-Net [50] introduces reinforce learning to realize the coarse-to-fine strategy by treating the coarse segmentation as environment. In this paper, we propose a simple yet effective coarse-to-fine segmentation framework by refining the segmentation with semantic constraint.

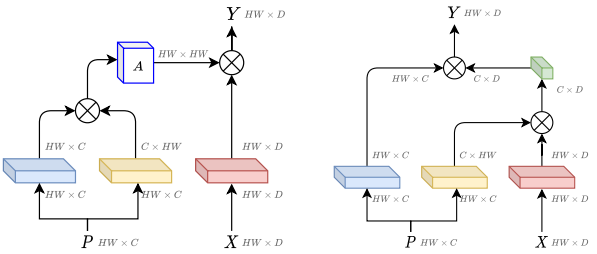
3. Methodology

We first introduce a general definition of non-local operation for contextual information aggregation in Sec. 3.1,



(a) point-wise attention (b) self-attention

Figure 2. Diagrams of two typical attention methods for contextual information aggregation. The output signal Y is aggregated by the attention map A.



(a) category attention (b) efficient category attention

Figure 3. Our proposed non-local blocks with category prior: (a) Category attention block with computation complexity $O(HW \times HW)$, (b) Efficient category attention block with computation complexity $O(HW)$.

then present an efficient CA block in Sec. 3.2. Next, we adaptively balance the non-local and local information by a category-wise learning weight in Sec. 3.3. Finally, we construct the SCARF network in Sec. 3.4.

3.1. General framework of non-local block

The general definition of the non-local operation for contextual information aggregation is:

$$\boldsymbol{y}_i = \mathcal{N}(\sum_j a_{ij} \cdot g(\boldsymbol{x}_j)), \tag{1}$$

with $\boldsymbol{x}_j \in \mathbb{R}^D$ the input signal and $\boldsymbol{y}_i \in \mathbb{R}^D$ the output signal. The unary function g computes the embedding of input signal. a_{ij} is a scalar representing the pairwise relationship between positions i and j. $\mathcal{N}$ is a normalization operator.

The general definition (1) includes two special cases of typical attention based methods, i.e., the point-wise attention and self-attention methods:

- For point-wise attention methods [47, 38] shown in Fig. 2(a), the unary function g is the identity embedding: $g(\boldsymbol{x}_j) = \boldsymbol{x}_j$, and the attention map is $A = f(X)$, where f is a series of convolution layers with batch normalization and activation function.
- For self-attention methods [36, 45, 12] shown in Fig. 2(b), the unary function $g(\boldsymbol{x}_j) = W_g \boldsymbol{x}_j$ is a linear embedding, where W_g is a weight matrix to be learned. The attention map $a_{ij} = e^{\theta(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j)}$, where $\theta(\boldsymbol{x}_i) = W_\theta \boldsymbol{x}_i$ and $\phi(\boldsymbol{x}_j) = W_\phi \boldsymbol{x}_j$ are embeddings. The output signal is normalized by a factor $\sum_j a_{ij}$.

Here $X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{HW}]^T$ is the input signal matrix, and $Y = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_{HW}]^T$ is the output signal matrix, with HW the image size and D the channel dimension. $A = (a_{ij})_{HW \times HW}$ is the spatial attention map.

The non-local block is significant for semantic segmentation by capturing the long-range contextual dependencies. However, it has two flaws. First, the relationship representation of the attention mechanism is less clear due to the lack of context prior on the attention map. Second, the non-local block needs to generate the spatial attention map to measure the impact between different positions, leading to high computational complexity $O(HW \times HW)$.

To overcome these two flaws, we apply the coarse-to-fine strategy with supervision to the coarse segmentation. To make the attention mechanism clear, we represent the pairwise relationship by the similarity of coarse segmentation probability vectors. In addition, the main computation and time cost of non-local block is from the attention map. We change the calculation order of non-local block to reduce the computation cost.

3.2. Efficient category attention

Attention with coarse segmentation. Here we first propose a category prior non-local block, called CA block, as follows. We set the unary function g to the identity embedding $g(\boldsymbol{x}_j) = \boldsymbol{x}_j$. As shown in Fig. 3(a), we construct the non-local block by representing the definition (1) as

$$\boldsymbol{y}_i = \frac{1}{\sum_j \boldsymbol{p}_i^T \boldsymbol{p}_j} \sum_j (\boldsymbol{p}_i^T \boldsymbol{p}_j) \boldsymbol{x}_j, \tag{2}$$

where $a_{ij} = \boldsymbol{p}_i^T \boldsymbol{p}_j$ is the pair-wise relationship, $\boldsymbol{p}_i \in \mathbb{R}^C$ is the coarse segmentation probability vector in position i, with C the category number, and $[\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_{HW}]^T$ constitutes the coarse segmentation map P. The inner product of category probability vectors inherently represents the similarity of two vectors. Then, the contextual aggregation from input to output signal is:

$$Y = NAX = NPP^T X. \tag{3}$$

Here N is the normalization matrix

$$N = \begin{pmatrix} \frac{1}{\boldsymbol{p}_1^T \sum_j \boldsymbol{p}_j} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{1}{\boldsymbol{p}_{HW}^T \sum_j \boldsymbol{p}_j} \end{pmatrix}_{HW \times HW} \tag{4}$$

which is efficiently generated and applied by the element-wise calculation.

Reduction of computational complexity. We improve the structure of non-local block from CA block to efficient CA block by changing the calculation order. In fact, the attention map A shown in Fig. 3(a) is the covariance matrix of coarse segmentation $A = PP^T$, which needs to be calculated generally. But in our case, it is not necessary due to the associative law. As depicted in Fig. 3(b), we employ the following calculation order

$$Y = NP(P^T X) \tag{5}$$

to avoid calculating the covariance matrix. Specifically, for CA block, the generation of covariance matrix $A = PP^T$ and the aggregation process $(PP^T)X$ result in the computational cost $\mathcal{O}(HW \times HW \times C)$ and $\mathcal{O}(HW \times HW \times D)$, respectively. The CA block thus has quadratic complexity in the image size HW, causing high computational cost since HW usually is very large. Differently, for efficient CA block, the two matrix multiplication process $P^T X$ and $P(P^T X)$ lead to computational cost $\mathcal{O}(HW \times D \times C)$ and $\mathcal{O}(HW \times D \times C)$, respectively. Therefore, the efficient CA block reduces the computational cost from quadratic to linear complexity in the number of HW, highly increasing the computational efficiency.

3.3. Non-local and local information

Balance of non-local and local information. The non-local operation (2) reveals that the output signal $\boldsymbol{y}_i$ is the expectation of input signal $\boldsymbol{x}_j$ related to the category similarity $\boldsymbol{p}_i^T \boldsymbol{p}_j$, eliminating the local noise by aggregating the global contextual information. This aggregation method, however, is not appropriate for some categories whose feature information varies in a wide range. Aggregation of such category may lead to the contextual information confusion, which means the method does not capture the useful class representation, and even provides the unreasonable information to mislead the feature representation. For instance, if the background category contains both the sky and building area, the aggregation process will result in the contextual information confusion since their feature information would be totally different. Therefore, we propose a balance operation as:

$$\boldsymbol{z}_i = \lambda_i \cdot \boldsymbol{y}_i + (1 - \lambda_i) \cdot \boldsymbol{x}_i. \tag{6}$$

Here $\boldsymbol{z}_i$ is the fusion of non-local signal $\boldsymbol{y}_i$ and local signal $\boldsymbol{x}_i$. $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \cdots, \lambda_{HW}]^T$ is an adaptively learning weight to balance the non-local and local information with $\lambda_i \in [0, 1]$ the weight in position i. Equation (6) degenerates to (2) when the entry $\lambda_i = 1$ for all i, which means only the non-local term contributes to the output.

Category-wise adaptively learning weight. In order to overcome the contextual information confusion caused by the imbalance within category, we set the weight

$$\lambda_i = \boldsymbol{p}_i^T \boldsymbol{w}, \tag{7}$$

where $\boldsymbol{w} = [w_1, w_2, \cdots, w_C]^T$ is a category-wise weight and the entry $w_c \in [0, 1]$ is the non-local weight for category c, which is adaptively updated by the deep network.

We employ the sigmoid function to restrict all entries of the category-wise weight $\boldsymbol{w}$ from 0 to 1 to ensure $\lambda_i \in [0, 1]$ for all i, and initialize all entries close to 1 to enhance the effect of non-local contextual information. By learning the weight $\boldsymbol{w}$ through back-propagation, our model adaptively balances the non-local and local information, overcoming the contextual information confusion problem. Then we construct the balance block as:

$$Z = \Lambda Y + (I - \Lambda)X, \tag{8}$$

where I is the identity matrix, and

$$\Lambda = \begin{pmatrix} \boldsymbol{p}_1^T \boldsymbol{w} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \boldsymbol{p}_{HW}^T \boldsymbol{w} \end{pmatrix}_{HW \times HW} \tag{9}$$

is the weight matrix efficiently generated and applied by the element-wise calculation.

Balance Category Attention block. By balancing the non-local and local information (8), we construct the Balance Category Attention (BCA) block as:

$$Z = \Lambda NPP^T X + (I - \Lambda)X. \tag{10}$$

The flowchart of BCA block (10) is shown in Fig. 4.

3.4. Semantic constrained attention refinement

Category attention residual module. Based on the BCA block, we propose a category prior coarse-to-fine module, called the Category Attention Residual (CAR) module. As shown in Fig. 4, CAR takes the coarse logit L_{coarse} (last feature map before the coarse segmentation P) as the input to generate the fine logit L_{fine}. Specifically, we first employ BCA block (10) to aggregate the long-range contextual information of input signal X with the coarse segmentation P. The output feature maps are given by concatenating the output signal Z of BCA block with

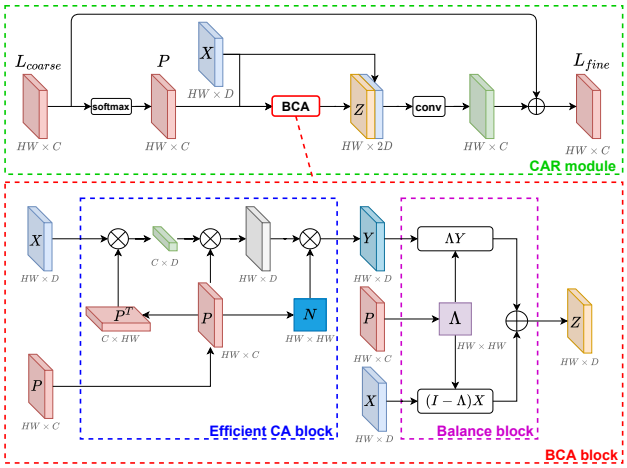


Figure 4. The Category Attention Residual (CAR) module.

the input signal X to avoid the information loss. Finally, several convolution layers are utilized to aggregate the feature information and generate the residual of coarse and fine logits.

Semantic constrained attention refinement network. Based on the CAR module, we construct a multi-scale coarse-to-fine network, called SCARF network. The overall architecture of our network is depicted in Fig. 5. We employ the pretrained ResNet-101 [17] backbone with decoder layers as our base network, and employ the CAR module to refine the segmentation iteratively across layers:

$$L^s = \mathrm{CAR}(L^{s+1}, X^s), \tag{11}$$

where L^s represents the logit of scale $s \in [1, 4]$, and CAR is the CAR module. For each scale s, the feature map X^s is obtained by combining high level feature X^{s+1} and low level feature F^s, using concatenation and two convolution layers. The CAR module takes the coarse logit L^{s+1} and the feature map X^s as the input to obtain the fine logit L^s, realizing a coarse-to-fine process from high to low level. By iteratively refining the logits, we get the output segmentation prediction L^1.

Loss function. Cross-entropy (CE) is utilized as the loss function. Besides, we apply deep supervision [23] to all logits to constrain the predicted segmentation probability maps. To minimize the difference of the network output to the ground truth, we weight the loss of different scales to a certain proportion. Suppose $\{\alpha_s\}_{s=1}^4$ are coefficients to balance the segmentation losses for different scales. The final loss is computed by

$$l = \sum_{s=1}^{4} \alpha_s \cdot \mathrm{CE}(L^s, \mathrm{GT}), \tag{12}$$

where GT is the ground truth, and the coefficient α_s decreases with the scale s. We empirically set $\alpha_1 = 1$ and the proportion between coefficients of adjacent scales $\frac{1}{2}$.

4. Experiments

In this section, we introduce the implementation details and show a series of evaluation on PASCAL VOC 2012 [9], PASCAL Context [27] and Cityscapes [8] datasets.

4.1. Implementation details

Network and loss. We employ ResNet-101 [17] pretrained on ImageNet with ASPP [6] as the baseline backbone, and follow [6, 44, 12] to apply the dilated strategy to the last two ResNet blocks. Synchronized batch normalization [44] is utilized in the training phase. In addition, we adopt deep supervision [23] to enhance the gradient flow of multi-scale models. Since deep supervision improves the performance of all the models, we employ it for all the experiments (including baseline with decoders). All the predictions are up-sampled by the bilinear interpolation to compute the segmentation loss.

Optimization. Following [6, 44, 38], we employ a poly learning rate scheduling $\gamma = \gamma_0 \cdot (1 - \frac{N_{iter}}{N_{total}})^{0.9}$, with γ_0 the base learning rate, N_{iter} the current iteration number, and N_{total} the total iteration number. We set the base learning rate to 0.001 for all datasets. Momentum and weight decay coefficients of Stochastic Gradient Descent (SGD) optimizer are set to 0.9 and 0.0001, respectively. Besides, the batch size is set to 8 for Cityscapes and 16 for other datasets. The training time is set to 80 epochs for PASCAL VOC augmented set, 50 epochs for fine-tuning on PASCAL VOC train+validation set, 80 epochs for PASCAL Context, and 240 epochs for Cityscapes. For Cityscapes dataset, we adopt the warm-up training strategy [17] with 5 epochs, the Online Hard Example Mining (OHEM) [42], and a hierarchy of grids (multi-grid) [7, 12] of different sizes (4, 8, 16) in the last ResNet block.

Data augmentation. In the training phase, we apply random horizontal flip to the input images, and then scale the images to the ratio of 0.5 to 2.0. Finally, we randomly crop the images into the training size (480×480 for PASCAL VOC 2012, 544×544 for PASCAL Context, 768×768 for Cityscapes).

Inference. We conduct comparison experiments with state-of-the-art algorithms on PASCAL VOC 2012, PASCAL Context and Cityscapes datasets. During inference, we crop images into training image size and feed them into the network for Cityscapes dataset. For other datasets, we get the prediction map by feeding the full image into the network. In addition, multi-scale inputs and left-right flip are employed during evaluation. We follow [44] to average the segmentation probability maps from multi scales for inference, where the scales is set to {0.75, 1.0, 1.25, 1.5, 1.75,

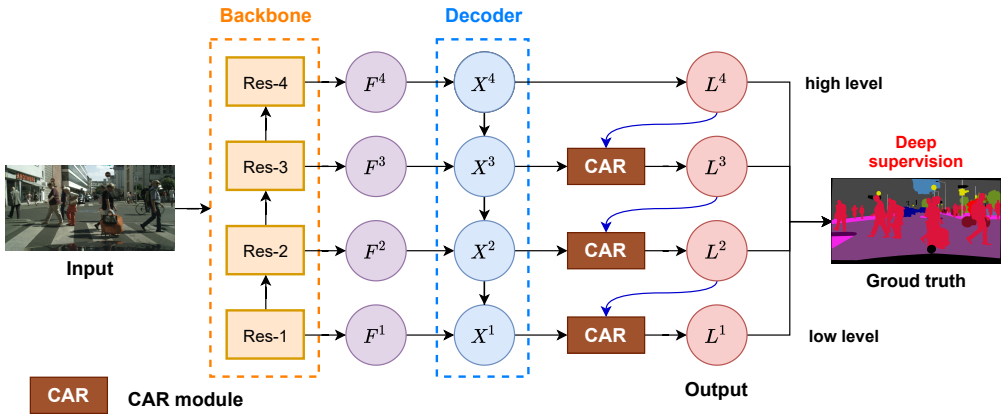


Figure 5. An overview of the SCARF network, with $\{F^s\}_{s=1}^4$ features of ResNet backbone, $\{X^s\}_{s=1}^4$ the decoder features and $\{L^s\}_{s=1}^4$ output logits (segmentation) of different scales. The model iteratively refines the segmentation to generate the output segmentation L^1.

2.0} for Cityscapes and {0.5, 0.75, 1.0, 1.25, 1.5, 1.75} for other datasets. We adopt mean intersection of union (mIoU) as the evaluation metric for PASCAL Context dataset. For PASCAL VOC 2012 and Cityscapes datasets, we utilize the public server for evaluation.

4.2. Evaluations on PASCAL VOC 2012 dataset

Dataset description. PASCAL VOC 2012 dataset is one of the gold-standard benchmarks for semantic segmentation, including 10,582 augmented, 1,464 train, 1,449 validation, and 1,456 test images. The performance is measured in terms of the mIOU across 21 classes.

Ablation studies of the CAR module to different layers. In order to validate the effectiveness of the CAR module, we apply it to the baseline model iteratively across layers. Besides, both non-dilated and dilated cases (dilated strategy for the backbone) are conducted for ablation studies. For fair comparisons, decoder layers are applied to both the baseline model and the SCARF network from high to low levels (level 4 to 1). As shown in Table 1, SCARF outperforms the baseline model on both non-dilated and dilated cases. In addition, the multi-scale refinement of CAR module highly improves the performance.

Method	4	3	2	1	Non-dilated	Dilated
Baseline	√				73.4	80.0
Baseline	√	√			78.7	80.0
Baseline	√	√	√		79.1	80.2
Baseline	√	√	√	√	78.8	80.4
SCARF	√	√			79.3	81.1
SCARF	√	√	√		79.7	81.2
SCARF	√	√	√	√	**80.7**	**81.6**

Table 1. Ablation results for CAR module to different layers on PASCAL VOC 2012 validation set (mIoU (%))

Ablation studies of BCA block. We also study the effect of the BCA block, as shown in Table 2. We apply CA block and balance block to the baseline model with full encoder layers (4321), respectively. SCARF network with CA block (3) achieves 0.8%/1.0% improvement of the model performance, and the network with both the CA and balance blocks (10) makes 1.9%/1.2% improvement.

Method	CA	Balance	Non-dilated	Dilated
Baseline			78.8	80.4
SCARF	√		79.6	81.4
SCARF	√	√	**80.7**	**81.6**

Table 2. Ablation results for BCA block on PASCAL VOC 2012 validation set (mIoU (%))

Variation of balance weight w for different categories. The category-wise weight $\boldsymbol{w}$ balances the non-local and local information. In our experiments, we employ sigmoid function to restrict the category-wise weight $\boldsymbol{w}$ from 0 to 1, and initialize all entries close to 1 to enhance the effect of non-local contextual information. Table 3 displays the value of weight $\boldsymbol{w}$ for all categories after training. As we assumed, the non-local term plays the main role of the balance block for almost all categories except background category. For background category, the balance weight decreases to 0.00%, which means the network adaptively switches the contextual information from non-local to local term, verifying the contextual information confusion of background category and the importance of local term.

Ablation studies of computation cost. We study the computation cost of the BCA block by applying it to the baseline model. We report the model performance, memory and time cost in the inference stage with the batch size 1. As shown in Table 4, the first and second rows illustrate the results of the baseline model without and

BG	plane	bike	bird	boat	bottle	bus
0.00	99.81	99.75	99.92	99.83	99.91	99.91
car	cat	chair	cow	table	dog	horse
99.75	99.88	99.95	99.86	99.91	99.86	99.92
motor	person	plant	sheep	sofa	train	TV
99.80	99.94	99.85	99.91	99.86	99.81	99.91

Table 3. Value of balance weight $\boldsymbol{w}$ after training, with the weight $\boldsymbol{w}$ initialized close to 1 for all categories. BG is the background category.

with decoder layers, respectively. For further comparison with typical non-local methods, we add the self-attention method [32] shown in the third row. Both the CA block and self-attention block lead to high memory cost 2062M. The efficient CA block highly reduces the memory cost of CA block to 24M, overcoming the high computation cost shortcoming of attention methods. In addition, the BCA block further improves the model performance from 81.4% to 81.6% without extra computation cost, which validates the effectiveness of the balance block.

Method	Memory	Time	mIoU
Baseline (4)	1880M	0.092s	80.0%
Baseline (1234)	1902M (+22)	0.097s	80.4%
Self-attention [32]	3942M (+2062)	0.112s	81.2%
CA	3942M (+2062)	0.108s	81.4%
Efficient CA	1904M (+24)	0.098s	81.4%
BCA	1904M (+24)	0.099s	**81.6**%

Table 4. Ablation of computation cost on PASCAL VOC 2012 validation set.

Comparison with state-of-the-art. We evaluate the performance of SCARF network on the PASCAL VOC 2012 segmentation dataset. The comparison results are shown in Table 5. SCARF achieves 85.0% mIoU on the test set, outperforming previous works without COCO pretraining.

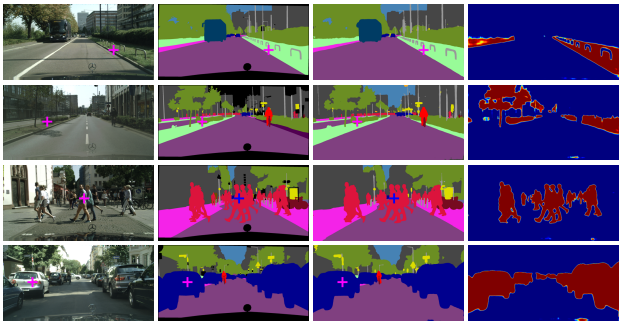
4.3. Evaluations on Cityscapes dataset

Dataset description. Cityscapes dataset is a high-resolution city street parsing dataset. 2,975, 500 and 1,525 fine annotated images captured from 50 different cities are provided for training, validation and testing. In our experiments, we only utilize the fine annotations including 19 categories for evaluation.

Visualization of the category attention maps. To understand the category contextual dependencies, we visualize the pairwise similarity a_{ij} between a given pixel i and other pixels for all j in the attention map A. As depicted in Fig. 6, CA block aggregates the contextual information from pixels with similar class information, enhancing both consistency of intra-class features and differences of inter-class features.

Methods	Reference	Backbone	mIoU (%)
PSPNet [46]	CVPR 2017	Res101	82.6
DFN [40]	CVPR 2018	Res101	82.7
EncNet [44]	CVPR 2018	Res101	82.9
DANet [12]	CVPR 2019	Res101	82.6
CFNet [45]	CVPR 2019	Res101	84.2
APCNet [16]	CVPR 2019	Res101	84.2
DMNet [15]	ICCV 2019	Res101	84.4
SANet [49]	CVPR 2020	Res101	83.2
SpyGR [24]	CVPR 2020	Res101	84.2
SCARF		Res101	**85.0**

Table 5. Quantitative evaluations on PASCAL VOC 2012 test set without pretraining on COCO dataset.



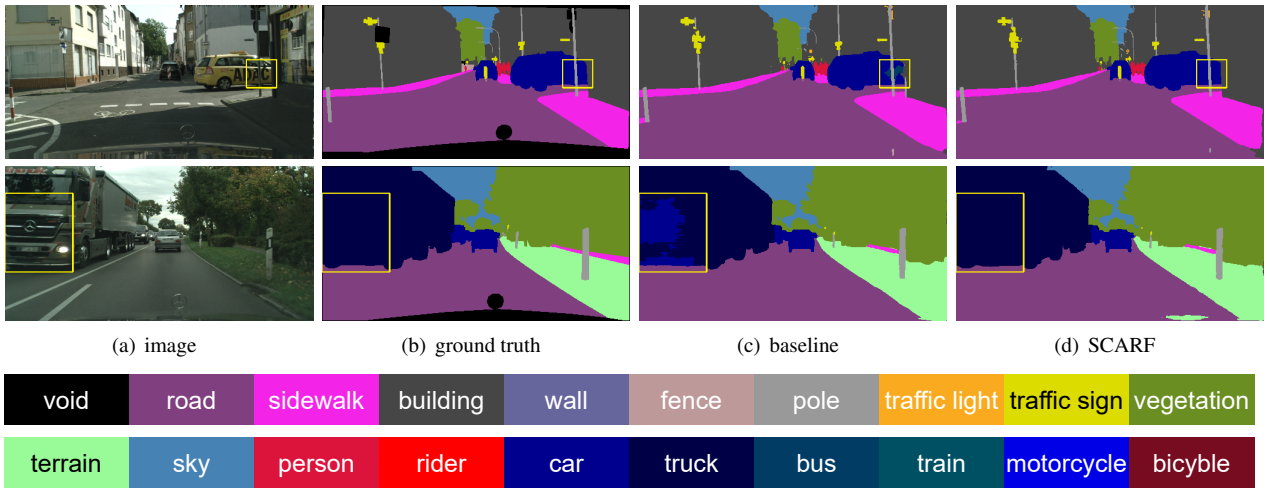
(a) image (b) ground truth (c) segmentation (d) attention map

Figure 6. Pairwise similarity visualization of all pixels to a given pixel. The selected pixel i is marked as $+$ in (a) raw image, (b) ground truth, and (c) segmentation. (d) Attention map shows the pairwise similarity A_{ij} of all pixels for all j to selected pixel i.

.

Visualizations of SCARF model. For further understanding of our model, we visualize the segmentation results of the baseline model and SCARF on Cityscapes datasets shown in Fig. 7. The SCARF model improves the classification accuracy since it eliminates the local noise by aggregating the category-prior contextual information, overcoming the contextual information confusion problem and enhancing the differences between different categories.

Comparison with state-of-the-art. We compare SCARF network with the existing methods on Cityscapes test set. As illustrated in Table 6, the SCARF network achieves 82.1% mIoU on the Cityscapes test set, outperforming other methods. Among previous works, ACFNet [43] utilizes the coarse segmentation to calculate the center features of different classes. OCR [41] calculates the relationship between pixel and category with the coarse segmentation. Different from them, our SCARF network exploits the coarse segmentation by efficiently capturing the pairwise relationship of different pixels and obtain better performance.



(a) image (b) ground truth (c) baseline (d) SCARF

Figure 7. Visualization results on Cityscapes validation set.

Method	Reference	Backbone	mIoU (%)
RefineNet [25]	CVPR 2017	Res101	73.6
PSPNet [46]	CVPR 2017	Res101	78.4
DFN [40]	CVPR 2018	Res101	79.3
PSANet [47]	ECCV 2018	Res101	80.1
CFNet [45]	CVPR 2019	Res101	79.6
SeENet [28]	ICCV 2019	Res101	81.2
DANet [12]	CVPR 2019	Res101	81.5
ACFNet [43]	ICCV 2019	Res101	81.8
CPNet [38]	CVPR 2020	Res101	81.3
SpyGR [24]	CVPR 2020	Res101	81.6
OCR [41]	ECCV 2020	Res101	81.8
SCARF		Res101	**82.1**

Table 6. Quantitative evaluations on Cityscapes test set. We list the methods training with the fine data.

Method	Reference	Backbone	mIoU (%)
RefineNet [25]	CVPR 2017	Res152	47.3
PSPNet [46]	CVPR 2017	Res101	47.8
EncNet [44]	CVPR 2018	Res101	51.7
DANet [12]	CVPR 2019	Res101	52.6
CFNet [45]	CVPR 2019	Res101	54.0
APCNet [16]	ICCV 2019	Res101	54.7
SpyGR [24]	CVPR 2020	Res101	52.8
CPNet [38]	CVPR 2020	Res101	53.9
SANet [49]	CVPR 2020	Res101	54.4
OCR [41]	ECCV 2020	Res101	54.8
SCARF		Res101	**55.0**

Table 7. Quantitative evaluations on PASCAL Context validation set. (Note: mIoU on 60 classes w/ background.)

4.4. Evaluations on PASCAL Context dataset

Dataset description. PASCAL Context dataset is a scene parsing dataset, including 4,998 training and 5,105 testing images. We conduct our experiments on the most frequent 59 classes with background class (60 classes in total).

Comparison with state-of-the-art. We conduct comparisons with state-of-the-art methods [44, 12] on PASCAL Context dataset. As shown in Table 7, SCARF network achieve 55.0% mIoU, highly outperforming other methods.

5. Conclusion

We have proposed the SCARF network. We first introduce the semantic constrained attention mechanism to capture contextual information from coarse segmentation. Then, we present the efficient CA block to capture semantic constrained context. Besides, we adaptively balance the non-local and local information by introducing a category-wise attention weight. Finally, our model refines the segmentation iteratively across layers with semantic constraint. Extensive evaluations demonstrate that our model can efficiently capture the long-range contextual information with semantic constraint layer-by-layer, enhancing the structure reasoning of the model.

Acknowledgement

This work was supported in part by the National Key R&D Program of China (2021YFE0203700), the National Natural Science Foundation of China under Grant 11771276, Grant NSFC/RGC (N_CUHK 415/19), Grants RGC (14300219, 14302920, 14301121), and CUHK Direct Grant for Research (4053405, 4053460), Shanghai Science and Technology Innovation Action Plan (18441909000, 21S31901000).

References

[1] Yali Amit, Donald Geman, and Xiaodong Fan. A coarse-to-fine strategy for multiclass shape detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(12):1606–1621, 2004.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

[3] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *European Conference on Computer Vision*, pages 25–36, 2004.

[4] Siddhartha Chandra, Nicolas Usunier, and Iasonas Kokkinos. Dense and low-rank Gaussian CRFs using deep embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5103–5112, 2017.

[5] Liangchieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision*, pages 833–851, 2018.

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017.

[7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[10] Rui Fan, Hengli Wang, Mohammud J Bocus, and Ming Liu. We learn better road pothole detection: from attention aggregation to adversarial domain adaptation. In *European Conference on Computer Vision*, pages 285–300, 2020.

[11] Francois Fleuret and Donald Geman. Coarse-to-fine face detection. *International Journal of Computer Vision*, 41(1-2):85–107, 2001.

[12] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019.

[13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012.

[14] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian Conference on Computer Vision*, pages 213–228, 2016.

[15] Junjun He, Zhongying Deng, and Yu Qiao. Dynamic multi-scale filters for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3562–3572, 2019.

[16] Junjun He, Zhongying Deng, Lei Zhou, Yali Wang, and Yu Qiao. Adaptive pyramid context network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7519–7528, 2019.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[18] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[19] Ping Hu, Fabian Caba, Oliver Wang, Zhe Lin, Stan Sclaroff, and Federico Perazzi. Temporally distributed networks for fast video semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8818–8827, 2020.

[20] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 603–612, 2019.

[21] Ma Jun, Wang Yixin, An Xingle, Ge Cheng, Yu Ziqi, Chen Jianan, Zhu Qiongjie, Dong Guoqiang, He Jian, He Zhiqiang, Ni Ziwei, and Yang Xiaoping. Towards efficient COVID-19 CT annotation: A benchmark for lung and infection segmentation. *arXiv preprint arXiv:2004.12537*, 2020.

[22] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *Advances in Neural Information Processing Systems*, pages 109–117, 2011.

[23] Chenyu Lee, Saining Xie, Patrick W Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *International Conference on Artificial Intelligence and Statistics*, pages 562–570, 2015.

[24] Xia Li, Yibo Yang, Qijie Zhao, Tiancheng Shen, Zhouchen Lin, and Hong Liu. Spatial pyramid based graph reasoning for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8950–8959, 2020.

[25] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1925–1934, 2017.

[26] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[27] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014.

[28] Yanwei Pang, Yazhao Li, Jianbing Shen, and Ling Shao. Towards bridging semantic gap to improve semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4230–4239, 2019.

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 234–241, 2015.

[30] Yuxiang Sun, Weixun Zuo, and Ming Liu. RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, 4(3):2576–2583, 2019.

[31] Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5229–5238, 2019.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[33] Raviteja Vemulapalli, Oncel Tuzel, Ming-Yu Liu, and Rama Chellapa. Gaussian conditional random field network for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3224–3233, 2016.

[34] Hengli Wang, Rui Fan, Peide Cai, and Ming Liu. Sne-roadseg+: Rethinking depth-normal translation and deep supervision for freespace detection. *arXiv preprint arXiv:2107.14599*, 2021.

[35] Hengli Wang, Rui Fan, Yuxiang Sun, and Ming Liu. Dynamic fusion module evolves drivable area and road anomaly detection: A benchmark and algorithms. *IEEE Transactions on Cybernetics*, 2021.

[36] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[37] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018.

[38] Changqian Yu, Jingbo Wang, Changxin Gao, Gang Yu, Chunhua Shen, and Nong Sang. Context prior for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12416–12425, 2020.

[39] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 325–341, 2018.

[40] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1857–1866, 2018.

[41] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. *Proceedings of the European Conference on Computer Vision*, 2020.

[42] Yuhui Yuan and Jingdong Wang. Ocnet: Object context network for scene parsing. *arXiv preprint arXiv:1809.00916*, 2018.

[43] Fan Zhang, Yanqin Chen, Zhihang Li, Zhibin Hong, Jingtuo Liu, Feifei Ma, Junyu Han, and Errui Ding. Acfnet: Attentional class feature network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6798–6807, 2019.

[44] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7151–7160, 2018.

[45] Hang Zhang, Han Zhang, Chenguang Wang, and Junyuan Xie. Co-occurrent features in semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 548–557, 2019.

[46] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2881–2890, 2017.

[47] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *Proceedings of the European Conference on Computer Vision*, pages 267–283, 2018.

[48] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1537, 2015.

[49] Zilong Zhong, Zhong Qiu Lin, Rene Bidart, Xiaodan Hu, Ibrahim Ben Daya, Zhifeng Li, Wei-Shi Zheng, Jonathan Li, and Alexander Wong. Squeeze-and-attention networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13065–13074, 2020.

[50] Yizhou Zhou, Xiaoyan Sun, Zheng-Jun Zha, and Wenjun Zeng. Context-reinforced semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4046–4055, 2019.