Context Prior for Scene Segmentation

Changqian Yu[1,2] Jingbo Wang[3] Changxin Gao[1*] Gang Yu[4] Chunhua Shen[2] Nong Sang[1]
[1]Key Laboratory of Image Processing and Intelligent Control,
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
[2]The University of Adelaide, Australia [3]The Chinese University of Hong Kong [4]Tencent
{changqian_yu,cgao,nsang}@hust.edu.cn

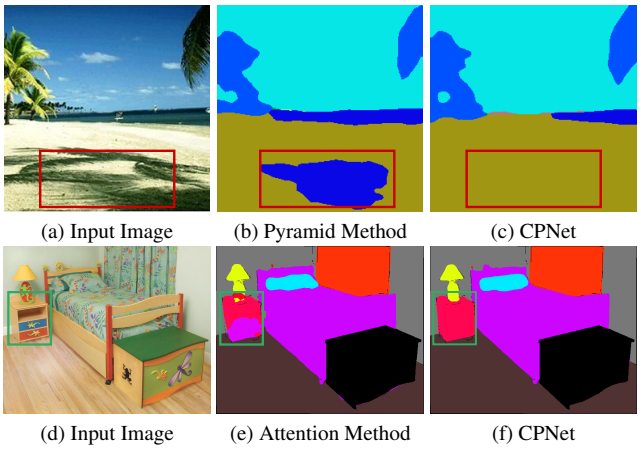
Abstract

Recent works have widely explored the contextual dependencies to achieve more accurate segmentation results. However, most approaches rarely distinguish different types of contextual dependencies, which may pollute the scene understanding. In this work, we directly supervise the feature aggregation to distinguish the intra-class and inter-class context clearly. Specifically, we develop a Context Prior with the supervision of the Affinity Loss. Given an input image and corresponding ground truth, Affinity Loss constructs an ideal affinity map to supervise the learning of Context Prior. The learned Context Prior extracts the pixels belonging to the same category, while the reversed prior focuses on the pixels of different classes. Embedded into a conventional deep CNN, the proposed Context Prior Layer can selectively capture the intra-class and inter-class contextual dependencies, leading to robust feature representation. To validate the effectiveness, we design an effective Context Prior Network (CPNet). Extensive quantitative and qualitative evaluations demonstrate that the proposed model performs favorably against state-of-the-art semantic segmentation approaches. More specifically, our algorithm achieves 46.3% *mIoU on ADE20K,* 53.9% *mIoU on PASCAL-Context, and* 81.3% *mIoU on Cityscapes. Code is available at* https://git.io/ContextPrior.

1. Introduction

Scene segmentation is a long-standing and challenging problem in computer vision with many downstream applications e.g., augmented reality, autonomous driving [8, 12], human-machine interaction, and video content analysis. The goal is to assign each pixel with a category label, which provides comprehensive scene understanding.

Benefiting from the effective feature representation of the Fully Convolutional Network (FCN), a few approaches have obtained promising performance. However, limited by the structure of convolutional layers, the FCN provides insufficient contextual information, leaving room for improvement. Therefore, various methods [1, 3, 5, 32, 49, 43, 45, 35, 19] explore the contextual dependencies to obtain more accurate segmentation results. There are mainly two paths to aggregate the contextual information: 1) Pyramid-based aggregation method. Several methods [49, 1, 3, 5] adopt pyramid-based modules or global pooling to aggre-



(a) Input Image (b) Pyramid Method (c) CPNet
(d) Input Image (e) Attention Method (f) CPNet

Figure 1. **Hard examples in scene segmentation.** In the first row, the central part of the *sand* in the red box is misclassified as the *sea*, because the shadow part has a similar appearance with the *sea*. With the pyramid-based aggregation method [3], aggregation of the confused spatial information may lead to undesirable prediction as visualized in (b). In the second row, the *table* in the green box has a similar appearance to the bottom part of the bed. The attention-based method [50] fails to effectively distinguish the confused spatial information without prior knowledge, leading to less correct prediction as shown in (e). In the proposed CPNet, we aggregate the contextual dependencies with clear distinguishment. Notably, the Context Prior models the intra-class and inter-class relationships as a context prior knowledge to capture the intra-class and inter-class contextual dependencies.

*Corresponding author. Part of the work was done when C. Yu was visiting The University of Adelaide.

gate regional or global contextual details regularly. However, they capture the homogeneous contextual relationship, ignoring the contextual dependencies of different categories, as shown in Figure 1(b). When there are confused categories in the scene, these methods may result in a less reliable context. 2) Attention-based aggregation method. Recent attention-based methods learn channel attention [45, 43], spatial attention [23], or point-wise attention [50, 11, 44] to aggregate the heterogeneous contextual information selectively. Nevertheless, due to the lack of explicit regularization, the relationship description of the attention mechanism is less clear. Therefore, it may select undesirable contextual dependencies, as visualized in Figure 1(e). Overall, both paths aggregate contextual information without explicit distinction, causing a mixture of different contextual relationships.

We notice that the identified contextual dependencies help the network understand the scene. The correlation of the same category (intra-class context) and the difference between the different classes (inter-class context) make the feature representation more robust and reduce the search space of possible categories. Therefore, we model the contextual relationships among categories as prior knowledge to obtain more accurate prediction, which is of great importance to the scene segmentation.

In this paper, we construct a **Context Prior** to model the intra-class and inter-class dependencies as the prior knowledge. We formulate the context prior as a binary classifier to distinguish which pixels belong to the same category for the current pixel, while the reversed prior can focus on the pixels of different classes. Specifically, we first use a fully convolutional network to generate the feature map and the corresponding prior map. For each pixel in the feature map, the prior map can selectively highlight other pixels belonging to the same category to aggregate the intra-class context, while the reversed prior can aggregate the inter-class context. To embed the prior into the network, we develop a **Context Prior Layer** incorporating an **Affinity Loss**, which directly supervises the learning of the prior. Meanwhile, Context Prior also requires spatial information to reason the relationships. To this end, we design an **Aggregation Module**, which adopts the fully separable convolution (separate on both the spatial and depth dimensions) [32, 7, 48, 29] to efficiently aggregate spatial information.

To demonstrate the effectiveness of the proposed Context Prior, we design a simple fully convolutional network called **Context Prior Network (CPNet)**. Based on the output features of the backbone network [1, 3, 36], the Context Prior Layer uses the Aggregation Module to aggregate the spatial information to generate a Context Prior Map. With the supervision of Affinity Loss, the Context Prior Map can capture intra-class context and inter-class context to refine the prediction. Extensive evaluations demonstrate that the proposed method performs favorably against several recent *state-of-the-art* semantic segmentation approaches.

The main contributions of this work are summarized as follows.

- We construct a Context Prior with supervision of an Affinity Loss embedded in a Context Prior Layer to capture the intra-class and inter-class contextual dependencies explicitly.
- We design an effective Context Prior Network (CPNet) for scene segmentation, which contains a backbone network and a Context Prior Layer.
- We demonstrate the proposed method performs favorably against *state-of-the-art* approaches on the benchmarks of ADE20K, Pascal-Context, and Cityscapes. More specifically, our single model achieves 46.3% on the ADE20K validation set, 53.9% on the PASCAL-Context validation set and 81.3% on the Cityscapes test set.

2. Related Work

Context Aggregation. In recent years, various methods have explored contextual information, which is crucial to scene understanding [1, 5, 32, 49, 43, 45, 44, 19, 26, 41]. There are mainly two paths to capture contextual dependencies. 1) PSPNet [49] adopts the pyramid pooling module to partition the feature map into different scale regions. It averages the pixels of each area as the local context of each pixel in this region. Meanwhile, Deeplab [1, 3, 5] methods employ atrous spatial pyramid pooling to sample the different range of pixels as the local context. 2) DANet [11], OCNet [44], and CCNet [18] take advantage of the self-similarity manner [37] to aggregate long-range spatial information. Besides, EncNet [45], DFN [43], and ParseNet [27] use global pooling to harvest the global context.

Despite the success of these attention mechanisms, they maybe capture undesirable contextual dependencies without explicitly distinguishing the difference of different contextual relationships. Therefore, in the proposed approach, we explicitly regularize the model to obtain the intra-class and inter-class contextual dependencies.

Attention Mechanism. Recent years have witnessed the broad application of the attention mechanism. It can be used for various tasks such as machine translation [34], image/action recognition [37, 6, 16], object detection [15] and semantic segmentation [43, 45, 42, 50, 11, 44].

For the semantic segmentation task, [4] learns an attention mechanism to weight the multi-scale features softly. Inspired by SENet [16], some methods such as EncNet [45], DFN [43], and BiSeNet [42] adopt the channel attention to select the desired feature map. Following [34, 37], DANet [11] and OCNet [44] use the self-attention to capture the long-range dependency, while PSANet [50] adap-

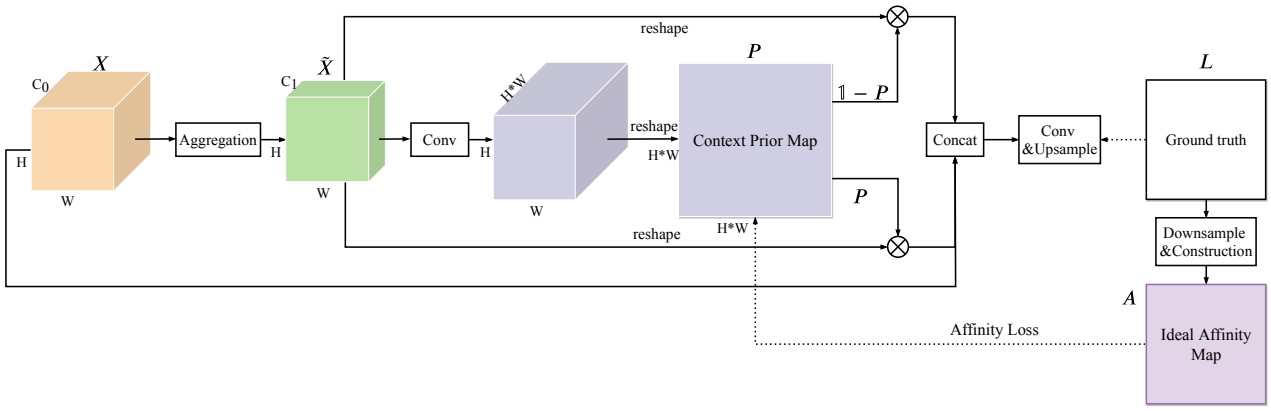


Figure 2. **Overview of the proposed Context Prior Layer**. The Context Prior Layer contains an Aggregation Module and a Context Prior Map supervised by Affinity Loss. With the extracted input features, the Aggregation Module aggregates the spatial information to reason the contextual relationship. We generate a point-wise Context Prior Map with the supervision of an Affinity Loss. The Affinity Loss constructs an Ideal Affinity Map which indicates the pixels of the same category to supervise the learning of the Context Prior Map. Based on the Context Prior Map, we can obtain the intra-prior ($\boldsymbol{P}$) and inter-prior ($\mathbb{1} - \boldsymbol{P}$). The original feature map is reshaped to $N \times C_1$ size, where $N = H \times W$. We conduct matrix multiplication on the reshaped feature map with $\boldsymbol{P}$ and $(\mathbb{1} - \boldsymbol{P})$ to capture the intra-class and inter-class context. Finally, we feed the representation of the Context Prior Layer into the last convolutional layer to generate a per-pixel prediction. (Notation: *Aggregation* Aggregation Module, *Conv* convolutional layer, $\otimes$ matrix multiplication, $\boldsymbol{P}$ Context Prior Map, *Concat* concatenate operation).

tively learns point-wise attention to harvest the long-range information. However, these effective methods lack the explicit regularization, maybe leading to an undesirable context aggregation. Therefore, in our work, we propose a Context Prior embedded in the Context Prior Layer with an explicit Affinity Loss to supervise the learning process.

3. Context Prior

Contextual dependencies play a crucial role in scene understanding, which is widely explored in various methods [49, 32, 27, 45, 3, 43]. However, these methods aggregate different contextual dependencies as a mixture. As discussed in Section 1, the clear distinguished contextual relationships are desirable to the scene understanding.

In our study, we propose a Context Prior to model the relationships between pixels of the same category (intra-context) and pixels of the different categories (inter-context). Based on the Context Prior, we propose a Context Prior Network, incorporating a Context Prior Layer with the supervision of an Affinity Loss, as shown in Figure 2. In this section, we first introduce the Affinity Loss, which supervises the layer to learn a Context Prior Map. Next, we demonstrate the Context Prior Layer, which uses the learned Context Prior Map to aggregate the intra-context and inter-context for each pixel. The Aggregation Module is designed to aggregate the spatial information for reasoning. Finally, we elaborate on our complete network structure.

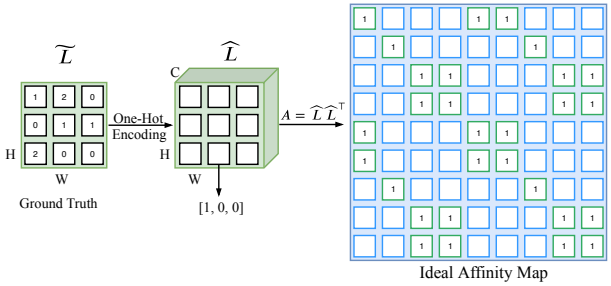


Figure 3. **Illustration of the construction of the Ideal Affinity Map**. The downsampled ground truth $\widetilde{\boldsymbol{L}}$ is first encoded with the one-hot encoding. The size of the ground truth $\widehat{\boldsymbol{L}}$ becomes $H \times W \times C$, where C is the number of the classes. Each vector in $\widehat{\boldsymbol{L}}$ is composed of a single high value (1) and all the others low (0). We conduct $\boldsymbol{A} = \widehat{\boldsymbol{L}}\widehat{\boldsymbol{L}}^{\top}$ to generate the Ideal Affinity Map. In this map, the green box and blue box represent 1 and 0, respectively.

3.1. Affinity Loss

In the scene segmentation task, for each image, we have one ground truth, which assigns a semantic category for each pixel. It is hard for the network to model the contextual information from isolated pixels. To explicitly regularize the network to model the relationship between categories, we introduce an Affinity Loss. For each pixel in the image, this loss forces the network to consider the pixels of the same category (intra-context) and the pixels among the different categories (inter-context).

Given a ground truth for an input, we can know the "context prior" of each pixel (i.e., which pixels belong to the same category and which pixels do not). Therefore, we

can learn a Context Prior to guiding the network according to the ground truth. To this end, we first construct an Ideal Affinity Map from the ground truth as the supervision. Given an input image $\boldsymbol{I}$ and the ground truth $\boldsymbol{L}$, we feed the input image $\boldsymbol{I}$ to the network, obtaining a feature map $\boldsymbol{X}$ of size $H \times W$. As shown in Figure 3, we first down-sample the ground truth $\boldsymbol{L}$ into the same size of the feature map $\boldsymbol{X}$, yielding a smaller ground truth $\widetilde{\boldsymbol{L}}$. We use a one-of-K scheme (one-hot encoding) to encode each categorical integer label in the ground truth $\widetilde{\boldsymbol{L}}$, leading to a matrix $\widehat{\boldsymbol{L}}$ of $H \times W \times C$ size, where C is the number of classes. Next, we reshape the encoded ground truth to $N \times C$ size, in which $N = H \times W$. Finally, we conduct the matrix multiplication: $\boldsymbol{A} = \widehat{\boldsymbol{L}}\widehat{\boldsymbol{L}}^{\top}$. $\boldsymbol{A}$ is our desired Ideal Affinity Map with size $N \times N$, which encodes which pixels belong to the same category. We employ the Ideal Affinity Map to supervise the learning of Context Prior Map.

For each pixel in the prior map, it is a binary classification problem. A conventional method for addressing this problem is to use the binary cross entropy loss. Given the predicted Prior Map $\boldsymbol{P}$ of size $N \times N$, where $\{p_n \in \boldsymbol{P}, n \in [1, N^2]\}$ and the reference Ideal Affinity Map $\boldsymbol{A}$, where $\{a_n \in \boldsymbol{A}, n \in [1, N^2]\}$, the binary cross entropy loss can be denoted as:

$$\mathcal{L}_u = -\frac{1}{N^2}\sum_{n=1}^{N^2}(a_n \log p_n + (1 - a_n)\log(1 - p_n)). \quad (1)$$

However, such a unary loss only considers the isolated pixel in the prior map ignoring the semantic correlation with other pixels. The pixels of each row of the Prior Map $\boldsymbol{P}$ is corresponding to the pixels of the feature map $\boldsymbol{X}$. We can divide them into intra-class pixels and inter-class pixels, the relationships of which are helpful to reason the semantic correlation and scene structure. Therefore, we can consider the intra-class pixels and inter-class pixels as two wholes to encode the relationships respectively. To this end, we devise the global term based on the binary cross entropy loss:

$$\mathcal{T}_j^p = \log \frac{\sum_{i=1}^{N} a_{ij} p_{ij}}{\sum_{i=1}^{N} p_{ij}}, \quad (2)$$

$$\mathcal{T}_j^r = \log \frac{\sum_{i=1}^{N} a_{ij} p_{ij}}{\sum_{i=1}^{N} a_{ij}}, \quad (3)$$

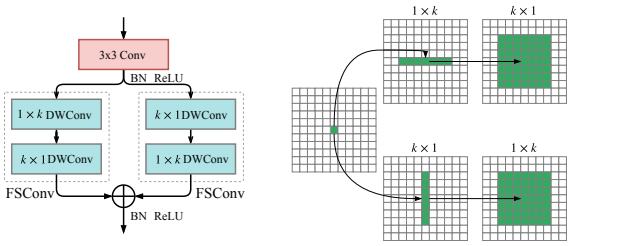
$$\mathcal{T}_j^s = \log \frac{\sum_{i=1}^{N} (1 - a_{ij})(1 - p_{ij})}{\sum_{i=1}^{N} (1 - a_{ij})}, \quad (4)$$

$$\mathcal{L}_g = -\frac{1}{N}\sum_{j=1}^{N}(\mathcal{T}_j^p + \mathcal{T}_j^r + \mathcal{T}_j^s), \quad (5)$$

where $\mathcal{T}_j^p$, $\mathcal{T}_j^r$, and $\mathcal{T}_j^s$ represent the intra-class predictive value (precision), true intra-class rate (recall), and true inter-class rate (specificity) at j^{th} row of $\boldsymbol{P}$, respectively. Finally, based on both the unary term and global term, the complete Affinity Loss can be denoted as follows:

$$\mathcal{L}_p = \lambda_u \mathcal{L}_u + \lambda_g \mathcal{L}_g, \quad (6)$$

where $\mathcal{L}_p$, $\mathcal{L}_u$, and $\mathcal{L}_g$ represent the affinity loss, unary loss (binary cross entropy loss), and global loss functions, respectively. In addition, λ_u and λ_g are the balance weights for the unary loss and global loss, respectively. We empirically set the weights as: $\lambda_u = 1$ and $\lambda_g = 1$.



(a) Aggregation Module (b) Receptive filed of Aggregation Module

Figure 4. **Illustration of the Aggregation Module and its receptive field**. (a) We use two asymmetric fully separable convolutions to aggregate the spatial information, the output of which has the same channels with the input features. (b) The Aggregation Module has the same size of receptive field with the standard convolution. However, our Aggregation Module leads to less computation. (Notation: *Conv* standard convolution, *DWConv* depthwise convolution *FSConv* fully separable convolution, k the filter size of the fully separable convolution, *BN* batch normalization, *ReLU* relu non-linear activation function.)

3.2. Context Prior Layer

Context Prior Layer considers an input feature $\boldsymbol{X}$ with the shape of $H \times W \times C_0$, as illustrated in Figure 2. We adopt an aggregation module to adapt $\boldsymbol{X}$ to $\widetilde{\boldsymbol{X}}$ with the shape of $H \times W \times C_1$. Given $\widetilde{\boldsymbol{X}}$, one 1×1 convolution layer followed by a BN layer [20] and a Sigmoid function is applied to learn a prior map $\boldsymbol{P}$ with the size $H \times W \times N (N = H \times W)$. With the explicit supervision of the Affinity Loss, Context Prior Map $\boldsymbol{P}$ can encode the relationship between intra-class pixels and inter-class pixels. The intra-class is given by $\boldsymbol{Y} = \boldsymbol{P}\widetilde{\boldsymbol{X}}$, where $\widetilde{\boldsymbol{X}}$ is reshaped into $N \times C_1$ size. In this operator, the prior map can adaptively select the intra-class pixels as the intra-class context for each pixel in the feature map. On the other hand, the reversed prior map is applied to selectively highlight the inter-class pixels as the inter-class context: $\overline{\boldsymbol{Y}} = (\mathbb{1} - \boldsymbol{P})\widetilde{\boldsymbol{X}}$, where $\mathbb{1}$ is an all-ones matrix with the same size of $\boldsymbol{P}$. Finally, we concatenate the original feature and both kinds of context to output the final prediction: $\mathcal{F} = Concat(\boldsymbol{X}, \boldsymbol{Y}, \overline{\boldsymbol{Y}})$. With both context, we can reason the semantic correlation and scene structure for each pixel.

3.3. Aggregation Module

As discussed in Section 1, the Context Prior Map requires some local spatial information to reason the semantic correlation. Therefore, we devise an efficient Aggregation Module with the fully separable convolution (separate on both the spatial and depth dimensions) to aggregate the spatial information. The convolution layer can inherently aggregate nearby spatial information. A natural method to aggregate more spatial information is to use the a large filter size convolutions. However, convolutions with large filter size are computationally expensive. Therefore, similar to [33, 32], we factorize the standard convolution into two asymmetric convolutions spatially. For a $k \times k$ convolution, we can use a $k \times 1$ convolution followed by a $1 \times k$ convolution as the alternative, termed spatial separable convolution. It can decrease $\frac{k}{2}$ computation and keep the equal size of receptive filed in comparison to the standard convolution. Meanwhile, each spatial separable convolution adopts the depth-wise convolution [7, 48, 14], further leading to the computation decrease. We call this separable convolution as Fully Separable Convolution with consideration both the spatial and depth dimensions. Figure 4 demonstrates the complete structure of the Aggregation Module.

3.4. Network Architecture

The Context Prior Network (CPNet) is a fully convolutional network composed of a backbone network and a Context Prior Layer, as shown in Figure 2. The backbone network is an off-the-shelf convolutional network [13, 48, 35], e.g., ResNet [13], with the dilation strategy [49, 50, 45]. In the Context Prior Layer, the Aggregation Module first aggregates some spatial information efficiently. Based on the aggregated spatial information, the Context Prior Layer learns a context prior map to capture intra-class context and inter-class context. Meanwhile, the Affinity Loss regularizes the learning of Context Prior, while the cross-entropy loss function is the segmentation supervision. Following the pioneering work [49, 50, 45], we employ the auxiliary loss on stage 4 of the backbone network, which is also a cross-entropy loss. The final loss function is as follows:

$$\mathcal{L} = \lambda_s \mathcal{L}_s + \lambda_a \mathcal{L}_a + \lambda_p \mathcal{L}_p, \tag{7}$$

where $\mathcal{L}_s$, $\mathcal{L}_a$, and $\mathcal{L}_p$ represent the main segmentation loss, auxiliary loss, and affinity loss functions, respectively. In addition, λ_s, λ_a, and λ_p are the weights to balance the segmentation loss, auxiliary loss, and affinity loss, respectively. We empirically set the weights as: $\lambda_s = 1$ and $\lambda_p = 1$. Similar to [49, 50, 45], we set the weight: $\lambda_a = 0.4$,

4. Experimental Results

In this section, we first introduce the implementation and training details of the proposed network. Next, we evaluate the proposed method and compare it with *state-of-the-art* approaches on three challenging scene segmentation datasets, including ADE20K [52], PASCAL-Context [30], and Cityscapes [8]. We implement the proposed model using PyTorch [31] toolbox.

4.1. Implementation Details

Network. We adopt the ResNet [13] as our pre-trained model with dilation strategy [1, 3, 5]. Then we adopt the bilinear interpolation to up-sample the prediction eight times to compute the segmentation loss. Following [49, 50, 45], we integrate the auxiliary loss on stage 4 of the backbone network. We set the filter size of the fully separable convolution in the Aggregation Module as 11.

Data Augmentation. In the training phase, we apply the mean subtraction, random horizontal flip and random scale, which contains $\{0.5, 0.75, 1.0, 1.5, 1.75, 2.0\}$, on the input images in avoiding of overfitting. Finally, we randomly crop the large image or pad the small image into a fix size for training (480×480 for ADE20K, 512×512 for PASCAL-Context and 768×768 for Cityscapes).

Optimization. We fine-tune the CPNet model using the stochastic gradient descent (SGD) algorithm [22] with 0.9 momentum, 10^{-4} weight decay and 16 batch size. Notably, we set the weight decay as 5×10^{-4} when training on the Cityscapes dataset. Following the pioneering work [2, 3, 43, 42], we adopt the "poly" learning rate strategy $\gamma = \gamma_0 \times (1 - \frac{N_{iter}}{N_{total}})^p$, where N_{iter} and N_{total} represent the current iteration number and total iteration number, and $p = 0.9$. We set the base learning rate γ_0 as 2×10^{-2} for the experiments on ADE20K, while 1×10^{-2} for the experiments on PASCAL-Context and Cityscapes. Meanwhile, we train the model for 80K iterations on ADE20K, 25K for PASCAL-Context and 60K for Cityscapes. We use the standard cross entropy loss when training on the ADE20K and PASCAL-Context dataset. While training on Cityscapes, similar to [38, 42, 44], we adopt the bootstrapped cross-entropy loss [38] to mitigate the class imbalance problem in this dataset.

Inference. In the inference phase, following [49, 32, 43, 45], we average the predictions of multiple scaled and flipped inputs to further improve the performance. We use the scales including $\{0.5, 0.75, 1.0, 1.5, 1.75\}$ for the ADE20K and PASCAL-Context datasets, while $\{0.5, 0.75, 1, 1.5\}$ for the Cityscapes dataset. In addition, we adopt the pixel accuracy (pixAcc) and mean intersection of union (mIoU) as the evaluation metrics.

4.2. Evaluations on the ADE20K Dataset

Dataset description. ADE20K is a challenging scene parsing benchmark due to its complex scene and up to 150 category labels. This dataset can be divided into 20K/2K/3K

model	*mIoU*	*pixAcc*
ResNet-50 (Dilation)	34.38	76.51
ResNet-50 + Aux (Baseline)	36.24	77.37
ResNet-50 + ASPP	40.39	79.71
ResNet-50 + PSP	41.49	79.61
ResNet-50 + NonLocal	40.96	79.98
ResNet-50 + PSA	41.92	80.17
ResNet-50 + Aggregation Module	41.51	79.93
ResNet-50 + IntraPrior (BCE)	42.34	80.15
ResNet-50 + InterPrior (BCE)	41.88	79.96
ResNet-50 + IntraPrior (AL)	42.74	80.30
ResNet-50 + InterPrior (AL)	42.43	80.21
ResNet-50 + ContextPriorLayer	43.92	80.77
ResNet-50 + ContextPriorLayer_MS	44.46	81.38
ResNet-101 + ContextPriorLayer	45.39	81.04
ResNet-101 + ContextPriorLayer_MS	**46.27**	**81.85**

Table 1. Ablative studies on the ADE20K [52] validation set in comparison to other contextual information aggregation approaches. (Notation: *Aux* auxiliary loss, *BCE* binary cross entropy loss, *AL* Affinity Loss, *MS* multi-scale and flip testing strategy.)

for training, validation and testing respectively. We report the results on the validation set using pixAcc and mIoU.

Ablation studies. To demonstrate the effectiveness of our Context Prior and CPNet, we conduct the experiments with different settings and compared with other spatial information aggregation module, as shown in Table 1.

First, we introduce our baseline model. We evaluate the FCN [28] model with dilated convolution [1] based on ResNet-50 [13] on the validation set. Following [49, 45, 50], we add the auxiliary loss on stage 4 of the ResNet backbone. This can improve mIoU by 1.86% ($34.38\% \rightarrow 36.24\%$) and pixAcc by 0.86% ($76.51\% \rightarrow 77.37\%$). We adopt this model as our baseline.

Based on the features extracted by FCN, various methods aggregate contextual information to improve the performance. The pyramid-based methods (e.g., PSP and ASPP) adopts pyramid pooling or pyramid dilation rates to aggregate multi-range spatial information. Recent approaches [44, 11] apply the self-attention [37] method to aggregate the long-range spatial information, while the PSA module [50] learns over-parametric point-wise attention. Table 1 lists our reimplement results with different spatial information aggregation modules. While these methods can improve the performance over the baseline, they aggregate the spatial information as a mixture of the intra-class and inter-class context, maybe making the network confused, as discussed in Section 1. Therefore, different from these methods, the proposed CPNet considers the contextual dependencies as a Context Prior to encoding the identified contextual relationship. Specifically, for each pixel, we capture the intra-class context and inter-class context with the Context Prior Layer. With the same backbone ResNet-50 and without other testing tricks, our method performs favorably against these methods.

k	3	5	7	9	11	13	15
w/o CP	42.06	41.86	41.87	42.32	41.51	**42.34**	42.23
w/ CP	42.26	42.81	43.38	43.14	**43.92**	42.54	42.59
Δ	0.2	0.95	1.51	0.82	**2.41**	0.2	0.36

Table 2. Experimental results (mIoU) w/ or w/o Context Prior based on different kernel sizes. (Notation: k the kernel size of the fully separable convolution, Δ the improvement of introducing the Context Prior, *CP* Context Prior.)

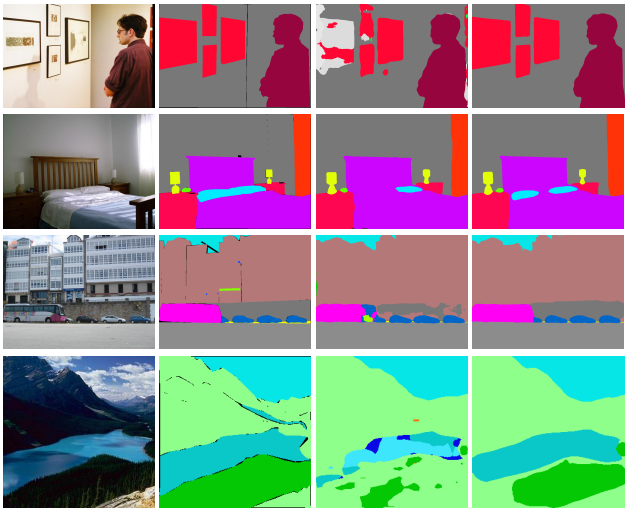
	PPM	ASPP	AM
w/o CP	41.49	40.39	**41.51**
w/ CP	42.55	42.69	**43.92**
Δ	$\uparrow$1.06	$\uparrow$2.3	$\uparrow$**2.41**

Table 3. Generalization to the PPM and ASPP module. The evaluation metric is mIoU (%). (Notation: *PPM* pyramid pooling module, *ASPP* atrous spatial pyramid pooling, *CP* Context Prior, *AM*: Aggregation Module.)

We also investigate the effectiveness of the Aggregation Module, IntraPrior branch, InterPrior branch and Affinity Loss in our CPNet model. We use the Aggregation Module with filter size 11 to aggregate the local spatial information. Similar to [50], the Aggregation Module generates an attention mask with the resolution of $N \times N (N = H \times W)$ to refine the prediction. As shown in Table 1, the Aggregation Module improves the mIoU and pixAcc by 5.27% / 2.56% over the baseline model. With the IntraPrior branch based on the binary cross entropy loss, our single scale testing results obtain 42.34% / 80.15% in terms of mIoU and pixAcc, surpassing the baseline by 6.1% / 2.78%. On the other hand, the InterPrior branch achieves 42.88% / 79.96% with the same setting. Both of the significant improvements demonstrate the effectiveness of the proposed Context Prior.

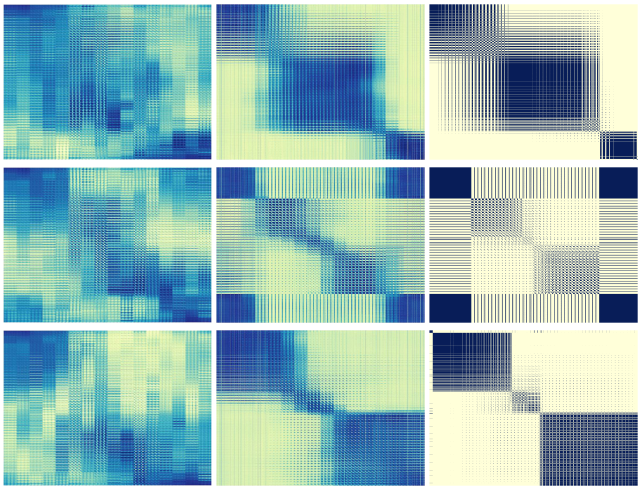
To further improve the quality of the context prior map, we devise an Affinity Loss. Table 1 indicates that the Affinity Loss can improve the mIoU and pixAcc by 0.4% / 0.15% based on IntraPrior branch, while boosting 0.55% / 0.25% based on InterPrior branch. We integrate both IntraPrior branch and InterPrior branch with the Affinity Loss to achieve 43.92% mIoU and 80.77% pixAcc, which demonstrates that both priors can be complementary. To further improve the performance, we apply the multi-scale and flipped testing strategy to achieve 44.46% mIoU and 81.38% pixAcc. Deeper network leading to better feature representation, our CPNet obtains 45.39% mIoU and 81.04% pixAcc with the ResNet-101. With the testing strategy, our model based on ResNet-101 achieves 46.27% mIoU and 81.85% pixAcc. Figure 5 provides some visualization examples.

Analysis and discussion. In Table 1, the proposed CPNet achieves considerable improvement on the ADE20K benchmark. Someone may argue that the large filter size of the Aggregation Module leads to the performance gain. Or



(a) Input Image (b) Ground Truth (c) FCN (d) CPNet (ours)

Figure 5. **Visual improvement** on validation set of ADE20K. Harvesting the intra-class context and inter-class context is helpful to the scene understanding.



(a) Attention Map (b) Learned Prior Map (c) Ideal Affinity Map

Figure 6. **Visualization of the Prior Map predicted by our CPNet**. (a) We only use the Aggregation Module to generate an attention map without the supervision of the Affinity Loss. (b) With the guidance of the Affinity Loss, the Context Prior Layer can capture the intra-class context and inter-class context. (c) The Ideal Affinity Map is constructed from the ground truth. Deeper color denotes higher response.

one may question whether the Context Prior can generalize to other algorithms. We thus provide more evidence to thoroughly understand the Context Prior. We conduct the discussion experiments on the ADE20K validation set with ResNet-50 backbone. The results reported in Table 2 and Table 3 are the single scale testing results.

(1) *The influence between the spatial information and Context Prior.* As discussed in Section 3, the distinguished contextual dependencies are helpful to scene understanding. Therefore, we propose a Context Prior to model the intra-context and inter-context. Meanwhile, the Context Prior requires some spatial information to reason the relationship. To this end, we integrate an Aggregation Module in the Context Prior Layer.

Table 2 indicates that with the increasing filter size, the models without Context Prior obtain the close results. However, with Context Prior, each model achieves improvements steadily. Meanwhile, the improvements gradually increase with the increasing filter size. When the filter size is 11, the performance (43.92% mIoU) and the relative gain (2.41%) reach the peak. If we continue to increase the filter size, the performance and the corresponding improvement both drop. In other words, Context Prior requries appropriate local spatial information to reason the relationships.

(2) *Generalization to other spatial information aggregation module.* To validate the generalization ability of the proposed Context Prior, we further replace the Aggregation Module with PPM or ASPP module to generate Context Prior Map with the supervision of Affinity Loss. As shown in Table 3, Context Prior can further improve the mIoU by 1.06% over the PPM without Context Prior, 2.3% over the ASPP module and 2.41% over our Aggregation Module. This improvement demonstrates the effectiveness and generalization ability of our Context Prior. Besides, without Context Prior, our Aggregation Module also achieves the highest performance comparing to the PPM and ASPP module.

Visualization of prior maps. To get a deeper understanding of our Context Prior, we randomly choose some examples from the ADE20K validation set and visualize the learned Context Prior Maps in Figure 6. We use the Aggregation Module to generate the attention map without the guidance of the Affinity Loss. Compared with the Ideal Affinity Map, we observe this attention map actually has a rough trend to learn this relationship. With the Affinity Loss, our Context Prior Layer can learn a prior map with more explicit structure information, which helps to refine the prediction.

Comparison with state-of-the-art. We conduct the comparison experiments with other *state-of-the-art* algorithms on Table 4. The proposed CPNet achieves 46.27% mIoU and 81.85% pixAcc, which performs favorably against previous *state-of-the-art* methods, even exceeds the winner entry of the COCO-Place Challenge 2017 based on ResNet-269. Our CPNet50 (with ResNet-50 as the backbone) achieves 44.46% mIoU and 81.38% pixAcc, even outperforms PSPNet [49], PSANet [50] and SAC [47] with deeper ResNet-101 and RefineNet with much deeper ResNet-152

model	reference	backbone	*mIoU*	*picAcc*
RefineNet [25]	CVPR2017	ResNet-101	40.2	-
RefineNet [25]	CVPR2017	ResNet-152	40.7	-
UperNet [39]	ECCV2018	ResNet-101	42.66	81.01
PSPNet [49]	CVPR2017	ResNet-101	43.29	81.39
PSPNet [49]	CVPR2017	ResNet-269	44.94	81.69
DSSPN [24]	CVPR2018	ResNet-101	43.68	81.13
PSANet [50]	ECCV2018	ResNet-101	43.77	81.51
SAC [47]	ICCV2017	ResNet-101	44.30	**81.86**
EncNet [45]	CVPR2018	ResNet-101	44.65	81.69
CFNet [46]	CVPR2019	ResNet-101	44.89	-
ANL [53]	ICCV2019	ResNet-101	45.24	-
CPNet50	-	ResNet-50	44.46	81.38
CPNet101	-	ResNet-101	**46.27**	81.85

Table 4. Quantitative evaluations on the ADE20K validation set. The proposed CPNet performs favorably against *state-of-the-art* segmentation algorithms.

model	reference	backbone	*mIoU*
FCN-8S [28]	CVPR2015	VGG16	37.8
CRF-RNN [51]	ICCV2015	VGG16	39.3
BoxSup [9]	ICCV2015	VGG16	40.5
Deeplabv2† [1]	ICLR2016	ResNet101	45.7
RefineNet [25]	CVPR2017	ResNet-152	47.3
PSPNet [49]	CVPR2017	ResNet-101	47.8
CCL [10]	CVPR2018	ResNet-101	51.6
EncNet [45]	CVPR2018	ResNet-101	51.7
DANet [11]	CVPR2019	ResNet-101	52.6
ANL [53]	ICCV2019	ResNet-101	52.8
CPNet101	-	ResNet-101	**53.9**

Table 5. Quantitative evaluations on the PASCAL-Context validation set. The proposed CPNet performs favorably against *state-of-the-art* segmentation methods. † means the method uses extra dataset.

as the backbone. This significant improvement manifests the effectiveness of our Context Prior.

4.3. Evaluations on PASCAL-Context

Dataset description. PASCAL-Context [30] is a scene understanding dataset which contains $10,103$ images from PASCAL VOC 2010. These images are re-annotated as pixel-wise segmentation maps with consideration of both the stuff and thing categories. This dataset can be divided into $4,998$ images for training and $5,105$ images for testing. The most common 59 categories are used for evaluation.

Comparison with state-of-the-art. Table 5 shows the performance comparison with other *state-of-the-art* approaches. Our algorithm achieves 53.9% mIoU on validation set and outperforms *state-of-the-art* EncNet by over 1.0 point. Similar to [1, 25, 49, 10, 45, 11], we evaluate the model with the multi-scale and flipped testing strategy. The scales contain $\{0.5, 0.75, 1, 1.5, 1.75\}$.

model	reference	backbone	*mIoU*
RefineNet [25]	CVPR2017	ResNet-101	73.6
GCN [32]	CVPR2017	ResNet-101	76.9
DUC [36]	WACV2018	ResNet-101	77.6
DSSPN [24]	CVPR2018	ResNet-101	77.8
SAC [47]	ICCV2017	ResNet-101	78.1
PSPNet [49]	CVPR2017	ResNet-101	78.4
BiSeNet [42]	ECCV2018	ResNet-101	78.9
AAF [21]	ECCV2018	ResNet-101	79.1
DFN [43]	CVPR2018	ResNet-101	79.3
PSANet [50]	ECCV2018	ResNet-101	80.1
DenseASPP [40]	CVPR2018	DenseNet-161	80.6
ANL [53]	ICCV2019	ResNet-101	**81.3**
CPNet101	-	ResNet-101	**81.3**

Table 6. Quantitative evaluations on the Cityscapes test set. The proposed CPNet performs favorably against *state-of-the-art* segmentation methods. We only list the methods training with merely the fine dataset.

4.4. Evaluations on Cityscapes

Dataset description. Cityscapes [8] is a large urban street scene parsing benchmark. It contains $2,975$ fine annotation images for training, 500 images for validation, $1,525$ images for testing and extra $20,000$ coarsely annotated images for training. We only use the fine annotation set in our experiments. It includes 19 categories for evaluation.

Comparison with state-of-the-art. Table 6 lists the performance results of other *state-of-the-art* methods and our CPNet. We adopt the multi-scale and flipped testing strategy on our experiments. Following the pioneering work [32, 43, 42], we train our model with both the train-fine set and val-fine set to improve the performance on the test set. Our CPNet achieves 81.3% mIoU on the Cityscapes test set only with the fine dataset, which outperforms the DenseASPP based on DenseNet-161 [17] by 0.9 point.

5. Concluding Remarks

In this work, we construct an effective Context Prior for scene segmentation. It distinguishes the different contextual dependencies with the supervision of the proposed Affinity Loss. To embed the Context Prior into the network, we present a Context Prior Network, composed of a backbone network and a Context Prior Layer. The Aggregation Module is applied to aggregate spatial information for reasoning the contextual relationship and embedded into the Context Prior Layer. Extensive quantitative and qualitative comparison shows that the proposed CPNet performs favorably against recent state-of-the-art scene segmentation approaches.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 61433007 and 61876210).

References

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *Proc. International Conference on Learning Representations (ICLR)*, 2015. 1, 2, 5, 6, 8

[2] 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. *arXiv*, 2016. 5

[3] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv*, 2017. 1, 2, 3, 5

[4] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proc. European Conference on Computer Vision (ECCV)*, pages 801–818, 2018. 1, 2, 5

[6] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. Aˆ 2-nets: Double attention networks. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 352–361, 2018. 2

[7] François Chollet. Xception: Deep learning with depthwise separable convolutions. 2017. 2, 5

[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 *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 5, 8

[9] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 1635–1643, 2015. 8

[10] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2393–2402, 2018. 8

[11] Jun Fu, Jing Liu, Haijie Tian, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 6, 8

[12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012. 1

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5, 6

[14] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv*, 2017. 5

[15] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[17] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017. 8

[18] Zilong Huang, Xinggang Wang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[19] Wei-Chih Hung, Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Scene parsing with global context embedding. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2

[20] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. International Conference on Machine Learning (ICML)*, pages 448–456, 2015. 4

[21] Tsung-Wei Ke, Jyh-Jing Hwang, Ziwei Liu, and Stella X Yu. Adaptive affinity fields for semantic segmentation. In *Proc. European Conference on Computer Vision (ECCV)*, pages 587–602, 2018. 8

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2012. 5

[23] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. *Proc. the British Machine Vision Conference (BMVC)*, 2018. 2

[24] Xiaodan Liang, Hongfei Zhou, and Eric P. Xing. Dynamic-structured semantic propagation network. *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 752–761, 2018. 8

[25] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks with identity mappings for high-resolution semantic segmentation. *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 8

[26] Huanyu Liu, Chao Peng, Changqian Yu, Jingbo Wang, Xu Liu, Gang Yu, and Wei Jiang. An end-to-end network for panoptic segmentation. *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6165–6174, 2019. 2

[27] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv*, 2016. 2, 3

[28] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proc.*

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015. 6, 8

[29] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proc. European Conference on Computer Vision (ECCV)*, pages 116–131, 2018. 2

[30] 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 *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 5, 8

[31] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *Neural Information Processing Systems Workshops*, 2017. 5

[32] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters–improve semantic segmentation by global convolutional network. *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 3, 5, 8

[33] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. 5

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2

[35] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. 2019. 1, 5

[36] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. *Proc. IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018. 2, 8

[37] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 6

[38] Zifeng Wu, Chunhua Shen, and Anton van den Hengel. High-performance semantic segmentation using very deep fully convolutional networks. *arXiv*, 2016. 5

[39] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proc. European Conference on Computer Vision (ECCV)*, pages 418–434, 2018. 8

[40] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3684–3692, 2018. 8

[41] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 5683–5692, 2019. 2

[42] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proc. European Conference on Computer Vision (ECCV)*, pages 325–341, 2018. 2, 5, 8

[43] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 3, 5, 8

[44] Yuhui Yuan and Jingdong Wang. Ocnet: Object context network for scene parsing. *arXiv*, 2018. 2, 5, 6

[45] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7151–7160, 2018. 1, 2, 3, 5, 6, 8

[46] Hang Zhang, Han Zhang, Chenguang Wang, and Junyuan Xie. Co-occurrent features in semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 548–557, 2019. 8

[47] Rui Zhang, Sheng Tang, Yongdong Zhang, Jintao Li, and Shuicheng Yan. Scale-adaptive convolutions for scene parsing. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 2031–2039, 2017. 7, 8

[48] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6848–6856, 2018. 2, 5

[49] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 3, 5, 6, 7, 8

[50] 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 *Proc. European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 5, 6, 7, 8

[51] 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 *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2015. 8

[52] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5, 6

[53] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 593–602, 2019. 8