Context Contrasted Feature and Gated Multi-scale Aggregation for Scene Segmentation

Henghui Ding[1] Xudong Jiang[1] Bing Shuai[1] Ai Qun Liu[1] Gang Wang[2]

[1]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[2]Alibaba Group, Hangzhou, China

Abstract

Scene segmentation is a challenging task as it need label every pixel in the image. It is crucial to exploit discriminative context and aggregate multi-scale features to achieve better segmentation. In this paper, we first propose a novel context contrasted local feature that not only leverages the informative context but also spotlights the local information in contrast to the context. The proposed context contrasted local feature greatly improves the parsing performance, especially for inconspicuous objects and background stuff. Furthermore, we propose a scheme of gated sum to selectively aggregate multi-scale features for each spatial position. The gates in this scheme control the information flow of different scale features. Their values are generated from the testing image by the proposed network learnt from the training data so that they are adaptive not only to the training data, but also to the specific testing image. Without bells and whistles, the proposed approach achieves the state-of-the-arts consistently on the three popular scene segmentation datasets, Pascal Context, SUN-RGBD and COCO Stuff.

1. Introduction

Scene segmentation has been an essential component of image understanding and is in intensely demand for automation devices, virtual reality, self-driving vehicles and etc. The goal of scene segmentation is parsing a scene image into a set of coherent semantic regions and labeling each pixel to one of classes including not only objects but also stuff (e.g. road, grass, sky). It implicitly involves image classification, object localization and boundary delineation. Thus, scene segmentation demands multi-scale and multi-level visual recognition.

The recent success of Deep Convolutional Neural Networks (DCNN) has greatly improved the performance of computer vision tasks [20], such as image classification [53, 55, 22, 39, 24] and object detection [45, 46, 15, 37, 16]. However, there are some limitations when applying DCNN to dense prediction tasks like scene segmentation [38, 51, 8, 29]. The success of DCNN is closely related with its inherent invariance to feature deformations [62]. This invariance lets the DCNN learn very abstract feature representation of the whole image, therefore the network can obtain information of dominated/salient objects at any position, which is desirable for image classification. But for scene segmentation, spatial information is essential and pixel-level discriminative features are desired. Most state-of-the-arts scene segmentation frameworks are based on image classification networks pre-trained on [49], but it remains an open question of how to better adopt DCNN on scene segmentation. Herein, we mainly consider two handicaps when applying DCNN on dense prediction tasks: the various forms of objects/stuff (e.g. salient or inconspicuous) and the existence of multi-scale objects.

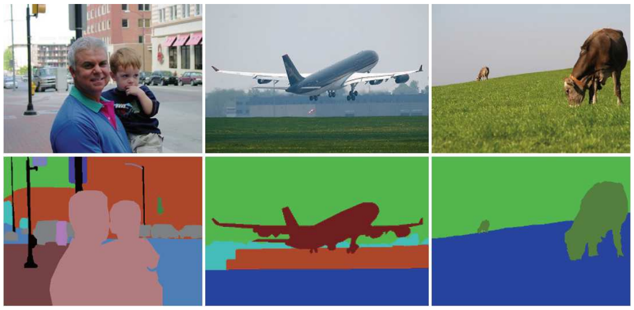

Figure 1: Scene segmentation refers to labeling each pixel including salient objects, inconspicuous objects and stuff. However, the various forms of objects/stuff (e.g salient or inconspicuous, foreground or background) and the existence of multi-scale objects (e.g the multi-scale cows in third image) make it challenging to parsing each pixel using DCNN.

First, different from object segmentation and image classification, scene segmentation aims to labeling every pixel to one of many classes including stuff and object classes, thus not only the dominated salient objects but also the stuff and inconspicuous objects should be parsed well. DCNN pre-trained on [49] prefers image-level abstract features, which is not equally discriminative for every spatial position. Meanwhile, due to the various forms of objects/stuff in scene segmentation, a pixel may belong to salient object, inconspicuous object or stuff. Therefore, when directly applying DCNN on scene segmentation, inconspicuous objects and stuff will be dominated by salient objects and its information will be somewhat weakened or even disregarded, which is contradictory with the goal of scene segmentation. To address this issue, locally discriminative features are desired. Context is essential for scene segmentation and lots of works devote to get informative context, e.g. [8, 51, 36, 61]. However, contexts often have smooth representation and are dominated by features of salient objects, which is harmful for labeling inconspicuous objects and stuff. Better features for scene segmentation are discriminative context aware local features, i.e., the features for pixel position p will not be dominated by other parts of image while being aware of the context information. For this purpose, we propose a context contrasted local feature, which benefits from both context and local information. Context contrasted local features could not only exploit the informative context but also spotlights the local information in contrast to the context. Further, we use a context contrasted local (CCL) model to obtain multi-scale and multi-level context contrasted local features.

Second, due to the huge scale variation of objects in scene segmentation, it is irrational to classify all individual pixels based on a single scale feature. There are several ways to address this issue. One way is to resize the input image to multiple resolutions and feed them to different (or a shared) networks, then fuse the corresponding features form multiple resolutions, such as [30, 12, 9, 44]. The aggregation ability for multi-scale features of this strategy is limited in practice due to expensive computation and the finite scales of input images. Another way makes use of features from middle layers, such as [38, 21, 48, 14]. The intention of this strategy is to exploit multi-scale features with multi-level information. We follow the way of FCN [38] to adopt skip layers to utilize multi-scale features, which is effective as well as economic. However, in previous works, such as [38, 21, 40, 51, 7, 43], the score maps of skip layers are integrated via a simple sum fusion and hence the different importance of different scales are ignored. To address this problem and find an optimal integration choice, we propose a network that controls the information flow of different scale features. It generates control signals to perform a gated sum of the score maps to aggregate multi-scale features selectively. As a selection mechanism is embedded in the multi-scale fusion, more skip layers can participate in the aggregation to provide rich information for selection. This also improves the aggregation ability of multi-scale features.

In summary, this paper makes the following contributions:

- We propose a novel context contrasted local feature which is tailored for scene segmentation and propose a context contrasted local (CCL) model to obtain multi-scale and multi-level context contrasted local features.
- We further propose a gated sum to selectively aggregate appropriate scale features for each spatial location, which is an efficient and effective way to address the issue of the existence of multi-scale objects.
- We achieve new state-of-the-art performance consistently on the three public scene segmentation benchmarks, Pascal Context, SUN-RGBD and COCO Stuff.

2. Related work

2.1. Contextual Modeling

One direction is to apply new layers to enhance high-level contextual aggregation. For example, Chen et al. [8] introduced an atrous spatial pyramid pooling (ASPP) to capture useful context information at multiple scales. Visin et al.[56], Shuai et al. [51] and Byeon et al.[4] adopted recurrent neural networks to capture long-range context. Zhao et al.[63] employed multiple pooling to exploit global information from different regions. Liu et al. [36] proposed to model the mean field algorithm with local convolution layers and incorporate it in deep parsing network (DPN). Yu et al. [61] attached multiple dilated convolution layers after class likelihood maps to exercise multi-scale context aggregation. Another way is to use Conditional Random Fields (CRF) [28] to model the context of score maps [7, 8, 64, 30, 36]. For example, Chen et al. [8] adopted CRF to post-process the unary predictions. Zheng et al. [64] proposed CRF-RNN to jointly train CRF with their segmentation networks.

Different with previous works, in this paper, we propose a context contrasted local feature to perform discriminative high-level feature modeling. Furthermore, a context contrasted local (CCL) model is proposed to collect multi-level context aware local features.

2.2. Multi-scale Aggregation

Due to the huge scale variation of objects in scene segmentation, it is difficult to achieve robust segmentation with single scale features' prediction. Multi-scale aggregation is a crucial way to deliver detailed parsing maps. There are several methods to achieve multi-scale aggregation. Farabet et al. [12] and Lin et al. [30] adopted multi-resolution input (image pyramid) approach and fuse the corresponding features from multiple resolution. Liu et al. [34] generated multi-scale patches and aggregated the results. Pinheiro et al. [44] inputted multi-size images at different layers of a recurrent convolutional neural networks. However, the above approaches are computational expensive and consume large GPU memory, thus their aggregation ability for multi-scale features is limited in practice. The seminal work FCN [38] introduced the skip layers to locally classify multi-scale feature maps and aggregate their predictions via sum fusion. This is an effective as well as efficient method to integrate different scale features and our work follows this way. Nonetheless, in previous works [38, 21, 40, 51, 7, 43], the score maps of skip layers are fused via a simple sum and hence the different importance of different scales are ignored. To address this issue, we propose a network that facilitates a gated sum to selectively aggregate different scale features. With gated sum fusion, the network can exploit more skip layers from richer scale features in DCNN and customize a suitable integration of different scale features. To the best of our knowledge, our gated sum is the first work to selectively aggregate appropriate scale features in a single network.

3. Segmentation Networks

Challenges of applying DCNN on scene segmentation are closely associate with the various forms of objects/stuff (e.g. salient or inconspicuous, foreground or background) and the existence of multi-scale objects. A robust segmentation network should be able to handle huge scale variation of objects and detect inconspicuous objects/stuff from images overwhelmed by other salient objects.

3.1. Overall Framework

The overall framework of our network is shown in Figure 2. Our baseline is FCN-like architecture with ResNet-101 (pre-trained on ImageNet [49]) as backbone network. We add more skip layers to fuse rich scale feature maps. The proposed context contrasted local (CCL) model in Figure 2 generates multi-level and multi-scale context aware local features. Furthermore, we propose a gated sum denoted by $g+$ in Figure 2 to selectively aggregate rich scale features in DCNN and CCL.

The proposed CCL and Gated Sum are presented in details in the following sections.

3.2. Context Contrasted Local Feature

Context information is known being essential for scene labeling that can greatly improve performance. In fact, DCNN has already generated relatively high-level context features for object recognition [53, 22], but these context features penchant for abstract feature representation of the whole image, which are not appropriate for scene segmentation where labeling for each pixel is required. First, these context features generated for object recognition focus on the dominated objects of the whole image and cannot ensure useful context for inconspicuous objects and stuff. Also, they are not discriminative at different spatial positions. Therefore, it is significant to design tailored high-level features for scene segmentation.

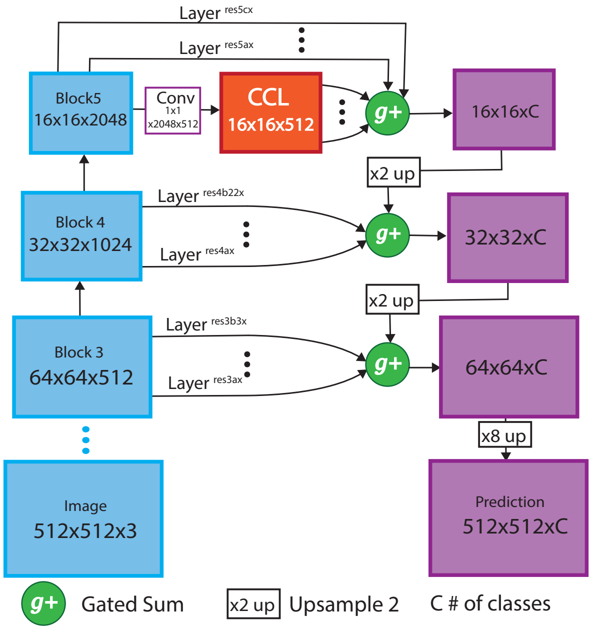

Figure 2: Overview of our network framework. The proposed context contrasted local (CCL) model generates multi-level and multi-scale context aware local features. Gated sum selectively aggregate rich scale features in DCNN and CCL.

Lots of previous works devote to obtain informative context for robust semantic segmentation, such as [8, 51, 57, 61]. Different from previous works, we introduce a context contrasted local feature to perform high-level feature modeling. Compared with object segmentation, there are richer categories and complex conjunctions between categories in scene segmentation. Due to the complexity of objects and stuff in scene segmentation, indiscriminately collecting context information will bring harmful noise, especially under clutter surroundings. For example, in Figure 4, compared with the two persons, the cars behind them are inconspicuous objects. The detailed local feature collects information around pixel $\mathbb{A}$ and is discriminative to other pixels, but it is not aware of global information such as road and building, thus could not obtain robust high level features for pixel $\mathbb{A}$. However, aggregating context will bring features of dominated objects like the men, thus the features of pixels at the car, like pixel $\mathbb{A}$, will be dominated by the features of the men. Some information of cars would be ignored in the final prediction, resulting wrong labeling for pixels at that location. Also,

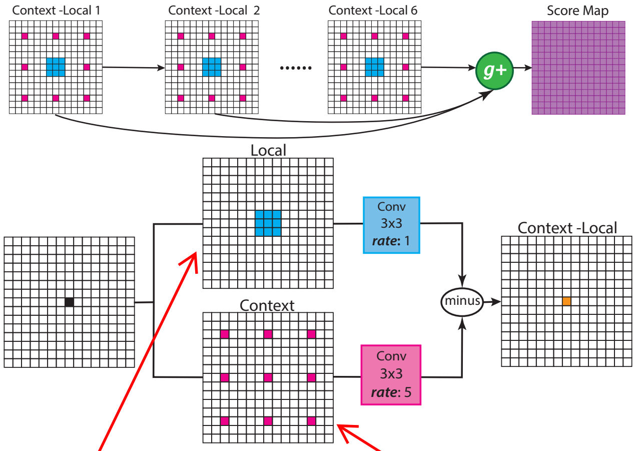

Figure 3: Context Contrasted Local (CCL) is a convolutional network integrating multi-level context aware local features. Each block of CCL consists of two parallel parts: coarse context and delicate local. The context aware local features are obtained via making a contrast between the context and local information. Several blocks are chained to make multi-level context contrasted local features.

contexts for different position are apt to a consistency representation of dominated features, thus are lacking in discrimination. Therefore, it's very hard to collect appropriate and discriminative high level features for pixel $\mathbb{A}$. To address this issue, we propose to generate local information and context separately and then fuse them via making contrast between these two:

$$\mathbb{CL} = \mathcal{F}_l(\mathbb{F}, \Theta_l) - \mathcal{F}_c(\mathbb{F}, \Theta_c) \tag{1}$$

where $\mathbb{F}$ is the input features, $\mathcal{F}_l$ is the function of local `Conv`, $\mathcal{F}_c$ is the function of context `Conv`, Θ_l and Θ_c are respective parameters, and $\mathbb{CL}$ is the desired context contrasted local features. They make a contrast between the separated context and local information, thus could not only exploit useful context but also foreground the local information in contrast to the context. Function of context-local forces the networks generating tailored features for scene segmentation. It is a mechanism that imitates human behavior. When our human beings look at one object we always collect discriminative-context for that object in a way that our eyes focus on that object in contrast to the blurred surroundings [13]. In other word, we concentrate on that object while we are aware of its surroundings.

Context Contrasted Local (CCL) Model. The architecture of CCL is shown in Figure 3. The CCL consists of several chained context-local blocks to make multi-level context contrasted local features. Gated sum (presented in the next section) is adopted in CCL to selectively aggregate different levels of context contrasted local features.

Comparison with State-of-the-art Context Models.

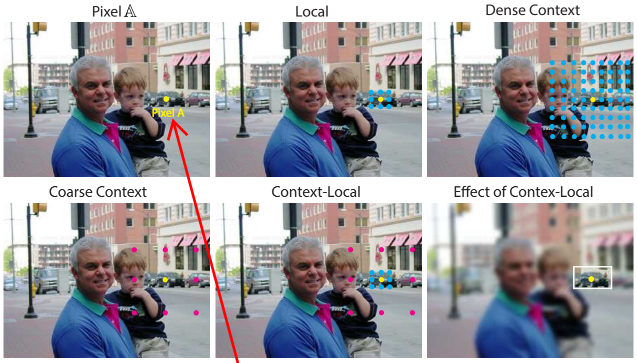

Figure 4: (Best viewed in color) Visualization of different feature information. The local information of pixel $\mathbb{A}$ could not aggregate useful contexts, such as road and other cars. However, its contexts will be dominated by the features of the men in the both schemes of dense context and coarse context. The context-local scheme injects blur context to local feature of pixel $\mathbb{A}$ to make discriminative context aware local feature.

ASPP[8] aggregates multi-scale contexts via combining score maps generated by different context aggregation branches, each of which uses dilated `Conv` kernels with different stride rates to incorporate different scale contexts. Compared with this type of context model, CCL first contextualizes contrasted features at every block to obtain context aware local features, which combines two different scales in the feature level and take advantage of both context and local information, then further aggregate multi-scale context contrasted local features in score level. Moreover, the score maps of CCL are fused via gated sum instead of the simple sum. DAG-RNN [51] performs contextual modeling by propagating local information in feature maps to encode long-range context. Different from DAG-RNN, CCL exploits multi-scale features for segmentation, and the context aware local features of CCL are different from those in DAG-RNN. CRF [28] is ordinarily applied to score maps and boosts consistency of low-level information like boundary, while CCL aims to discriminative high-level features. In fact, CRF can also be used as a post-processing step to promote performance of our segmentation network. We compare these context models in a controlled experiment and summarize their performance on Pascal Context in Table 1. The proposed CCL noticeably outperforms others, which demonstrates the significance of CCL.

3.3. Gated Multi-scale Aggregation

In this section, we discuss how to select different scale of features. One of the challenges in applying DCNN to scene segmentation is that it is difficult to use a single scale to obtain appropriate information for all pixels because of the existence of objects at multiple scales. An efficient and effective way to address this challenge is to add skip layers from the middle layers of DCNN. Based on the encoder-decoder architecture FCN [38], skip layers being as classifiers are used to exploit multi-scale features in DCNN to generate corresponding segmentation score maps.

However, in previous works such as [38, 21, 40, 51, 7], the score maps of skip layers are mainly integrated via sum fusion that does not take into account the individual differences of these inputs. Sum fusion can only non-selectively collect the score maps from different skip layers, but some of them may not be appropriate or even be harmful. If these score maps are aggregated indiscriminately, the inapposite or incorrect scores will harm the final prediction. To address this problem, we propose an aggregation scheme called gated sum to select different scale features. There are inherent position-wise gates in this scheme to control the information flow of skip layers. The primary motivation of gated sum is that we need to adaptively decide the desirable receptive field of each pixel in the image based on its scale, contextual support, etc. A by-pass and simple approximated solution is to pick different scale features for different pixel in FCN framework, where skip layers are aimed to capture multi-scale features. With gated sum fusion, the network can customize a suitable aggregation choice of score maps according to the information of images, corresponding to choose which scale of feature is better and more desirable. More importantly, with gated sum fusion, we can add more skip layers to extract richer scale information without posing problem of inapposite results.

The proposed scheme of gated sum is shown in Figure 5 where the values of gates are not directly learned from the training data but are generated from the testing image by a network learnt from the training data. In this way, the values of the gates are adaptive to the different testing input images. In order to obtain the information to control the gates, such as scale and contextual support, info-skip layers consisting of `Conv+Sigmoid` are introduced to extract the information from corresponding feature maps and generate information maps with size $H\times W\times 1$, where $H\times W$ is the spatial size of feature maps. Since these information maps and score maps of skip layers are generated from a same DCNN, the sequence relationship, e.g. from low level to high level, among feature maps of DCNN should also be considered. Recurrent Neural Networks (RNN) [17, 18, 19, 33] is effective and efficient to learn such sequence relationship, thus all of the information maps are feeded to RNN in sequence to learn the relationship of these information maps. Based on RNN, these information maps can be aware of neighbourhood maps and acquire the sequence relationship among all of the information maps.

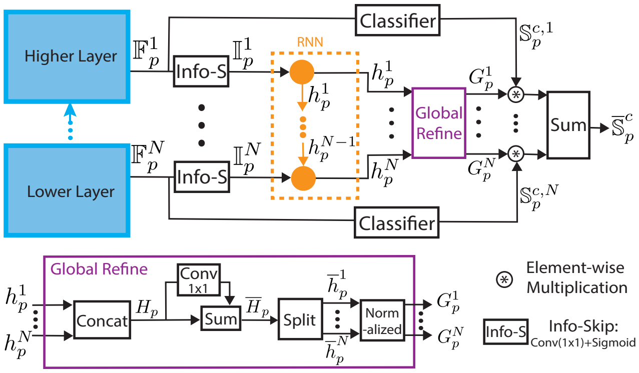

Figure 5: Gated Sum could control the information flow of skip layers via its inherent gates. The gates G_p^n could adjust its value according to the input images. All the maps in gated sum have the same spatial size of $H\times W$.

In details, we hypothesize that the information maps from higher layers have already grasped the information of lower layers due to the effect of DCNN, thus the RNN begin with information map of the last layer of our segmentation network. Suppose there are N score maps $\mathbb{S}_p^{c,n}$ generated by N skip layers from different scale features $\mathbb{F}_p^n$, i.e. $\mathbb{S}_p^{c,n} = \mathcal{F}_s^n(\mathbb{F}_p^n, \Theta_s^n)$, where p is the spatial position, $n \in 1, 2..., N$, $c \in 1, 2..., C$ and C is the number of class labels, $\mathcal{F}_s^n$ is the classifier function of nth skip layer and Θ_s^n is its parameters, $\mathbb{F}_p^n$ is the input feature with the dimensionality of $H\times W\times \#channels$. For each skip layer, we first generate an information map $\mathbb{I}_p^n$ of size $H\times W\times 1$ from corresponding feature:

$$\mathbb{I}_p^n = \mathcal{F}_i^n(\mathbb{F}_p^n, \Theta_i^n) \tag{2}$$

where $\mathcal{F}_i^n$ is the function of nth info-skip layer `Conv+Sigmoid` and Θ_i^n is its parameters. Then these information maps $\mathbb{I}_p^n$ are inputted to RNN in sequence to learn their relationships:

$$h_p^n = \tanh\left(W^n\left(\begin{array}{c} h_p^{n-1} \\ \mathbb{I}_p^n \end{array}\right)\right) \tag{3}$$

where h_p^n is the nth output of RNN. To make our network efficient, all positions are processed parallely and W^n is shared for all spatial positions. To ensure every information map be aware of global information, the outputs of RNN are concatnated , $H_p = (h_p^1 ... h_p^N)^T$, and refined with global information:

$$\overline{H}_p = \mathcal{F}_g\left(H_p, \Theta_g\right) + H_p \tag{4}$$

where $\mathcal{F}_g$ is a $1\times 1\times N\times N$ `Conv` and Θ_g is its parameters. Next, $\overline{H}_p$ is splitted, $\overline{H}_p = (\overline{h}_p^1 ... \overline{h}_p^N)^T$, and used to generate the gates G_p^n for gated sum:

$$G_p^n = N \cdot \frac{e^{\overline{h}_p^n}}{\sum_{i=1}^N e^{\overline{h}_p^i}} \tag{5}$$

the sum of G_p^n for each position p is normalized to N. Finally, N score maps are selectively fused via gated sum:

$$\overline{\mathbb{S}}_p^c = \sum_{n=1}^{N} G_p^n \mathbb{S}_p^{c,n} \tag{6}$$

where $\overline{\mathbb{S}}_p^c$ is the output of gated sum.

The gates of gated sum control the information flow of skip layers, i.e. how much can the $\mathbb{S}_p^{c,n}$ pass the gates depends on the value of G_p^n. A larger G_p^n means a better feature, for labeling of position p, is used for nth skip layer. While a smaller G_p^n means that for position p, the parsing results generated by the nth skip layer is not desirable and should be inhibited. More importantly, G_p^n is neither fixed value nor directly learned from training data. It is generated from the testing image by the proposed networks learned from the training data. Thus, G_p^n is adaptive to different testing images. The values of G_p^n not only depend on the training data, but also depend on the testing input images and vary according to the feature maps. Therefore, we call them "gates" to differentiate them from the simple fixed or learned "weights". With gated sum, the network adaptively (to different testing images) selects appropriate score maps from richer scales of features.

- **Sum** is a special case of the gated sum where all the gates are fixed to "1". Sum fusion dose not take into account the individual characteristic of different inputs and could only indiscriminately fuse all the inputs.
- **Gated sum** selectively aggregates appropriate score maps for each position's parsing via its inherent gates. The gate G_p^n adjusts its value adaptive to the testing input features to control the information flow of skip layers.

4. Experiments

We evaluate our segmentation framework on 3 public scene segmentation datasets, Pascal Context, SUN-RGBD and COCO Stuff.

4.1. Implementation Details

We use truncated ResNet-101 [22] (pre-trained on ImageNet [49]) as our fine-tune model. In detail, `pool5` and layers after it are discarded and a convolutional adaption layer that decrease the feature channels from 2048 to 512 is placed on the top of truncated ResNet-101 to reduce parameters. The number of blocks in CCL can be modified according to inputs, ours is six. We upsample the score maps with deconvolution (transpose convolution).

Our Network is trained end-to-end with SGD with fixed momentum 0.9 and weight decay 0.0005. Following [8], we employ the "poly" learning rate, $Lr_c = Lr_i \times (1 - \frac{iter}{max\_iter})^{power}$, where the Lr_c is current learning rate and Lr_i is the initial learning rate. The initial learning rate is set to be 10^{-3} and the power is set to 0.9. The iteration number is set to 15K for Pascal Context, 13K for SUN-RGBD and 20K for COCO Stuff. Batch size is 10 during training and the statistics of batch normalization layer is updated after the final iteration. The parameters of new layers are randomly initialized with Gaussian distribution (variance 10^{-2}) and trained with higher learning rate ($\times 3$). For batch processing, all images are resized to have maximum extent of 512 pixels and padded with zero to 512 $\times$ 512 pixels during training. We randomly flip the images horizontally to augment the training data.

Networks	CA	IoU
Baseline	None	42.5%
Baseline + CRF[28]	CRF	43.2%
Baseline + DAG-RNN [51]	DAG-RNN	44.1%
Baseline + ASPP [8]	ASPP	44.9%
Baseline + CCL	CCL	**48.3**%

Table 1: Segmentation networks are adapted to encode-decode architecture with rich skip layers, the stride rates (dilation factors) of the four branches in ASPP are revised to {1, 3, 4, 6} respectively. For fair comparisons, gated sum is not adopted, and they only differentiate each other in terms of context aggregation (CA).

Method	GPA	ACA	IoU
Baseline	73.5%	53.9%	42.5%
Baseline+LA	75.8%	57.6%	45.9%
Baseline+LA^d	75.7%	56.6%	45.4%
Baseline+CCL	**76.6**%	**61.1**%	**48.3**%

Table 2: Ablation experiments of CCL on Pascal Context. LA is local aggregation generated by removing the context part of CCL. LA^d doubles the hidden dimensionality of LA from 512 to 1024, thus its parameter quantity is the same as CCL. Other settings are all the same.

We evaluate our network with three performance metrics: Global Pixel Accuracy (**GPA**), Average Class Accuracy (**ACA**) and Mean Intersection-over-Union (**IoU**). Mathematical definitions please refer to [38].

4.2. Multi-scale Context Contrasted Local Features

In section 3.2 we introduced context contrasted local (CCL) model to integrate multi-level context aware local features. To evaluate the key principle (i.e. multi-scale context contrasted local features) of CCL, we simplify our context-local network architecture CCL to LA, and LA^d. LA abandons the context parts (dilated `Conv`) of CCL and LA^d doubles the hidden dimensionality of LA. The performance of these models are listed in Table 2. Their performance gap clearly demonstrates the benefit brought by the proposed CCL model.

First, compared with LA that is conventional convolutional feature, CCL aggregates specialized context contrasted local features that not only leverages the informative context but also exploits the discriminative local information in contrast to the context. In consequence, CCL outperforms LA by a noticeable margin, which clearly shows the significance of the context contrasted local features for scene segmentation.

It's crucial to introduce new parameters that fill the domain gap during the fine-tuning of segmentation networks from classification networks. However, we believe that the network architecture outweighs the magnitude of parameters for boosting the performance. To validate this claim, we increase the hidden dimension of LA from 512 to 1024, which is denoted by LA^d in Table 2. The parameter quantity of LA^d is then the same as CCL, but LA^d does not improve the performance of LA and even slightly make it worse. This convinces us that the noticeable performance boost is mainly contributed by the architecture of the context contrasted local features, not from simple increase of the network parameters.

4.3. Embed Gated Sum into Encoder-Decoder Architecture

Gated sum is a selection mechanism to pick appropriate features. But for the encoder-decoder architecture, the spatial sizes of distinct blocks are not the same, e.g. 16×16 for block 5 and 32×32 for block 4 in Figure 2. This causes difficulty of aggregating all the score maps. The most straightforward solution is upsampling all the score maps to the same resolution. However, this will consume a large amount of resources. Therefore, in this work, we embed the gated sum into the encoder-decoder architecture. For this purpose, we adopt gated sum within each block where the feature maps possess with the same spatial resolution. Then the output of gated sum is upsampled to higher resolution to participate in the gated sum in block with higher resolution. Meanwhile, to pass the information maps form block to block, the last output of RNN is also upsampled to generate the gates for the upsampled score map.

We present an ablation experiment of the gated sum in Table 3. As shown in Table 3, the gated sum improves the performance visibly. Comparing ResNet-101 to ResNet-50 and comparing the networks with CCL to those without CCL, we see that the performance gain brought by the gated sum will be higher if there are more score maps for selecting.

4.4. Results On Scene Segmentation

Pascal Context [41] contains 10103 images from Pascal VOC 2010, and these images are re-annotated as pixel-wise segmentation maps. There are 4998 images for training and 5105 images for testing in Pascal Context. We use the most common 59 categories in this dataset for evaluation. A few examples on validation set of Pascal Context are shown in Figure 6. Compared with the baseline, our segmentation network performers better at global information, salient objects, stuff and inconspicuous objects and has a robust adaptability to multi-scale objects. Quantitative results of Pascal Context are shown in Table 4. It shows that our segmentation network outperforms the state-of-the-arts by a large margin for all the three evaluation metrics.

Baseline Model	Gated Sum	CCL	IoU
ResNet-50	no	no	40.7%
ResNet-50	yes	no	41.5%
ResNet-50	no	yes	46.3%
ResNet-50	yes	yes	48.1%
ResNet-101	no	no	42.5%
ResNet-101	yes	no	43.9%
ResNet-101	no	yes	48.3%
ResNet-101	yes	yes	**51.6%**

Table 3: Ablation experiments of Gated Sum on Pascal Context.

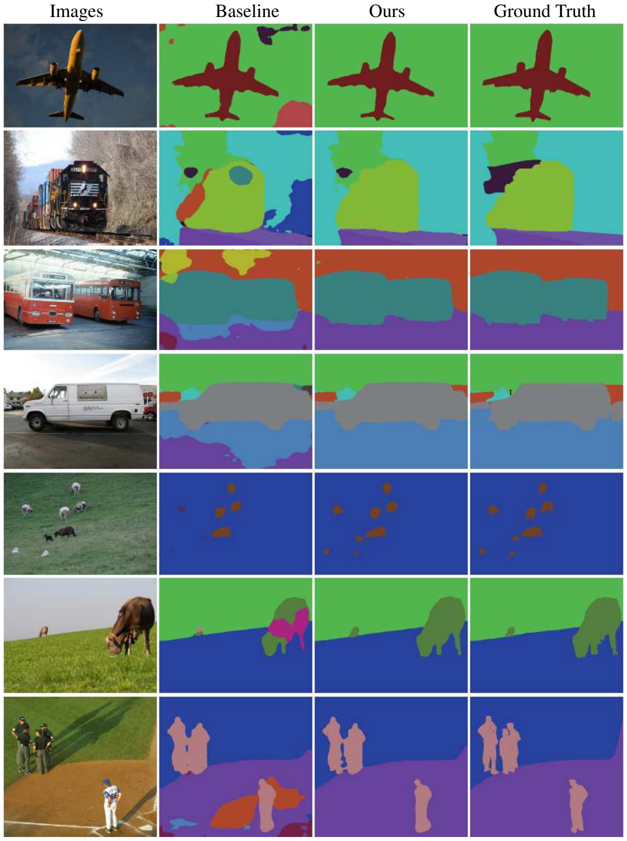

Figure 6: Qualitative segmentation result comparisons on Pascal Context. Our segmentation network performers well at salient objects, stuff (e.g. road, grass, sky) and inconspicuous objects. Further, our network has a robust adaptability to multi-scale objects.

SUN-RGBD [54] provides pixel-wise labeling for 37 categories. It has 10335 indoor images which are from

Methods	GPA	ACA	IoU
O2P[6]	-	-	18.1%
CFM [11]	-	-	34.4%
FCN-8s [50]	67.5%	52.3%	39.1%
CRF-RNN [64]	-	-	39.3%
ParseNet [35]	-	-	40.4%
BoxSup [10]	-	-	40.5%
ConvPP-8 [60]	-	-	41.0%
HO-CRF [1]	-	-	41.3%
PixelNet [3]	-	51.5%	41.4%
Context-CRF [30]	71.5%	53.9%	43.3%
DAG-RNN + CRF [51]	73.6%	55.8%	43.7%
FCRN [58]	72.9%	54.8%	44.5%
DeepLab-v2+CRF†[8]	-	-	45.7%
Hu et al.[23]	73.5%	56.7%	45.8%
Global-Context[25]	73.8%	-	46.5%
RefineNet-Res101 [29]	-	-	47.1%
RefineNet-Res152 [29]	-	-	47.3%
PSPNet-Res101 [63]	76.0%	60.6%	47.8%
Ours	**78.4**%	**63.9**%	**51.6**%

Table 4: Pascal Context testing accuracies. Our network outperforms all existing methods by a large margin across all evaluation metrics. Methods trained with extra data are marked with †.

Methods	GPA	ACA	IoU
Liu et al. [32]	-	10.0%	-
Ren et al. [47]	-	36.3%	-
FCN-8s [38]	68.2%	38.4%	27.4%
DeconvNet [42]	66.1%	33.3%	22.6%
Kendall et al. [27]	71.2%	45.9%	30.7%
SegNet [2]	72.6%	44.8%	31.8%
DeepLab [8]	71.9%	42.2%	32.1%
Context-CRF [30]	78.4%	53.4%	42.3%
RefineNet-Res101 [29]	80.4%	57.8%	45.7%
RefineNet-Res152 [29]	80.6%	58.5%	45.9%
Ours	**81.4**%	**60.3**%	**47.1**%

Table 5: SUN-RGBD (37 classes) segmentation results. We do not use the depth information for training. Our segmentation network outperforms existing methods consistently across all the three evaluation metrics.

SUN3D [59], NYUDv2 [52], Berkeley B3DO [26] and the newly captured images. The training set has 5285 images and the test set contains 5050 images. We only use the RGB modality as input for training. Quantitative results of SUN-RGBD are reported in Table 5. It shows that our segmentation network outperforms the previous state-of-the-arts consistently across all evaluation metrics.

COCO Stuff [5] contains 10000 images from Microsoft COCO dataset [31], out of which 9000 images are for training and 1000 images for testing. The unlabeled stuff pixels in original images of Microsoft COCO are further annotated with additional 91 classes in COCO Stuff. Herein, this dataset contains 171 categories including objects and stuff annotated to each pixel. Quantitative results of COCO Stuff are shown in Table 6. Our scene segmentation network outperforms the existing methods by a large margin across all evaluation metrics.

Networks	GPA	ACA	IoU
FCN [5]	52.0%	34.0%	22.7%
DeepLab [7]	57.8%	38.1%	26.9%
DAG-RNN[51]	62.2%	42.3%	30.4%
RefineNet-Res101 [29]	65.2%	45.3%	33.6%
Ours	**66.3**%	**48.8**%	**35.7**%

Table 6: Parsing performance of different networks on COCO Stuff dataset. Our segmentation network outperforms the state-of-the-arts by a large margin across all evaluation metrics.

5. Conclusion

In this paper, we address the challenging task of scene segmentation. Scene segmentation aims at parsing an image into a set of coherent semantic regions and classifying each pixel to one of classes, and hence the context and multi-scale aggregation are crucial to achieve good segmentation. However, DCNN designed for image classification tends to extract abstract features of dominated objects, thus some essentially discriminative information for inconspicuous objects and stuff are weakened or even disregarded. To address this issue, we propose a novel context contrasted local feature to leverage the useful context and spotlight the local information in contrast to the context. The proposed context contrasted local feature greatly improves the parsing performance, especially for inconspicuous objects and stuff. Adding skip layers is a common way to exploit multi-scale features, but the existing approaches indiscriminately fuse the score maps of skip layers via a simple summation. To achieve an optimal multi-scale aggregation, we propose a scheme of gated sum to selectively aggregate multi-scale features. The values of gates are generated from the testing image by the proposed networks learnt from the training data. Thus, they are adaptive not only to the training data, but also to the specific testing image. Without bells and whistles, our segmentation network achieves state-of-the-arts consistently on the 3 popular scene segmentation datasets used in the evaluation, Pascal Context, SUN-RGBD and COCO Stuff.

Acknowledgement

This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. This research was supported by Singapore Ministry of Education Academic Research Fund Tier 1 RG 123/15. The ROSE Lab is supported by the National Research Foundation, Singapore, under its Interactive Digital Media (IDM) Strategic Research Programme.

References

[1] A. Arnab, S. Jayasumana, S. Zheng, and P. H. Torr. Higher order conditional random fields in deep neural networks. In *European Conference on Computer Vision*. Springer, 2016.

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015.

[3] A. Bansal, X. Chen, B. Russell, A. Gupta, and D. Ramanan. Pixelnet: Towards a general pixel-level architecture. *arXiv preprint arXiv:1609.06694*, 2016.

[4] W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki. Scene labeling with lstm recurrent neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[5] H. Caesar, J. Uijlings, and V. Ferrari. Coco-stuff: Thing and stuff classes in context. *arXiv preprint arXiv:1612.03716*, 2016.

[6] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. *Computer Vision–ECCV 2012*, 2012.

[7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015.

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016.

[9] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[10] J. Dai, K. He, and J. Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[11] J. Dai, K. He, and J. Sun. Convolutional feature masking for joint object and stuff segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[12] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1915–1929, 2013.

[13] R. Garland-Thomson. *Staring: How we look*. Oxford University Press, 2009.

[14] G. Ghiasi and C. C. Fowlkes. Laplacian pyramid reconstruction and refinement for semantic segmentation. In *European Conference on Computer Vision*. Springer, 2016.

[15] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2015.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014.

[17] A. Graves. Offline arabic handwriting recognition with multidimensional recurrent neural networks. In *Guide to OCR for Arabic scripts*, pages 297–313. Springer, 2012.

[18] A. Graves, A.-r. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE, 2013.

[19] J. Gu, G. Wang, J. Cai, and T. Chen. An empirical study of language cnn for image captioning. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017.

[20] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, et al. Recent advances in convolutional neural networks. *Pattern Recognition*, 2017.

[21] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[23] H. Hu, Z. Deng, G.-T. Zhou, F. Sha, and G. Mori. Labelbank: Revisiting global perspectives for semantic segmentation. *arXiv preprint arXiv:1703.09891*, 2017.

[24] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[25] W.-C. Hung, Y.-H. Tsai, X. Shen, Z. Lin, K. Sunkavalli, X. Lu, and M.-H. Yang. Scene parsing with global context embedding. *arXiv preprint arXiv:1710.06507*, 2017.

[26] A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko, and T. Darrell. A category-level 3d object dataset: Putting the kinect to work. In *Consumer Depth Cameras for Computer Vision*. Springer, 2013.

[27] A. Kendall, V. Badrinarayanan, and R. Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015.

[28] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011.

[29] G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[30] G. Lin, C. Shen, A. van dan Hengel, and I. Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[31] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 2014.

[32] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):978–994, 2011.

[33] J. Liu, A. Shahroudy, G. Wang, L.-Y. Duan, and A. C. Kot. Ssnet: Scale selection network for online 3d action prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[34] S. Liu, X. Qi, J. Shi, H. Zhang, and J. Jia. Multi-scale patch aggregation (mpa) for simultaneous detection and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[35] W. Liu, A. Rabinovich, and A. C. Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015.

[36] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang. Semantic image segmentation via deep parsing network. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[37] Z. Liu, G. Lin, S. Yang, J. Feng, W. Lin, and G. Wangling. Learning markov clustering networks for scene text detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[38] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[39] Z. Lu, X. Jiang, and A. C. Kot. Deep coupled resnet for low-resolution face recognition. *IEEE Signal Processing Letters*, 2018.

[40] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich. Feedforward semantic segmentation with zoom-out features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[41] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[42] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[43] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun. Large kernel matters – improve semantic segmentation by global convolutional network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[44] P. Pinheiro and R. Collobert. Recurrent convolutional neural networks for scene labeling. In *International Conference on Machine Learning*, 2014.

[45] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[46] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[47] X. Ren, L. Bo, and D. Fox. Rgb-(d) scene labeling: Features and algorithms. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012.

[48] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015.

[49] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3), 2015.

[50] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 2016.

[51] B. Shuai, Z. Zuo, B. Wang, and G. Wang. Scene segmentation with dag-recurrent neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[52] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*. Springer, 2012.

[53] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[54] S. Song, S. P. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[55] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[56] F. Visin, M. Ciccone, A. Romero, K. Kastner, K. Cho, Y. Bengio, M. Matteucci, and A. Courville. Reseg: A recurrent neural network-based model for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2016.

[57] P. Wang, P. Chen, y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell. Understanding convolution for semantic segmentation. *arXiv preprint arXiv:1702.08502*, 2017.

[58] Z. Wu, C. Shen, and A. v. d. Hengel. Bridging category-level and instance-level semantic image segmentation. *arXiv preprint arXiv:1605.06885*, 2016.

[59] J. Xiao, A. Owens, and A. Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of the IEEE International Conference on Computer Vision*, 2013.

[60] S. Xie, X. Huang, and Z. Tu. Convolutional pseudo-prior for structured labeling. *arXiv preprint arXiv:1511.07409*, 2015.

[61] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[62] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer, 2014.

[63] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[64] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.