Semantic Correlation Promoted Shape-Variant Context for Segmentation

Henghui Ding[1] Xudong Jiang[1] Bing Shuai[2] Ai Qun Liu[1] Gang Wang[3]

[1]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[2]Amazon, Seattle, United States [3]Alibaba Group, Hangzhou, China

Abstract

Context is essential for semantic segmentation. Due to the diverse shapes of objects and their complex layout in various scene images, the spatial scales and shapes of contexts for different objects have very large variation. It is thus ineffective or inefficient to aggregate various context information from a predefined fixed region. In this work, we propose to generate a scale- and shape-variant semantic mask for each pixel to confine its contextual region. To this end, we first propose a novel paired convolution to infer the semantic correlation of the pair and based on that to generate a shape mask. Using the inferred spatial scope of the contextual region, we propose a shape-variant convolution, of which the receptive field is controlled by the shape mask that varies with the appearance of input. In this way, the proposed network aggregates the context information of a pixel from its semantic-correlated region instead of a predefined fixed region. Furthermore, this work also proposes a labeling denoising model to reduce wrong predictions caused by the noisy low-level features. Without bells and whistles, the proposed segmentation network achieves new state-of-the-arts consistently on the six public segmentation datasets.

1. Introduction

Semantic segmentation or scene parsing is aimed at classifying every pixel in scene images to one of the predefined categories (e.g., person, car, etc.). It has been a critical element in artificial intelligence and can be applied in many practice applications, such as automatic parking system. The recent success of Deep Neural Networks has greatly improved the performance of semantic segmentation [9, 43, 75, 11]. Most of state-of-the-arts segmentation networks are based on Convolutional Neural Networks (CNNs) [37, 63, 64, 26, 28] pre-trained on ImageNet [58], in which CNNs are employed as the local feature extractor. To achieve robust semantic segmentation, informative high-level context is necessary. Context provides surrounding information of the object, which helps a better discrimination of the object.

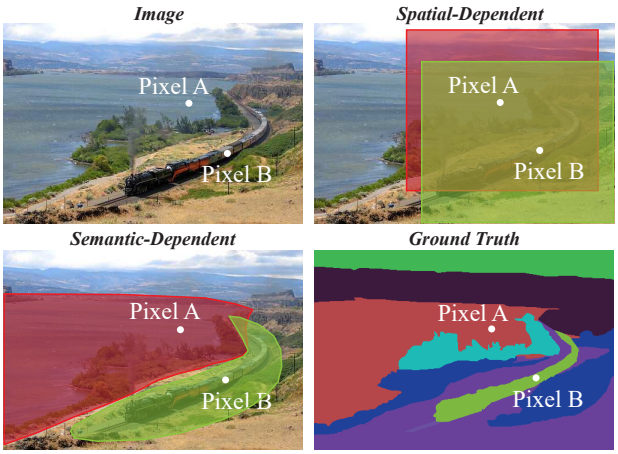

Figure 1. Most existing methods model spatial-dependent context with predefined window (e.g., the red rectangle region for pixel A in the second image), which may not provide the specific context information and thus weaken the semantic shape layout. In this work, we propose to aggregate the context information from the semantic-dependent region instead of the spatial-dependent one.

However, due to the diverse shapes (including contours, scales, etc) and the complex layout of objects in scene images, the common context aggregated within a predefined fixed region weakens the semantic shape layout and might bring unnecessary irrelevant information. For example in Figure 1, contexts of pixel A (lake) and pixel B (train) should be different but the predefined receptive fields to collect their contexts largely overlap, which decrease their discriminative power. Meanwhile, not all information in a predefined surrounding region (rectangle region in the second image) is beneficial for their final parsing. Information collected in the semantic correlated region is more helpful to identify the object while that in an irrelevant region, though spatially close, may result in parsing error and hence should be suppressed or even disregarded. For pixel A in Figure 1, information of pixels belonging to the lake and its shore, which are semantic

correlated, might be more beneficial than information of other pixels in the predefined fixed window. And for pixel B, the desirable shape of context would align with the train and railroad track. Besides, uniformly integrating smooth global information would degrade the location identity and the local discriminative features [69]. Therefore, for better scene parsing, diverse shapes of semantic consistent context should be customized. Most existing methods tend to model statistical average representation among a fixed rectangle region [9, 18, 73, 71] or the global region [72, 74, 30]. In this work, by taking into account the semantic correlation and the shape layout of objects, we propose a shape-variant context model to aggregate the surrounding information of each pixel from their semantic-correlated region inferred according to the appearance of input image.

To this end, we propose first to learn the relation between a target pixel and others by a novel paired convolution followed by a Gaussian mapping function. The learnt network produces higher value for two pixels with stronger semantic correlation and lower value for weaker correlation. Thus, the proposed network will generate a shape mask indicating a semantic correlated region for each pixel. With the shape mask specifying the size and shape of desirable receptive field, we further propose a shape-variant convolution to aggregate context from semantic-dependent region. The shape-variant convolution is specified by a set of learnable location-invariant convolution parameters and the location-variant shape masks. Thus, the parameters are dependent variables of the semantic correlated region of input image, which change with varying shapes and scales of objects. Furthermore, since the shape-variant context is implicitly scale-variant, we can model not only multi-shape but also multi-scale information in a single layer instead of paralleled [9, 75, 71] or stacked [73, 18] multi-layers. From a macro perspective, the proposed approach helps control the information flow within network through learning the semantic and spatial relationship of features and determine the information passing or suppression.

The proposed scale- and shape-variant context model enhances the discriminative power of the high level features. Higher level features are more robust to noise than lower level features at a price of lower spatial position sensitivity. Thus, many segmentation networks also aggregate low-level features to improve the position accuracy of the segmentation [51, 11, 18, 25]. However, aggregating low-level features, though helps to recover spatial information, may bring some debatable noise sensitive information that leads to a wrong classification of some pixels. Thus, we propose a model that utilizes the higher level features to attenuate the noisy information of the low-level features before aggregating them, i.e. signals denoising. In such a way, the network could better exploit the advantages of low-level features by alleviating their problems.

In summary, this paper makes the following contributions: 1) we propose a novel paired convolution to infer the semantic correlation of two pixels and based on that to generate a semantic-correlated region for each pixel; 2) we propose a shape-variant context aggregated within the semantic-correlated region to model the diverse shapes and scales of contexts, which greatly enhances the modeling ability of network; 3) we propose a labeling denoising model to reduce the labeling errors caused by the noisy low-level features; 4) we achieve new state-of-the-art performance consistently on six public semantic segmentation datasets.

2. Related work

Recently, Deep Neural Networks has achieved great success on computer vision [26, 23, 22, 47, 49, 24, 50, 65]. Based on the Fully Convolutional Network (FCN) [51], in which the fully connected layers in original CNNs are converted to convolutional layers, a lot of approaches, e.g., [9, 21, 55, 32, 48, 76, 5, 61], are proposed to improve the performance of semantic segmentation.

Contextual features modeling plays an important role in scene parsing. [52] shows that global spatial information helps enforce the features consistency. DeepLab [9] proposes atrous spatial pyramid pooling (ASPP) to aggregate multi-scale image representations from parallel branches with different dilated rate. DilatedNet [73] appends several dilated convolution layers after the score maps to perform multi-scale context aggregation. DAG-RNN [62] and Byeon [6] propose to model long-range context through recurrent neural networks. Zoom-out [53] proposes a feed-forward architecture to extract hierarchical zoom-out features. CRF-RNN [77] uses recurrent layers for jointly end-to-end training the dense CRFs [36] with their segmentation networks. Piecewise [44] formulates CNN-based pairwise potential functions to capture patch-patch context and designs image pyramid input for patch-background context. PSPNet [75] introduces pyramid spatial pooling (PSP) to perform different-region-based global information aggregation. Recently, CCL [18] proposes a context contrasted local model to parallelly collect local and its surrounding information. EncNet [74] encodes semantic context to network and stress class-dependent feature maps.

Different from previous methods, in this work, we try to aggregate context information from semantic-closer region but suppress irrelevant information even in the spatial-closer region. We propose a shape-adaptive convolutional layer to learn diverse shaped contexts whose shapes are determined by the object shape, scale and its surrounding support of the input image. The proposed approach is aimed at not only retaining the location identity and layout information but also building the effective semantic correlation shown in the training images.

Figure 2. (Best viewed in color) We propose a novel semantic correlation dependent shape-variant context, which boosts the semantic-correlated features (magenta color) while suppresses the others (white color).

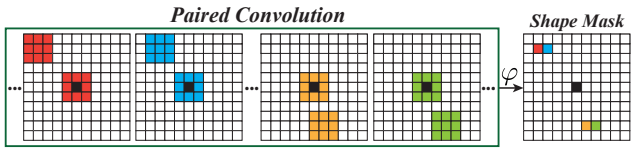

Figure 3. (Best viewed in color) A shape mask is inferred by the Paired Convolution and Gaussian mapping function φ, which are designed to learn the semantic correlation between the target pixel and other pixels within the shape mask. Here we show an example of 4 values in a 11×11 mask of the target pixel (dark), in which 4 mask values are generated by the 4 filters of the same color.

Label variety is another challenging problem in semantic segmentation. PSPNet [75] observes the confusion categories and demonstrates that the PSPNet can better address the confusion labels than FCN[51]. Geng et al. [20] propose to infer the discriminative confusing groups from a prior confusion matrix. DFN [72] introduces a smooth and border network to tackle confusing classes. Davis et al. [17] propose to refine the parsing results using Bayesian strategy with confusion probabilities and label priors. Huang et al. [29] propose a LabelReplacement network to correct the error predictions. Different from these methods, we propose a learnable Labeling Denoising (LD) model to solve the problems of confusion labels by utilizing the robust higher level features to attenuate the noise in lower level features.

3. The Proposed Approach

3.1. Semantic Correlation Dependent Context

Semantic Segmentation needs to simultaneously deal with object recognition and localization, and hence should build the dense feature connections among large region as well as retain the location identity. Meanwhile, due to the diverse shapes and complicated layout of objects in scene images, the scales and shapes of contexts for different objects are supposed to have very large variation.

Many existing context modeling methods tend to aggregate surrounding information with a fixed size of rectangle window across all locations, which weaken the location identity and might not be able to effectively represent the diverse shapes and scales of objects in scene images. Different from previous works, we put forward that the more desirable context region should be shape-variant according to the shape of the object and its background that support the object. For example, for pixels belong to the train in Figure 2, the more beneficial context should be the surrounding information along the railroad track (magenta color), which are closer in semantics than in space. In a word, for different pixel positions, surrounding information should be collected from semantic correlated region that supports the existence of the correct class of this pixel. Therefore, in this work, we propose a semantic correlation dependent shape-variant context (SVC) to model diverse shape/scale contexts with location identity. In the SVC, context aggregation is controlled by a semantic correlation mask, specifying where the information should be collected with what extent. With the semantic mask, features in semantic correlated regions are boosted and other irrelevant ones are suppressed. Thus, better context information for parsing of each pixel is aggregated within the specific shape region that supports the existence of the correct class of the object.

Learn the Semantic Correlation

The shape masks that represent semantic correlation trim the context shapes and decide where the information should be collected with what extent. Next we discuss how to learn the semantic correlation, i.e. how to generate the semantic shape masks. Each value in a shape mask represents the correlation of the corresponding pixel to the target pixel (the center pixel of the mask). Thus, the semantic relationship of each pixel with the target pixel need be learned and injected to the corresponding position in the shape mask. To this end, we introduce a paired convolution, as shown in Figure 3, in which a pair of 3×3 local convolution with specific relative position are employed to learn the semantic and spatial correlation of the corresponding pixel with the target pixel (the central dark pixel in Figure 3). In each filter of the paired convolution, there are a central convolution for the target pixel and another convolution whose position corresponds to the position in shape mask for the corresponding pixel. We have observed and hence assumed that the feature appearances of pixels belonging to the same object and its context will show strong correlation because they frequently coexist in the training images. Therefore, the difference of the two convolution outputs $\mathcal{D}^{i,j}_{m,n}$ can be minimized for two pixels belonging to the same object and its context by learning the convolution parameters from the training images.

$$\mathcal{D}^{i,j}_{m,n} = \mathcal{F}^{i,j} * \Theta^{m,n}_{0,0} - \mathcal{F}^{i-m,j-n} * \Theta^{m,n}_{m,n} \qquad (1)$$

where $*$ donates the local convolution operator, $\mathcal{D}^{i,j}_{m,n}$ represents the convolution output discrepancy of $(i-m, j-n)$ to target position (i, j), $\mathcal{F}^{i,j}$ and $\mathcal{F}^{i-m,j-n}$ are local

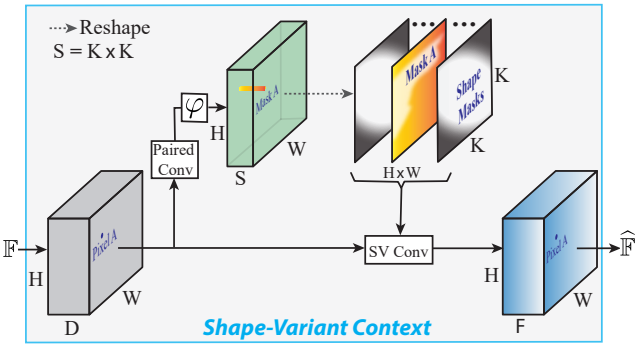

Figure 4. Semantic correlation-dependent shape-variant context aggregates surrounding information according to the semantic correlation and hence customizes an effective contextual region. It helps control the information flow within network via deciding what information to be passed or suppressed.

features at position (i, j) and $(i-m, j-m)$, $\Theta_{0,0}^{m,n}$ and $\Theta_{m,n}^{m,n}$ are corresponding parameters of two local kernels in the paired convolution. As there might be negative or positive difference, we map it to the value of our context shape mask by a Gaussian function:

$$\mathcal{M}_{m,n}^{i,j} = \varphi(\mathcal{D}_{m,n}^{i,j}) \tag{2}$$

where $\varphi(a)=exp(-\frac{a^2}{\sigma^2})$, which maps the convolution output discrepancy to a semantic correlation value. A smaller discrepancy generates a higher semantic correlation value. $\mathcal{M}_{m,n}^{i,j}$ is the mask value at position (m, n) in the semantic shape mask of pixel (i, j). Note that the result is not sensitive to σ as the parameters of the two convolutions are learnt. We use $\sigma = 3$ in our experiments.

Shape-Variant Context

The goal of shape-variant context is to customize a desirable shape/scale of context for each pixel instead of a simple smooth context information. To achieve this, we further propose a shape-variant convolution (SV Conv) to adaptively collect the surrounding information. The parameters of shape-variant convolution consists of location-invariant learnable convolution parameters and semantic shape mask inferred by the proposed paired convolution. The shape mask is used to control the receptive field of the convolution process for each position according to the semantic correlation. Such shape mask crops the convolution kernels into different shapes/scales and leads to a shape-variant convolution operation. In such a way, the proposed method greatly enhances the network modeling ability of diverse shape context.

The proposed shape-variant context is shown in Figure 4. There are two branches, the bypass is designed to learn the semantic correlation, whose outputs are then input to the shape-variant convolution (SV Conv) to provide the semantic shape mask. In detail, the side branch employs paired convolution to learn from the local feature $\mathbb{F}$ from a pre-trained CNN the correlations of each pixel with each of all other pixels within the kernel of size $K\times K$ centered at this pixel, as described in Eq. (1) and Eq. (2). The number of the output channels is $S{=}K\times K$ where $K\times K$ is the kernel size of the proposed shape-variant convolution. The semantic shape masks inferred form input features are employed to weight the normal learnable convolutions (F filters) parameterized by $\theta_{m,n}^{d,f}$ of the main branch by:

$$\hat{\theta}_{m,n}^{i,j,d,f} = \mathcal{M}_{m,n}^{i,j}\theta_{m,n}^{d,f} \tag{3}$$

where $\theta_{m,n}^{d,f}{\in}\Theta$ is the convolution parameter for d^{th} input channel at position (m, n) of the f^{th} normal location-invariant learnable filter, $d{\in}(1, 2..., D)$ and $f{\in}(1, 2..., F)$. The filter kernel size is $K\times K$ and (i, j) is the index of feature map position across all the $H\times W$ positions. By Eq. (3), the receptive field of the F normal learnable convolutions parameterized by $\theta_{m,n}^{d,f}$ is transformed from a fixed size of $K\times K$ to effectively different sizes and shapes for different pixels (i, j) determined by the proposed semantic shape masks $\mathcal{M}_{m,n}^{i,j}$. The resulting F shape-variant filters are employed to generate variegated shape context for each spatial position (i, j):

$$\widehat{\mathcal{F}}^{i,j,f} = \sum_{d=1}^{D}\sum_{m=-K_1}^{K_1}\sum_{n=-K_1}^{K_1}\hat{\theta}_{m,n}^{i,j,d,f}\mathcal{F}^{i-m,j-n,d} \tag{4}$$

where $K_1{=}(K-1)/2$ and $\mathcal{F}^{i-m,j-n,d} \in \mathbb{F}$. $\widehat{\mathcal{F}}^{i,j,f} \in \widehat{\mathbb{F}}$ is one target contextual feature map. In such a way, $\mathcal{M}_{m,n}^{i,j}$ instructs how to collect $\mathcal{F}^{i-m,j-n}$ for $\widehat{\mathcal{F}}^{i,j}$ during convolution. All of these functions are differentiable and its back-propagation is easy to derive.

The standard convolutional operation is location invariant and do not vary with testing images after training. Thus, it could not customize different shapes/scales of context information for different objects of input images. The proposed SV Conv consists of a learnable location-invariant convolution and a location-variant semantic shape mask inferred from the input image. The former is to model the statistical average of the spatial-channel distribution and the latter is to determine the size and shape of the convolution receptive field. They together function as a shape-variant operator to better model the shaped context.

Modeling Diverse Shapes in a Single Layer

Due to the diverse shapes/scales of contextual regions and the shape constraint of convolutional kernels, it is difficult to use a single normal convolutional layer to effectively model shape-variant context because the scope of the context, including its scale and contour, dramatically changes for different objects of input images. With the proposed shape-variant semantic correlation masks,

convolution regions with diverse shapes and scales are specified and thus we could model multi-shape and multi-scale information in a single layer.

Comparison with State-of-the-art Context Models

Different from the previous context methods that tend to model statistical average representation within a predefined rectangle region, e.g., [9, 73, 75, 18, 71, 56], the proposed approach utilizes the semantic correlation and intentionally picks up the relevant information according to the semantic shape mask inferred by the feature appearance of testing image. Thus, it could not only retain the shape and location identity but also effectively build the beneficial connection among correlated features for classification. Comparing with the deformable convolution [16], the objective of the proposed approach is similar to it in terms of collecting the relevant information in the convolution. However, different from [16] that tries to achieve this via deforming the sampling locations, the proposed SVC finds out the semantic correlation to enhance or attenuate the corresponding information, explicitly leading to the shape- and scale-variant modeling. The criteria or the methods to find the relevant information in the two approaches are also different. Furthermore, our method models diverse shapes of semantic-dependent context in a single layer without stacking layers in [16], and avoids the "atrous" in deformable convolution that may lose some detailed information.

3.2. Labeling Denoising

Due to the label variety and the complicated correlation among labels in segmentation datasets, regular errors can be found in the results of most state-of-the-arts segmentation networks [9, 75, 17, 11, 43]. These regular errors could be categorized into "in-context" error and "out-context" error, as show in Figure 5. The "in-context" error is mainly caused by inaccurate positioning and inter-context influence while the "out-context" error is mainly caused by inaccurate classification. The proposed shape-variant context aggregates information from specific semantic correlated region, which helps mitigate the in-context and out-context labeling errors. To get elaborate spatial information, lower-level features from middle layers of CNNs are important in the encoder-decoder architecture [51, 11, 18, 25] as they contain more information about where these objects are [21, 51]. But these low-level features also bring debatable noisy information that results in out-context errors. In contrast, the high-level features, e.g., the shape-variant context in this work, though less sensitive to the spatial location, are more robust to noise and more aware of what categories existing in a scene image. To better combine "what" and "where", we propose a labeling denoising (LD) model in this work that attenuates the noise information when extracting low-level spatial information from middle layers.

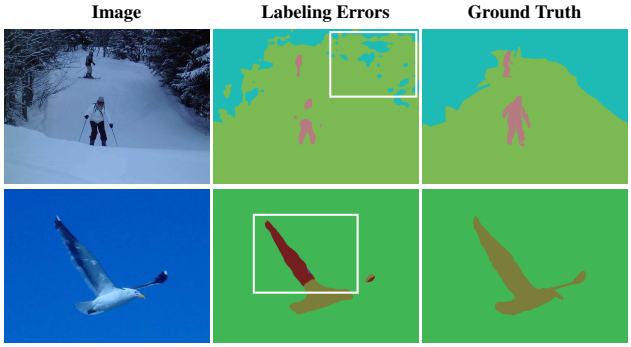

Figure 5. In-context error, e.g., the first row, refers to incorrect labeling within the label set of the image. Out-context error, e.g., the second row, refers to incorrect labeling outside the label set of the image.

The labeling denoising model first infers the existence potential of each category from a higher-level block and learns penalty scores from the existence potentials. Then, the score maps generated from a lower-level block are charged by the penalty scores. Using the penalty scores learnt from the higher-level block, the scores of nonexistent categories of an input image generated by the lower-level block are greatly suppressed. First, the existence potential is inferred by the score maps from a higher-level block by:

$$\mathcal{E}_k = F_g(F_{sf}(\mathcal{S}_k)) \tag{5}$$

where $\mathcal{S}_k$ is the score maps from a higher-level (level k) block, F_{sf} is softmax and F_g is global max pooling. $\mathcal{E}_k = (e_k^1, ..., e_k^c, ..., e_k^C)$ and e_k^c is the existent potential for class c inferred by level k. Then, the penalty $\mathcal{P}^c$ is learnt by:

$$\mathcal{P}_k^c = \text{ReLU}(\mathcal{T} - e_k^c)\Delta_k^c \tag{6}$$

where $\mathcal{T}$ is a penalty threshold and Δ_k^c is a learnable penalty parameter. The penalty threshold and function ReLU are used to keep the score distribution of existent classes unchanged. The penalty $\mathcal{P}_k^c$ is used to modify the score map of its next lower-level block before aggregating it into the upsampled score map of its next higher-level block as:

$$\mathcal{S}_{k-1}^c = \text{ReLU}(\hat{\mathcal{S}}_{k-1}^c - \mathcal{P}_k^c) + \mathcal{S}_k^c \tag{7}$$

where $\hat{\mathcal{S}}_{k-1}^c$ is the score map of class c directly from the lower-level block. $\mathcal{S}_{k-1}^c$ is the denoised and aggregated score map from the highest level to the level $k-1$, which is further used to modify and be aggregated to the score maps of the lower-level blocks as Eqs (5), (6), and (7). The proposed labeling denoising (LD) model is shown in Figure 6. In such a way the network could take advantage of both high-level features and low-level features, i.e. better combine "what" and "where". For the skip layers from low-level features, the scores of nonexistent categories in

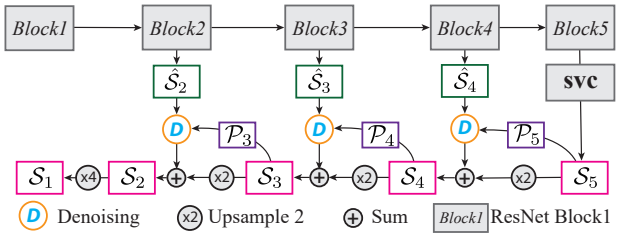

Figure 6. Network Architecture. We use ResNet-101 as our base model for fine-tuning and FCN-4s as the backbone segmentation framework. LD is used in decode process for denoising.

an input image are attenuated and those of the existent categories are retained and supplemented to score maps for positioning enhancement. As this mechanism is included in the end-to-end training process, less noisy scores are taken into account during training, and gradients for training such noisy information could be saved for other things such as positioning. The proposed approach could also be viewed as some kind of dropout, which applies dropout to connections that reach some conditions.

4. Experiments

We evaluate the proposed approach on six public benchmark, COCO-Stuff, SIFT-Flow, CamVid, PASCAL-Person-Part, PASCAL-Context, and Cityscapes. We use ResNet-101[26] pre-trained on ImageNet [58] as our base model for fine-tuning and FCN-4s as the backbone framework. During training, the proposed Network is trained end-to-end using standard SGD with batch size 8, fixed momentum 0.9 and weight decay 0.0005. Data augmentations like random flipping, random resize between 0.8 and 1.2 and mean subtraction are used in training. Inspired by [9], we use the "poly" learning rate and set the initial learning rate to 5×10^{-3} for newly initialized parameters and 10^{-4} for parameters of pretrained layers, the power is set to 0.9. Batch Normalization [31] is used in new added layers to accelerate training process. The performance is evaluated by standard pixel accuracy (pixel acc.), mean class accuracy (mean acc.) and mean Intersection-over-Union (mean IoU). Please refer to [51] for mathematical definitions.

To model diverse semantic shapes in a single layer, larger kernel is required due to the dramatic changed shapes/scales of objects. But very large kernel is resource-intensive and difficult to converge. To address this issue, we modify Eq. (4) of the proposed SVC similar to the depthwise separable convolution [13]. The simplified computation of Eq. (4) allows us using large kernel size to model the diverse shapes in spatial space followed by a pointwise convolution to learn the cross-channel correlation. And in labeling denoising model, we use ascending penalty thresholds $\mathcal{T} = t, 2t, 4t$, from the highest to the lowest blocks, where $t = \frac{1}{C}$ and C is the number of classes.

Methods	PASCAL-Context	COCO-Stuff
Baseline	42.7	31.5
Baseline+SVC	52.4	38.5
Baseline+SVC+LD	**53.2**	**39.6**

Table 1. Ablation study of the proposed approach in terms of IoU.

Kernel Size	SFC	SVC
0×0	42.7	42.7
7×7	45.6	48.5
11×11	46.6	49.4
15×15	47.1	51.2
19×19	47.0	52.1
23×23	46.7	**52.4**
27×27	46.5	52.3

Table 2. Ablation study of the proposed shape-variant context (SVC) approach by comparing it with the shape-fixed context (SFC) on different kernel size. It also shows that the performance gain is not simply brought by the increased number of parameters.

4.1. Ablation Study

In this section we do ablation studies of the proposed shape-variant context (SVC) and labeling denoising (LD). As shown in Table 1, comparing the performance gain brought by the proposed LD on PASCAL-Conext (59 classes) and COCO-Stuff (171 classes), we can conclude that the LD could mitigate noisy prediction and it works better on datasets with more semantic categories. This is not a surprise because more categories cause heavier prediction noise and hence the LD works more effectively. Table 1 shows the significant performance gain (7 percent on COCO-Stuff and almost 10 percent on PASCAL-Context) from the baseline by applying the proposed SVC.

To further study where the performance gain of the proposed SVC comes from, we compare it with the shape-fixed context (SFC) that is implemented by setting a constant shape mask, i.e., $\mathcal{M}^{i,j}_{m,n}$=1 in Eq. (3). We compare them on PASCAL-Context with different kernel sizes shown in Table 2, where the 0×0 means the baseline without the context layer. As we employ just a single layer to capture the context information, all kernels used in Table 2 are larger than convolutional kernels of most other work so that diverse shapes and scales of contextual information could be modelled in a single layer. Table 2 shows that with the increase of kernel size, the segmentation performance improves up to a certain level then slightly drops with further increase of the kernel size. This is because the network lose too much locality information in overlarge kernel situation. It also shows that a simple increase of the network parameters may not always improve the performance. Table 2 shows that the proposed SVC significantly outperforms the SFC at all different kernel sizes. It is also not a surprise that the best performance of the SVC is achieved at the kernel size (23×23), much larger

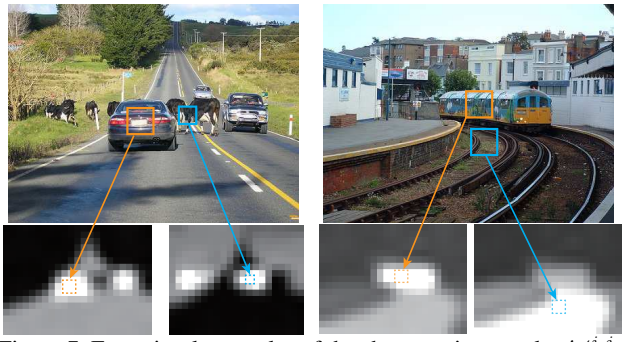

Figure 7. Four visual examples of the shape-variant masks $\mathcal{M}_{m,n}^{i,j}$ generated at four different locations of two testing images by the learned network. The mask center (i, j) is indicated by the small square and its value at (m, n) within the image is shown by the gray level.

than that of the SFC (15×15) because the proposed SVC provides diverse shape context, which is implicitly multi-scale with better location identity.

4.2. Visualization of the semantic shape mask

As Table 1 and Table 2 show the significant performance gains brought by the proposed shape-variant context (SVC) that is determined by the proposed shape mask $\mathcal{M}_{m,n}^{i,j}$, it is worthwhile to further study how the mask captures the shape of the context by visualizing it. Four examples of the shape mask $\mathcal{M}_{m,n}^{i,j}$ generated at four different locations of two testing images by the learned network are shown in Figure 7. The first is the mask of the center of a car. It has higher values at pixels of cars and road as they contains the context information of the center of the car. The second is the mask of some pixel of the cow in the middle of a road. It has higher values at pixels of cows and grass, though they are far away from the target cow and separated by the road. Values of the second mask are low in the area of road as it does not show correlation with the cow in the training database. Consistently, the semantic correlation of the context is also shown by the third and fourth masks respectively for a train and a railway in the second testing image.

4.3. Comparison with the State-of-the-Arts

The proposed semantic segmentation network is named as SVCNet and we compare it with the state-of-the-arts on six public benchmark, COCO-Stuff, SIFT-Flow, CamVid, PASCAL-Person-Part, PASCAL-Context, and Cityscapes. Before the quantitative comparison, some qualitative results of the proposed SVCNet are shown in Figure 8.

COCO-Stuff [7] provides dense pixel-wise annotations for 171 semantic categories. There are 9000 images used for training and 1000 images used for testing. Quantitative results of COCO-Stuff are shown in Table 3. The proposed SVCNet outperforms the previous state-of-the-arts across all evaluation metrics.

Methods	pixel acc.	mean acc.	mean IoU
FCN [7]	52.0	34.0	22.7
DeepLab [8]	57.8	38.1	26.9
FCN-8s [51]	60.4	38.5	27.2
DAG-RNN+CRF [62]	63.0	42.8	31.2
DC+FCN+ [27]	65.5	44.6	33.6
Deeplab-V2 [9]	65.1	45.5	34.4
CCL-ResNet101 [18]	66.3	48.8	35.7
DSSPN [41]	68.5	48.1	36.2
SVCNet (ours)	**69.2**	**51.5**	**39.6**

Table 3. COCO-Stuff testing accuracies.

Methods	pixel acc.	mean acc.	mean IoU
Liu et al. [46]	76.7	-	-
Tighe et al. [66]	75.6	41.1	-
Farabet et al. [19]	78.5	29.6	-
Pinheiro et al. [57]	77.7	29.8	-
Sharma et al. [59]	79.6	33.6	-
Yang et al. [70]	79.8	48.7	-
FCN-8s [60]	85.9	53.9	41.2
DAG-RNN+CRF [62]	87.8	57.8	44.8
Piecewise [44]	88.1	53.4	44.9
SVCNet (ours)	**89.1**	**58.2**	**46.3**

Table 4. SIFT-Flow testing accuracies.

Methods	mean IoU
DeconvNet [55]	48.9
SegNet [2]	50.2
DeepLab [8]	54.7
DilatedNet [73]	65.3
Dilation+FSO [38]	66.1
FC-DenseNet [33]	66.9
G-FRNet [32]	68.0
DenseDecoder [3]	70.9
SVCNet (ours)	**75.4**

Table 5. CamVid.

Methods	mean IoU
Attention [10]	56.4
HAZN [68]	57.5
LG-LSTM [40]	58.0
Graph LSTM [39]	60.2
DeepLab [8]	62.8
DeepLab-V2 [9]	64.9
RefineNet [43]	68.6
DenseDecoder [3]	68.6
SVCNet (ours)	**73.9**

Table 6. PASCAL-Person-Part.

SIFT-Flow [45] contains 2688 images annotated with 33 semantic classes. There are 2488 training images and 200 testing images. Quantitative results are shown in Table 4. The proposed SVCNet outperforms previous state-of-the-arts across all evaluation metrics.

CamVid [4] is a road scene image segmentation dataset which provides dense pixel-wise annotations for 11 semantic categories. There are 367 training images, 101 validation images and 233 testing images. The testing results are shown in Table 5. It shows that the proposed SVCNet outperforms previous state-of-the-arts by a large margin.

PASCAL-Person-Part [12] provides pixel-level labels for six person parts. There are 1717 training/validation images and 1818 testing images. Quantitative results of PASCAL-Person-Part are reported in Table 6. It shows that the proposed SVCNet outperforms the previous state-of-the-arts by a large margin on this small dataset, which indicates

Methods	mean IoU
FCN-8s [60]	39.1
CRF-RNN [77]	39.3
BoxSup [15]	40.5
HO-CRF [1]	41.3
Piecewise [44]	43.3
FCRN [67]	44.5
EFCN [61]	45.0
DeepLab-V2[9]	45.7
Global-Context[30]	46.5
RefineNet-ResNet152 [43]	47.3
DenseDecoder [3]	47.8
MSCI [42]	50.3
CCL-ResNet101 [18]	51.6
EncNet [74]	51.7
SVCNet (ours)	**53.2**

Table 7. PASCAL-Context testing accuracies.

Methods	mean IoU
Deeplab-v2 [9]	70.4
RefineNet-Res101 [43]	73.6
DSSPN-Universal [41]	76.6
GCN [56]	76.9
DepthSet [35]	78.2
PSPNet [75]	78.4
AAF [34]	79.1
DFN [72]	79.3
PSANet [76]	80.1
DenseASPP-DenseNet161 [71]	80.6
SVCNet (ours)	**81.0**

Table 8. Cityscapes testing accuracies.

that the the proposed approach could be trained very well even on small dataset.

PASCAL-Context [54] provides pixel-wise segmentation annotation for 59 classes. There are 4998 training images and 5105 testing images. Quantitative results of Pascal Context are shown in Table 7. It shows that the proposed SVCNet outperforms the state-of-the-arts by a large margin.

Cityscapes [14] contains 5000 street scene images with pixel-level fine annotations and 19 classes are considered for evaluation. There are 2975 training images, 500 validation images and 1525 testing images. The test results are shown in Table 8.

5. Conclusion

In this work, we propose to aggregate the context information based on the semantic correlation rather than the predefined spatial-dependent window to collect more effective and discriminative surrounding information for semantic segmentation. The semantic-correlated information even at a far away spatial location will be enhanced and the semantic-uncorrelated information even at a close spatial location will be suppressed in collecting the context information. To this end, we first propose a novel paired convolution to learn the feature semantic-correlation from the training images and to infer it of the query image. This generates a semantic shape mask at each position of the image. Based on it, we propose a shape-variant convolution, in which the receptive field of the convolution is specified by different semantic shape masks at different positions of different query images. The semantic shape masks form diverse scales and shapes of the convolution receptive field to aggregate discriminative context information effectively. Furthermore, to ease the labeling errors, we propose a labeling denoising model, which utilizes more robust higher-level features to attenuate the prediction errors caused by noisier lower-level features. Without bells and whistles, the proposed segmentation network achieves new state-of-the-arts consistently on the six public semantic segmentation datasets, COCO-Stuff, SIFT-Flow, CamVid, PASCAL-Person-Part, PASCAL-Context, and Cityscapes.

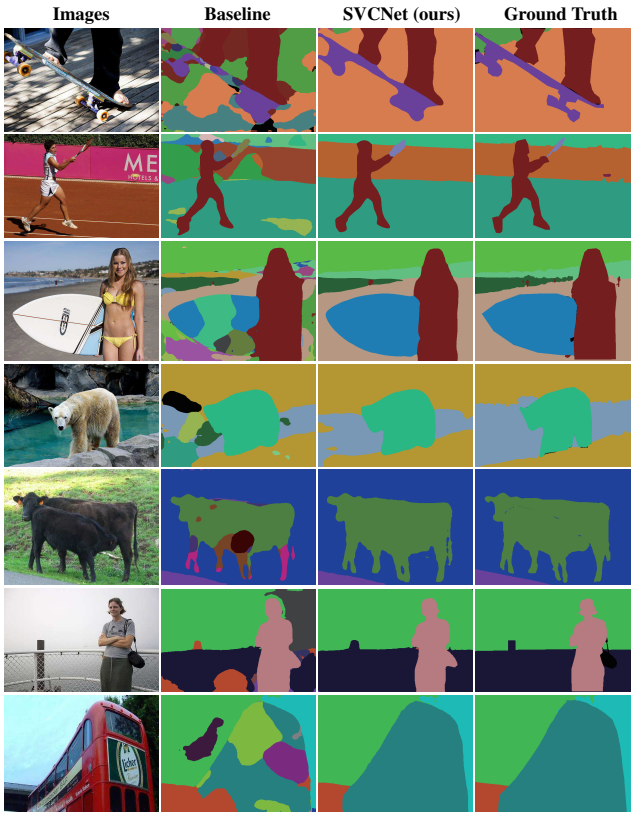

Figure 8. Qualitative segmentation examples on COCO-Stuff (1st-4th rows) and PASCAL-Context (5th-7th rows).

Acknowledgement

This research is supported by Singapore Ministry of Education Academic Research Fund Grant no: 2015-T1-002-140, MOE Tier 1 RG 123/15. It is also supported by the BeingTogether Centre, a collaboration between Nanyang Technological University (NTU) Singapore and University of North Carolina (UNC) at Chapel Hill. The BeingTogether Centre is supported by the National Research Foundation, Prime Ministers Office, Singapore under its International Research Centres in Singapore Funding Initiative.

References

[1] Anurag Arnab, Sadeep Jayasumana, Shuai Zheng, and Philip HS Torr. Higher order conditional random fields in deep neural networks. In *ECCV*, 2016.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *TPAMI*, 2017.

[3] Piotr Bilinski and Victor Prisacariu. Dense decoder shortcut connections for single-pass semantic segmentation. In *CVPR*, 2018.

[4] Gabriel J Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla. Segmentation and recognition using structure from motion point clouds. In *ECCV*, 2008.

[5] S Rota Bulo, Gerhard Neuhold, and Peter Kontschieder. Loss max-pooling for semantic image segmentation. In *CVPR*, 2017.

[6] Wonmin Byeon, Thomas M. Breuel, Federico Raue, and Marcus Liwicki. Scene labeling with lstm recurrent neural networks. In *CVPR*, 2015.

[7] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *CVPR*, 2018.

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015.

[9] 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. *TPAMI*, 2018.

[10] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, 2016.

[11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv:1802.02611*, 2018.

[12] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, 2014.

[13] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2017.

[14] 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 *CVPR*, 2016.

[15] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, 2015.

[16] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017.

[17] James W Davis, Christopher Menart, Muhammad Akbar, and Roman Ilin. A classification refinement strategy for semantic segmentation. *arXiv:1801.07674*, 2018.

[18] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *CVPR*, 2018.

[19] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *TPAMI*, 35(8), 2013.

[20] Qichuan Geng, Xinyu Huang, Zhong Zhou, and Ruigang Yang. A network structure to explicitly reduce confusion errors in semantic segmentation. *arXiv:1808.00313*, 2018.

[21] Golnaz Ghiasi and Charless C Fowlkes. Laplacian pyramid reconstruction and refinement for semantic segmentation. In *ECCV*, 2016.

[22] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, and Gang Wang. Unpaired image captioning by language pivoting. In *ECCV*, 2018.

[23] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. Recent advances in convolutional neural networks. *Pattern Recognition*, 2018.

[24] Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. Scene graph generation with external knowledge and image reconstruction. In *CVPR*, 2019.

[25] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *CVPR*, 2015.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[27] Hexiang Hu, Zhiwei Deng, Guang-Tong Zhou, Fei Sha, and Greg Mori. Labelbank: Revisiting global perspectives for semantic segmentation. *arXiv:1703.09891*, 2017.

[28] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[29] Yu-Hui Huang, Xu Jia, Stamatios Georgoulis, Tinne Tuytelaars, and Luc Van Gool. Error correction for dense semantic image labeling. *arXiv:1712.03812*, 2017.

[30] Wei-Chih Hung, Yi-Hsuan Tsai, Xiaohui Shen, Zhe L Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Scene parsing with global context embedding. In *ICCV*, 2017.

[31] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[32] Md Amirul Islam, Mrigank Rochan, Neil DB Bruce, and Yang Wang. Gated feedback refinement network for dense image labeling. In *CVPR*, 2017.

[33] Simon Jégou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *CVPRW*, 2017.

[34] Tsung-Wei Ke, Jyh-Jing Hwang, Ziwei Liu, and Stella X Yu. Adaptive affinity fields for semantic segmentation. In *ECCV*, 2018.

[35] Shu Kong and Charless C Fowlkes. Recurrent scene parsing with perspective understanding in the loop. In *CVPR*, 2018.

[36] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011.

[37] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[38] Abhijit Kundu, Vibhav Vineet, and Vladlen Koltun. Feature space optimization for semantic video segmentation. In *CVPR*, 2016.

[39] Xiaodan Liang, Xiaohui Shen, Jiashi Feng, Liang Lin, and Shuicheng Yan. Semantic object parsing with graph lstm. In *ECCV*, 2016.

[40] Xiaodan Liang, Xiaohui Shen, Donglai Xiang, Jiashi Feng, Liang Lin, and Shuicheng Yan. Semantic object parsing with local-global long short-term memory. In *CVPR*, 2016.

[41] Xiaodan Liang, Hongfei Zhou, and Eric Xing. Dynamic-structured semantic propagation network. In *CVPR*, 2018.

[42] Di Lin, Yuanfeng Ji, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Multi-scale context intertwining for semantic segmentation. In *ECCV*, 2018.

[43] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, 2017.

[44] G. Lin, C. Shen, A. van dan Hengel, and I. Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *CVPR*, 2016.

[45] Ce Liu, Jenny Yuen, and Antonio Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *CVPR*, 2009.

[46] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *TPAMI*, 2011.

[47] Jun Liu, Henghui Ding, Amir Shahroudy, Ling-Yu Duan, Xudong Jiang, Gang Wang, and Alex Kot Chichung. Feature boosting network for 3d pose estimation. *TPAMI*, 2019.

[48] Ziwei Liu, Xiaoxiao Li, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Semantic image segmentation via deep parsing network. In *ICCV*, 2015.

[49] Zichuan Liu, Guosheng Lin, Sheng Yang, Jiashi Feng, Weisi Lin, and Wang Ling Goh. Learning markov clustering networks for scene text detection. In *CVPR*, 2018.

[50] Zichuan Liu, Guosheng Lin, Sheng Yang, Fayao Liu, Weisi Lin, and Wang Ling Goh. Towards robust curve text detection with conditional spatial expansion. In *CVPR*, 2019.

[51] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[52] Aurelien Lucchi, Yunpeng Li, Xavier Boix, Kevin Smith, and Pascal Fua. Are spatial and global constraints really necessary for segmentation? In *ICCV*, 2011.

[53] Mohammadreza Mostajabi, Payman Yadollahpour, and Gregory Shakhnarovich. Feedforward semantic segmentation with zoom-out features. In *CVPR*, 2015.

[54] 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 *CVPR*, 2014.

[55] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *ICCV*, 2015.

[56] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters – improve semantic segmentation by global convolutional network. In *CVPR*, 2017.

[57] Pedro Pinheiro and Ronan Collobert. Recurrent convolutional neural networks for scene labeling. In *ICML*, 2014.

[58] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

[59] Abhishek Sharma, Oncel Tuzel, and Ming-Yu Liu. Recursive context propagation network for semantic scene labeling. In *NIPS*, 2014.

[60] Evan Shelhamer, Jonathon Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *TPAMI*, 2016.

[61] Bing Shuai, Henghui Ding, Ting Liu, Gang Wang, and Xudong Jiang. Toward achieving robust low-level and high-level scene parsing. *TIP*, 2019.

[62] Bing Shuai, Zhen Zuo, Bing Wang, and Gang Wang. Scene segmentation with dag-recurrent neural networks. *TPAMI*, 2018.

[63] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[64] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[65] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap. Aanet: Attribute attention netowrk for person re-identification. In *CVPR*, 2019.

[66] Joseph Tighe and Svetlana Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *CVPR*, 2013.

[67] Zifeng Wu, Chunhua Shen, and Anton van den Hengel. Bridging category-level and instance-level semantic image segmentation. *arXiv:1605.06885*, 2016.

[68] Fangting Xia, Peng Wang, Liang-Chieh Chen, and Alan L Yuille. Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In *ECCV*, 2016.

[69] Zhicheng Yan, Hao Zhang, Yangqing Jia, Thomas Breuel, and Yizhou Yu. Combining the best of convolutional layers and recurrent layers: A hybrid network for semantic segmentation. *arXiv:1603.04871*, 2016.

[70] Jimei Yang, Brian Price, Scott Cohen, and Ming-Hsuan Yang. Context driven scene parsing with attention to rare classes. In *CVPR*, 2014.

[71] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *CVPR*, 2018.

[72] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. In *CVPR*, 2018.

[73] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv:1511.07122*, 2015.

[74] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *CVPR*, 2018.

[75] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.

[76] 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 *ECCV*, 2018.

[77] 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 *ICCV*, 2015.