Adaptive Pyramid Context Network for Semantic Segmentation

Junjun He [1,2] Zhongying Deng [1] Lei Zhou [1] Yali Wang [1] Yu Qiao* [1,3]
[1] Shenzhen Key Lab of Computer Vision and Pattern Recognition, SIAT-SenseTime Joint Lab,
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
[2] Shanghai Jiao Tong University [3] The Chinese University of Hong Kong

Abstract

Recent studies witnessed that context features can significantly improve the performance of deep semantic segmentation networks. Current context based segmentation methods differ with each other in how to construct context features and perform differently in practice. This paper firstly introduces three desirable properties of context features in segmentation task. Specially, we find that Global-guided Local Affinity (GLA) can play a vital role in constructing effective context features, while this property has been largely ignored in previous works. Based on this analysis, this paper proposes Adaptive Pyramid Context Network (APCNet) for semantic segmentation. APCNet adaptively constructs multi-scale contextual representations with multiple well-designed Adaptive Context Modules (ACMs). Specifically, each ACM leverages a global image representation as a guidance to estimate the local affinity coefficients for each sub-region, and then calculates a context vector with these affinities. We empirically evaluate our APCNet on three semantic segmentation and scene parsing datasets, including PASCAL VOC 2012, Pascal-Context, and ADE20K dataset. Experimental results show that APCNet achieves state-of-the-art performance on all three benchmarks, and obtains a new record 84.2% on PASCAL VOC 2012 test set without MS COCO pre-trained and any post-processing.

1. Introduction

Semantic segmentation, aiming at assigning a category label for each pixel, is a fundamental yet important problem in computer vision, with wide applications in scene understanding, medical imaging, robot vision etc [27] [28]. The challenge of semantic segmentation comes from the inner content, shape, and scale variations of the same object/stuff, as well as the easily confused and fine boundaries among different objects/stuff. Current state-of-the-art semantic segmentation methods heavily exploit deep convolutional neural networks (CNNs), e.g. Fully Convolutional Network (FCN) [22], U-Net [28], to extract dense semantic representations from input images and predict pixel-level labels. When trained properly, deep CNNs can capture rich scene information with multilayer convolutional operations and nonlinear pooling/activation functions. Due to the convolutional nature of CNN, however, local convolutional features usually have limited receptive fields. Moreover, even with a large receptive field, these features mainly describe the core region and largely ignore the context around boundary [23]. On the other hand, local regions from different category may share near features, e.g. wood table and chair may exhibit similar local textures. The precise semantic segmentation always requires context information from different scales and large regions to release the ambiguity caused by local regions.

To address this problem, a number of recent works [4, 40, 20, 10, 13] aggregate context vector to local convolutional feature to boost the segmentation performance. These methods differ with each other in the way to construct context vector, and perform differently on different datasets. So there is a natural question, *what are the optimal contexts for semantic segmentation.* This paper tries to address this question by investigating the desirable properties that the optimal context vector should exhibit. In principle, the optimal context vector should describe segmentation-relevant image contents which are complementary to the local features, meanwhile, this vector should be compact with irrelevant information as less as possible. Specifically, we summarize three key properties as follows.

Property 1-Multi-scale. For semantic segmentation, holistic objects/stuff regions yield important cues to determine the sematic labels of local pixels. Since objects usually have different sizes and positions, it is necessary to construct multi-scale representations to capture image contents from different scales. As shown in the first row of Figure 1, method without multi-scale contexts can only capture objects in single scale and lose details in other scales.

Property 2-Adaptive. Not all areas in input image con-

*Yu Qiao is the corresponding author. The author emails are hejunjun@sjtu.edu.cn, {zy.deng1, lei.zhou, yl.wang, yu.qiao}@siat.ac.cn.

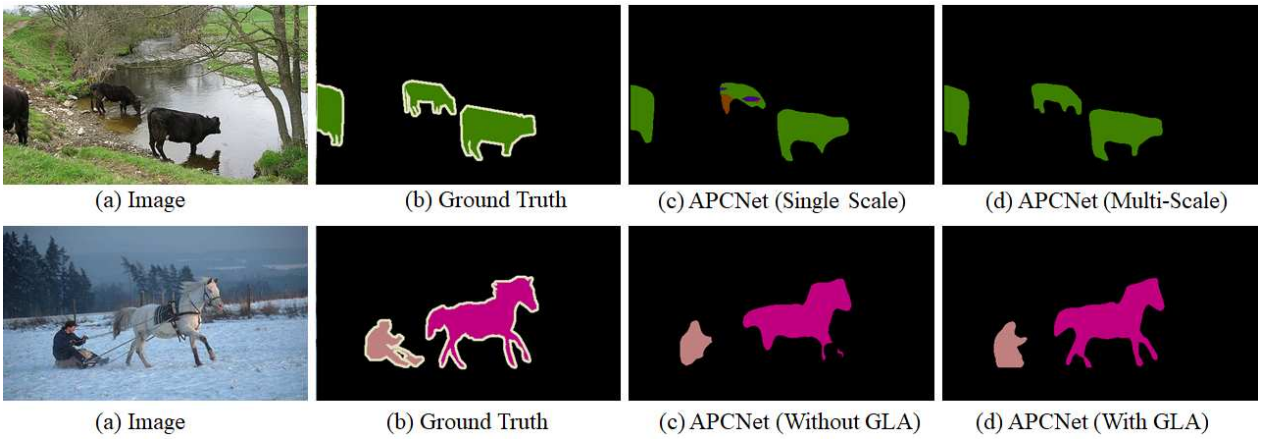


Figure 1. Illustration of Multi-scale and Global-guided Local Affinity properties. The first row: Multi-scale context can capture objects in different scales. The second row: Global-guided Local Affinity is benefit to segment complete and consist object.

tribute equally to determine the sematic label of a given pixel. Areas which contain related objects can yield useful information, while others may contribute very few. In practice, the relevant regions/pixels may exist near around the given pixel and also can be far away from it, highly depend on the contents and layout of input images. Therefore, it is important to adaptively identify these important regions for constructing optimal context vectors.

Property 3-Global-guided Local Affinity (GLA). To construct effective context vector, one need to aggregate the features from related pixels or regions. In practice, this can be implemented by summarizing their features in a weighted way. So there is a problem of estimating the affinity weights for aggregation. These weights indicate how different areas contribute to predict the semantic label of a local pixel. Previous works [20, 10, 13] mainly estimate these adaptive weights with local representations of pixels and regions, ignoring the global context. Unlike these works, here our insight is that both local and global representations are necessary to estimate robust affinity weights. As shown in the second row of Figure 1, the legs of horse are small and exhibit similar texture with snow which belongs to background class and dominates the whole scene. It is prone to classify the legs to background class. Clearly segmentation task can benefit from global representation. We call this property as Global-guided Local Affinity (GLA), as the local affinity weights are guided with global representation.

In the next, we make comparison of current context based semantic segmentation methods from the perspective of the properties mentioned above. DeepLab [4], ParseNet [20], and PSPNet [40] utilize ASPP (atrous spatial pyramid pooling), GAP (global average pooling), and PPM (pyramid pooling module) to obtain context at different scales, respectively. All these context vectors, however, only describe contents at fixed locations and are NOT adaptive. More recently, DANet [10] encodes global context with well designed self-attention mechanism. PSANet [13] learns an adaptive pixel-wise position sensitive spatial attention mask for aggregating contextual features. OCNet [37] embeds self-attention mechanism to PPM and ASPP to exploit multi-scale property. But these methods ignore the Global-guided Local Affinity property discussed in the above.

As summarized in Table 1, the previous methods can only account for some of the three properties. Partly inspired by this fact, this paper proposes Adaptive Pyramid Context Network (APCNet) for sematic segmentation, which effectively constructs the contextual representations with all the three properties. Specifically, APCNet designs pyramid Adaptive Context Modules to capture multi-scale global representations. The main contributions are as follows.

Method	MS	Adaptive	GLA
DeepLab[4]	✓		
PSPNet [20]	✓		
ParseNet [20]			
PSANet [13]		✓	
DANet [10]		✓	
OCNet [37]	✓	✓	
Ours	✓	✓	✓

Table 1. Comparison of different deep context based semantic segmentation methods. MS: multi-scale, GLA: global-guided local affinity.

- We summarize three desirable properties of context vectors for semantic segmentation, and compare recent deep context based semantic segmentation methods from the perspective of these properties.

- We propose Adaptive Context Modules which exploit GLA property by leveraging local and global representation to estimate affinity weights for local regions. These affinities further allow us to construct adaptive and multi-scale contextual representations for segmentation task.

- Our method achieves state-of-the-art performance on three widely used benchmarks, including PASCAL VOC 2012, Pascal-Context and ADE20K dataset, and obtains a new record 84.2% on PASCAL VOC 2012 test set without MS COCO pre-trained and any post-processing.

2. Related Work

Recently, FCN [22] based approaches have achieved promising performance on scene parsing and semantic segmentation task, through encoding contextual information. But most approaches only consider some properties as mentioned in table 1.

Multi-scale context. Multi-scale context plays a key role in semantic segmentation, especially for objects/stuff with vast variation of scales. Image Pyramid is a common way to obtain multi-scale context. [9] uses Laplacian pyramid to scale the input image of a DCNN [14] and merge the feature maps. SegNet [2], UNet[28], and [5] design Encoder-Decoder architecture to fuse low-level and high-level feature map from encoder and decoder, respectively. PSPNet [40] and DeepLab [4] propose PPM (pyramid pooling module) and ASPP (atrous spatial Pyramid pooling) module to encoding multi-scale context, respectively. These two modules are effective and efficiency to some extent, but they treat all images regions equally, not in an adaptive way.

Global context. Global context is particularly import for comprehensive complex scene understanding. ParseNet [20] proposes a simple but effective method to encoding global context through GAP (global average pooling) for semantic segmentation. PSPNet [40] exploits pyramid region based context aggregation to construct global context utilizing PPM. These methods cannot encode global context adaptively for every specific pixel. DANet [10] and OCNet [37] adopt self-attention to capture long-range global context, which calculate pixel-wise similarity map based on the pairs semantic features. While PSANet [13] aggregates global context through learning a pixel-wise position sensitive spatial attention mask to guide information flow. Calculated pixel-wise similarity map and learned pixel-wise attention map are adaptive to every specific pixel, but these pairs pixel relations which obtained by calculating pixel-wise similarity or convolving on a specific pixel position are lack of global information. While our method learns relations guided by local and global information.

Different from all previous work, our proposed method can generate more powerful multi-scale and global contexts through aggregating multi-scale features with learned adaptive affinities guided by local and global information.

3. Method

Context information is essentially important for complex scene parsing and semantic segmentation. Global context is useful to capture long-range dependency and provide a comprehensive understanding of the whole scene, while segmentation of objects with different sizes can benefit from multi-scale contextual features. In the next, we describe the proposed Adaptive Pyramid Context Network which adaptively constructs multi-scale context vectors with the guidance of global image representation.

3.1. Formulation

To begin with, we describe the mathematical formulation of our problem as follows. Given an image I for segmentation, we calculate a dense 3D convolution feature cube $\mathbf{X}$ with a backbone CNN, where $\mathbf{X}_i$ denotes the convolutional feature vector at position i. And $\mathbf{x}_i$ denotes the reduced convolutional feature vector at position i for efficient computation. The segmentation task can be reduced to predict a semantic label of a pixel, take i for example. One direct idea toward this problem is to estimate the semantic label just with the local feature $\mathbf{X}_i$. However, this idea ignores the relevant contents in other areas and limit the segmentation performance. To address this problem, context features have been successfully exploited to boost segmentation performance in previous works [4, 40, 20, 10, 13]. Mathematically, we introduce $\mathbf{z}_i = F_{\text{context}}(\mathbf{X}, i)$ to denote the context feature vector for $\mathbf{X}_i$, where F_{context} represents the function to extract $\mathbf{z}_i$ from input feature cube at position i. Previous context segmentation methods differ with each other in how to define F_{context}.

As discussed in the Section 1, this paper aims to design a novel context which satisfied the three properties, 1) Multi-scale, 2) Adaptive, and 3) Global-guided Local Affinity. Toward this objective, we first transform $\mathbf{X}$ into multi-scale pyramid representations. Then we adaptively construct context vectors for each scale separately. Here we just take one scale s as an example and the other scales can be processed in a similar way. For this scale, we divide the feature map $\mathbf{X}$ of image I into $s \times s$ subregions, thus transform $\mathbf{X}$ into a set of subregion representations, $\mathbf{Y}^s = [\mathbf{Y}^s_1, \mathbf{Y}^s_2, ..., \mathbf{Y}^s_{s\times s}]$, according to this division. For each subregion $\mathbf{Y}^s_j$, we summarize its contents with a feature vector $\mathbf{y}^s_j$ by average pooling and one convolution operation. We introduce affinity coefficient $\alpha^s_{i,j}$ to denote the degree of how subregion $\mathbf{Y}^s_j$ contributes to estimate the sematic label of $\mathbf{X}_i$. Then, the

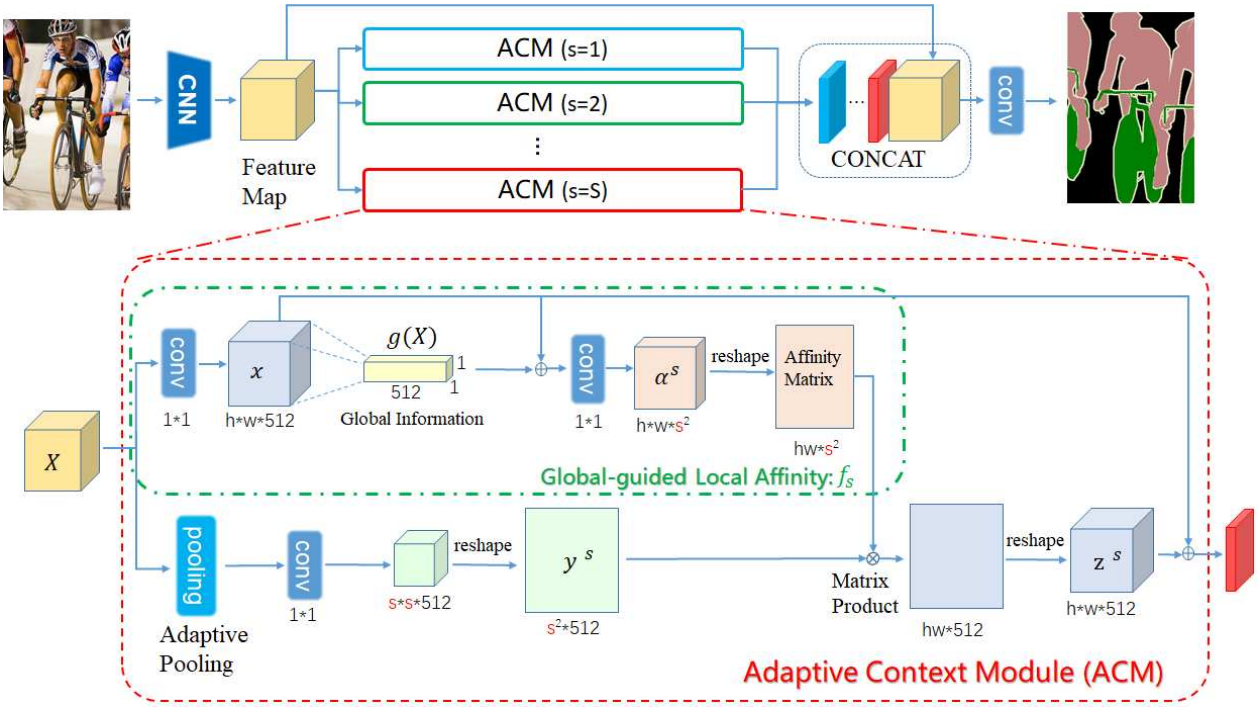


Figure 2. The pipeline of Adaptive Pyramid Context Network (APCNet). The input image is fed into a backbone CNN to obtain convolutional feature cube $\mathbf{X}$. $\mathbf{X}$ is decomposed into multi-scale pyramid representation. The representation of each scale is feed into Adaptive Context Module (ACM) to estimate adaptive context vectors for each local position. APCNet consists of multiple ACMs organized in parallel. Each ACM consists of two branches with one branch to estimate GLA affinity coefficients and the other branch to obtain subregion representations. The output of these two branches are multiplied to obtain adaptive context vectors. Finally, APCNet concatenates context vectors from different scales and the original feature cube $\mathbf{X}$ for predicting the semantic labels of the input pixels

adaptive context vector can be calculated as,

$$\mathbf{z}_i^s = \sum_{j=1}^{s\times s} \alpha_{ij}^s \mathbf{y}_j^s, \tag{1}$$

Here the key problem is how to calculate coefficient $\alpha_{i,j}^s$. Ideally, $\alpha_{i,j}^s$ should satisfy the GLA property by accounting for both local feature from $\mathbf{x}_i$ and global representation from $\mathbf{X}$ given scale s and position j. Let $g(\mathbf{X})$ denote the global information representation vector of $\mathbf{X}$ and g is a global information extractor. In this paper, we calculate $\alpha_{i,j}^s = f_s(\mathbf{x}_i, g(\mathbf{X}), j)$. Then Eq. 1 evolves to

$$\mathbf{z}_i^s = \sum_{j=1}^{s\times s} f_s(\mathbf{x}_i, g(\mathbf{X}), j)\mathbf{y}_j^s. \tag{2}$$

The above Eq. 2 plays a key role in our design of Adaptive Pyramid Context Network.

3.2. Adaptive Context Module

The Adaptive Context Module (ACM) is a key component in our Adaptive Pyramid Context Network. In principle, ACM aims to calculate a context vector for each local position by leveraging Global-guided Local Affinity. Essentially, ACM implements Eq. 2 with the network architecture shown in Fig. 2. ACM consists of two branches. The first branch aims to calculate affinity coefficients α^s while the second approach processes single-scale representation $\mathbf{y}^s$. Details are given in the below.

In the first branch, we first process $\mathbf{X}$ with a 1×1 convolution to get the reduced feature map $\mathbf{x}$, and then obtain global information representation vector $g(\mathbf{X})$ by applying spatial global average pooling and one 1×1 convolutional transform on $\mathbf{x}$. In the next, we integrate both local features $\{\mathbf{x}_i\}$ and global vector $g(X)$ to calculate an Global-guided Local Affinity vector for each local positions i. This is implemented with a 1×1 convolution followed by a sigmoid activation function in our design. One may argue to exploit large spatial convolution. But this leads to poor performance in experiments, partly due to the complexity of large filters. Each affinity vector has a dimension $s \times s$, corresponding the number of subregions in this scale. Totally, we have $h * w$ affinity vectors, which can be reshaped to an affinity map with size $hw \times s^s$. The second branch applies adaptive average pooling and a 1×1 convolution on $\mathbf{X}$ to obtain $\mathbf{y}^s \in \mathbb{R}^{s\times s\times 512}$. Then we reshape $\mathbf{y}^s$ into size of

$s^2 \times 512$ to match that of the affinity map. Then we multiply them together and reshape the results to obtain the adaptive context matrix $\mathbf{z}^s$ composed of $\{\mathbf{z}_i^s\}$. Residual learning is adopted to ease the training process, and thus we add $\mathbf{x}$ to $\mathbf{z}^s$.

3.3. Adaptive Pyramid Context Network

Next, we will describe the proposed Adaptive Pyramid Context Network (APCNet) for semantic segmentation, whose architecture is shown in Figure 2. APCNet takes a backbone CNN e.g. ResNet or InceptionNet to calculate a convolutional feature cube $\mathbf{X} \in \mathbb{R}^{h \times w \times c}$, where h, w, c represent width, height and channel number respectively. Then APCNet transforms $\mathbf{X}$ into pyramid representations with S scales in total. Specifically, for each scale s, we adopt adaptive average pooling and one 1×1 convolution to transform $\mathbf{X}$ to a specific spatial size $s \times s$ and obtain $\mathbf{y}^s \in \mathbb{R}^{s \times s \times c}$. Then each $\mathbf{y}^s$ together with original $\mathbf{X}$ is processed with an Adaptive Context Module (ACM) to obtain an adaptive context vector $\mathbf{z}_i^s$ for each spatial position. Totally, APCNet includes multiple ACMs organized in parallel. In the next, we can concatenate $\{\mathbf{z}_i^s\}$ obtained from different scales into the final adaptive context vector $\mathbf{z}_i = [\mathbf{z}_i^1, \mathbf{z}_i^2, ..., \mathbf{z}_i^S]$. Finally, we exploit both local features $\{\mathbf{X}_i\}$ and their associate context vectors $\{\mathbf{z}_i\}$ to predict the semantic label of each pixel.

3.4. Relation to Other Approaches

In this subsection, we make comparison between our Adaptive Pyramid Context Network and other context approaches for semantic segmentation. ParseNet [20] aggregates global context through global average pooling, which can be seen as an extreme case of our model if we just set $\alpha_{i,j}^s = 1$, S=1, and $\mathbf{y}_j^s = g(\mathbf{X})$. In PSPNet [40], α^s are set as a fixed bilinear interpolation coefficients for $\mathbf{y}^s$. In contrast our APCNet estimates α^s in an adaptive way with Eq. 2. Recent methods PSANet [13], DANet [10], OCNet [37] also alleviate this problem by introducing adaptive weights. These methods calculate pair-wise similarity or learn pixel-wise attention map. But they all neglect the importance of global guidance from $g(\mathbf{X})$. Unlike these works, our APCNet not only takes Global-guided Local Affinity into account with f_s to estimate α^s from both local and global representations, but also exploits multi-scale representation with feature pyramid.

4. Experiments

We conduct extensive experiments on three challenging semantic segmentation and scene parsing datasets to evaluate our proposed method, including PASCAL VOC 2012 [7], Pascal-Context [24], and ADE20K dataset [42].

4.1. Implementation Details

We adopt ResNet [12] as our backbone which is pre-trained on ImageNet [29]. Following [36, 4, 38], we remove stride and set dilation rate 2 and 4 to the last two stages of backbone networks respectively, and the output feature map is 1/8 size of the input image [4, 38, 35]. The output predictions are bilinear interpolated to target size for predicting semantic labels of each pixel. We use poly learning rate policy $lr = initial\_lr \times (1 - \frac{iter}{total\_iter})^{power}$ [4, 5, 38]. The initial learning rate is 0.01 for PASCAL VOC 2012 [7] and ADE20K dataset [42], 0.001 for Pascal-Context dataset [7], and the power is 0.9 [38]. Stochastic gradient descent (SGD) [3] with momentum 0.9 and weight decay 0.0001 is chosen as optimizer. We train networks for 80 epochs on PASCAL VOC 2012 [7] and Pascal-Context dataset [24], and 120 epochs on ADE20K dataset [42]. In practice, appropriately larger crop size can obtain better performance, so we set crop size to 512 on PASCAL VOC 2012 and Pascal-Context dataset, and 576 on ADE20K as the average image size of ADE20K dataset is larger than other two datasets [4, 40, 38]. We randomly flip and scale the input image from 0.5 to 2 as our data augmentation. Our evaluation metric is mean of class-wise intersection over union (mIoU). For multi-scale and flip evaluation, we resize the input image to multiple scales and horizontally flip them. The predictions are averaged as final predictions [20, 40, 30, 34]. All experiments are implemented based on PyTorch [26].

4.2. PASCAL VOC 2012

PASCAL VOC 2012 [7] is a benchmark dataset of semantic segmentation, which originally contains 1,464 images for training, 1,449 for validation, and 1,456 for test. Totally, there are 20 foreground object classes and one background class in the original PASCAL VOC 2012 dataset [7]. The original dataset is augmented to 10,582 images for training by [11]. Following [4, 38, 5], we use this augmented training set in our experiments.

We conduct experiments with different settings to evaluate the effectiveness of our proposed modules. Our baseline is dilated ResNet based FCN [4, 22] as mentioned above.

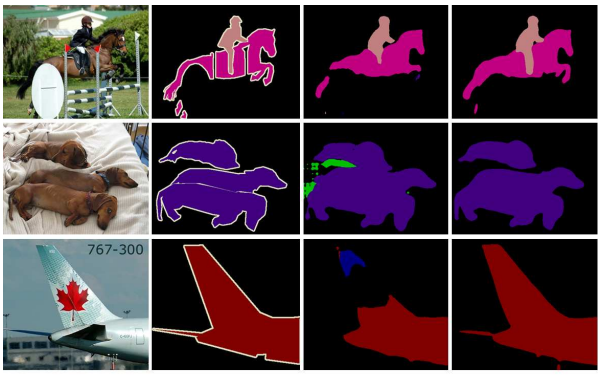
Pyramid scales. ResNet50-based FCN [22] with dilated network is adopted as our baseline. We investigate the performance of APCNet with different setting of pyramid scales (PS). Results are listed in Table 2. From Table 2, we have the following observations. Firstly, compared with baseline FCN (1st row), all pyramid scales settings improve the performance significantly. Secondly, pyramid scales of {1,2,3,6} achieve the best result, which improves the performance of baseline FCN by 8.37% (from 69.83% to 78.20%). We can infer that properly designed pyramid scales can help to effectively capture the features of objects with varied scale. In all the following experiments, we will

adopt pyramid scales of {1,2,3,6}. Finally, deeper backbone network, e.g. ResNet101, can further improve the result.

Backbone	PS	mIoU%
ResNet50	None	69.83
ResNet50	{1}	77.89
ResNet50	{1,2}	77.48
ResNet50	{1,2,3}	77.60
ResNet50	{1,2,3,6}	**78.20**
ResNet50	{1,2,3,6,32}	77.29
ResNet101	{1,2,3,6}	**80.71**

Table 2. Investigation of different pyramid scales and backbones. Baseline is ResNet50-based FCN with dilated network (PS in none). PS: pyramid scales, $\{1, 2, 3, 6, 32\}$: bin sizes of pooled feature, 1×1, 2×2, 3×3, 6×6, 32×32. The results are evaluated on the validation set of PASCAL VOC 2012 dataset, with the single-scale input.

Figure 3 shows the visualization results of our APCNet and baseline model FCN. It is obvious that APCNet keeps more details (1st row) due to its pyramid scale. And it also introduces less mislabeled pixel (2nd and 3rd row), which leads to better performance than FCN.



(a) Image (b) Ground Truth (c) FCN (d) APCNet

Figure 3. Comparison with baseline method.

To further illustrate the effectiveness of pyramid scales, we visualize the improvement of different scales in Figure 4. It can be observed from the figure that APCNet with single scale is inferior to multi-scale APCNet since single scale APCNet can hardly segment the objects with large scale variations. More specifically, in the first row of Figure 4, APCNet with single scale lacks detailed information of the boat and fails to segment the person on the boat. With multi-scale setup, APCNet not only preserves most detailed information of the boat but also correctly segments the person.

Global-guided Local Affinity (GLA). We conduct experiments w/o GLA with different backbones, to validate the essential importance of GLA in our APCNet. Table 3 lists the performance of different backbones w/o GLA on the validation set of PASCAL VOC 2012 dataset. It is obvious that GLA consistently increases the performance of different backbones.

Backbone	GLA	mIoU%
ResNet50		77.68
ResNet50	✓	**78.20**
ResNet101		80.17
ResNet101	✓	**80.71**

Table 3. Investigation on the importance of GLA with different backbone networks, and PS is {1,2,3,6}. GLA: Global-guided Local Affinity. The results are evaluated on the validation set of PASCAL VOC 2012 dataset, with the single-scale input.

Also, we visualize the segmentation results to show the improvement of GLA in Figure 5. The 1st row shows that APCNet with GLA can lead to more accurate segmentation (for the dog near the person). The 2nd and 3rd show that APCNet with GLA can alleviate the problem of segmenting an object into different classes. This verifies that global information introduced by GLA can help better understanding of complex context and more consistent segmentation of a certain object.

Training and evaluation strategies. The results of different training and evaluation strategy are shown in Table 4. We can observe that 1) deep supervision can optimize the learning process and further improve the performance, 2) scaling the input image to multiple scales and flipping the images left-right for evaluation are useful, 3) fine-tuning the trained model with original training set boosts the result to 82.67% mIoU on PASCAL VOC 2012 validation set, without MS COCO pre-trained.

Backbone	DS	Flip	MS	FT	mIoU%
ResNet101					80.71
ResNet101	✓				80.93
ResNet101	✓	✓			81.33
ResNet101	✓	✓	✓		81.93
ResNet101	✓	✓	✓	✓	**82.67**

Table 4. Influence of different setting in training and evaluation strategies, and PS is {1,2,3,6}. DS: deep supervised [40], Flip: horizontally flip input image for evaluation, MS: multi-scale evaluation, FT: fine tune the trained model on PASCAL VOC 2012 original training set. The results are evaluated on the validation set of PASCAL VOC 2012 dataset.

Adaptive. Our proposed model can be reduced as PSPNet if removing adaptive and GLA modules. So we reimplemented PSPNet with our experimental settings (add deep supervised) as our baseline with backbone ResNet101 which gets 79.79% mIoU on PASCAL VOC validation set (single scale). With adaptive and GLA modules, the performance is improved clearly as shown in Table 5.

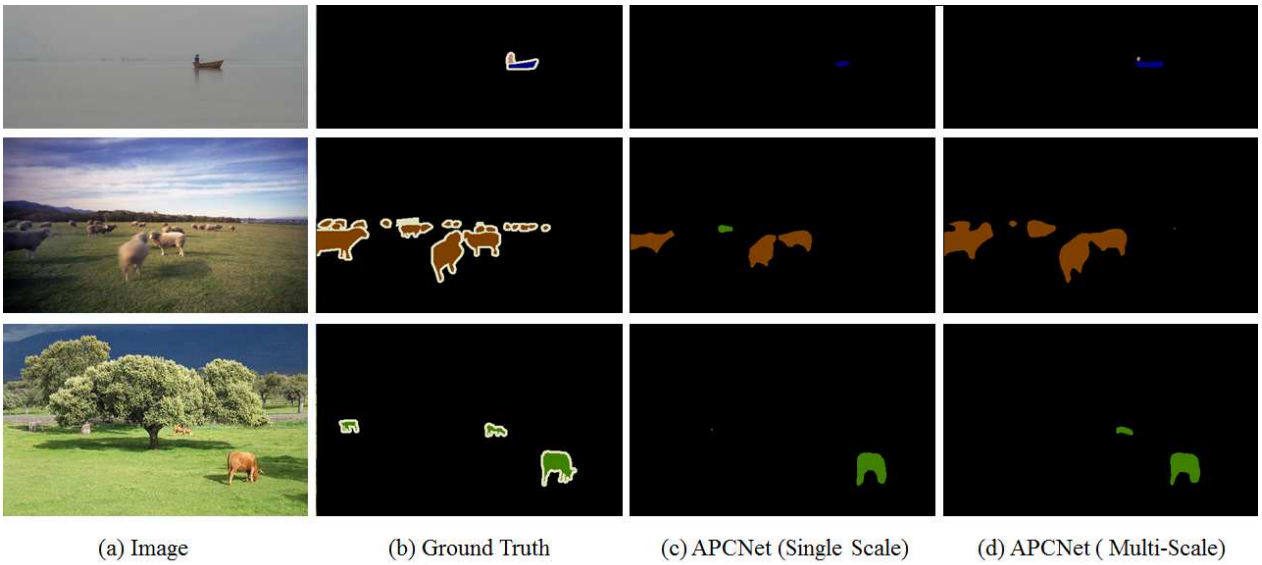


Figure 4. Visualization of segmentation results of single scale and multi-scale.

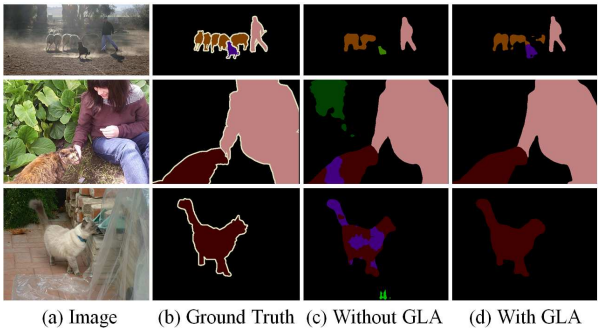


Figure 5. Visualization of segmentation results with/without Global-guided Local Affinity (GLA).

Adaptive	GLA	mIoU (%)
✓		80.19
✓	✓	80.93

Table 5. The improved performance based on PSPNet with adaptive and GLA module. PSPNet gets 79.79% mIoU. The results are evaluated on the validation set of PASCAL VOC 2012 dataset.

For evaluation on PASCAL VOC 2012 [19] test set, we set pyramid scales to {1,2,3,6} and adopt deep supervised strategy [40] to train the backbone model on augmented training set. The backbone model is ResNet101 pre-trained on ImageNet [29]. Then, we fine tune the trained model on original training and validation set. After training, multi-scale and flip are adopted for testing. Final results are submitted to official server for evaluation and the comparison to the state-of-the-art methods is shown in Table 6. Clearly, our APCNet significantly outperforms other methods on almost all categories of PASCAL VOC 2012. Note that APCNet can distinguish categories that look very similar, e.g. cow (93.7%) and horse (95%). This may owe to the GLA properties of our methods which take both global and local information into consideration. Without pre-trained on MS COCO dataset [16], APCNet achieves state-of-the-art performance of 84.2% mIoU, which demonstrates the effectiveness of our proposed method. With MS COCO pre-trained, our proposed method also achieves the best performance of 87.13% mIoU among the methods based on backbone ResNet101.

4.3. Pascal-Context

Pascal-Context dataset [24] is annotated additionally for PASCAL VOC 2010 [8] with whole scene label. Following [38, 17], we train our model on the training set of 4,998 images and evaluate on the test set of 5,105 images, and report our result on 60 classes including 59 foreground classes and one background class. Table 7 compares the performance of the state-of-the-art methods. With the same backbone model, our APCNet surpasses DeepLab-v2 [4], EncNet [38], and DANet [10] in a large margin. Moreover, our APCNet achieves the state-of-the-art performance on Pascal-Context dataset and thus demonstrates its effectiveness for semantic segmentation.

4.4. ADE20K

ADE20K dataset [42] is a challenge scene parsing dataset providing 150 classes dense labels, which consists of 20K/2K/3K images for training, validation and test, respectively. Due to the diverse and complex scene in this dataset, it is hard to achieve subtle improvements. Results

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mIoU%
FCN [22]	76.8	34.2	68.9	49.4	60.3	75.3	74.7	77.6	21.4	62.5	46.8	71.8	63.9	76.5	73.9	45.2	72.4	37.4	70.9	55.1	62.2
DeepLabv2 [4]	84.4	54.5	81.5	63.6	65.9	85.1	79.1	83.4	30.7	74.1	59.8	79	76.1	83.2	80.8	59.7	82.2	50.4	73.1	63.7	71.6
CRF-RNN [41]	87.5	39.0	79.7	64.2	68.3	87.6	80.8	84.4	30.4	78.2	60.4	80.5	77.8	83.1	80.6	59.5	82.8	47.8	78.3	67.1	72.0
DeconvNet [25]	89.9	39.3	79.7	63.9	68.2	87.4	81.2	86.1	28.5	77.0	62.0	79.0	80.3	83.6	80.2	58.8	83.4	54.3	80.7	65	72.5
DPN [21]	87.7	59.4	78.4	64.9	70.3	89.3	83.5	86.1	31.7	79.9	62.6	81.9	80.0	83.5	82.3	60.5	83.2	53.4	77.9	65	74.1
Piecewise [18]	90.6	37.6	80.0	67.8	74.4	92	85.2	86.2	39.1	81.2	58.9	83.8	83.9	84.3	84.8	62.1	83.2	58.2	80.8	72.3	75.3
ResNet38 [32]	94.4	72.9	94.9	68.8	78.4	90.6	90.0	92.1	40.1	90.4	71.7	89.9	93.7	91.0	89.1	71.3	90.7	61.3	87.7	78.1	82.5
PSPNet [40]	91.8	71.9	94.7	71.2	75.8	95.2	89.9	95.9	39.3	90.7	71.7	90.5	94.5	88.8	**89.6**	72.8	89.6	**64.0**	85.1	76.3	82.6
EncNet [38]	94.1	69.2	**96.3**	**76.7**	**86.2**	96.3	**90.7**	94.2	38.8	90.7	73.3	90.0	92.5	88.8	87.9	68.7	92.6	59.0	86.4	73.4	82.9
Ours	**95.8**	**75.8**	84.5	76.0	80.6	**96.9**	90.0	**96.0**	**42.0**	**93.7**	**75.4**	**91.6**	**95.0**	**90.5**	89.3	**75.8**	**92.8**	61.9	**88.9**	**79.6**	**84.2**

Table 6. Per-class results on PASCAL VOC 2012 test set. Our method outperforms all previous state-of-art methods and achieves 84.2% without MS COCO dataset pre-trained.

Method	Backbone	mIoU%
FCN-8S [22]		37.8
CRF-RNN [41]		39.3
ParseNet [20]		40.4
BoxSup [6]		40.5
HO_CRF [1]		41.3
Piecewise [18]		43.3
VeryDeep [31]		44.5
DeepLab-v2 [4]	ResNet101-COCO	45.7
RefineNet [17]	ResNet152	47.3
MSCI [16]	ResNet152	50.3
EncNet [38]	ResNet101	51.7
DANet [10]	ResNet101	52.6
Ours	ResNet101	**54.7**

Table 7. Segmentation results on PASCAL-Context dataset of 60 classes with background. Our method outperforms all previous state-of-art methods with a large margin.

Method	Backbone	mIoU%
FCN [22]		29.39
SegNet [2]		21.64
DilatedNet [35]		32.31
CascadeNet [42]		34.90
RefineNet [17]	ResNet152	40.7
PSPNet [40]	ResNet101	43.29
PSPNet [40]	ResNet269	44.94
EncNet [38]	ResNet101	44.65
SAC [39]	ResNet101	44.30
PSANet [13]	ResNet101	43.77
UperNet [33]	ResNet101	42.66
DSSPN [15]	ResNet101	43.68
OCNet [37]	ResNet101	45.08
Ours	ResNet101	**45.38**

Table 8. Segmentation results on ADE20K validation set. Our method outperforms all previous methods.

on ADE20K validation set of different methods are summarized in Table 8. Our result outperforms other state-of-the-art results, even with a shallower backbone networks. We also submit the test set segmentation results of our method to official evaluation server. The pixel accuracy is 72.94%, mIoU is 38.39%, and score is 55.67%, which ranks top on the leaderboard.

4.5. Summary

Comparing to ParseNet [20] and PSPNet [40], our method achieves better results on PASCAL VOC 2012, Pascal-Context and ADE20K dataset. These results demonstrate the APCNet to adaptively aggregate multi-scale context with the guidance of global representation. Different from PSANet [13], OCNet [37] and DANet [10] which construct semantic context by calculating semantic correlation on every pair pixels or convolving on a specific pixel, our global-guided local affinity is more reasonable and leads to higher performance.

5. Conclusion

In this paper, we discuss the properties of context features, and propose APCNet to adaptively construct multi-scale context representation for semantic segmentation and scene parsing. APCNet introduces Adaptive Context Modules which generate local affinity coefficients with our elaborately designed Global-guided Local Affinity. Extensive experiments show that APCNet can capture different scales objects, and the predictions of objects are more completely and consistently. APCNet can not only be embedded to any FCN based semantic segmentation networks, but also any layer of the networks which independent of the input feature map size. APCNet may extend to other scene understanding tasks, according to the properties and flexibility.

Acknowledgements. This work is partially supported by National Natural Science Foundation of China (61876176, U1613211, U1713208), Shenzhen Research Program (JCYJ20150925163005055, CXB201104220032A), the Joint Lab of CAS-HK.

References

[1] Anurag Arnab, Sadeep Jayasumana, Shuai Zheng, and Philip HS Torr. Higher order conditional random fields in deep neural networks. In *European Conference on Computer Vision*, pages 524–540. Springer, 2016.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture

for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015.

[3] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

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

[6] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1635–1643, 2015.

[7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[9] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1915–1929, 2013.

[10] Jun Fu, Jing Liu, Haijie Tian, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. *arXiv preprint arXiv:1809.02983*, 2018.

[11] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 447–456, 2015.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. 2018.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[15] Xiaodan Liang, Hongfei Zhou, and Eric Xing. Dynamic-structured semantic propagation network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 752–761, 2018.

[16] Di Lin, Yuanfeng Ji, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Multi-scale context intertwining for semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 603–619, 2018.

[17] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian D Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Cvpr*, volume 1, page 5, 2017.

[18] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3194–3203, 2016.

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[20] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015.

[21] Ziwei Liu, Xiaoxiao Li, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Semantic image segmentation via deep parsing network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1377–1385, 2015.

[22] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[23] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in neural information processing systems*, pages 4898–4906, 2016.

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

[25] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015.

[26] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[27] Mengye Ren and Richard S Zemel. End-to-end instance segmentation with recurrent attention. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA*, pages 21–26, 2017.

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[29] 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. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[30] Gabriel Schwartz and Ko Nishino. Material recognition from local appearance in global context. *arXiv preprint arXiv:1611.09394*, 2016.

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

[32] Zifeng Wu, Chunhua Shen, and Anton van den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *arXiv preprint arXiv:1611.10080*, 2016.

[33] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. *arXiv preprint arXiv:1807.10221*, 2018.

[34] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. *arXiv preprint arXiv:1804.09337*, 2018.

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

[36] Fisher Yu, Vladlen Koltun, and Thomas A Funkhouser. Dilated residual networks. In *CVPR*, volume 2, page 3, 2017.

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

[38] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[39] Rui Zhang, Sheng Tang, Yongdong Zhang, Jintao Li, and Shuicheng Yan. Scale-adaptive convolutions for scene parsing. In *Proc. 26th Int. Conf. Comput. Vis.*, pages 2031–2039, 2017.

[40] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017.

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

[42] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 4. IEEE, 2017.