ORDNet: Capturing Omni-Range Dependencies for Scene Parsing

Shaofei Huang, Si Liu, *Member, IEEE*, Tianrui Hui, Jizhong Han, *Member, IEEE*, Bo Li, Jiashi Feng, *Member, IEEE*, and Shuicheng Yan, *Fellow, IEEE*

***Abstract*— Learning to capture dependencies between spatial positions is essential to many visual tasks, especially the dense labeling problems like scene parsing. Existing methods can effectively capture long-range dependencies with self-attention mechanism while short ones by local convolution. However, there is still much gap between long-range and short-range dependencies, which largely reduces the models' flexibility in application to diverse spatial scales and relationships in complicated natural scene images. To fill such a gap, we develop a Middle-Range (MR) branch to capture middle-range dependencies by restricting self-attention into local patches. Also, we observe that the spatial regions which have large correlations with others can be emphasized to exploit long-range dependencies more accurately, and thus propose a Reweighed Long-Range (RLR) branch. Based on the proposed MR and RLR branches, we build an Omni-Range Dependencies Network (ORDNet) which can effectively capture short-, middle- and long-range dependencies. Our ORDNet is able to extract more comprehensive context information and well adapt to complex spatial variance in scene images. Extensive experiments show that our proposed ORDNet outperforms previous state-of-the-art methods on three scene parsing benchmarks including PASCAL Context, COCO Stuff and ADE20K, demonstrating the superiority of capturing omni-range dependencies in deep models for scene parsing task.**

***Index Terms*— Scene parsing, omni-range dependencies, self-attention.**

I. Introduction

SCENE parsing [1]–[4] aims to divide the entire scene into different segments and predict the semantic category for each of them. It is a fundamental task in computer vision and image processing, challenged by the complexity of natural scenes that usually contain multiple elements of various categories, including discrete objects (e.g., person, cat) and stuff (e.g., sky, river, grass). The elements within a scene may be spatially dependent upon each other. For example, a ship usually appears on the sea rather than on the road. Such dependencies of spatial positions can be exploited to boost the prediction accuracy further.

Mainstream scene parsing models built on Fully Convolutional Networks [5] incorporate carefully designed modules to exploit spatial context information. For example, Deeplabv2 [6] uses an Atrous Spatial Pyramid Pooling (ASPP) module to sample feature maps in parallel with different atrous rates to enlarge the receptive field of filters; PSPNet [1] performs pooling operations at multiple grid scales for the same goal. With a larger receptive field, these networks are able to extract broader scales of spatial context information. However, these methods model the dependencies between positions in an implicit way.

Self-attention mechanism [7] is able to capture long-range dependencies between positions explicitly and has been applied to scene parsing [8], [9] with a remarkable performance boost. The key idea behind self-attention is that the response at a certain position is a weighted sum of features at all the positions. In this way, all the positions are related to each other and provide the network with a global receptive field. Long-range dependencies captured by self-attention can be combined with short-range ones captured by local convolution, leading to rich context information for dense labeling problems.

However, long-range dependencies do not always work well for scene parsing tasks since a position is often less correlated with the positions far away from it, compared with those which are nearer. Moreover, information from distant positions may not be beneficial to building discriminative features. In Fig. 1(b), we visualize the attention map between a certain position and all the positions in the image computed in the self-attention process, where brighter color represents higher attention weight. The specified position is denoted as a green cross in Fig. 1(a). We find that in the conventional self-attention mechanism, the feature of this position would aggregate information from a wide area of the input image. For example, in the second row of (b), the position on the cat receives information from distant ones of the image, like those on the window and curtain. However, there is no apparent correlation between the window, curtain, and cat. Thus the long-range dependencies captured from these positions are not useful for the model to classify a certain position on the

Manuscript received October 13, 2019; revised May 5, 2020 and June 19, 2020; accepted July 17, 2020. Date of publication August 5, 2020; date of current version August 13, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2020AAA0130200; in part by the National Natural Science Foundation of China under Grant 61876177; and in part by the Beijing Natural Science Foundation under Grant 4202034. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Christophoros Nikou. *(Corresponding author: Si Liu.)*

Shaofei Huang, Tianrui Hui, and Jizhong Han are with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China (e-mail: huangshaofei@iie.ac.cn; huitianrui@iie.ac.cn; hanjizhong@iie.ac.cn).

Si Liu and Bo Li are with the Institute of Artificial Intelligence, Beihang University, Beijing 100191, China (e-mail: liusi@buaa.edu.cn; boli@buaa.edu.cn).

Jiashi Feng is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119077 (e-mail: elefjia@nus.edu.sg).

Shuicheng Yan is with YITU Technology, Guangzhou 201103, China (e-mail: shuicheng.yan@yitu-inc.com).

Digital Object Identifier 10.1109/TIP.2020.3013142

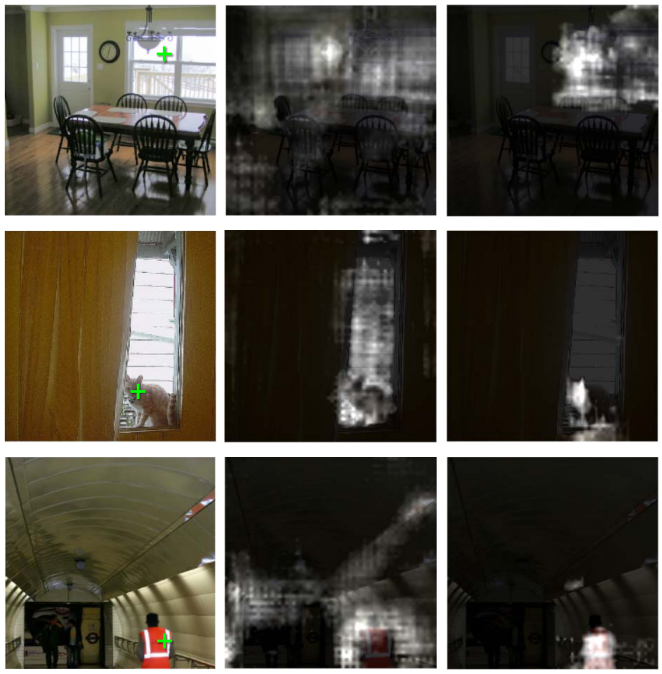


Fig. 1. Visualization of attention maps. (a) Original images. (b) Attention maps for a certain position (green cross in (a)) computed with conventional self-attention. (c) Attention maps for a certain position (green cross in (a)) when self-attention is restricted in local patches. The attention maps of self-attention in local patches focus on surrounding areas that closely correlate with the specified position in (a).

cat. When we restrict self-attention to a local patch of the image, the visualized attention map is more concentrated on the surrounding area of the specified position. For example, in the second row of Fig. 1 (c), self-attention is restricted to the bottom-right $1/4$ patch of the image. The attention map between the specified position and all the positions within the patch mainly focuses on the head and body parts of the cat, indicating that useful middle-range dependencies among positions are captured. Benefiting from the close correlations within the same category, context information aggregated from these dependencies can severe as more valid guidance for the model to classify the specified position. Based on this observation, we devise a novel Middle-Range (MR) branch which restricts the self-attention mechanism to local patches of the input feature in order to fill the gap between long-range and short-range dependencies for complete context extraction.

Furthermore, we analyze the attention map generated with conventional self-attention and find that each position contributes different attention weights to others for context aggregation. For each position, the total value of attention weights that it contributes to others reveals its correlation with other positions as well as its importance to the global context. A larger value implies that the position has stronger correlations with most of the other positions. Thus, the features of positions contributing higher attention weights to others encode the common patterns of the whole image, including main elements appearing in the scene, large-area continuous background, etc. These patterns contain useful global context information which is crucial to scene understanding. By emphasizing features of the positions with larger contributions, long-range dependencies can be captured more accurately and adaptively to complicated scene elements, which can enhance the aggregation of global context by self-attention. We instantiate this idea with a Reweighed Long-Range (RLR) branch to modulate feature responses according to the attention weights contributions of each position.

With the newly proposed MR and RLR branches, we build an Omni-Range Dependencies Network (ORDNet) in which short-range, middle-range, and reweighed long-range dependencies collaborate seamlessly to achieve adaptability to diverse spatial region contents and relationships in natural scene images. The ORDNet is general and can be applied to any FCN backbone for learning more discriminative feature representations.

Our main contributions are summarized as follows:

- We devise a Middle-Range (MR) branch which explicitly captures middle-range dependencies within local patches of scene image, filling the gap between long-range and short-range dependencies.
- We also propose a Reweighed Long-Range (RLR) branch to emphasize features of the positions which encode common patterns, so that more accurate and adaptive long-range dependencies could be captured.
- With the above two branches, we develop a novel Omni-Range Dependencies Network (ORDNet) which effectively integrates short-range, middle-range and reweighed long-range dependencies to extract comprehensive context information for accurate scene parsing. Our ORDNet outperforms previous state-of-the-art methods on three popular scene parsing benchmarks, including PASCAL-Context [10], COCO Stuff [11] and ADE20K [2] datasets, which well demonstrates its effectiveness.

II. Related Work

A. Semantic Segmentation

The goal of semantic segmentation is to assign category labels to the pixels of foreground objects and stuff in the scene, rather than segmenting the entire scene as scene parsing does. By expanding the set of pre-defined categories on which the model to segment, semantic segmentation serves as a basic technology of scene parsing. In recent years, remarkable progress has been achieved based on Fully Convolutional Networks [5] (FCNs). FCN replaces the fully-connected layers of the image classification network (e.g., VGG16 [12]) with convolution layers and introduces transposed convolution and skip layers to predict pixel-level labels. The many pooling layers in FCN increase the receptive field of convolution filters, and meanwhile reduce the resolution of feature maps, leading to inaccurate semantic masks. In order to maintain the resolution of feature maps while enjoying the increased receptive field of convolution filters, DeepLab [13] integrates atrous convolution into CNN, which boosts segmentation performance largely and becomes the de-facto component of latter segmentation methods.

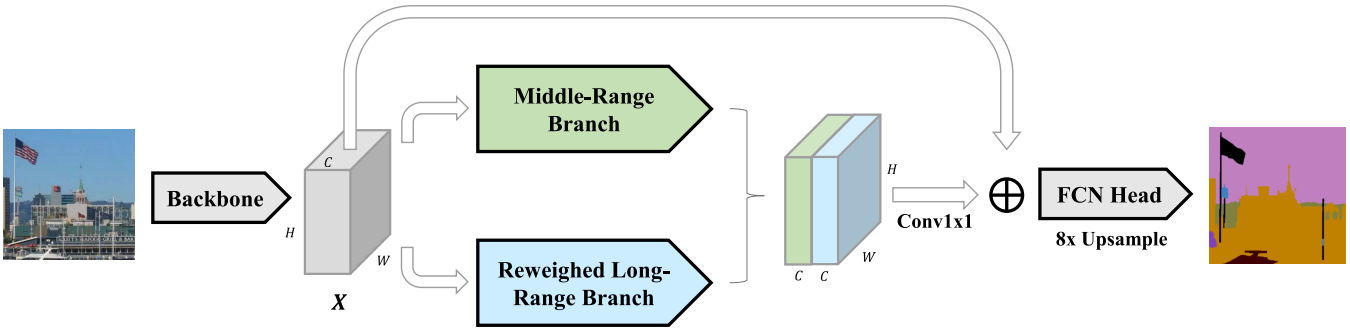


Fig. 2. The pipeline of proposed Omni-Range Dependencies Network (ORDNet). Given an input image, the extracted feature X of a CNN backbone is fed into a Middle-Range (MR) branch and a Reweighed Long-Range (RLR) branch to capture the middle-range and reweighed long-range dependencies respectively. The outputs of these two branches are then concatenated along the channel dimension and fused by a 1×1 convolution. An identity skip connection of X is added to ease optimization. The fused feature is fed into an FCN Head to predict the logit map and then upsampled 8 times to obtain the final parsing mask.

Many deep models [14]–[16] have proposed various approaches to aggregate local spatial context information to refine the feature representations and achieved great performances. Later works further propose to aggregate the important multi-scale spatial context information based on the last feature map of an FCN backbone. For example, DeepLab v2 [6] employs parallel atrous convolutions with different atrous rates called ASPP to capture context information of multiple receptive fields. DeepLab v3 [17] further integrates image-level features into ASPP to get a global receptive field. PSPNet [1] performs pooling operations at multiple grid scales in order to aggregate multi-scale contextual information. In addition to extracting context information from feature maps, EncNet [18] uses a Context Encoding Module which exploits semantic category prior information of the scenes to provide global contexts. Recent DANet [8] and CFNet [9] exploit the self-attention mechanism to effectively capture long-range dependencies, which outperform previous multi-scale context aggregation methods in semantic segmentation and scene parsing. InterlacedSSA [19] proposes a factorized self-attention approach to approximately capture long-range dependencies with low computational costs, which achieves comparable performance with DANet and CFNet. Different from the above methods, in this article we further propose to capture middle-range dependencies and reweighed long-range dependencies to provide richer semantic information than vanilla self-attention. Our Omni-Range Dependencies Network (ORDNet) can fill the semantic gap between original long-range and short-range dependencies, and also capture more accurate long-range dependencies by feature reweighing, achieving more comprehensive scene understanding.

B. Attention Mechanism

Attention mechanism is first introduced in [20] for neural machine translation, and later widely applied to various tasks like machine translation [21], VQA [22]–[24] and image captioning [25]. [26] is the first work to apply self-attention for capturing long-range dependencies within input sentences, achieving noticeably boosted performance in machine translation. In [7], self-attention mechanism is further extended to vision tasks and a non-local network is proposed to capture long-range dependencies. For image tasks, self-attention methods compute the response at a position as a weighted sum of the features at all positions in the input feature maps, in which way the receptive field for the current position can go beyond local convolution kernels to cover the whole feature map. Self-attention mechanism is widely adopted in vision tasks such as semantic segmentation [8] [27], GANs [28] and image de-raining [29]. Inspired by [7], we propose a new self-attention architecture to capture omni-range dependencies of positions, where the Middle-Range (MR) branch restricts self-attention to patches to model middle-range dependencies and the Reweighed Long-Range (RLR) branch further emphasizes features of positions which encode common patterns of the image to obtain more accurate long-range dependencies. Compared with previous works, our method conforms better to practical spatial relations between semantic regions and achieves higher performance on several benchmarks.

In addition to self-attention, researchers also explore other attention methods to refine feature maps by adjusting their scales. SENet [30] utilizes a squeeze-and-excitation process to recalibrate feature channels with signals pooled from the entire feature maps. The first squeeze operator conducts global average pooling to generates a channel descriptor as the global information embedding. The second excitation operator maps the channel descriptor to a set of channel-specific weights with two successive fully connection layers. Finally, the channel-specific weights are multiplied with original features to rescale the channel responses. CBAM [31] and BAM [32] apply SE operation to both channel and spatial dimensions. Our proposed reweighed long-range branch serves as a spatial recalibration module to some extent. Compared with the spatial branch in CBAM, which is based on the single response of the current position, our RLR branch reweighs the feature responses according to correlations among all positions of the entire feature map so that positions encoding common patterns and main elements of the scene can be emphasized to form a more discriminative feature representation.

III. METHOD

A. Revisiting Self-Attention

Self-attention mechanism computes the response at a position as a weighted sum of the features at all positions in the

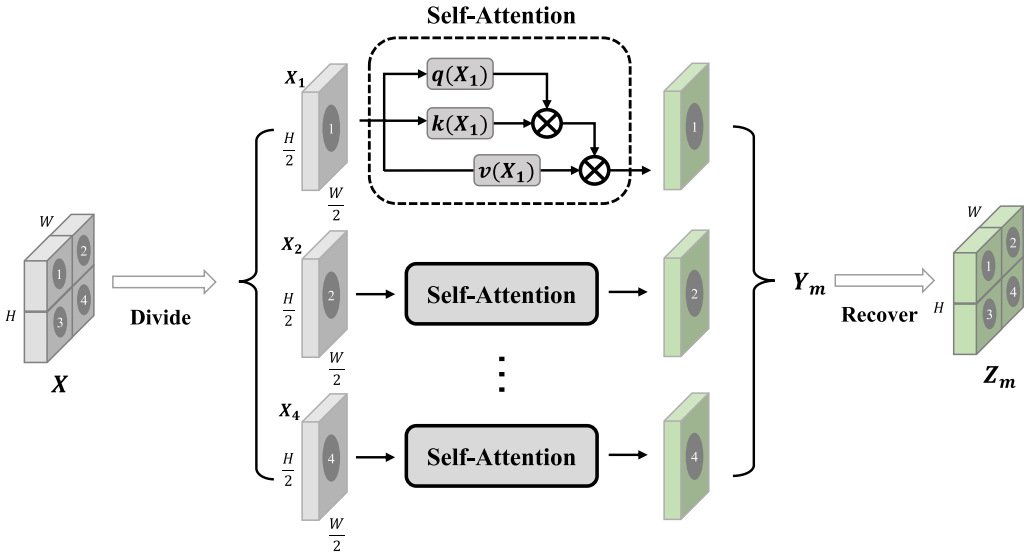


Fig. 3. The pipeline of proposed Middle-Range (MR) branch. This branch contains 3 steps. First, The input feature $X \in \mathbb{R}^{H\times W\times C}$ is divided into 2×2 patches i.e. $[X_1, X_2, X_3, X_4]$, which are ordered by rows. Second, each patch is enhanced by a self-attention module separately to get $Y_m \in \mathbb{R}^{4\times \frac{H}{2}\times \frac{W}{2}\times C}$ as the intermediate output. The operation of self-attention is as same as that described in III-A. Third, Y_m is recovered to the same size of input feature to obtain Z_m as final output.

input feature maps. In this way, each position of the input feature can interact with others regardless of their spatial distance. The network can then effectively capture long-range dependencies among all the spatial positions. The overall workflow of self-attention is illustrated in the top area of Fig. 3. Given an input feature $X = [x^1; x^2; \ldots; x^{HW}] \in \mathbb{R}^{HW\times C}$ and an output feature $Y = [y^1; y^2; \ldots; y^{HW}] \in \mathbb{R}^{HW\times C}$, which are both reshaped to the matrix form, self-attention mechanism computes the output as

$$y^i = \frac{1}{N(X)} \sum_{j=1}^{HW} attn^{ij}\, v(x^j), \tag{1}$$

where i is the index of a position of the output feature Y and j is the index enumerating all the positions of the input feature X. H, W and C are the height, width, and channel dimensions of X. $N(X)$ serves as a normalization factor which is set as HW. $v(\cdot)$ is the value transform function implemented as a 1×1 convolution, $v(x^j) \in \mathbb{R}^{C_v}$ is the transformed feature. $attn^{ij}$ indicates the attention weight contributed by position j to position i, which is defined as

$$attn^{ij} = q(x^i)^\top k(x^j), \tag{2}$$

where $q(\cdot)$ and $k(\cdot)$ are the query and key transform functions, $q(x^i) \in \mathbb{R}^{C_q}$ and $k(x^j) \in \mathbb{R}^{C_k}$ are transformed features respectively. In this work, we implement both $q(\cdot)$ and $k(\cdot)$ as 1×1 convolutions. Eq. (1) computes the response at query position i as a weighted sum of the features of all positions.

Conventional self-attention, as described above, can effectively capture long-range dependencies, which well complements convolution layers that capture short-range dependencies within a local region. However, exploiting context information from too distant positions is sometimes problematic since a position is often less correlated with those far away from it. A position tends to have stronger correlations with those are moderately near from it, where middle-range dependencies are contained within this context. Therefore, it is necessary to fill the gap between short-range and long-range dependencies and capture various correlations among entities in the scene. Also, during the process of self-attention operation, the features of the positions which contribute significant attention weights to others usually encode common patterns contained in the scene. These patterns are beneficial to the comprehensive understanding of sophisticated scenes and deserve appropriate emphasis for better capturing long-range dependencies. Based on the two observations, we propose a novel Omni-Range Dependencies Network (ORDNet) which consists of a Middle-Range (MR) branch and a Reweighed Long-Range (RLR) branch to mine middle-range dependencies and refine long-range dependencies respectively for better scene understanding (*see Fig. 2*). We will elaborate on the MR branch in III-B and the RLR branch in III-C. The overall architecture of ORDNet will be described in detail in III-D.

B. Middle-Range Branch

Conventional self-attention exploits all positions of a feature map to update the feature of each query position. By analyzing correlation patterns among ground-truth mask patches in Fig. 4, we observe that a query position tends to have stronger correlations with the positions near to it compared with those which are distant. To demonstrate this, we randomly select $1,000$ images from PASCAL-Context [10] dataset and divide the ground-truth masks into 2×2 and 4×4 patches along the height and width dimensions. We further calculate correlations inside and between patches by taking the average of corresponding similarity values:

$$Corr(p^m, p^n) = \sum_{i\in\Omega_{p^m}, j\in\Omega_{p^n}} sim(l^i, l^j)/(|\Omega_{p^m}||\Omega_{p^n}|), \tag{3}$$

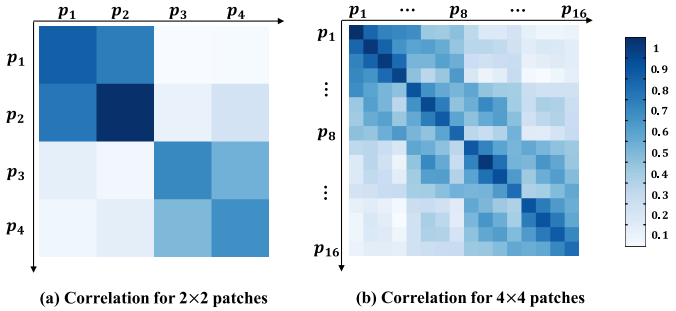


Fig. 4. Visualization of intra-patch and inter-patch correlations. (a) Correlation values computed on 2×2 patches. (b) Correlation values computed on 4×4 patches. The patches are ordered along rows. Darker color denotes a larger correlation value. The values of top left and bottom right elements as well as their surroundings are much larger than those of the rest regions. This means the pixels within the same or closer patches tend to have the same label. Furthermore, intra-patch correlation is usually stronger than inter-patch correlation.

where $sim(\cdot, \cdot)$ computes the similarity between two positions, which is either 1 or 0. We define the similarity between a pair of positions as whether their semantic labels are the same:

$$sim(l^i, l^j) = \begin{cases} 1 & l^i = l^j \\ 0 & l^i \neq l^j. \end{cases} \tag{4}$$

Here l^i and l^j denote the ground-truth category labels of position i and j; p^n and p^m represent the n-th and m-th patch, and Ω_{p^n} and Ω_{p^m} denote the set of all the positions belonging to them respectively; $|\cdot|$ means cardinality of their sets.

The visualized correlation matrices are shown in Fig. 4 (a) and (b). We can observe that the values of the top left and bottom right elements, as well as their surroundings, are much larger (darker color) than those of the rest regions, which indicates intra-patch correlations are stronger than inter-patch ones. Moreover, pixels within the same or near patches tend to share the same label. Visualization of the attention map in Fig. 1 also demonstrates that conducting self-attention in local patches can capture middle-range dependencies among nearby similar positions instead of original long-range dependencies among all the positions.

Middle-range dependencies captured from local patches are able to provide more relevant context information than long-range dependencies, considering the higher intra-patch correlations than inter-patch ones. We then develop a Middle-Range (MR) branch to capture such more informative middle-range dependencies to complement with long-range dependencies. As illustrated in Fig. 3, our MR branch explicitly divides the input feature maps into 2×2 patches and conducts self-attention operation within each patch separately. The output Y_m is then recovered to the original spatial dimensions. We narrow the self-attention range from the entire feature map to patch level so that the local nature of the feature can be exploited. Benefited from the complementarity among short-range, middle-range and long-range dependencies, the network can adapt to diverse spatial relationships between different scene elements. We use 2×2 patches for all the experiments. We also have tried to divide features into 4×4 patches but found diminishing return, possibly due to limited receptive field. Experimental results are shown in Tab. II.

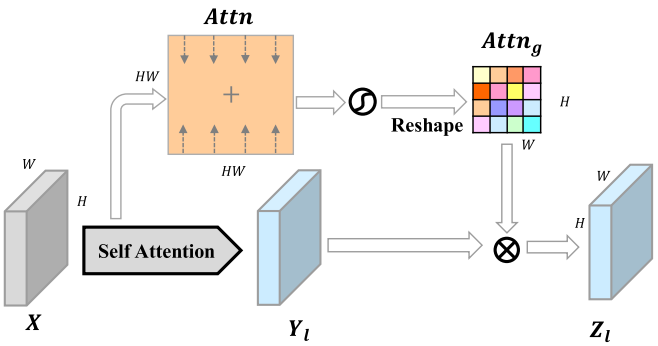


Fig. 5. Our proposed Reweighed Long-Range (RLR) branch. A self-attention module takes backbone feature as input X and outputs attended feature Y_l along with the attention map $Attn \in \mathbb{R}^{HW \times HW}$. Then $Attn$ is summed up along each column and fed into a sigmoid function to get the attention contribution vector, which is then reshaped to $H \times W$ to obtain the global attention weight contribution map $Attn_g \in \mathbb{R}^{H \times W}$. The output feature Z_l is attained by multiplying $Attn_g$ with Y_l elementwisely. Note that the difference between $Attn$ and $Attn_g$ is that each element of $Attn$ denotes the attention weights between a pair of positions in X, whereas each element of $Attn_g$ denotes the summation of contribution of current position to all the positions.

C. Reweighed Long-Range Branch

During the self-attention process, the multiplication between query and key features will generate an attention map where each element indicates the attention weight between each pair of positions. We observe that some positions contribute larger attention weights to other positions, which implies that there are stronger correlations between these positions and other ones. Features of these positions usually encode common patterns like main elements appearing in the scene and large-area continuous background. These positions may be crucial to the global context during the self-attention process. By emphasizing features of these essential positions, the long-range dependencies modeled by self-attention is able to be more accurate. Therefore, we propose a Reweighed Long-Range (RLR) branch to selectively enhance features of these positions which contribute large attention weights to others.

As illustrated in Fig 5, given an input feature $X \in \mathbb{R}^{H \times W \times C}$, we first feed it to a self-attention module to output an attended feature $Y_l \in \mathbb{R}^{H \times W \times C}$ via Equation (1) and the attention map $Attn \in \mathbb{R}^{HW \times HW}$ via Equation (2). $attn^{ji} \in Attn$ can be viewed as an attention weight contributed by position i to position j. We compute the global attention weight contribution at position i by summing up all the attention weights it contributes to other positions and further employ a simple gated mechanism as below:

$$attn_g^i = \sigma\left(\sum_{j=1}^{HW} attn^{ji}\right). \tag{5}$$

Here $\sigma(\cdot)$ is the sigmoid function and is applied to normalize the scale of $Attn_g$. The spatial dimensions of the global attention weight contribution map $Attn_g = [attn_g^1; attn_g^2; \ldots; attn_g^{HW}]$ are $H \times W$ after reshaping operation. Through Equation (5) we obtain $attn_g^i$, i.e., the normalized attention weight contribution of position i to all the

TABLE I
ABLATION STUDIES ON PASCAL-CONTEXT TEST SET. MIOU IS CALCULATED ON 59 CATEGORIES W/O BACKGROUND

Method	Backbone	mIoU(%)	pixAcc(%)
Dilated FCN [38]	ResNet50	45.30	75.34
Basic SA [7]	ResNet50	49.45	78.61
Basic SA + MR	ResNet50	50.26	78.95
Basic SA + RLR	ResNet50	50.28	78.89
ORDNet	ResNet50	**50.67**	**79.35**
ORDNet	ResNet101	**53.03**	**80.24**

positions, which measures the effect of position i on the global context of self-attended features. We finally multiply $Y_l \in \mathbb{R}^{H\times W\times C}$ with $Attn_g \in \mathbb{R}^{H\times W}$ elementwisely using the broadcasting rule to get $Z_l \in \mathbb{R}^{H\times W\times C}$ as the output of this branch:

$$Z_l = Y_l * Attn_g. \tag{6}$$

By multiplied with the global attention weight contribution $Attn_g$, the feature at each position is reweighed according to its contribution to other positions. Features of positions which encode common patterns of the scene could be emphasized for better representation. Experimental results in Table I shows that our RLR branch improves the performance without introducing extra parameters.

D. Omni-Range Dependencies Network

We here explain our proposed Omni-Range Dependencies Network (ORDNet) in detail. The architecture of our ORDNet is illustrated in Fig. 2. We use ResNet101 [33] pretrained on ImageNet [34] as the backbone network to extract visual features. To enlarge the receptive field as well as maintain feature resolution, we replace the strided convolutions in the last two stages of ResNet with atrous convolutions, with stride set as 1 and dilation rates set as 2 and 4 respectively. We also follow [18] to replace the first 7×7 convolution of ResNet with 3 consecutive 3×3 convolutions. The resolution of the output feature from backbone network $X \in \mathbb{R}^{H\times W\times C}$ is $1/8$ of the original image. X is then fed into our proposed two branches to capture middle-range and reweighed long-range dependencies in visual feature. After getting the output features Z_m and Z_l from these two branches, we concatenate them along the channel dimension. Then a 1×1 convolution is applied on the concatenated feature to reduce its channel dimensions to the same number of the input feature. We also add a shortcut connection from input feature X to the fused output of two branches to ease optimization. The fused output feature is then passed into an FCN head for final mask prediction, and we further upsample the prediction result by 8 times to match the original resolution. In practice, our proposed two branches can be easily plugged into a segmentation network due to the residual nature and further enhance feature representations.

The concurrent work InterlacedSSA [19] proposes a factorized self-attention method similar to our MR branch for semantic segmentation. The motivation of InterlacedSSA is to decrease the computation/memory cost of self-attention mechanism by factorizing it into two consecutive self-attention processes occurred in patches. However, the motivation of our MR branch is to demonstrate the effectiveness of capturing middle-range dependencies by restricting self-attention in feature patches. Moreover, the factorized self-attention in InterlacedSSA still aims to capture long-range dependencies among positions by stepwise information propagation. However, the factorized self-attention in our MR branch aims to explicitly capture middle-range dependencies among positions to fill the semantic gap between long-range and short-range dependencies for more comprehensive scene understanding. It serves as an additional information source to complement with reweighed self-attention and normal convolutions. In summary, our ORDNet can make self-attention more comprehensive in aggregating information, while InterlacedSSA can reduce the computational budget of self-attention. InterlacedSSA makes a step forward over our middle-range branch and the contributions of us and InterlacedSSA are complementary.

E. Loss Functions

Our full loss function $\mathcal{L}_{Full}$ contains two parts namely standard cross-entropy loss $\mathcal{L}_{CE}$ and Lovasz-hinge loss [35] $\mathcal{L}_{IoU}$, which are formulated as follows:

$$\begin{aligned}&\mathcal{L}_{CE}(y^*, \tilde{y})\\ &= -\sum_{k=1}^{K} y_k^* \log \tilde{y}_k,\end{aligned} \tag{7}$$

$$\begin{aligned}&\mathcal{L}_{IoU}(y^*, \tilde{y})\\ &= -Jaccard(y^*, \tilde{y})\\ &= -\sum_{k=1}^{K} \frac{|(\arg\max(y^*) == k) \bigcap (\arg\max(\tilde{y}) == k)|}{|(\arg\max(y^*) == k) \bigcup (\arg\max(\tilde{y}) == k)|},\end{aligned} \tag{8}$$

$$\begin{aligned}&\mathcal{L}_{Full}(y^*, \tilde{y})\\ &= \alpha_1 \mathcal{L}_{CE}(y^*, \tilde{y}) + \alpha_2 \mathcal{L}_{IoU}(y^*, \tilde{y}),\end{aligned} \tag{9}$$

where y^* denotes ground-truth mask, $\tilde{y}$ denotes predicted logits, α_1 and α_2 are weights for different loss terms, K denotes the number of semantic categories.

IV. EXPERIMENTS

A. Experimental Setup

1) Training : We conduct all the experiments using PyTorch [36] on three scene parsing benchmarks, including PASCAL-Context [10], COCO Stuff [11] and ADE20K [2]. We also evaluate our model on PASCAL VOC 2012 dataset [37] for semantic segmentation task. We choose dilated FCN [38] as the baseline model and plug our MR and RLR branches between the ResNet backbone and FCN head. The output prediction is bilinearly upsampled by 8 times to match the input size. We initialize the backbone with an ImageNet [39] pretrained model and other layers including MR branch, RLR branch and the FCN head are randomly initialized. We adopt the SGD optimizer with momentum set to 0.9 and weight decay set to 0.0001 to train the network. We use the polynomial learning rate scheduling $lr = baselr * (1 - \frac{iter}{total\_iter})^{power}$. The base learning rate is set to 0.004 for ADE20K dataset and 0.001 for other datasets.

The channel dimension C of the input feature X for our MR branch and RLR branch is 2048. Self-attention contains three linear layers to transform the input feature, namely query (q), key (k) and value (v). To reduce the memory cost, we set the output channel dimensions of query and key layers $C_q = C_k = 256$ and set $C_v = 512$ for all the self-attention modules we used. Our MR branch adopts 2×2 patches to capture middle-range dependencies. We conduct all the experiments on 4 NVIDIA TITAN RTX GPU cards. For data augmentation, We apply random flipping, random cropping and random resize between 0.5 and 2 for all datasets. When comparing with other methods, we adopt both standard cross-entropy loss and Lovasz-hinge loss [35] as our full loss function to train the network. Loss weights α_1 and α_2 are both set as 1.0. For the whole scene parsing dataset ADE20K, we follow [18] and the standard competition benchmark [2] to calculate mIoU by ignoring background pixels. When training our ORDNet on PASCAL-Context dataset, we also ignore the pixels of background category following [8], [18].

2) Evaluation: As prior works [8], [9], [18] show that employing multi-scale testing during evaluation is able to bring significant performance gain in semantic segmentation and scene parsing, we follow the best practice in [18] to average the predictions of different scales as the final results. During evaluation, the input image is first resized according to a set of different scales {0.5, 0.75, 1.0, 1.25, 1.5, 1.75}, then cropped to a pre-defined image size for training. The cropped image is randomly flipped and fed into the segmentation network. The output logits are cropped and averaged across above scales as final prediction. For the consideration of fairness, we adopt multi-scale testing when comparing with state-of-the-art methods. Mean Intersection of Union (mIoU) and pixel accuracy (pixAcc) are adopted as evaluation metrics.

B. Ablation Studies

We conduct both quantitative and qualitative ablation experiments on the test set of PASCAL-Context dataset [10] to verify the effectiveness of our MR and RLR branches and their variants. We train our model for 50 epochs with batch size of 16. Following [9], we use the most common 59 categories of PASCAL-Context without the background category for ablation studies.

1) MR and RLR Branches: Experimental results of our proposed two branches are illustrated in Table I. The dilated FCN baseline yields mIoU of 45.30%. After combining with basic self-attention (Basic SA) [7], the mIoU can increase significantly by 4.15%, which forms a strong baseline for our method. Upon the basic self-attention mechanism (Basic SA), adding our Middle-Range branch (Basic SA + MR) is able to achieve 0.81% improvement in mIoU, which demonstrates that more comprehensive scene context information can be extracted by integrating middle-range dependencies into the segmentation network to complement with long-range dependencies captured by original self-attention. Adding our Reweighed Long-Range branch (Basic SA + RLR) can also bring 0.81% mIoU gain over the Basic SA baseline, indicating that emphasizing features of positions which encode the common patterns of scenes is able to capture more accurate long-range dependencies to achieve better understanding of scenes. Furthermore, the RLR branch introduces no extra parameters over the strong Basic SA baseline while outperforming it by a large margin. When incorporating our proposed two branches together, our Omni-Range Dependencies Network (ORDNet) is capable of obtaining a further improvement with 1.22% mIoU gain and 0.74% pixAcc gain due to the effective complementarity between short-range (captured by convolutions), middle-range and reweighed long-range dependencies. After utilizing a deeper backbone network, our ORDNet can further achieve the performance boost to 3.58% and 1.63% gains in mIoU and pixAcc, respectively.

2) Qualitative Results: Qualitative comparison with baseline Basic SA [7] are shown in Fig 6. Our ORDNet obtains better parsing results in both global and local parts. For example, in the 4-th row of (d) and (e), Basic SA produces muddled result with nonexistent categories, while our ORDNet is able to understand the entire scene correctly and generate coherent parsing map in the assistance of omni-range dependencies. Also in the 1-st and 6-th rows of (d) and (e), our ORDNet can fix local prediction errors in Basic SA, e.g., missing tiny sheeps and confusing ear, which indicates the superiority of integrating middle-range and reweighed long-range dependencies into the segmentation network. There are also some failure cases in Fig 6. For instance, our model fails to identify the "background" category (two pillars, snowboard) in the 3-rd and 4-th row. The reason is that we ignore the pixels of "background" category when training our ORDNet on PASCAL-Context dataset following [8], [18]. Therefore, our ORDNet cannot identify the "background" category and predicts other labels for these pixels, which has no harm to the performance. We also observe other poor results like inaccurate boundaries in the last row of Fig 6. We suppose the reason is that our ORDNet cannot aggregate enough boundary details from low resolution feature maps. Exploiting low level features from the CNN backbone may alleviate this problem.

3) Number of Patches in MR Branch: We also conduct analysis on 2×2 patches and 4×4 patches in our MR branch only. Experimental results are shown in Table II. Given the same input size, conducting self-attention over 2×2 patches reduce FLOPs from 94.50G to 62.27G and reduce mIoU from 49.45% to 48.75%. Conducting self-attention over 4×4 patches could reduce FLOPs from 94.50G to 51.17G given the same input size. However, it does not work as well as original self-attention (MR_nopatch) or 2×2 patches MR branch. Enlarging crop size to 640 has limited improvement but results in larger FLOPs which is larger than 2×2 patches (Second row). We suppose that dividing feature map into too many patches for self-attention will bring about fragmented receptive field, leading to inferior performance. Note that we only compare variants of MR branch without integrating with Basic SA, thus the mIoU and PixAcc of 2×2 patches MR branch in Table II are lower than those in Table I.

4) Variants of RLR Branch: We evaluate different versions of RLR branch. Experiment results are shown in Table III. All the models are based on ResNet50. Attention-in means that

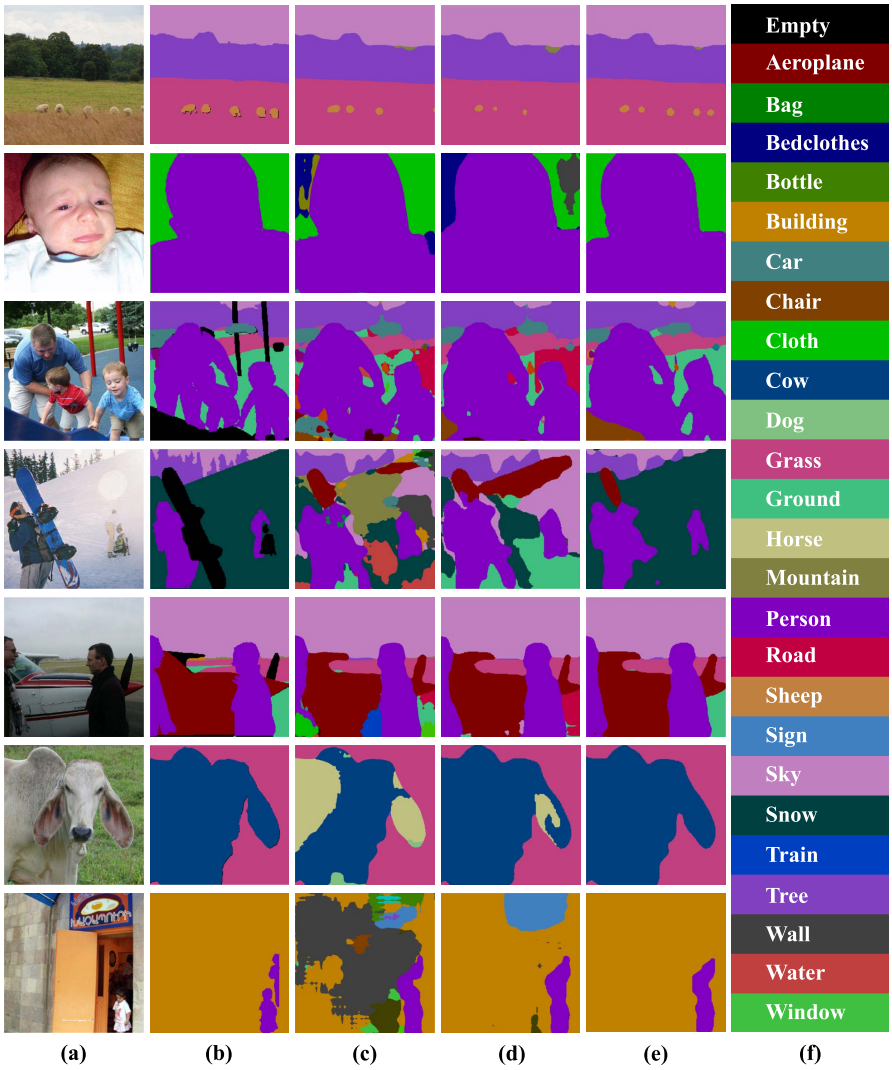


Fig. 6. Qualitative comparison with baseline Dilated FCN and Basic SA on PASCAL-Context test set. (a) Original image. (b) Ground-truth masks. (c) Results of Dilated FCN. (d) Results of Basic SA. (e) Results of our ORDNet. (f) Legend of semantic categories.

TABLE II

RESULTS OF DIFFERENT VERSIONS OF MR BRANCH ON PASCAL-CONTEXT TEST SET. ALL THE MODELS ARE BASED ON RESNET50 BACKBONE

Method	Crop Size	mIoU(%)	PixAcc(%)	GFLOPs
MR_nopatch	480 × 480	**49.45**	**78.61**	94.50
MR_2x2patch	480 × 480	48.75	76.83	62.27
MR_4x4patch	480 × 480	46.19	75.94	51.17
MR_4x4patch	640 × 640	46.92	76.08	78.97

TABLE III

RESULTS OF DIFFERENT VERSIONS OF RLR BRANCH ON PASCAL-CONTEXT TEST SET. ALL THE MODELS ARE BASED ON RESNET50 BACKBONE

Attention Matrix	Normalizing Method	mIoU(%)	PixAcc(%)
Attention-in	Softmax	49.80	78.58
Attention-in	Sigmoid	49.97	78.61
Attention-out	Softmax	50.13	**78.92**
Attention-out	Sigmoid	**50.28**	78.89

$Attn_g$ is obtained by summing up $Attn$ along each row so that each element of $Attn_g$ denotes the contribution of all positions to the current position. Attention-out means that $Attn_g$ is obtained by summing up $Attn$ along each column so that each element of $Attn_g$ denotes the contribution of current position to all others. We also try different normalization methods for $Attn_g$, including Softmax and Sigmoid functions. Results in Table III show that calculating $Attn_g$ as attention-out attains better performance than attention-in and combining attention-out with sigmoid normalization achieves the best performance on PASCAL-Context dataset.

TABLE IV

RESULTS OF DIFFERENT FUSING METHODS FOR OUTPUTS OF MR BRANCH AND RLR BRANCH. ALL THE MODELS ARE BASED ON RESNET50 BACKBONE

Fusing method	mIoU(%)	PixAcc(%)
Element-wise Summation	50.19	78.94
Concat + 1 × 1 Convolution	**50.67**	**79.35**
Attention to Scale [40]	49.73	78.81
Channel Selection [41]	49.60	78.69

5) Fusing Methods for MR and RLR Branches: We compare different approaches to fuse the outputs of our proposed MR and RLR branches on the test set of PASCAL-Context. Experiment results are presented in Table IV. All the models

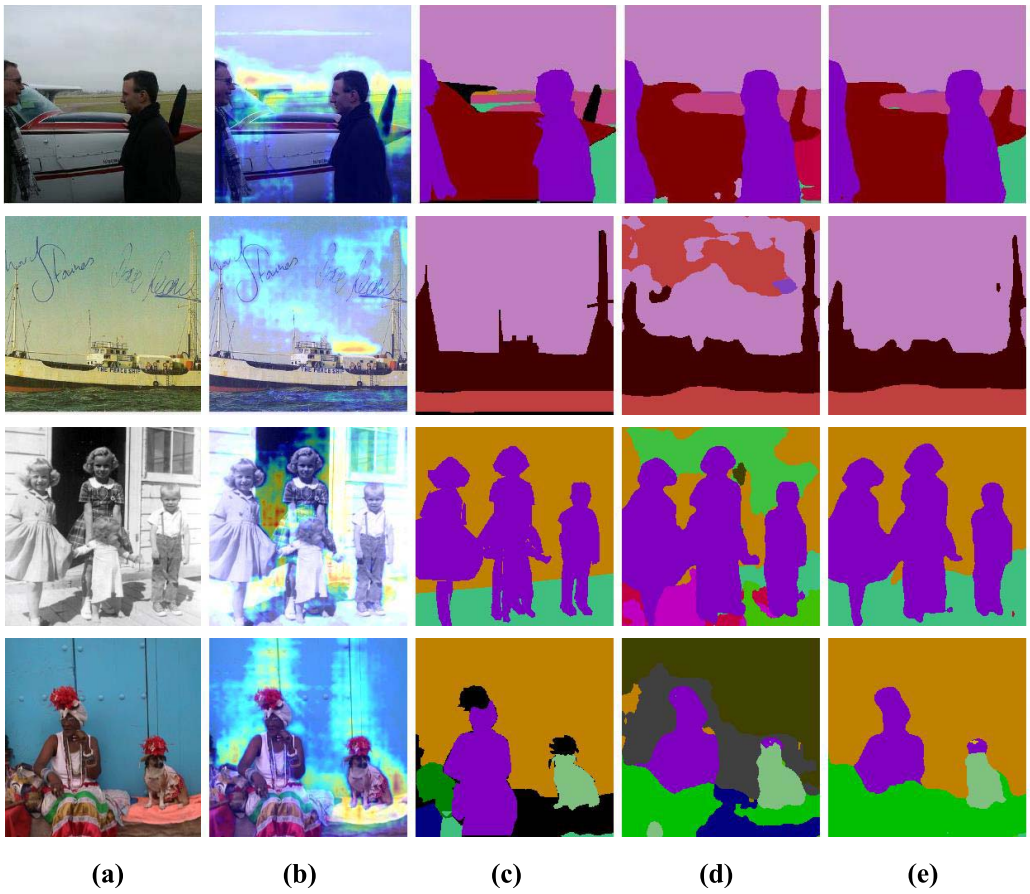


Fig. 7. Visualization of attention maps and parsing results of self-attention method and our RLR branch. (a) Original images. (b) Visualization of global attention weight contribution map $Attn_g$, which is calculated by summing up attention weights that each position contributes to other positions. (c) Ground-truth label maps. (d) Parsing results of basic self-attention. (e) Parsing results of RLR branch. We can observe from (d) that large-area continuous background usually contributes more attention weights to other positions, e.g. sky in the 2nd row, wall in the 4th row. By emphasizing these regions, our RLR branch could correct the prediction errors made by basic-attention and assign proper labels to these regions.

are based on ResNet50. Besides elementwise summation and concatenation followed by a 1×1 convolution, we also explore Attention to Scale [40] and Channel Selection [41] for feature fusion. Attention to scale obtains a selection weight map for each branch and each position of the fused feature will receive a weighted summation of the features at the same positions from the two branches. Instead of spatial dimension, Channel Selection obtains the selection weight vector for each branch and conduct weighted summation along channel dimension. Experiment results indicate that simply concatenating the two features and fusing them with a 1×1 convolution achieves the best performance. A possible reason is that for features from our two branches, normal convolution fusing them along both spatial and channel dimensions, while Attention to Scale and Channel Selection perform fusion only along spatial dimension or channel dimension, respectively.

6) Visualization of RLR Branch: To further illustrate the proposed RLR branch, we visualize the attention maps and parsing results of our model with RLR branch only in Fig 7. Column (b) visualizes of global attention weight contribution map $Attn_g$, i.e., summation of attention weights contributed by each position to all the positions. We can observe that areas with larger attention weight contribution (brighter color) usually represent common patterns of the scene, e.g., sky, grass and wall which serve as large-area continuous background. As shown in column (d) and (e), by emphasizing common pattern regions, our RLR branch is able to correct erroneous predictions made by basic self-attention and make the network understand the scene contents more comprehensively.

C. Results on PASCAL-Context

PASCAL-Context dataset [10] contains 4,998 images for training and 5,105 images for testing. All images are densely annotated with 60 categories in total including background. We train our model for 80 epochs with batch size of 16 when comparing with state-of-the-art methods. The image crop size is set to 480. We use the total 60 categories including the most frequent 59 categories and the background category to evaluate our method as prior works [9], [18], [52] do.

Compared methods and results are shown in Table V. Our ORDNet achieves 54.5% mIoU which outperforms previous state-of-the-art methods. SGR+ [55] and DeepLab-v2 [6] utilize additional COCO Stuff [11] and COCO [59] data to pretrain their models. RefineNet-152 [52] and MSCI [54] adopt deeper backbone network to boost performance. Recent CFNet [9] integrates the Context Encoding module from EncNet [18] and obtains 54.0% mIoU. InterlacedSSA [19]

TABLE V

COMPARISON WITH STATE-OF-THE-ART METHODS ON PASCAL-CONTEXT TEST SET. MIOU IS CALCULATED ON 60 CATEGORIES W/ BACKGROUND

Method	Backbone	mIoU(%)
FCN-8s [5]	-	37.8
CRF-RNN [42]	-	39.3
ParseNet [43]	-	40.4
BoxSup [44]	-	40.5
ConvPP-8 [45]	-	41.0
HO_CRF [46]	-	41.3
PixelNet [47]	-	41.4
Piecewise [48]	-	43.3
DAG-RNN + CRF [49]	-	43.7
VeryDeep [50]	-	44.5
DeepLab-v2 [6]	ResNet101 + COCO	45.7
LabelBank [51]	ResNet101	45.8
RefineNet-101 [52]	ResNet101	47.1
RefineNet-152 [52]	ResNet152	47.3
PSPNet [1]	ResNet101	47.8
Model A2, 2 conv [53]	-	48.1
MSCI [54]	ResNet152	50.3
SGR [55]	ResNet101	50.8
CLL [56]	ResNet101	51.6
EncNet [18]	ResNet101	51.7
SGR+ [55]	ResNet101 + COCO Stuff	52.5
DUpsampling [57]	Xception-71	52.5
DANet [8]	ResNet101	52.6
SVCNet [58]	ResNet101	53.2
CFNet [9]	ResNet101	54.0
InterlacedSSA [19]	ResNet101	54.1
ORDNet (ours)	ResNet101	**54.5**

further boosts the performance to 54.1% mIoU via factorized self-attention similar to our MR branch. Our ORDNet achieves better performance than above methods without using extra pretraining data, deeper backbone network or other context aggregation modules. It is demonstrated that capturing omni-range dependencies is more effective in providing richer semantic information for scene parsing.

D. Results on COCO Stuff

COCO Stuff dataset [11] contains 10,000 images from MSCOCO dataset [59] with dense annotations of 80 thing (e.g. book, clock) and 91 stuff categories (e.g. flower). We use 9,000 images for training and the rest for testing. We adopt batch size of 16 and train our model for 80 epochs. The image crop size is set to 520. Mean IoU results calculated on all the 171 categories are shown in Table VI. Among the compared methods, DAG-RNN [49] utilizes chain-RNNs to model rich spatial dependencies. CCL [56] adopts a gating mechanism in the decoder stage to improve inconspicuous objects and background stuff segmentation. SGR [55] uses a knowledge graph to convert image features into symbolic nodes and conducts graph reasoning on them. SVCNet [58] generates a scale- and shape-variant semantic mask for each pixel to confine its contextual region for more adaptive context aggregation. DANet [8] employs spatial and channel-wise self-attention to further improve performance. Recent InterlacedSSA [19] proposes a factorized approach similar to our MR branch to accelerate self-attention. Our method outperforms these methods with a large margin and achieves a new state-of-the-art result of 40.5% mIoU with no external knowledge used.

TABLE VI

COMPARISON WITH STATE-OF-THE-ART METHODS ON COCO STUFF TEST SET

Method	Backbone	mIoU(%)
FCN [11]	-	22.7
FCN-8s [5]	-	27.2
DAG-RNN + CRF [49]	ResNet101	31.2
RefineNet [52]	ResNet152	33.6
LabelBank [51]	ResNet101	34.3
DeepLab-v2 [6]	ResNet101	34.4
CLL [56]	ResNet101	35.7
DSSPN [60]	ResNet101	36.2
SGR [55]	ResNet101	39.1
InterlacedSSA [19]	ResNet101	39.2
SVCNet [58]	ResNet101	39.6
DANet [8]	ResNet101	39.7
ORDNet (ours)	ResNet101	**40.5**

TABLE VII

COMPARISON WITH STATE-OF-THE-ART METHODS ON ADE20K VALIDATION SET

Method	Backbone	mIoU(%)	pixAcc(%)
SegNet [61]	-	21.64	71.00
FCN [5]	-	29.39	71.32
DilatedNet [38]	-	32.31	73.55
CascadeNet [2]	-	34.90	74.52
DeepLabv2 [6]	ResNet101	38.97	79.01
RefineNet-101 [52]	ResNet101	40.20	-
RefineNet-152 [52]	ResNet152	40.70	-
DSSPN [60]	ResNet101	42.03	81.21
PSPNet-101 [1]	ResNet101	43.29	81.39
PSPNet-269 [1]	ResNet269	44.94	81.69
Model A2, 2c [53]	-	43.73	81.17
SGR [55]	ResNet101	44.32	81.43
EncNet [18]	ResNet101	44.65	**81.69**
CFNet [9]	ResNet101	44.89	-
InterlacedSSA [19]	ResNet101	45.04	-
ORDNet (ours)	ResNet101	**45.39**	81.48

This result indicates that capturing omni-range dependencies is more effective than merely modeling long-range dependencies in conventional self-attention.

E. Results on ADE20K

ADE20K [2] is a large scene parsing benchmark with 150 categories including stuff and objects. It contains 20,210 images for training and 2,000 for validation. We train our model for 120 epochs on the training set with batch size of 16 and report mIoU and pixAcc results on the validation set. As the average image size of ADE20K dataset is larger than others, we adopt image crop size of 576 on ADE20K. Comparison with previous state-of-the-art methods is shown in Table VII. PSPNet-269 [1] uses a much deeper backbone network than other methods. EncNet [18] and CFNet [9] exploit prior information of semantic categories appearing in the scene to improve performance. InterlacedSSA [19] introduces factorized self-attention similar to our MR branch and obtain 45.04% mIoU. Our method achieves 45.39% mIoU which outperforms previous methods without using deeper backbone, category prior or external knowledge like [55]. As mentioned in the Section 4.1 of InterlacedSSA, the 0.2% improvement of their method is not neglectable considering the improvements on ADE20K is very challenging.

TABLE VIII
RESULTS ON ADE20K TEST SET. EVALUATION PROVIDED BY THE CHALLENGE ORGANIZERS

Method	Ensemble models	Train/Trainval	Backbone	Final Score
Dense Relation Network	-	-	-	56.35
DRANet101_SingleModel	No	Trainval	ResNet101	56.72
Adelaide	Yes	Trainval	-	56.73
SenseCUSceneParsing	No	Train	-	55.38
SenseCUSceneParsing	Yes	Trainval	-	**57.21**
ORDNet(Ours)	No	Train	ResNet101	56.67
ORDNet(ours)	No	Trainval	ResNet101	**56.86**

Therefore, results of our method demonstrate capturing omni-range dependencies is also effective in challenging and complicated scenes.

To further demonstrate the effectiveness of our ORDNet, we evaluate our model on the test set. The experiment results are shown in Table VIII. Our model without finetuning on validation set achieves 56.67 final score and surpasses most other methods. The final score denotes the average of PixAcc and mIoU. For fair comparison, we further finetune our model on the train+val set of ADE20K for 20 epochs, with the same training scheme except that the initial learning rate is set to $1e-4$. Our ORDNet achieves a final score of 56.86 on the test set with a single model and ranks at the 2-nd place on the leaderboard of MIT Scene Parsing Benchmark. Our single model surpasses the 3-rd place, Adelaide, which ensembles multiple models. The 1-st place, SenseCUSceneParsing achieves 57.21 final score by ensembling multiple models as well. Its single model trained only on the training set achieves 55.38 while our ORDNet achieves 56.67 with the same setting.

F. Results on PASCAL VOC 2012

We also evaluate the proposed ORDNet on PASCAL VOC 2012 dataset [37] with 21 categories for semantic segmentation task. We adopt the augmented dataset [62] which contains $10,582$ images for training, $1,449$ images for validation and $1,456$ images for testing. We first train on the augmented train + val set for 60 epochs with initial learning rate of $1e-3$. Then we finetune the model on the original PASCAL VOC training set for another 20 epochs with learning rate of $1e-4$. We adopt ResNet101 as backbone network and the image crop size is set to 480. Our ORDNet achieves 83.3% mIoU on PASCAL VOC 2012 test set without using COCO pretraining or additional context aggregation modules. It is demonstrated that our method can also adapt to foreground object segmentation task by capturing omni-range dependencies.

G. Analysis of Computational Overhead

As shown in TABLE IX, all the models are run on a single NVIDIA TITAN XP GPU card to report computational overhead. InterlacedSSA is superior to other methods including ours in terms of speed and memory cost. It is a natural result since the main idea of InterlacedSSA is to decrease the computational budget of self-attention via feature factorization. While the main idea of our ORDNet is to demonstrate the effectiveness of capturing omni-range dependencies by our MR and RLR branches, which outperforms

TABLE IX
EFFICIENCY COMPARISON GIVEN INPUT FEATURE MAP OF SIZE [2048 × 128 × 128] IN INFERENCE STAGE

Methods	Memory (MB)	GFLOPs	Time (ms)
InterlacedSSA [19]	**252**	**386**	**45**
Basic SA [7]	2168	619	77
Ours	2192	624	83
DANet [8]	2339	1110	121

InterlacedSSA on all the datasets we adopted in this article. Reducing computational complexity is not one of our claims. Comparing with Basic SA and DANet, our ORDNet actually has a moderate computational overhead since we reduce the channel dimensions of query, key and value layers in our MR and RLR branches to 256, 256 and 512 respectively, but our ORDNet achieves higher performances. We will explore how to reduce its time and memory costs in the future work.

V. CONCLUSION AND FUTURE WORK

In this article, we address the scene parsing problem which requires the model to segment the entire scene instead of foreground objects. We propose a novel Omni-Range Dependencies Network (ORDNet) which restricts the scope of self-attention to local patches to capture middle-range dependencies and meanwhile selectively emphasizes spatial regions contributing significant attention weights to others to model more accurate long-range dependencies. By integrating middle-range, reweighed long-range and short-range dependencies captured by local convolutions together, our ORDNet can aid models in adapting to various spatial scales and relationships in the complicated natural images, thus strengthening local and global feature representations. Extensive experiments on four scene parsing and segmentation benchmarks demonstrate its superior performance. Furthermore, our ORDNet can be applied to other visual tasks for capturing omni-range dependencies due to its generality and plug-and-play property. In the future, we hope to apply the ORDNet to other visual tasks and study how to further reduce its computing budget.

REFERENCES

[1] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.
[2] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 633–641.
[3] W.-C. Hung *et al.*, "Scene parsing with global context embedding," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2631–2639.

[4] R. Zhang, S. Tang, Y. Zhang, J. Li, and S. Yan, "Scale-adaptive convolutions for scene parsing," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2031–2039.

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[7] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[8] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[9] H. Zhang, H. Zhang, C. Wang, and J. Xie, "Co-occurrent features in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 548–557.

[10] R. Mottaghi *et al.*, "The role of context for object detection and semantic segmentation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 891–898.

[11] H. Caesar, J. Uijlings, and V. Ferrari, "COCO-stuff: Thing and stuff classes in context," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1209–1218.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[13] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*. [Online]. Available: http://arxiv.org/abs/1412.7062

[14] S. Kuanar, K. Rao, M. Bilas, and J. Bredow, "Adaptive cu mode selection in HEVC intra prediction: A deep learning approach," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 1, pp. 144–155, Jan. 2019.

[15] S. Kuanar, V. Athitsos, N. Pradhan, A. Mishra, and K. R. Rao, "Cognitive analysis of working memory load from eeg, by a deep recurrent neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018.

[16] S. Kuanar, V. Athitsos, D. Mahapatra, K. R. Rao, Z. Akhtar, and D. Dasgupta, "Low dose abdominal CT image reconstruction: An unsupervised learning based approach," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1351–1355.

[17] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587

[18] H. Zhang *et al.*, "Context encoding for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7151–7160.

[19] L. Huang, Y. Yuan, J. Guo, C. Zhang, X. Chen, and J. Wang, "Interlaced sparse self-attention for semantic segmentation," 2019, *arXiv:1907.12273*. [Online]. Available: http://arxiv.org/abs/1907.12273

[20] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: http://arxiv.org/abs/1409.0473

[21] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2015, *arXiv:1508.04025*. [Online]. Available: http://arxiv.org/abs/1508.04025

[22] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. NeurIPS*, 2016, pp. 289–297.

[23] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7W: Grounded question answering in images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4995–5004.

[24] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 21–29.

[25] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. ICML*, 2015, pp. 2048–2057.

[26] A. Vaswani *et al.*, "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 5998–6008.

[27] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.

[28] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," 2018, *arXiv:1805.08318*. [Online]. Available: http://arxiv.org/abs/1805.08318

[29] G. Li, X. He, W. Zhang, H. Chang, L. Dong, and L. Lin, "Non-locally enhanced encoder-decoder network for single image de-raining," in *Proc. ACM Multimedia Conf. Multimedia Conf. MM*, 2018, pp. 1056–1064.

[30] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[31] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, Sep. 2018, pp. 3–19.

[32] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "BAM: Bottleneck attention module," in *Proc. BMVC*, 2018, pp. 1–14.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[35] J. Yu and M. Blaschko, "Learning submodular losses with the lovász hinge," in *Proc. ICML*, 2015, pp. 1623–1631.

[36] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. NeurIPS*, 2019.

[37] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[38] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: http://arxiv.org/abs/1511.07122

[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NeurIPS*, 2012, pp. 1097–1105.

[40] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3640–3649.

[41] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.

[42] S. Zheng *et al.*, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1529–1537.

[43] W. Liu, A. Rabinovich, and A. C. Berg, "ParseNet: Looking wider to see better," 2015, *arXiv:1506.04579*. [Online]. Available: http://arxiv.org/abs/1506.04579

[44] J. Dai, K. He, and J. Sun, "BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1635–1643.

[45] S. Xie, X. Huang, and Z. Tu, "Top-down learning for structured labeling with convolutional pseudoprior," 2015, *arXiv:1511.07409*. [Online]. Available: http://arxiv.org/abs/1511.07409

[46] A. Arnab, S. Jayasumana, S. Zheng, and P. H. Torr, "Higher order conditional random fields in deep neural networks," in *Proc. ECCV*, 2016, pp. 524–540.

[47] A. Bansal, X. Chen, B. Russell, A. Gupta, and D. Ramanan, "PixelNet: Towards a general pixel-level architecture," 2016, *arXiv:1609.06694*. [Online]. Available: http://arxiv.org/abs/1609.06694

[48] G. Lin, C. Shen, A. van den Hengel, and I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3194–3203.

[49] B. Shuai, Z. Zuo, B. Wang, and G. Wang, "Scene segmentation with DAG-recurrent neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1480–1493, Jun. 2018.

[50] Z. Wu, C. Shen, and A. van den Hengel, "Bridging category-level and instance-level semantic image segmentation," 2016, *arXiv:1605.06885*. [Online]. Available: http://arxiv.org/abs/1605.06885

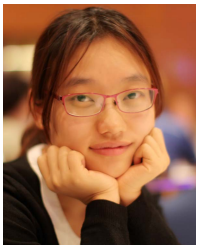
[51] H. Hu, Z. Deng, G.-T. Zhou, F. Sha, and G. Mori, "LabelBank: Revisiting global perspectives for semantic segmentation," 2017, *arXiv:1703.09891*. [Online]. Available: http://arxiv.org/abs/1703.09891

[52] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1925–1934.

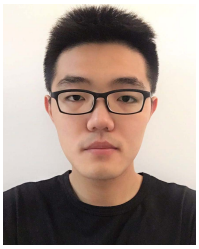
[53] Z. Wu, C. Shen, and A. van den Hengel, "Wider or deeper: Revisiting the ResNet model for visual recognition," 2016, *arXiv:1611.10080*. [Online]. Available: http://arxiv.org/abs/1611.10080

[54] D. Lin, Y. Ji, D. Lischinski, D. Cohen-Or, and H. Huang, "Multi-scale context intertwining for semantic segmentation," in *Proc. ECCV*, Sep. 2018, pp. 603–619.
[55] X. Liang, Z. Hu, H. Zhang, L. Lin, and E. P. Xing, "Symbolic graph reasoning meets convolutions," in *Proc. NeurIPS*, 2018, pp. 1853–1863.
[56] H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang, "Context contrasted feature and gated multi-scale aggregation for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2393–2402.
[57] Z. Tian, T. He, C. Shen, and Y. Yan, "Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3126–3135.
[58] H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang, "Semantic correlation promoted shape-variant context for segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8885–8894.
[59] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *Proc. ECCV*, 2014, pp. 740–755.
[60] X. Liang, E. Xing, and H. Zhou, "Dynamic-structured semantic propagation network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 752–761.
[61] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for scene segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[62] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 991–998.

Shaofei Huang received the B.S. degree from Peking University. She is currently pursuing the master's degree with the Institute of Information Engineering, Chinese Academy of Sciences. Her research interests include scene parsing and instance segmentation.



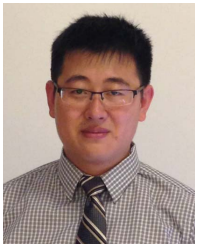
Si Liu (Member, IEEE) received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences. She is currently an Associate Professor with Beihang University. She has been a Research Assistant and a Postdoctoral Researcher with the National University of Singapore. She has published over 40 cutting-edge articles on the human-related analysis including the human parsing, face editing, and image retrieval. Her research interests include computer vision and multimedia analysis. She was a recipient of the Best Paper Award of ACM MM 2013 and the Best Demo Award of ACM MM 2012. She was the Champion of CVPR 2017 Look Into Person Challenge and the Organizer of ECCV 2018 Person in Context Challenge.



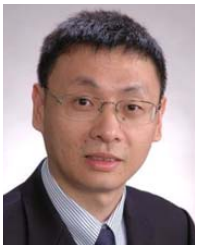
Tianrui Hui received the B.Eng. degree from Sun Yat-sen University. He is currently pursuing the master's degree with the Institute of Information Engineering, Chinese Academy of Sciences. His research interests include semantic segmentation and referring segmentation.

Jizhong Han (Member, IEEE) received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences. He is currently a Professor with the Institute of Information Engineering, Chinese Academy of Sciences. He has published over 60 articles and held over ten domestic patents. He is also the principal investigator or participant of several National 973 or 863 Programs. His research interests include big data storage and intelligent information processing.

Bo Li is currently a Changjiang Distinguished Professor with the School of Computer Science and Engineering, Beihang University. He is currently the Dean of the AI Research Institute, Beihang University. He is the Chief Scientist of National 973 Program and the Principal Investigator of the National Key Research and Development Program. He has published over 100 articles in top journals and conferences and held over 50 domestic and foreign patents. He was a recipient of The National Science Fund for Distinguished Young Scholars.



Jiashi Feng (Member, IEEE) received the B.Eng. degree from the University of Science and Technology, China, in 2007, and the Ph.D. degree from the National University of Singapore in 2014. He was a Postdoctoral Researcher with the University of California from 2014 to 2015. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering with the National University of Singapore. His current research interests include machine learning and computer vision techniques for large-scale data analysis. Specifically, he has done work in object recognition, deep learning, machine learning, high-dimensional statistics, and big data analysis.



Shuicheng Yan (Fellow, IEEE) is currently the Chief Technology Officer of YITU Technology and the Dean's Chair Associate Professor with the National University of Singapore. He is a Fellow of IAPR and the Academy of Engineering, Singapore. His research include machine learning, computer vision, and multimedia, and he has authored/coauthored hundreds of technical articles over a wide range of research topics, with Google Scholar citation over 60 000 times and H-index 104. He is an ISI Highly-Cited Researcher of 2014, 2015, and 2016, and an ACM Distinguished Scientist. His team received seven times winner or honorable-mention prizes in PASCAL VOC and ILSVRC competitions, along with more than ten times best (student) paper prizes.