Semantic Image Segmentation by Scale-Adaptive Networks

Zilong Huang, Chunyu Wang, Xinggang Wang, Wenyu Liu, Jingdong Wang

***Abstract*—Semantic image segmentation is an important yet unsolved problem. One of the major challenges is the large variability of the object scales. To tackle this scale problem, we propose a Scale-Adaptive Network (SAN) which consists of multiple branches with each one taking charge of the segmentation of the objects of a certain range of scales. Given an image, SAN first computes a dense scale map indicating the scale of each pixel which is automatically determined by the size of the enclosing object. Then the features of different branches are fused according to the scale map to generate the final segmentation map. To ensure that each branch indeed learns the features for a certain scale, we propose a scale-induced ground-truth map and enforce a scale-aware segmentation loss for the corresponding branch in addition to the final loss. Extensive experiments over the PASCAL-Person-Part, the PASCAL VOC 2012, and the Look into Person datasets demonstrate that our SAN can handle the large variability of the object scales and outperforms the state-of-the-art semantic segmentation methods.**

***Index Terms*—Semantic Object Parsing, Human Parsing, Scale Adaptive.**

I. Introduction

SEMANTIC image segmentation is the task of assigning semantic class labels to every pixel in the image and has been actively studied in recent papers [1]–[9]. Many applications can be classified to this task depending on the pre-defined class label set such as person re-identification [10], human part segmentation [2], action segmentation [11], clothing parsing [12] and pose estimation [13].

Deep Convolutional Neural Networks have significantly advanced the image segmentation problem due to the powerful end-to-end learned features. For example, [5] proposes a fully convolutional network (FCN) which predicts dense outputs from arbitrary-sized input images. Without additional machinery, the approach exceeds its previous state-of-the-arts and becomes a cornerstone of modern semantic segmentation methods. Considering that the pixels in the images are not totally independent, [14] proposes to build a fully connected conditional random field (CRF) on top of the CNN outputs. The experiment results show that it can obtain more consistent segmentations. To obtain a dense output, [15] proposes the dilated convolutions to support the exponential expansion of the receptive field without loss of resolution. And, [16] proposes a decoder network to map the low-resolution encoder feature maps to full input resolution feature maps for pixel-wise classification. [17] proposes a hybrid dilated convolution alleviates the "gridding issue" caused by the standard dilated convolution operation later on.

In spite of the significant progress made by the CNN based methods, they have notable drawbacks of having fixed receptive field. Consequently, they can only perfectly segment the objects of a single scale and have degraded performance for objects which are much larger or smaller. Similar observations have been made in [18]. More specifically, for large objects, because the approach only observes local information, the enclosing pixels may have inconsistent labels; in contrast, smaller objects are often ignored and classified as background.

To address the scale issues, DeepLab-MSc-LargeFOV [14] utilizes a skip-net architecture that exploits features from different levels of the network to obtain multi-scale features. [19] employs an object detector and zooms the detected image regions into their proper scales to refine the parsing. The attention-based method [2] and the Deeplabv2 [20] both feed multi-scale inputs into CNNs to generate multi-scale predictions. Scale-Adaptive Convolution [21] and Deformable Convolutional Networks [22] improve the convolutional layer that effectively have dynamic and learnable receptive field.

In this work, we propose a Scale-Adaptive Network (SAN) to address the problem. In the training stage, SAN first quantizes the object scales (sizes) into T sets based on the areas of the bounding boxes in the training datasets. For a training image, the ground truth annotation consists of not only a class label but also a scale label which takes values from 1 to T. The scale label of a pixel is determined or approximated by the scale of the enclosing object.

SAN consists of a shared fully convolutional network followed by T branches. See Fig. 1 for an overview of the structures. We visualize three branches in the figure. Each branch takes charge of the segmentation of the objects of a particular scale. So in the training stage, each branch will predict class labels for the pixels of its corresponding scale, the output of each branch is named scale-induced segmentation map. See the ground truth annotations visualized in the gray boxes. In addition to the class labels, SAN also predicts a scale label for each pixel thus produces a scale mask map for the whole image. The scale mask map encodes the probability of each pixel belonging to each scale. The output features maps of the T branches are fused according to the scale mask to generate the final class label map.

Fig. 2 shows an example of semantic human part segmentation results and the intermediate results by SAN. The middle column indicates that each branch can actually make

Zilong Huang, Xinggang Wang and Wenyu Liu are with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China. Chunyu Wang and Jingdong Wang are with Microsoft Research Asia, Beijing 100080, China. Corresponding author: Xinggang Wang (xgwang@hust.edu.cn). This work was mainly done when Zilong was an intern at Microsoft Research Asia.

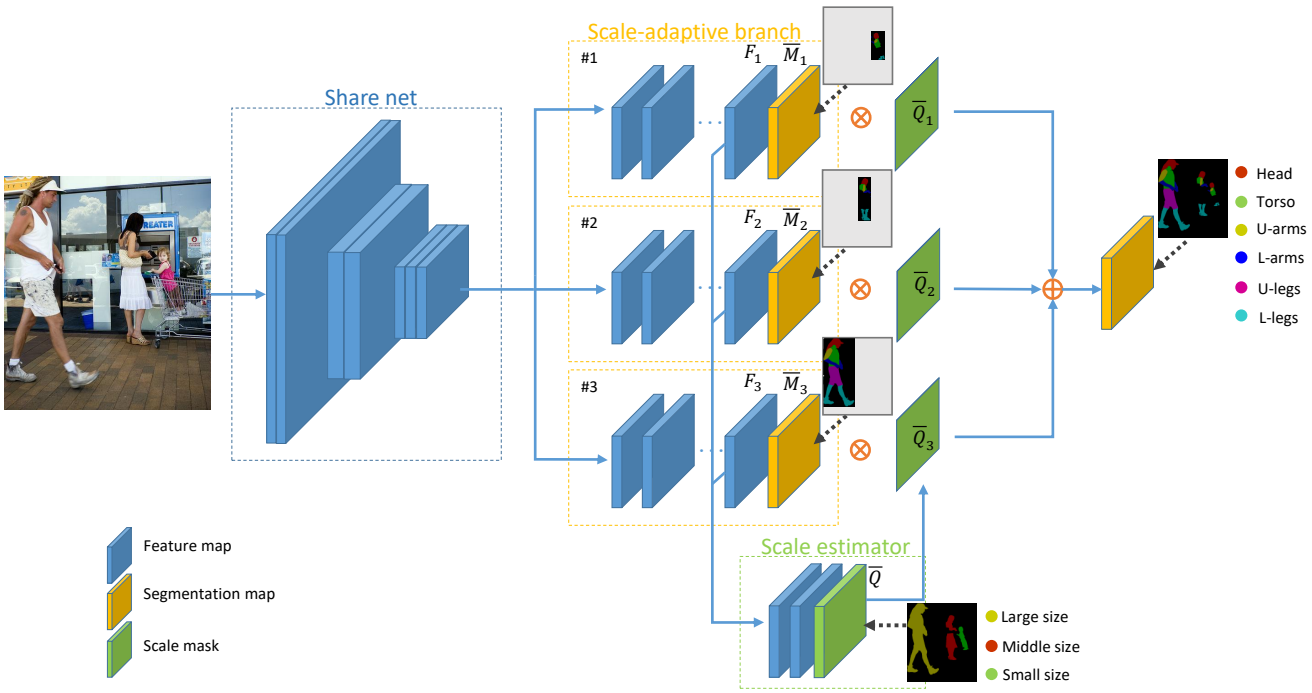


Fig. 1. Overview of the proposed Scale-adaptive Network. On top of the convolutional network is called a shared net, we put multiple ($T = 3$) scale-adaptive branches that consist of several convolutional layers to segment the objects of the corresponding scale. The scale estimator takes as input the concatenated feature maps generated by the branches and generates a scale mask. Finally, the scale mask is used to select and fuse high-quality results generated by scale-adaptive branches into a final segmentation map. The entire network is trained under multi-scale supervision (dashed lines).

predictions for the pixels of a particular scale, the right column indicates that our approach can accurately predict the probability map of scale mask. The final output in the left column is produced by the sum of the product of each scale-induced segmentation map and the corresponding scale mask. This scale-induced fusion is indeed better than simply summing the feature maps, because each branch may make mistakes to segment object with non-corresponding scale, simply summing the feature maps may result in bad results. Beside the pixel-wise class label, the proposed method also needs bounding box annotation which is cheaper and more effective to obtain. Compared with the pixel-wise class label, the addition annotation, i.e. object bounding box, the cost is low.

There are many datasets for object segmentation such as Pascal-Person-Part [23], LIP [24], Fashionista [25], and Penn-Fudan pedestrians [26], among which Pascal-Person-Part and LIP have the largest variation in scale. Thus, we choose the Pascal-Person-Part and LIP datasets to evaluate our approach with extensive experiments on human part segmentation. The experiment results show that our SAN outperforms the previous state-of-the-art methods which justifies that our method can handle the variability of object scale. Meanwhile, to validate the generalization capability of our method, we conduct experiments on the PASCAL VOC 2012 and Cow-Horse-Sheep dataset and also present competitive performance over alternative methods.

Our main contributions are summarized below:

- We propose a scale-adaptive network, which is composed of a shared net, scale-adaptive branches, scale estimator, and scale-based fusion network that generates the final segmentation map. The scale-adaptive network can handle the large variability of object scale.
- The proposed scale-adaptive branches can segment the objects of the corresponding scale with no need of feature pyramid [27] or image pyramid [2].
- We achieve outstanding performance using the scale-

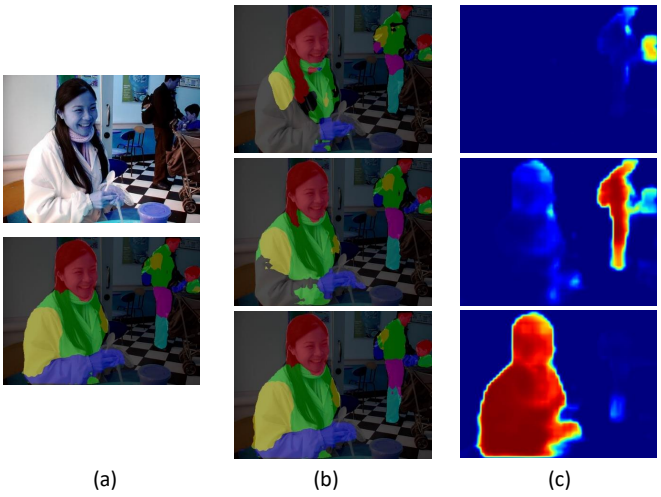


Fig. 2. An example of semantic human part segmentation results and the intermediate results by SAN. (a) the input image and final human parsing result. (b) the scale-induced segmentation maps (from top to bottom corresponds to three scales: small, middle, large). (c) the scale masks (from top to bottom corresponds to three scales: small, middle, large). Final human parsing result is produced by the sum of each scale-induced segmentation map produces the corresponding scale mask.

adaptive network trained on PASCAL-Person-Part, PASCAL VOC 2012, Cow-Horse-Sheep, and LIP, and obtain the best accuracies using a single model.

The rest of this paper is organized as follows. We first review related work in Section II and describe the architecture of our network in Section III. In Section IV, the detailed procedure to learn a scale-adaptive network is discussed and experimental results are analyzed. Section V presents our conclusion and future work.

II. Related work

The last years have seen a renewal of interest on semantic object parsing. [28] performs probabilistic inference in a generative model for parts-based object segmentation, [29] constructs an efficient fully connected conditional random field (FCRF) [30] to jointly predict the final object and part labels simultaneously. [31] proposes Graph LSTM to model the spatial relations on superpixels for semantic object parsing. Our work pays close attention to scale problem in the segmentation object parsing task.

A. Approaches to scale variation

The traditional approaches [13], [28], [32] to semantic object parsing are to perform inference under constrained conditions with pre-suppose known scales, which are limited when applied to parsing human instances in the wild, since humans in real-world images often vary in poses, scales, and may be occluded or highly deformed.

There are many works to address the scale problem to improve object detection or semantic segmentation. [19] divided and conquered the problem by employing a general object detector and performing object part segmentation for each detection. Once an object is detected, the scale of the object is obtained, then it can be zoomed into its proper scale to refine its parsing. These top-down approaches directly leverage existing techniques of objection detection for semantic object parsing. But the framework relies heavily on the performance of object detector, which means that if the object detector fails, there is no chance of recovery.

A skip-net architecture that exploits features from different levels of the network is also a common approach in semantic segmentation and object detection. For example, DeepLab-MSc-LargeFOV [14] attached two convolution layers to the input image and the output of each of the first four max-pooling layers. The network concatenated feature maps generated by forementioned convolutional layers to the main networks last layer feature map and generated segmentation maps. But this is not an effective solution for large variations of objects size and the performance gain is not significant. Another common approach is to feed multi-scale inputs to the fully convolutional network. For example, [2] resized the input image into three scales to result in three-scale features and used an attention mechanism that learns to softly weight the multi-scale features at each pixel location to generate the final segmentation map. [33] applied the multi-scale convolutional net that contains multiple copies of a single network(all sharing the same weights) to different scales of a Laplacian pyramid version of the input image, and fused the features from all the scales. Scale-Adaptive Convolution [21] added a new scale regression layer to dynamically infer the position-adaptive scale coefficients which are adopted to resize the convolutional patches. And Deformable Convolutional Networks [22] added another convolutional layer to learn 2D offset for the regular grid sampling locations in the standard convolution.

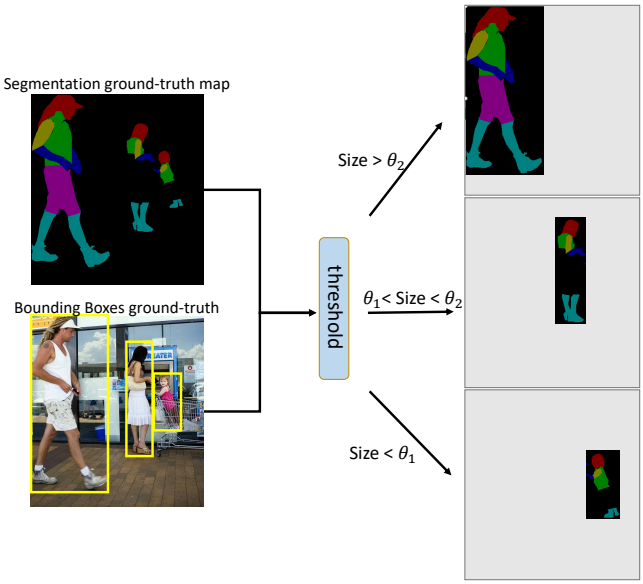


Fig. 3. The way to generate a scale-induced ground-truth map to supervise the scale-adaptive branches to learn multi-scale features. Supposing the network has three branches ($T = 3$), three scale-induce ground-truth maps are generated and gray areas in the maps indicate where there is no need for backward computation. We consider only the pixels which lie in the bounding box.

Different from the above methods, we propose a scale-adaptive network which takes the single scale image as input and uses scale-adaptive branches to generate multi-scale features. This method is followed by a scale-based fusion to generate the final segmentation map. Instead of using skip-net to generate multi-scale feature, the scale-adaptive branches learn more abstract multi-scale structure under the supervision of scale-induced ground-truth map as shown in Fig. 3. More details will be given in Section III

B. Fusion methods

To merge the predictions from multi-scale features, there are three common approaches: average-pooling ([34]) over scales, max-pooling ([35]) over scales or using attention model ([2]) that learns to softly weight the multi-scale features. Motivated by [2], we propose to jointly learn a scale estimator to generate a scale mask. The scale mask indicates which branch is responsible for each scale and position. The final output of our model is produced by the weighted sum of segmentation maps across all scales.

III. Scale-adaptive network

A. Problem Formulation

Semantic segmentation is to predict the class of each pixel, and produce a segmentation map. Formally, given an image

$\mathbf{I}$ with the width and the height being W and H, pixel-wise class labels, $\mathbf{M} \in \mathbb{R}^{W \times H}$, in which each value m_{xy} in the map indicates whether the pixel p_{xy} belongs to the class $c \in \{1, ..., C\}$ where C is the number of classes of interest. We also have bounding box annotations, $\mathbf{B}$. The purpose is to output a segmentation map. We also need a scale mask $\mathbf{Q} \in \mathbb{R}^{W \times H}$, in which each value q_{xy} indicates the scale of the object the pixel p_{xy} belongs to. In this paper, we avoid tedious annotation to achieve the scale mask and simply estimate it according to the area of the bounding box of each object, which is then quantized into a T discrete scales. Considering the scale mask, we decompose the segmentation map $\mathbf{M}$ into T maps, $\{\mathbf{M}_1, \ldots, \mathbf{M}_T\}$, where each map $\mathbf{M}_t$ corresponds to the segmentation map with the scale of the objects being t. As shown in Fig. 3, suppose $T = 3$, first of all, sorting the objects by their square roots of the bounding box area. Next, finding two thresholds to split the scale space into three subspaces. Each scale subspace has the same amount of object instances. Then, each bounding box has the category c and scale t by comparing with the thresholds. The pixels within the bounding box $b_{c,t} \in \mathbf{B}$ having the same class c in pixel-wise class labels M will be labeled as scale t. Thus, the pixel-wise scale map and pixel-wise scale-induced map are obtained.

B. Network Architecture

The network architecture is given in Fig. 1. It consists of a shared net, three scale-adaptive branches, a scale estimator, and a scale-based fusion subnet which generates the final segmentation map. An input image passes through a shared net, and T scale-adaptive branches, then, produces T feature maps, $\{\mathbf{F}_1, \ldots, \mathbf{F}_T\}$, which are next fed into scale-adaptive segmentation map generator. There are T separate segmentation map generators, and the input of each generator is a single feature map $\mathbf{F}_t$. The output segmentation maps are $\{\bar{\mathbf{M}}_1, \ldots, \bar{\mathbf{M}}_T\}$. The T feature maps $\{\mathbf{F}_1, \ldots, \mathbf{F}_T\}$ are concatenated together as the input of the scale estimator. The output of the scale estimator is a soft scale mask $\bar{\mathbf{Q}} \in \mathbb{R}^{W \times H \times T}$ where the entry $\bar{q}_{xyt}$ indicates the scale of the pixel at position (x, y) (the object the pixel belongs to) is t. We denote the final segmentation map $\bar{\mathbf{M}}$ to be the weighted sum of score maps for all scales,

$$\bar{\mathbf{M}} = \sum_{t=1}^{T} \bar{M}_t \odot \bar{\mathbf{Q}}_t. \tag{1}$$

T is a number of discrete scales. The scale-adaptive branch produces the score map for scale t, denoted as $\bar{M}_t$. $\odot$ denotes element-wise multiplication. By dividing the scale space, each branch could handle the smaller variability of scale. At the same time, the scale mask selects out the finer segmentation regions of branches. The proposed method finally employs bilinear interpolation to upsample the segmentation map of the final layer to original image resolution. In this way, our network achieves a great performance.

C. Shared Network

FCNs [5] have proven successful in semantic image segmentation [36]–[38]. In this subsection, we briefly review the DeepLab [14] model, which is as a shared network in our method. DeepLab adopts the 16-layer architecture of the state-of-the-art classification network of [39] (i.e., VGG-16 net). The network is modified to be fully convolutional [5], producing dense feature maps. In particular, the last fully connected layers of the original VGG-16 net are turned into convolutional layers (e.g., the last layer has a spatial convolutional kernel with size 1x1). The spatial decimation factor of the original VGG-16 net is 32 due to the presence of multiple max-pooling and striding (downsampling). DeepLab reduces it to 8 by using the atrous (with holes) algorithm [40].

D. Scale-adaptive Branches

Each scale-adaptive branch consists of several convolutional layers, generating scale-induced segmentation map. Unlike the *skip net* architecture which generates multi-scale features by utilizing features from different level layers or feeding multi-scale input into FCNs. The T scale-adaptive branches share the same input feature which generated by the shared net. We observe that our scale-adaptive branches can learn much more abstract multi-scale structures under the supervision of scale-induced ground-truth map, even if they have the same network architecture and the same initialized weights.

E. Scale Estimator

The proposed scale estimator model takes the concatenated of T features map F as input and it consists of two convolutional layers: the first layer has 512 filters with kernel size 3x3 and the second layer has $(T + 1)$ filters with kernel size 1x1; then, they are passed through a SoftMax layer to generate soft probability map $\bar{\mathbf{H}} \in \mathbb{R}^{W \times H \times (T+1)}$, with $(T + 1)$ channels: background and T discrete scales. Note that the number of channels of $\bar{\mathbf{H}}$ are different from the number of branches. To make them match and reserve the background information simultaneously, we convert the soft probability map $\bar{\mathbf{H}}$ into soft scale mask $\bar{\mathbf{Q}}$,

$$\bar{\mathbf{Q}}_t = \frac{\bar{\mathbf{H}}_1}{T} + \bar{\mathbf{H}}_{t+1} \quad \text{for } t = 1, 2, ..., T \tag{2}$$

where $\bar{\mathbf{H}}_1$ is probability of background channel, T is the number of discrete scales.

F. Loss Settings

As illustrated in Fig. 1, it is a multi-task learning network. Rather than merely formulating the loss over the final segmentation map, we introduce two kinds of novel losses: three scale-induced segmentation losses and one scale estimation loss. Herein, we discuss that how these losses help guide our model to generate segmentation maps $\{\bar{\mathbf{M}}_1, \ldots, \bar{\mathbf{M}}_K\}$ and scale masks $\bar{\mathbf{Q}}$, which bring significant improvements on object parsing task.

Segmentation loss: Our segmentation loss function is the sum of cross-entropy terms for each spatial position in the CNN output map, it can be written as:

$$L_m = \ell(\mathbf{M}, \bar{\mathbf{M}}) \tag{3}$$

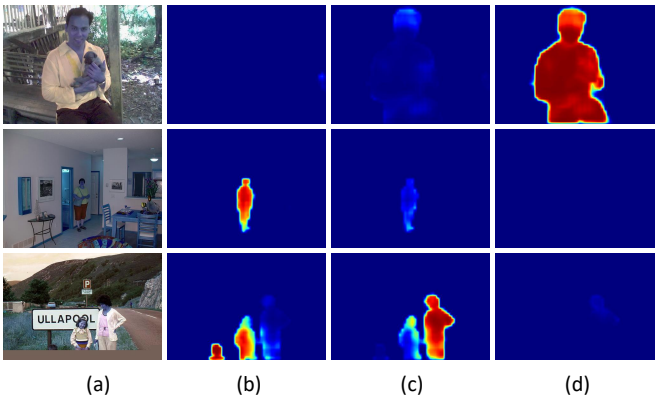


Fig. 4. Some examples of scale masks generated by our model and we have $T = 3$ scale-adaptive branches. (a) the input image. (b) the scale mask captures small-scale person. (c) the scale mask concentrates on middle-scale person. (d) the scale mask catches the large-scale person.

where ℓ is a cross-entropy loss. This loss guides the components of our network to generate the final segmentation map in cooperation.

Scale-induced segmentation loss: The scale-induced segmentation loss is formed over the estimated scale-adaptive segmentation map. Since the goal of each scale-adaptive branch is to segment out the objects of the corresponding scale, the loss is formed to penalize the difference of the estimation from the scale-induced ground-truth map (instead of the whole ground-truth map like deep supervision ([41]) and extra supervision ([2])). The losses are denoted as $\{\ell_1(M_1, \bar{M}_1), \ldots, \ell_T(M_T, \bar{M}_T)\}$, which are summed up,

$$L_{sm} = \sum_t \ell_t(M_t, \bar{M}_t) \tag{4}$$

$$\ell_t = -\frac{1}{N} \sum_{\langle i,j \rangle \in \mathcal{B}_t} \sum_{c=1}^{C} [p_{i,j}^c \log \hat{p}_{i,j}^c + (1 - p_{i,j}^c) \log(1 - \hat{p}_{i,j}^c)] \tag{5}$$

The terms $p_{i,j}^c$ and $\hat{p}_{i,j}^c$ are ground-truth and predicted probability which belongs to class c at position x, y, respectively. We consider only the pixels which lie in the bounding box $\mathcal{B}_t$ to keep a balance between background and foreground. The scale-induced segmentation loss directly guides the branches to learn features for different scales. By controlling the flow of information of different scales, this loss empowers the multiple branches to segment out the objects and parts in their corresponding scales.

Scale estimation loss: The scale loss is formed from the ground-truth scale,

$$L_s = \ell(\mathbf{Q}, \bar{\mathbf{Q}}) \tag{6}$$

the examples of scale mask are generated by our scale estimator as shown in Fig. 4. The scale estimator does a similar work with human segmentation but each classifier just responses to human with corresponding scale. The scale mask is used to select out high-quality results generated by scale-adaptive branches, and fuse them into a final segmentation map.

Overall loss: Combining the segmentation map estimation loss, scale-induced segmentation loss and scale estimation loss, the overall loss L can be written as:

$$L = L_m + L_{sm} + L_s. \tag{7}$$

We use stochastic gradient descent (SGD) algorithm with mini-batch to optimize the objective function mentioned above.

IV. Experiments

This section first describes our implementation details and experiment setup. Then, we analyze and evaluate the proposed network in various aspects. Extensive experiments are performed on public datasets such as Pascal-Person-Part dataset, Cow-Horse-Sheep dataset, LIP dataset, and Pascal VOC 2012..

Implementation details: Our scale-adaptive network is based on the publicly available deep learning models and has two forms: VGG-16 [39] based SAN and Resnet-101 [42] based SAN. We fine tune the model weights of the ImageNet [43] pre-trained VGG-16 and ResNet-101 networks to adapt them to the semantic segmentation task following the procedure of FCN [5]. **VGG-16 based SAN**: We replace the 1000-way ImageNet classifier in the last layer of VGG-16 with a classifier with targets of the same number of semantic classes of our task. Following [15] we remove the last two pooling layers and the convolutional filters in all subsequent layers were dilated by a factor of 2 for each pooling layer. We take the top 10 convolution layers (from conv1_1 to conv4_3) as a shared net. Each scale-induced branch net consists of 6 convolutional layers after conv4_3 layer. The proposed scale estimator takes the convolutional fc7 features as inputs. **Resnet-101 based SAN**: the modifications to Resnet-101 is similar to VGG-16 based model. We take the first 100 convolution layers as a shared net. Each scale-induced branch consists of 6 convolutional layers: the first layer has 1024 filters with kernel size 1x1; the second layer has 1024 filters with kernel size 3x3, dilation 12 to get large field of view; the third, fourth and fifth layer have 1024 filters with kernel size 3x3; the sixth layer has K (number of semantic classes of our task) filters with kernel size 1x1. The proposed scale estimator takes the output of the third layer in scale-induced branch net as input.

Training: The SGD with mini-batch is used for training. The initial learning rate is 0.001 (0.01 for the newly added convolution layer) and we employ a "poly" learning rate policy (the learning rate is multiplied by $1 - (\frac{iter}{max\_iter})^{power}$) with power = 0.9. We use the momentum of 0.9 and a weight decay of 0.0005. The training images are augmented by randomly scaling (from 0.5 to 2.0), then randomly cropping out the high-resolution patches (505×505) from the resulting images. We employ batch size = 1, 60K iterations for PASCAL-Person-Part dataset; batch size = 1, 12K iterations for Cow-Horse-Sheep dataset; batch size = 1, 300K iterations for LIP dataset.

Evaluation metric: The standard intersection over union (IOU) criterion and pixel-wise accuracy are adopted for evaluation on PASCAL-Person-Part dataset, Cow-Horse-Sheep dataset, LIP dataset, and Pascal VOC 2012.

TABLE I
PART PARSING ACCURACY (%) ON PASCAL-PERSON-PART IN TERMS OF MEAN IOU. WE COMPARE OUR TWO SAN MODELS WITH OTHER STATE-OF-THE-ART METHODS.

Method	bg	head	torso	u-arms	l-arms	u-legs	l-legs	mIOU
DeepLab-LargeFOV-CRF [14]	93.52	80.13	55.56	36.43	38.72	35.50	30.82	52.95
DeepLab-MS-LargeFOV [14]	93.64	79.55	57.96	40.21	39.14	36.37	33.04	54.27
Multi-Scale Averaging [2]	93.43	79.89	57.40	40.57	41.14	37.66	34.31	54.91
Multi-Scale Attention [2]	93.65	81.47	59.06	44.15	42.50	38.28	35.62	56.39
HAZN [19]	93.78	80.76	60.50	45.65	43.11	41.21	37.74	57.54
LG-LSTM [45]	88.63	82.72	60.99	45.40	45.40	42.33	37.96	57.97
Part-Net [46]	94.12	81.92	60.24	46.32	45.07	43.38	38.46	58.50
Graph-LSTM [31]	94.59	82.69	62.68	46.88	47.71	45.66	40.93	60.16
Attention + SSL [24]	94.68	83.26	62.40	47.80	45.58	42.32	39.48	59.36
Deeplabv2 [20]	-	-	-	-	-	-	-	64.94
SAN(VGG-16)	94.12	83.17	63.43	50.42	50.10	42.21	39.36	**60.40**
SAN(Resnet-101)	96.01	86.12	73.49	59.20	56.20	51.39	49.58	**67.42**

TABLE II
COMPARISON WITH OTHER STATE-OF-ART METHODS ON PASCAL-PERSON-PART DATASET. EMPLOYING VGG-16 AND RESNET-101 FOR SCALE-ADAPTIVE NETWORK ON PASCAL-PERSON-PART DATASET. **AUG**: DATA AUGMENTATION BY RANDOMLY RESCALING INPUTS AND RANDOMLY MIRROR FLIPS. $\mathbf{L_s}$: ADDING THE SCALE ESTIMATOR. $\mathbf{L_{sm}}$: ADDING SCALE-INDUCED SUPERVISION ON BRANCHES. **COCO**:MODELS PRETRAINED ON MS-COCO. **CRF**: USING FULLY-CONNECTED CONDITIONAL RANDOM FIELD (CRF) [30] AS POST-PROCESSING STEP

Method	Aug	$\mathbf{L_s}$	$\mathbf{L_{sm}}$	COCO	CRF	mIOU
VGG-16 based						
Baseline	√					53.16
SAN	√	√				58.50
SAN	√		√			59.04
SAN	√	√	√			59.89
SAN	√	√	√		√	60.40
ResNet-101 based						
Baseline	√					60.57
SAN	√	√				63.53
SAN	√		√			65.67
SAN	√	√	√			65.96
SAN	√	√	√	√		66.73
SAN	√	√	√	√	√	67.42

Reproducibility: The proposed scale-adaptive network is implemented by extending the Caffe [44] framework. All networks are trained on a single NVIDIA GeForce GTX TITAN X GPU with 12GB memory. The source code is available at https://github.com/speedinghzl/Scale-Adaptive-Network.

A. PASCAL-Person-Part

Dataset: We conduct experiments on human part parsing using the PASCAL-Person-Part ([23]) dataset which is a subset of the PASCAL VOC 2010 dataset. Specifically, the dataset contains detailed part annotations for every person, including eyes, mouse, etc. We merge the annotations into the background and six person part categories: Head, Torso, Upper/Lower Arms, and Upper/Lower Legs. We only use those images containing persons for training (1716 images) and validation (1817 images).

Comparison with state-of-the-arts: As shown in Table I, we compare the performance of our SAN with previous approaches based on two different shared nets. On the Pascal-Person-Part test dataset, it achieves the highest mean intersection-over-union score. The denseCRF [30] method is used as a post-processing step only on Pascal-Person-Part test dataset for fair comparison.

We provide these results of other approaches for reference but it should be emphasized that their results should not be simply compared with our method, because these methods are trained on different (and larger) training sets or different basic network. Deeplabv2 [20] utilizes Resnet-101 as basic network and is pretrained on the MS-COCO [47] dataset, and other methods make use of VGG-16 as basic network and without using additional datasets. For a fair comparison, we take VGG-16 and Resnet-101 as our shared nets and build two models: SAN(VGG-16) and SAN(Resnet-101), meanwhile, they are trained in the same setup, i.e. pre-train SAN(Resnet-101) on MS-COCO dataset is identical to that of Deeplabv2.

It is important to note that the first four baselines which represent three different approaches to handle the variation of object scale. DeepLab-MS-LargeFOV [14] employs *skip net* architecture which adds a post-processing step to DeepLab-LargeFOV by the means of a fully-connected Conditional Random Field (CRF)[30]. Multi-Scale Attention [2] which feeds the DeepLab-LargeFOV model with images resized to three fixed scales (0.5, 1.0 and 1.5) and then takes a scale attention model to handle the scale variations in object parsing. Attention + SSL [24] imposes human pose supervision into Attention method [2]. HAZN [19] employs detection-segmentation cascade network, once an object is detected, the scale of the object is obtained, then zooms image regions into their proper scales to refine the parsing. Our SAN(VGG-16) model surpasses these methods and achieves a better result, significantly improving the segmentation accuracy in all parts. In addition, Deeplabv2 utilizes Resnet-101 as basic network and employs multi-scale input policy like Multi-Scale Attention. Our SAN (Resnet-101) model also surpasses this

TABLE III
THE DIFFERENT SETTINGS OF SUPERVISION FOR BRANCHES.

Supervision	FOV[14] of branches	mean IOU
full	{112,224,336}	57.81
	{224,224,224}	56.92
scale-induced	{112,224,336}	58.53
	{224,224,224}	59.04

method and achieves a better result.

LG-LSTM [45] and Graph-LSTM [31] both model the spatial relations on superpixels for semantic object parsing. Part-Net [46] adopted encoder-decoder framework to parse images. Our method still achieved better results.

The effect of L_s: We report the results in Table II. The baseline network consists of a shared net and a branch which has similar architectures to [14]. We find that the proposed scale estimator with loss brings 5.4% and 2.9% improvements in VGG-16 based model and Resnet-101 based model. The L_s guides the model to estimate the scale of object used for the fusion of the scale-induced segmentation maps. Meanwhile, it indirectly controls the information of different scales that flow into the branches respectively in back propagation process.

The effect of L_{sm}: As shown in Table II, the scale-induced segmentation map estimation loss can bring 5.88% and 5.1% improvements in VGG-16 based model and Resnet-101 based model. Based on L_s effect, the scale-induced segmentation loss still brings 1.4% and 2.4% improvements in VGG-16 based model and Resnet-101 based model. It directly guides the branches to learn features with the different scales and brings more obvious improvements. We think L_{sm} and L_s have the same effect to controls the flow of information of different scales and guides the branches to learn multi-scale features.

In order to further prove the effect of L_{sm}, we conduct an extra experiment with different settings of supervision for branches. In Table III, the **full** supervision and **scale-induced** supervision denote using the whole ground-truth map and scale-induced ground-truth map respectively. To avoid interference, we remove the L_s loss and sum of all branches output as a final segmentation map. When the three branches have the same Field-of-View(FOV) [14] and the same initialization, which uses scale-induced supervision, obtains the performance of 59.04% mean IOU, which is 2.12% better than full supervision. While the three branches have the different Field-of-View with a prior, the one that uses scale-induced supervision obtains better performance by 0.82% than full supervision. At the same time, we noticed an interesting phenomenon that under scale-induced supervision, branches that adopt the same Field-of-View obtain better performance than the different Field-of-View with a prior. But under full supervision, the conclusion is opposite. This is because the scale-adaptive branches can learn multi-scale structure under the supervision of scale-induced ground-truth map, even if they have the same network architecture and the same initialization. It may not match the real scale distribution when we set branches to different Field-of-View with a prior. Under full supervision, the prior does work.

TABLE IV
PART PARSING ACCURACY W.R.T. SIZE OF HUMAN INSTANCE (%) ON PASCALPERSON-PART IN TERMS OF MEAN IOU.

Method	Size XS	Size S	Size M	Size L
DeepLab-LargeFOV	32.5	44.5	50.7	50.9
DeepLab-LargeFOV-CRF	31.5	44.6	51.5	52.5
Multi-Scale Averaging	33.7	45.9	52.5	54.7
Multi-Scale Attention	37.6	49.8	55.1	55.5
HAZN	**47.1**	55.3	56.8	56.0
SAN(VGG-16)	42.5	**55.7**	**58.9**	**57.3**

TABLE V
PART PARSING ACCURACY W.R.T. DEGREES OF DIVERSITY IN OBJECT SIZES (%) ON PASCAL-PERSON-PART IN TERMS OF MEAN IOU.

Method	uniform	diverse	diff
DeepLab-LargeFOV	53.6	50.3	3.3
Multi-Scale Attention	56.2	55.0	1.2
SAN(VGG-16)	**59.7**	**58.9**	**0.8**

Part parsing accuracy w.r.t. size of human instance: It is necessary to check the performance of our model with respect to the change of human size in images. Following [19], we categorize all the ground truth human instances into four different sizes according to the bounding box area of each instance a_b (the square root of the bounding box area). The four sizes are defined as follows: (1) Size XS: $a_b \in [0, 80]$; (2) Size S: $a_b \in [80, 140]$; (3) Size M: $a_b \in [140, 220]$; (4) Size L: $a_b \in [220, 520]$. Then we calculate the mean IOU (within the bounding box) for each of these four scales. The results are given in Table IV, the baseline DeepLab-LargeFOV performs badly at size S or M, while our SAN model improves significantly by 11.1% for size S and 7.4% for size M. It shows that SAN is particularly good for the object with various scale. It is noteworthy that the way to split scale space is different from the settings to generate the scale-induced ground-truth map. As shown in the Fig. 3, we categorize all the ground truth human instances into three different sizes, setting θ_1, θ_2 to 112,224 respectively.

Part parsing accuracy w.r.t. degrees of diversity in object sizes: First of all, we quantize the objects into a $T = 3$ discrete scales. Then, we use Shannon's diversity index to measure the degree of diversity for each image. According to the degree which ranges from 0 to $\ln 3$, images are categorized into 2 groups: uniform ($[0, 0.5]$), diverse ($(0.5, ln3]$). Table V shows mean IOU on Pascal-Person-Part dataset. The results show that the proposed method improves the performance both on the uniform and diverse images. Meanwhile, our method can reduce the performance difference between uniform and diverse images.

How to choose the branches & #branches: The branches should be deep and have big Field-of-View(FOV) [14] to capture the structure of the whole object with the different scale, which contributes to the higher accuracy and finer part segmentation result. Meanwhile, it's important to choose appropriate #branches to improve performance. From Table VI

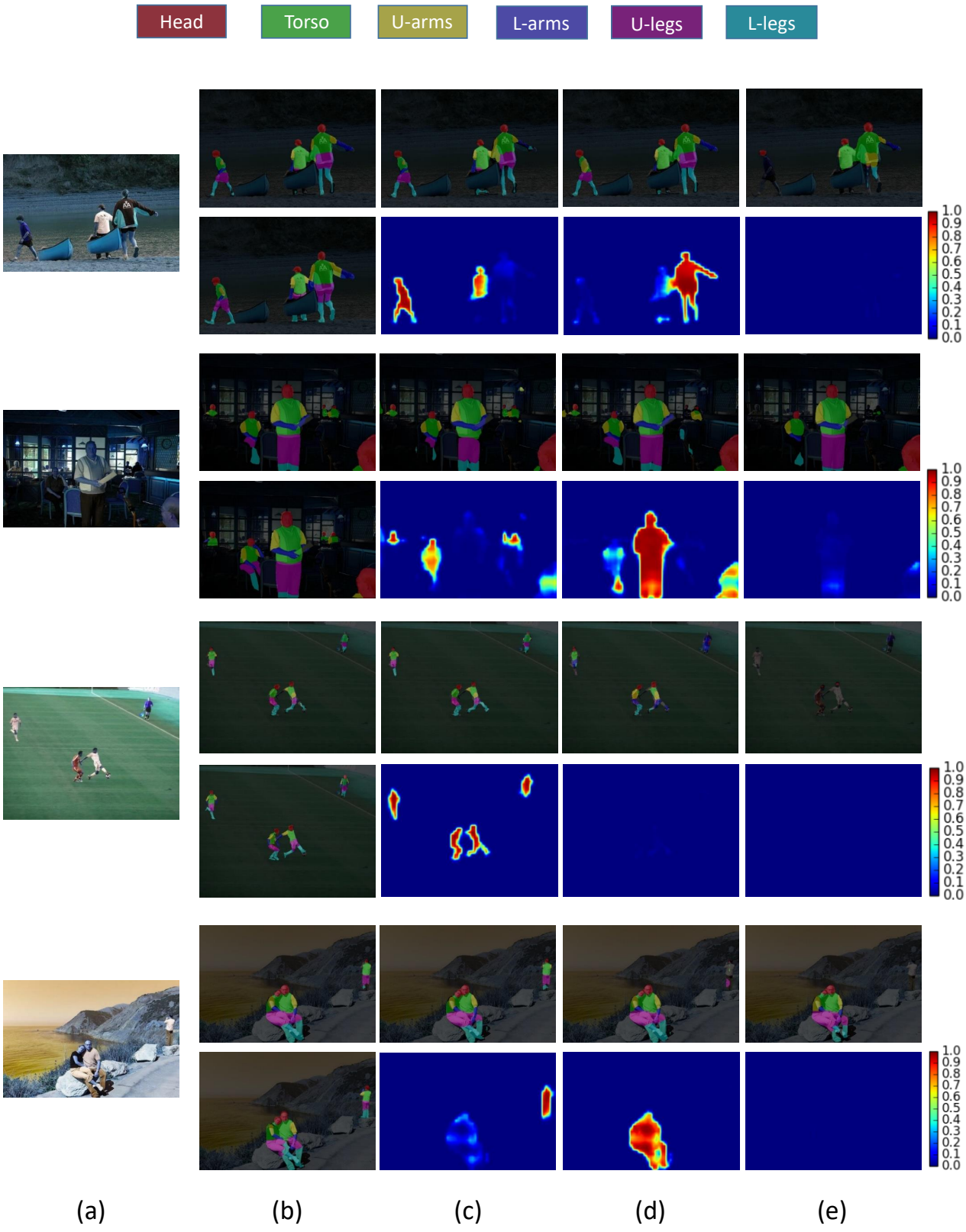


Fig. 5. Examples of semantic human parsing results and intermediate results by the proposed scale-adaptive network model. (a) the input image. (b) the human parsing result and ground truth. (c)(d)(e) the scale-induced segmentation map and scale mask for objects with small, middle and large scale.

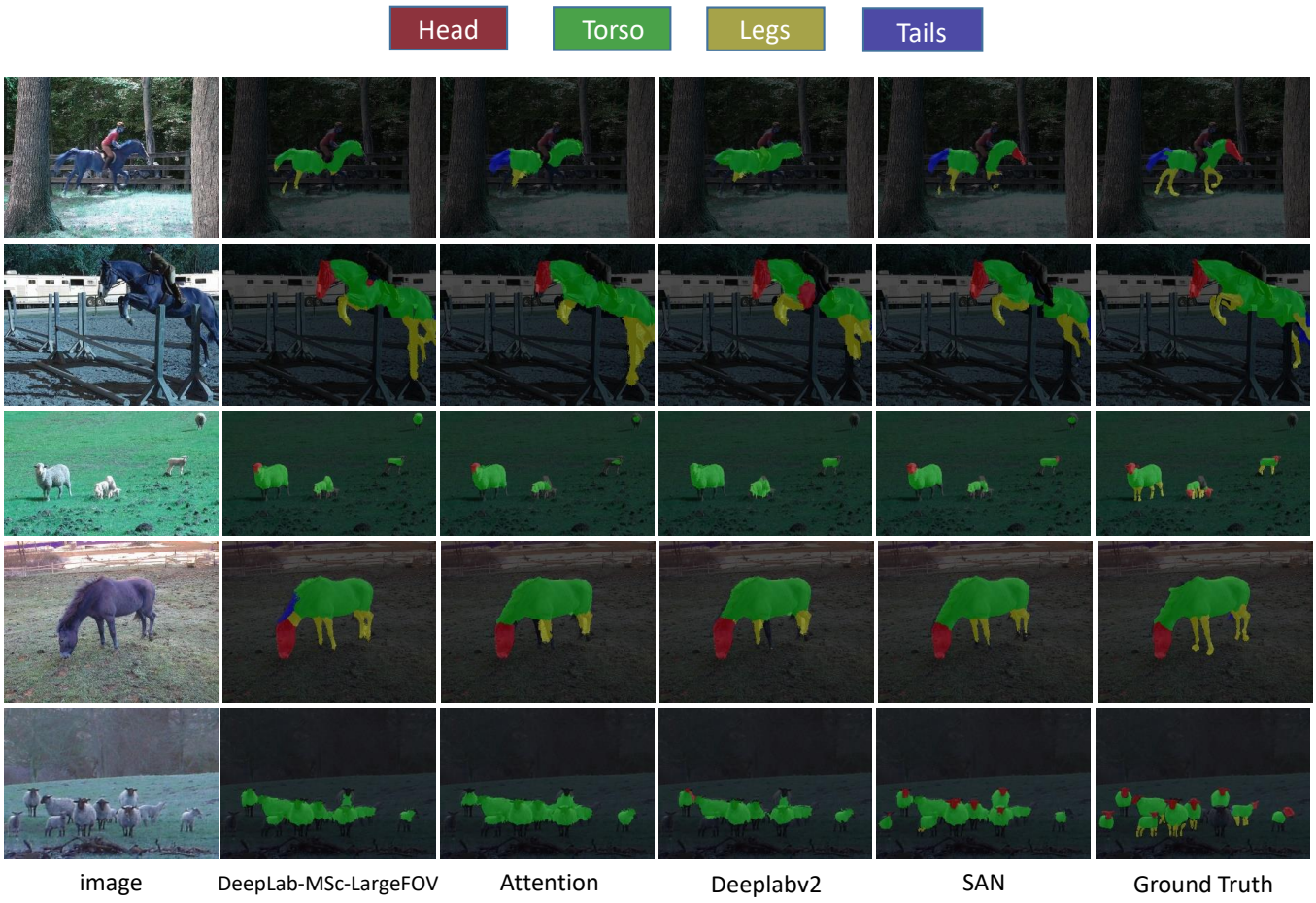


Fig. 6. Qualitative comparison on the Cow-Horse-Sheep Dataset.

TABLE VI
THE DIFFERENT SETTINGS OF BRANCH OF SAN(VGG-16).

Depth	FOV of branches	Mean IOU
3	{224,224,224}	57.21
6	{224,224,224}	59.89
6	{112,112,112}	56.74
6	{336,336,336}	59.9
6	{112,224,336}	59.43

TABLE VII
THE DIFFERENT NUMBER OF BRANCH OF SAN(VGG-16).

Number of scales	1	2	3	4
mIOU	53.16	58.9	59.89	59.5

TABLE VIII
THE DIFFERENT WAY TO SPLIT SCALE SPACE.

Method	mIOU
equal instance	59.89
equal pixels	53.23
clustering	56.77

and Table VII, we have explored different settings of scale-adaptive branches when training SAN on Pascal-Person-Part dataset.

- The depth of branch increases from 3 to 6, bringing about 2.6% improvement.
- Setting Field-of-View to be 224 is large enough for the branch to get context information. We try to set different FOV for branches corresponding to the different scale, but it does not bring any improvement. In fact, the scale-induce information guides the kernels of the branch to learn structure with the different scales even if the FOV of branches are the same.
- The performance increases along with the number of scales from 1 to 3, because the variance of scale subspace decreases. When the number of scales increases from 3 to 4, the performance will have a slight drop, which is due to the diversity diminution of the training samples. Finally, we select 3 for the number of scales on Pascal-Person-Part dataset.

After determining the number of scales, we have tried three methods to split scale space. 1. Sorting the objects by the square root of the bounding box area, then splitting the list into 3 sublists. Each sublist has the same amount of object instances. This method is donated as **equal instances** 2. Sorting the objects by the square root of the bounding box area, then splitting the list into 3 sublists. Each sublist has

TABLE IX
PERFORMANCE COMPARISON IN TERMS OF PER-CLASS IOU WITH FOUR STATE-OF-THE-ART METHODS ON LIP VALIDATION SET.

Method	hat	hair	gloves	sunglasses	u-clothes	dress	coat	socks	pants	jumpsuits	scarf	skirt	face	l-arm	r-arm	l-leg	r-leg	l-shoe	r-shoe	Bkg	Avg
SegNet [16]	26.60	44.01	0.01	0.00	34.46	0.00	15.97	3.59	33.56	0.01	0.00	0.00	52.38	15.30	24.23	13.82	13.17	9.26	6.47	70.62	18.17
FCN-8s [5]	39.78	58.96	5.32	3.08	49.08	12.36	26.82	15.66	49.41	6.48	0.00	2.16	62.65	29.78	36.63	28.12	26.05	17.76	17.70	78.02	28.29
DeepLabV2 [20]	57.94	66.11	28.50	18.40	60.94	23.17	47.03	34.51	64.00	22.38	14.29	18.74	69.70	49.44	51.66	37.49	34.60	28.22	22.41	83.25	41.64
Multi-Scale Attention [2]	58.87	66.78	23.32	19.48	63.20	29.63	49.70	35.23	66.04	24.73	12.84	20.41	70.58	50.17	54.03	38.35	37.70	26.20	27.09	84.00	42.92
Attention+SSL [24]	59.75	67.25	28.95	21.57	65.30	29.49	51.92	38.52	68.02	24.48	14.92	24.32	71.01	52.64	55.79	40.23	38.80	28.08	29.03	84.56	44.73
SAN(VGG-16)	59.87	66.81	27.65	22.42	65.38	29.17	53.42	36.51	69.10	26.85	16.46	25.00	68.73	54.32	55.43	38.71	36.01	29.19	30.36	84.91	**44.81**

the same amount of object pixels. This method is donated as **equal pixels**. 3. Using k-means clustering method to split the scale space into 3 subspace. This method is donated as **clustering**. The Table VIII shows mean IOU on Pascal-Person-Part dataset. The **equal instances** surpasses the other methods.

Qualitative results: We visually show several example results from the PASCAL-Person-Part dataset in Fig. 5. We can observe that our model can capture the scale information of the object and use the scale masks to select out the finer segmentation result from scale-induced segmentation maps to generate the final segmentation result.

B. Look into Person

Look into Person(LIP) [24] is a large-scale dataset focusing on semantic understanding of human bodies which has several appealing properties. The images in the LIP dataset are cropped person instances from COCO [47] training and validation sets. And, LIP is annotated with elaborated pixel-wise annotations with 19 semantic human part labels and one background label. In total, the dataset consists of 30,462 training and 10,000 validation images with publicly available annotations.

Comparison with state-of-the-arts: We report the results and the comparisons with five state-of-the-art methods on LIP validation set in Tab IX. The proposed architecture can give a huge boost in average IoU: 3.17% better than DeepLabV2 [20] and 1.89% better than Multi-Scale Attention [2]. This superior performance achieved by our method demonstrates the effectiveness of our proposed method. FCN-8s [5] was the first one to adopted fully Convolution network for semantic segmentation. SegNet [16] adopted encoder-decoder framework to parse images. Attention + SSL [24] imposes human pose supervision into Multi-Scale Attention method [2]. For fair comparison, the denseCRF [30] is not used as a post-processing step here. Our SAN also surpasses these methods and achieves a better result.

C. Cow-Horse-Sheep

Dataset: To show the generality of our method to object part parsing, we conduct experiments on animal part parsing by selecting 953 images containing cow, horse or sheep from PASCAL-Part [23] dataset. Like person annotation, the dataset contains detailed part annotations for cow, horse, and sheep, including eyes, nose, etc. We merge the annotations into the background and four animal part classes: Head, Torso, Legs, and Tail. We use 634 images for training and 319 images for testing. The denseCRF [30] method is not used as a post-processing step here for fair comparison.

Comparison with state-of-the-arts: For other methods, We conduct experiments on Cow-Horse-Sheep using the open source code provided by authors and the evaluation results are given in Table X. All the experiments are conducted under the same conditions. It shows that the DeepLab-LargeFOV-CRF [14] has already achieved competitive results, while our SAN model also provides a roughly 5.0% mIOU improvement for animal part. It is noticeable for small parts, e.g. the improvement of segmenting horse/cow/sheep tails is more than 10%. It shows that our method can be effectively generalized to other objects for part parsing.

Qualitative results: We also provide qualitative evaluations in Fig. 6, comparing our SAN model with four state-of-the-art methods. It's observed that our model has a good performance on the small objects or small parts such as legs and tails. Meanwhile, our model can obtain finer boundary of all parts.

D. General object segmentation on PASCAL VOC 2012

We apply our approach to general object segmentation. There are large variabilities of object scale, position and pose in PASCAL VOC dataset. In Table XI, we report the results on PASCAL VOC [48] 2012 validation set. The denseCRF [30] method is not used as a post-processing step here for fair comparison.

Effectivity: Compared with the baseline DeepLab-LargeFOV, our approach still brings about 5% improvement on PASCAL VOC 2012 test dataset. The performance improvement comes from the ability to hand the large variability of object scale.

Faster: Although Multi-Scale Attention [2] achieves better performance than our approach, SAN with a frame rate of 8fps (including all steps) on a single GPU, is faster than Multi-Scale Averaging and Multi-Scale Attention.

Limitation: There is a limitation to our approach. Splitting the objects into different scale spaces will lose the contextual information among objects with the different scale in the same image. But this will not happen in object parsing because the scale of all parts of the object is the same.

TABLE X
MEAN IOU (MIOU) OVER THE COW-HORSE-SHEEP DATASET.

Method	bg	head	torso	leg	tail	mIOU
DeepLab-LargeFOV-CRF [14]	93.96	66.06	69.03	41.63	30.51	60.24
DeepLab-MS-LargeFOV [14]	94.64	67.57	70.02	46.44	26.53	61.04
Multi-Scale Attention [2]	95.08	**71.04**	70.43	46.23	36.59	63.87
Deeplab-ASPP [20]	94.44	66.82	69.94	44.85	33.49	61.90
SAN(VGG-16)	**95.19**	69.47	**71.08**	**48.78**	**40.71**	**65.04**

TABLE XI
MEAN IOU (MIOU) OVER THE PASCAL VOC2012 VALIDATION/TEST SET.

Method	val	test	Rate
DeepLab-LargeFOV [14]	62.25	65.1	12 fps
DeepLab-MSc-LargeFOV [14]	64.21	67.0	9 fps
Multi-Scale Averaging [2]	67.98	70.5	5 fps
Multi-Scale Attention [2]	69.08	71.5	5 fps
SAN(VGG-16)	68.30	70.3	8 fps

V. CONCLUSION AND FUTURE WORK

We propose a Scale-adaptive Network to parse objects in natural images and demonstrate that our approach outperforms previous state-of-the-art methods under the same experimental conditions. We also identify the effectiveness of embedding scale information into DCNNs. Our experiments show that handling the variability of object scale can dramatically improve the performance of object part segmentation/parsing. As for the future work, we will shorten the test time and take spatial relations among parts into account.

VI. ACKNOWLEDGEMENTS

This work was supported by NSFC (No. 61733007, No. 61876212 and No. 61572207), China Scholarship Council, Hubei Scientific and Technical Innovation Key Project and National Key R&D Program of China (No. 2018YFB1402600).

REFERENCES

[1] L. Wang, G. Hua, J. Xue, Z. Gao, and N. Zheng, "Joint segmentation and recognition of categorized objects from noisy web image collection," *IEEE Transactions on Image Processing*, vol. 23, no. 9, pp. 4070–4086, 2014.

[2] 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 Recogn.*, 2016, pp. 3640–3649.

[3] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," *arXiv preprint arXiv:1703.08448*, 2017.

[4] L. Ran, Y. Zhang, and G. Hua, "Cannet: Context aware nonlocal convolutional networks for semantic image segmentation," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4669–4673.

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

[6] X. Dong, J. Shen, L. Shao, and L. Van Gool, "Sub-markov random walk for image segmentation," *IEEE Transactions on Image Processing*, vol. 25, no. 2, pp. 516–527, 2016.

[7] T. Ruan, T. Liu, Z. Huang, Y. Wei, S. Wei, and Y. Zhao, "Devil in the details: Towards accurate single and multiple human parsing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 4814–4821.

[8] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," 2019.

[9] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *arXiv preprint arXiv:1908.07919*, 2019.

[10] A. Bhuiyan, A. Perina, and V. Murino, "Person re-identification by discriminatively selecting parts and features," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 147–161.

[11] P. Yu, J. Wang, and Y. Wu, "Human action segmentation using 3d fully convolutional network," in *BMVC*, 2017.

[12] X. Liang, L. Lin, W. Yang, P. Luo, J. Huang, and S. Yan, "Clothes co-parsing via joint image segmentation and labeling with application to clothing retrieval," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1175–1186, 2016.

[13] J. Dong, Q. Chen, X. Shen, J. Yang, and S. Yan, "Towards unified human parsing and pose estimation," in *in Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2014, pp. 843–850.

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

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

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

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

[18] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1520–1528.

[19] F. Xia, P. Wang, L.-C. Chen, and A. L. Yuille, "Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 648–663.

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

[21] R. Zhang, S. Tang, Y. Zhang, J. Li, and S. Yan, "Scale-adaptive convolutions for scene parsing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2031–2039.

[22] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," *arXiv preprint arXiv:1703.06211*, 2017.

[23] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille, "Detect what you can: Detecting and representing objects using holistic models and body parts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2014, pp. 1971–1978.

[24] K. Gong, X. Liang, X. Shen, and L. Lin, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," *arXiv preprint arXiv:1703.05446*, 2017.

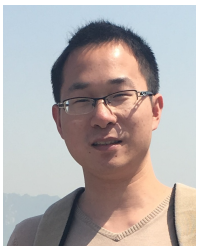
[25] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *in Proc. IEEE Conf. Comput. Vis. Pattern Recogn.* IEEE, 2012, pp. 3570–3577.

[26] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: a database and web-based tool for image annotation," *International journal of computer vision*, vol. 77, no. 1-3, pp. 157–173, 2008.

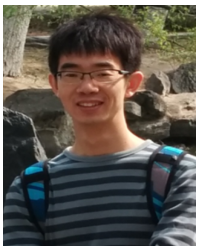
[27] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," *arXiv preprint arXiv:1612.03144*, 2016.

[28] S. Eslami and C. Williams, "A generative model for parts-based object segmentation," in *Adv. Neural Inf. Process. Syst.*, 2012, pp. 100–107.

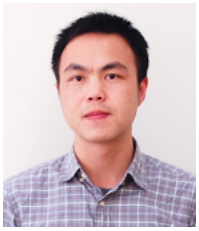
[29] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille, "Joint object and part segmentation using deep learned potentials," in *Int. J. Comput. Vis.*, 2015, pp. 1573–1581.
[30] V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," *Adv. Neural Inf. Process. Syst*, vol. 2, no. 3, p. 4, 2011.
[31] X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan, "Semantic object parsing with graph lstm," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 125–143.
[32] Y. Bo and C. C. Fowlkes, "Shape-based pedestrian parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.* IEEE, 2011, pp. 2265–2272.
[33] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, 2013.
[34] D. Ciregan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.* IEEE, 2012, pp. 3642–3649.
[35] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, 2010.
[36] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1529–1537.
[37] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *arXiv preprint arXiv:1612.01105*, 2016.
[38] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks with identity mappings for high-resolution semantic segmentation," *arXiv preprint arXiv:1611.06612*, 2016.
[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[40] S. Mallat, *A wavelet tour of signal processing*. Academic press, 1999.
[41] C.-Y. Lee, S. Xie, P. W. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets." in *AISTATS*, vol. 2, no. 3, 2015, p. 5.
[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2016, pp. 770–778.
[43] 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 Recogn.* IEEE, 2009, pp. 248–255.
[44] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.
[45] X. Liang, X. Shen, D. Xiang, J. Feng, L. Lin, and S. Yan, "Semantic object parsing with local-global long short-term memory," *arXiv preprint arXiv:1511.04510*, 2015.
[46] G. L. Oliveira, C. Bollen, W. Burgard, and T. Brox, "Efficient and robust deep networks for semantic segmentation," *The International Journal of Robotics Research*, p. 0278364917710542, 2017.
[47] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *in Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 740–755.
[48] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2015.



Zilong Huang is a Ph.D. student in the School of Electronics Information and Communications, Huazhong University of Science and Technology (HUST). He received his B.S. degree from HUST in 2015. His research interests include computer vision and machine learning. In particular, he focuses on semantic segmentation and object parsing.

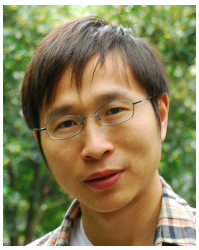


Chunyu Wang is a Researcher in Microsoft Research Asia. He received his Ph.D in computer science from Peking University in 2016. His research interests are in computer vision, artificial intelligence and machine learning.



Xinggang Wang is an Associate Professor in the School of Electronic Information and Communications in Huazhong University of Science and Technology (HUST). His research interests are computer vision, deep learning and machine learning. He received his B.S. degree in communication and information system and Ph.D. degree in computer vision both from HUST. From May 2010 to July 2011, he was with the Department of Computer and Information Science, Temple University, Philadelphia, PA., as a visiting scholar. From February 2013 to September 2013, he was with the University of California, Los Angeles, as a visiting graduate researcher. He is a reviewer of IEEE Transaction on Cybernetics, pattern recognition, computer vision and image understanding, n, CVPR, ICCV and ECCV etc. His research interests include computer vision and machine learning.

Wenyu Liu received the B.S. degree in Computer Science from Tsinghua University, Beijing, China, in 1986, and the M.S. and Ph.D. degrees, both in Electronics and Information Engineering, from Huazhong University of Science & Technology (HUST), Wuhan, China, in 1991 and 2001, respectively. He is now a professor and associate dean of the School of Electronic Information and Communications, HUST. His current research areas include computer vision, multimedia, and machine learning. He is a senior member of IEEE.



Jingdong Wang is a Senior Researcher at the Visual Computing Group, Microsoft Research Asia. He received the M.Eng. and B.Eng. degrees in Automation from the Department of Automation, Tsinghua University, Beijing, China, in 2001 and 2004, respectively, and the PhD degree in Computer Science from the Department of Computer Science and Engineering, the Hong Kong University of Science and Technology, Hong Kong, in 2007. His areas of interest include computer vision, machine learning, pattern recognition, and multimedia computing. In particular, he has worked on kernel methods, semi-supervised learning, data clustering, image segmentation, and image and video presentation, management search. At present, he is mainly working on the CNN architecture design, large-scale indexing, human understanding, and person re-identification.