

# Decoupled Spatial Neural Attention for Weakly Supervised Semantic Segmentation

Tianyi Zhang, Guosheng Lin, Jianfei Cai, Tong Shen, Chunhua Shen, Alex C. Kot,

**Abstract**—Weakly supervised semantic segmentation receives much research attention since it alleviates the need to obtain a large amount of dense pixel-wise ground-truth annotations for the training images. Compared with other forms of weak supervision, image labels are quite efficient to obtain. In our work, we focus on the weakly supervised semantic segmentation with image label annotations. Recent progress for this task has been largely dependent on the quality of generated pseudo-annotations. In this work, inspired by spatial neural-attention for image captioning, we propose a decoupled spatial neural attention network for generating pseudo-annotations. Our decoupled attention structure could simultaneously identify the object regions and localize the discriminative parts which generates high-quality pseudo-annotations in one forward path. The generated pseudo-annotations lead to the segmentation results which achieve the state-of-the-art in weakly-supervised semantic segmentation.

**Index Terms**—Semantic Segmentation, Deep Convolutional Neural Network (DCNN), Weakly-Supervised Learning

## I. INTRODUCTION

Semantic segmentation is a task to assign semantic label to every pixel within an image. In recent years, Deep convolutional neural networks (DCNNs) [1]–[3] have brought great improvement in semantic segmentation performance. Training DCNNs in a fully-supervised setting with pixel-wise ground-truth annotation achieves state-of-the-art semantic segmentation accuracy. However, the main limitation of such fully-supervised setting is that it is labor-intensive to obtain a large amount of accurate pixel-level annotations for training images. On the other hand, datasets with only image-level annotations are much easier to obtain. Therefore, weakly-supervised semantic segmentation supervised only with image labels has received much attention.

The performance of weakly supervised semantic segmentation with image-level annotation has been remarkably improved by introducing efficient localization cues [4]–[6]. The most widely used pipeline in weakly supervised semantic segmentation is to first estimate pseudo-annotations for the training images based on localization cues and then utilize the pseudo-annotations as the ground-truth to train the segmentation DCNNs. Clearly the quality of pseudo-annotations

T. Zhang is with the Interdisciplinary Graduate School, Nanyang Technological University (NTU), Singapore 639798 (e-mail: zh0023yi@e.ntu.edu.sg).

G. Lin and J. Cai are with the School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore 639798 (e-mail: gslin@ntu.edu.sg; asjfcai@ntu.edu.sg).

S. Tong and C. Shen are with School of Computer Science, The University of Adelaide, Adelaide, SA 5005, Australia (e-mail: tong.shen@adelaide.edu.au; chunhua.shen@adelaide.edu.au).

A. C. Kot is with the School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore 639798 (e-mail: eackot@ntu.edu.sg).

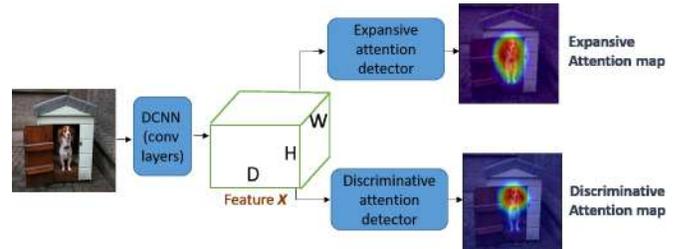


Fig. 1: The brief introduction of our proposed decoupled attention structure. It simultaneously generates two attention maps, namely Expansive attention map and Discriminative attention map. The Expansive attention map is to identify the object regions while the Discriminative attention map is to mine the discriminative parts.

directly affects the final segmentation performance. In our work, we follow the same pipeline and mainly focus on the first step, which is to generate high-quality pseudo-annotations for the training images with only image-level labels. In recent years, top-down neural saliency [7]–[9] performs well in weakly-supervised localization tasks and consequently has been widely applied in generating pseudo-annotations for semantic segmentation supervised with image-level labels. However, as is pointed out by previous works [6], such top-down neural saliency is good at identifying the most discriminative regions of the objects instead of the whole extent of the objects. Thus the pseudo-annotations generated by these methods are far from the ground-truth annotations. To alleviate this problem, some works consist of multiple ad-hoc processing steps (e.g., iterative training), which are difficult to implement. Some works introduce external information (e.g., web data) to guide the supervision, which greatly increase data and computation load. On the contrary, our work proposes a principal pipeline which is simple and effective to implement.

Our aim is to generate pseudo-annotations for weakly supervised semantic segmentation efficiently and effectively. Inspired by the spatial neural attention mechanism which has been widely used in VQA [10] and image captioning [11], we introduce spatial neural attention into our pseudo-annotation generation pipeline and propose a decoupled spatial neural attention structure which simultaneously localizes the discriminative parts and estimates object regions in one end-to-end framework. Such structure helps to generate effective pseudo-annotations in one forward pass. The brief description of our decoupled attention structure is illustrated in Fig. 1.

Our major contributions can be summarized as follows:

- We introduce spatial neural attention and propose a decoupled attention structure for generating pseudo-annotations for weakly supervised semantic segmentation.
- Our decoupled attention model outputs two attention maps which focus on identifying object regions and mining the discriminative parts respectively. These two attention maps are complimentary to each other to generate high-quality pseudo-annotations.
- We employ a simple and effective pipeline without heuristic multi-step iterative training steps, which is different from most existing methods of weakly-supervised semantic segmentation.
- We perform detailed ablation experiments to verify the effectiveness of our decoupled attention structure. We achieve state-of-the-art weakly supervised semantic segmentation results on Pascal VOC 2012 image segmentation benchmark.

## II. RELATED WORK

### A. Weakly Supervised Semantic Segmentation

In recent years, the performance of semantic segmentation has been greatly improved with the help of Deep Convolutional Neural Networks (DCNNs) [1], [3], [12]–[18]. Training DCNNs for semantic segmentation in a fully-supervised pipeline requires pixel-wise ground-truth annotations, which is very time-consuming to obtain.

Thus weakly-supervised semantic segmentation receives research attention to alleviate the workload of pixel-wise annotation for the training data. Among the weakly-supervised settings, image-level labels are the easiest annotations to obtain. As for the semantic segmentation with image-level labels, some early works [19], [20] tackle this problem as multiple instance learning (MIL) problem which views each image being positive if at least one pixel/superpixel within it is positive, and negative if all of the pixels are negative. Other early works [21] apply Expectation-Maximization (EM) procedure, which alternates between predicting pixel labels and optimizing DCNNs parameters. However, due to the lack of effective location cues, the performance of early works is not satisfactory enough.

The performance of semantic segmentation with image-level labels was significantly improved after introducing location information to generate localization seeds/pseudo-annotations for segmentation DCNNs. The quality of pseudo-annotations directly influences the segmentation results. There are several categories of methods to estimate pseudo-annotations. The first category is Simple-to-Complex (STC) strategy [22]–[24]. The methods in this category assume that the pseudo-annotations of simple images (*e.g.*, web images) can be accurately estimated by saliency detection [22] or co-segmentation [24]. Then the segmentation models trained on the simple images are utilized to generate pseudo-annotations for the complex images. The methods in this category usually require a large amount of external data which consequently increase the data and computation load. The second category is region-mining based methods. The methods in this category rely on region-mining

methods [7]–[9] to generate discriminative regions as localization seeds. Since such localization seeds mainly sparsely lie in the discriminate parts instead of the whole extent of the objects, which is far from the ground-truth annotation, many works try to alleviate this problem by expanding the localization seeds to the size of objects. Kolesnikov et al. [4] expand the seeds by aggregating the pixel-scores by global weighted rank-pooling. Wei et al. [6] apply an adversarial-erasing approach which iterates between suppressing the most discriminative image region and training the region mining model. It gradually localizes the next most discriminative regions through multiple iterations and merges all the mined discriminative regions into final pseudo-annotations. Similarly Two-phase [25] captures the full extent of the objects by suppressing and mining processing in two phases. Some works [5], [23] utilize external dependencies such as fully-supervised saliency method [26] trained on additional saliency datasets to facilitate estimating object scales.

To generate high-quality pseudo-annotations, the first category focuses on the quality of training data while the second category focuses on post-processing the localization seeds which is independent of the region-mining model structure. Different from previous methods, we focus on designing a region-mining model structure which is likely to highlight the object region. We aim to generate pseudo-annotations for weakly-supervised semantic segmentation in a single forward path without external data or external prior for efficiency and simplicity purpose.

### B. Mining Discriminative Regions

In this section we introduce some region-mining methods, which have been widely used in generating pseudo-annotations for semantic segmentation with image-level labels. Recent works of top-down neural saliency [7]–[9] perform well in weakly supervised localization tasks. Such works identify the discriminative regions respect to each individual class based on image classification DCNNs. Zhang et al. [8] propose Excitation Backprop to back-propagate in the network hierarchy to identify the discriminative regions. Zhou et al. [7] propose a technique called Class Activation Mapping (CAM) for identifying discriminative regions by replacing fully-connected layers in image classification CNNs with convolutional layers and global average pooling. Grad-CAM [9] is a strictly generalization of CAM [7] without the need of modifying DCNN structure. Among the methods listed above, CAM [7] is the most widely used one in weakly-supervised semantic segmentation [4], [6], [25] for generating pseudo-annotations.

### C. Spatial Attention Mechanism

Spatial neural attention is a mechanism to assign different weights to different feature spatial regions depending on their feature content. It automatically predicts the weighted heat map to enhance the relevant features and block the irrelevant features during the training process for specific tasks. Intuitively, such weighted heat map could be applied to our purpose of pseudo-annotation generation.

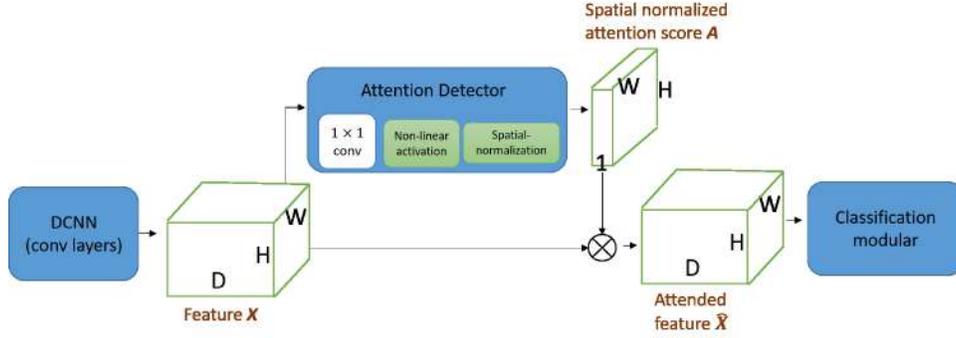


Fig. 2: The illustration of the structure of conventional spatial neural attention model.

Spatial neural attention mechanism has been proven to be beneficial in many tasks such as image captioning [11], [27], machine translation [28], multi-label classification [29], human pose estimation [30] and saliency detection [31]. Different from the previous works, we are the first to apply the attention mechanism to weakly supervised semantic segmentation to the best of our knowledge.

### III. APPROACH

First we give a brief review to the conventional spatial neural attention model in Sec. III-A. Second we introduce our decoupled attention structure in Sec. III-B. Then we introduce how to further generate pseudo-annotations based on this decoupled attention model in Sec. III-C. Finally, the generated pseudo-annotations are utilized as the ground-truth annotation to train segmentation DCNN networks.

#### A. Conventional Spatial Neural Attention

In this section, we will give a brief introduction to the conventional spatial neural attention mechanism. The conventional spatial neural attention structure is illustrated in Fig. 2. The attention structure consists of two modular branches: the main branch modular and attention detector modular. The attention detector modular is jointly trained with the main stream modular end-to-end.

Formally we denote the output features of some convolutional/pooling layers by  $\mathbf{X} \in \mathbb{R}^{W \times H \times D}$ . The attention detector takes feature map  $\mathbf{X} \in \mathbb{R}^{W \times H \times D}$  as the input and outputs spatial-normalized attention weights map  $\mathbf{A} \in \mathbb{R}^{W \times H}$ .  $\mathbf{A}$  is applied on  $\mathbf{X}$  to get attended feature  $\hat{\mathbf{X}} \in \mathbb{R}^{W \times H \times D}$ .  $\hat{\mathbf{X}}$  is fed into the classification modular to output image classification score vector  $\mathbf{p} \in \mathbb{R}^C$ , where  $C$  is the number of classes.

For notation simplicity, we denote the 3-D feature output of DCNNs with upper-case letters and feature at one spatial location with its corresponding lower-case letters. For example,  $\mathbf{x}_{ij} \in \mathbb{R}^D$  is the feature at the position  $(i, j)$  of  $\mathbf{X}$ . Attention detector outputs attention weights map  $\mathbf{A}$  which acts as a spatial regularizer to enhance the relevant regions and suppress the non-relevant regions for feature  $\mathbf{X}$ . Thus we are motivated to utilize the output of attention detector to generate pseudo-annotations for the task of semantic segmentation with image-level labels. The details of the attention detector modular are as

follows: Attention detector modular consists of a convolutional layer, a non-linear activation layer (Eq. (1)) and a spatial-normalization (Eq. (2)) as is shown in Fig. 2:

$$z_{ij} = F(\mathbf{w}^T \mathbf{x}_{ij} + \mathbf{b}), \quad (1)$$

$$a_{ij} = \frac{z_{ij}}{\sum_{i,j} z_{ij}}, \quad (2)$$

where  $F$  is a non-linear activation function, such as exponential function in [11], [30].  $\mathbf{w} \in \mathbb{R}^D$  and  $\mathbf{b} \in \mathbb{R}^1$  are the parameters of the attention detector model, which is a  $1 \times 1$  convolution layer. The attended feature  $\hat{\mathbf{x}}_{ij} \in \mathbb{R}^D$  is calculated as

$$\hat{\mathbf{x}}_{ij} = \mathbf{x}_{ij} a_{ij}, \quad (3)$$

The classification modular consists of a spatial average pooling layer and  $1 \times 1$  convolutional layer as the image classifier. We denote the  $\mathbf{v}_c$  and  $h_c$  as the parameters of the classifier for class  $c$ , thus the classification score for class  $c$  is calculated as:

$$p_c = \left( \sum_{i,j} \hat{\mathbf{x}}_{ij} a_{ij} \right)^T \mathbf{v}_c + h_c, \quad (4)$$

where  $p_c$  is the score for  $c$ -th class.

#### B. Decoupled Attention Structure

As described in Sec. III-A, the output of conventional spatial neural attention detector is class-agnostic. However, in semantic segmentation task each image may contain objects of multi-classes. Thus conventional class-agnostic attention map is not applicable to generate pseudo-annotations for such multi-class case since we need to predict pixel-wise label for each semantic class instead of just foreground/background. On the other hand, the output of conventional spatial neural attention detector is aimed to assist the task of image classification which may not necessarily generate desired pseudo-annotations for weakly-supervised semantic segmentation. Inspired by [29], we propose our decoupled attention structure especially for the task of generating pseudo-annotations for weakly supervised semantic segmentation to alleviate these problems.

We illustrate our structure in Fig. 3. Such structure extend the conventional attention structure to multi-class cases. Moreover, it generates two different types of attention maps

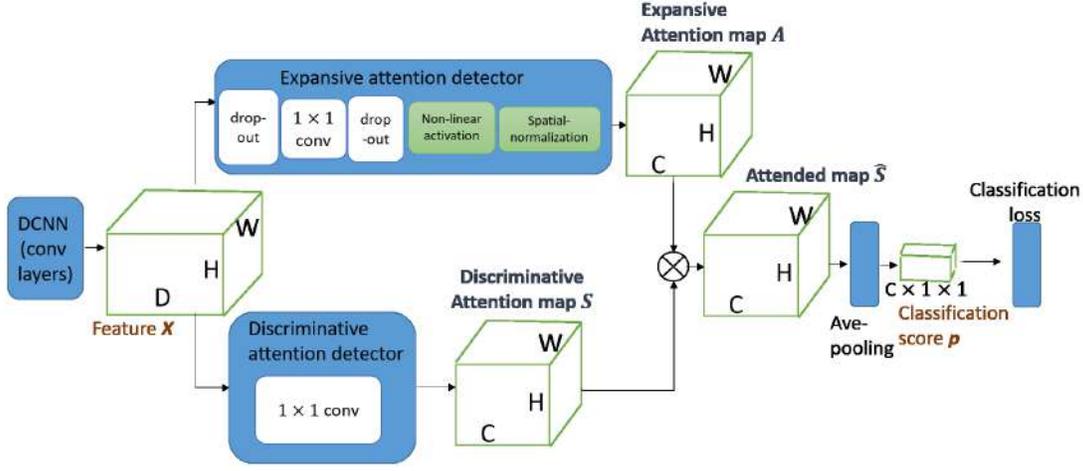


Fig. 3: Our proposed decoupled attention structure. The heat map generated by Expansive attention detector is named *Expansive attention map*, which identifies the object regions. The heat map generated by the Discriminative attention detector is named *Discriminative attention map*, which localizes the discriminative parts.

which identifies object regions and predicts the discriminative parts respectively. In Fig. 3, the attention map  $\mathbf{A} \in \mathbb{R}^{W \times H \times C}$  generated by the Expansive attention detector in the top branch is named **Expansive attention map**, which aims at identifying object regions. The attention map  $\mathbf{S} \in \mathbb{R}^{W \times H \times C}$  generated by the Discriminative attention detector in the bottom branch is named **Discriminative attention map**, which aims at mining the discriminative parts. Such two attention maps have different properties which are complimentary to each other to generate pseudo-annotations for weakly supervised semantic segmentation.

The details of the structure in Fig. 3 are described as follows:

Expansive Attention detector (E-A detector) modular consists of a convolutional layer, a non-linear activation layer (Eq. (5)) and a spatial-normalization step (Eq. (6)):

$$z_{ij}^c = F(\mathbf{w}_c^\top \mathbf{x}_{ij} + b^c), \quad (5)$$

$$a_{ij}^c = \frac{z_{ij}^c}{\sum_{i,j} z_{ij}^c}, \quad (6)$$

where the superscript/subscript  $c$  denotes the value of  $c$ -th channel/class.

As we mentioned in Sec. I, for the task of pseudo-annotation generation we aim to estimate the whole range of objects instead of only obtaining the most discriminative parts. Thus we design the E-A detector details as follows: we set

$$F(x) = \log(1 + \exp(x)) + \epsilon, \quad (7)$$

where  $\epsilon = 0.1$ , similar to [32]. Besides the  $1 \times 1$  convolutional layer in the attention detector, we add one drop-out layer right before and after it respectively. Such drop-out layers randomly zero-out the elements in the training features and consequently the attention detector will highlight more relevant features instead of just the most relevant one for successful classification.

The Discriminative attention detector (D-A detector) consists of  $1 \times 1$  convolutional layer whose parameters are denoted

as  $\mathbf{v}_c$  and  $h_c$  same as the classifier in Sec. III-A. The D-A detector takes feature map  $\mathbf{X}$  as the input and outputs class-specific attention map  $\mathbf{S} \in \mathbb{R}^{W \times H \times C}$ .

The attended feature  $\hat{\mathbf{S}} \in \mathbb{R}^{W \times H \times C}$  is calculated as:

$$\hat{s}_{ij}^c = s_{ij}^c a_{ij}^c, \quad (8)$$

$\hat{\mathbf{S}}$  is fed into a spatial average pooling layer to generate image classification score  $\mathbf{p} \in \mathbb{R}^C$ . Hence  $p_c$  is calculated as:

$$p_c = \sum_{i,j} (\mathbf{x}_{ij}^\top \mathbf{v}_c + h_c) a_{ij}^c. \quad (9)$$

Compared with Eq. (4), in our decoupled model the D-A detector is to predict the class score for dense pixel position instead of predicting image label score. Thus our model remains the spatial information for the classification map which is more suitable for semantic segmentation tasks. The multi-label classification loss for each image is formulated as:

$$L_{oss} = - \sum_c \left[ y_c \log\left(\frac{1}{1 + e^{-p_c}}\right) + (1 - y_c) \log\left(\frac{e^{-p_c}}{1 + e^{-p_c}}\right) \right], \quad (10)$$

where  $y_c$  is the binary image label corresponding to  $c$ -th class.

We show examples of Expansive attention map and Discriminative attention map in Fig. 4, which illustrates that Expansive attention maps perform well at identifying large object regions while Discriminative attention maps perform well at mining the small discriminative parts. Such two attention maps show different and complimentary properties. Thus we merge such two attention maps using Eq. (11)

$$\mathbf{T}_c = \hat{p}_c * \mathbf{A}_c + (1 - \hat{p}_c) * \mathbf{S}_c, \quad (11)$$

where  $\mathbf{A}_c$  is the normalized Expansive attention map and  $\mathbf{S}_c$  is the normalized Discriminative attention map corresponding to  $c$ -th class.  $\mathbf{T}_c$  is the result merged attention map.  $\hat{\mathbf{p}}$  is the softmax normalization result of the image classification score  $\mathbf{p}$ . Such weighted combination is intuitive: the small prediction score usually correspond to the difficult objects of small size so more weights should be put on the task of mining the discriminative regions.

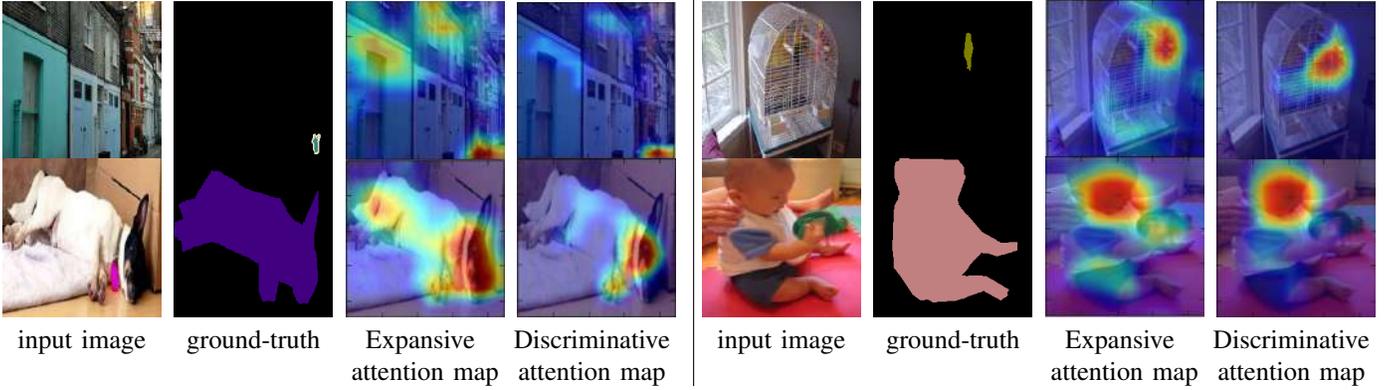


Fig. 4: The examples of Expansive attention map and Discriminative attention map. Expansive attention map and Discriminative attention map have different properties which are suitable for different situations.

### C. Pseudo-annotation Generation

In this section we describe how to generate pseudo-annotations according to the attention maps. We generate pseudo-annotations by simple thresholding following the similar practice used in [4], [6], [7]. Given an attention map of an image with image label set  $\mathcal{L}$  (excluding the background), for each class  $c$  in  $\mathcal{L}$ , we generate foreground regions as follows: first we perform min-max spatial normalization on the attention map corresponding to class  $c$  to [0,1] range. Then we threshold the normalized attention map by  $>0.2$ . Inspired by [33], we sum the feature map  $\mathbf{X}$  in the channel dimension and then perform min-max spatial normalization on it. Then we threshold this normalized map by  $<0.3$  to generate background regions. Since the regions are generated independently for each class, there may exist label conflicts at some pixel positions. We choose foreground label with smallest region size for the conflict regions following the practice of [4]. We assign the unclassified pixels with void label denoted as *void* which represents the label is unclear at this position and will not be considered in the training process.

The generated hard annotations  $M$  are coarse and have much unclear area. We can further apply denseCRF [34] on  $M$  to generate refined annotations. We first describe how to generate the class probability vector for each spatial locations. For an image  $I$ , the labels which are present in  $I$  are denoted as  $\mathcal{C}_I$ . All the class labels in the target dataset is denoted as  $\mathcal{C}$ .  $m_i$  is the mask label of  $M$  for pixel  $i$ . We calculate the class probability vector  $z_{i,c}$  for class  $c \in \mathcal{C}$  at pixel  $i$  as follows:

If  $m_i = \text{void}$ ,

$$z_{i,c} = \begin{cases} \frac{1}{|\mathcal{C}_I|}, & \text{if } c \in \mathcal{C}_I \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

Else:

$$z_{i,c} = \begin{cases} \tau, & \text{if } c = m_i \\ \frac{1-\tau}{|\mathcal{C}_I|-1}, & \text{if } c \neq m_i \text{ and } c \in \mathcal{C}_I \\ 0, & \text{otherwise} \end{cases} \quad (13)$$

where  $\tau \in (0.5, 1)$  is the manually fixed parameter.  $|\mathcal{C}_I|$  represents the number of labels that are present in image  $I$ .

The unary potential is calculated for the probability vector. We apply denseCRF [34] with the unary potential and take the result mask as the refined annotations.

## IV. EXPERIMENTS

### A. Experimental Set-up

**Dataset and Evaluation Metric** We evaluate our method on the PASCAL VOC 2012 image segmentation benchmark [35], which has 21 semantic classes including background. The original dataset has 1464 training images (*train*), 1449 validation images (*val*) and 1456 testing images (*test*). The dataset is augmented to 10582 training images by [36] following the common practice. In our experiments, we only utilize image-level class labels of the training images. We use the *val* images to evaluate our method. As for the evaluation measure for segmentation performance, we use the standard PASCAL VOC 2012 segmentation metric: mean intersection-over-union (mIoU).

**Training/Testing Settings** We train the proposed decoupled attention network for pseudo-annotation estimation. Based on the generated pseudo-annotations we train the state-of-the-art semantic segmentation network to predict the final segmentation result.

We build the proposed decoupled attention model based on VGG-16 model. We transfer the layers from VGG-16 from the first layer to layer *relu5\_3* as the starting convolutional layers (as is shown in Fig. 3) which outputs  $\mathbf{X} \in \mathbb{R}^{14 \times 14 \times 512}$ . We use a mini-batch size of 15 images with the data augmentation methods such as random flip and random scale. We set the 0.01 as initial learning rate for the VGG-16 transferred layers and 0.1 as the initial learning rate for the attention detector layers. We decrease the learning rate by a factor of 10 after 10 epochs. Training terminates after 20 epochs.

We train the DeepLab-LargeFOV (VGG-16 based) of [2] as our segmentation model. The input image crops for the network are of size  $321 \times 321$  and outputs segmentation mask are of size  $41 \times 41$ . The initial base learning rate is 0.001 and it is decreased by a factor of 10 after 8 epochs. Training terminates after 12 epochs. We use the public available pytorch implementation of Deeplab-largeFOV<sup>1</sup>. In the inference phase

<sup>1</sup>[https://github.com/BardOfCodes/pytorch\\_deeplab\\_large\\_fov](https://github.com/BardOfCodes/pytorch_deeplab_large_fov)

of segmentation, we use multi-scale inference to combine output score at different scales, which is common practice as in [3], [37]. The final outputs are post-processed by denseCRF [34].

TABLE I: Quantitative evaluation of attention maps.  $mS_{prec}$  is the criteria about localizing the concentrated interior object parts,  $mS_{rec}$  is the criteria about covering over the object regions and  $mS_{IoU}$  is the criteria for identifying the whole object regions. The results show that Expansive attention map (expan-atten) performs well at covering over and identifying the whole object regions and Discriminative attention map (disc-atten) performs well at localizing interior partial object regions. The merged attention map (merged-atten) achieves the highest  $mS_{IoU}$ , which means it performs best at identifying the whole object regions.

	expan-atten	disc-atten	merged-atten
$mS_{prec}$	52.0	<b>61.2</b>	—
$mS_{rec}$	<b>71.9</b>	54.4	—
$mS_{IoU}$	42.1	39.4	<b>44.0</b>

### B. Ablation Study of Decoupled Attention Model

1) *Properties of attention maps:* As described in Sec. III-B, our decoupled attention model outputs Expansive attention map to identify object regions and Discriminative attention map to localize the discriminative parts. We qualitatively and quantitatively compare these two attention maps generated by our decoupled attention structure and demonstrate their differences.

First we provide visual examples of the Expansive attention maps and Discriminative attention maps of *train* images in Fig. 5 (column 2 and column 3). In Fig. 5 we provide some examples including images with single/multiple objects and single/multiple class labels. We observed that for simple cases (*e.g.*, large objects), Expansive attention map is likely to cover whole region of objects, while Discriminative attention map is likely to locate the most discriminative part. We also observed that for difficult cases (*e.g.*, small objects), Expansive attention map is likely to identify a broad region, while Discriminative attention map is likely to precisely localize the object. Thus we draw the conclusion that these two attention maps have different properties and are suitable for different situations. They are potentially complimentary to each other. Thus the combination of the two maps will result in attention maps that are applicable to different cases which lead to object annotations of high quality.

We also quantitatively compare these two attention maps. We apply our decoupled attention model on *val* images and generate annotation masks without denseCRF refinement for Expansive attention map and Discriminative attention map. We evaluate the generated estimated masks regarding to the groundtruth of *val* images.

We propose three evaluation measures for the estimated masks to demonstrate the different properties of the two attention maps:

1) We use the commonly used IoU to evaluate the performance of identifying the whole objects, which are denoted as:

$$S_{IoU} = \frac{true\ pos.}{true\ pos. + false\ pos. + false\ neg.} \quad (14)$$

2) We propose a criteria to evaluate the localization precision within the object region, which are denoted as:

$$S_{prec} = \frac{true\ pos.}{true\ pos. + false\ pos.} \quad (15)$$

3) We propose a criteria to evaluate the localization recall over the object region, which are denoted as:

$$S_{rec} = \frac{true\ pos.}{true\ pos. + false\ neg.} \quad (16)$$

$S_{prec}$  emphasizes localizing the concentrate and small regions within the objects which not necessarily cover the whole object regions.  $S_{rec}$  emphasizes the expansion of the highlighted regions by measuring whether it includes the whole range of the objects which not necessarily localize within object regions.  $S_{IoU}$  emphasizes the accuracy of identifying the whole object regions which is the final criteria to indicate the quality of the attention map for generating pseudo-annotations. We use the average score over all classes, which are denoted as  $mS_{IoU}$ ,  $mS_{prec}$  and  $mS_{rec}$  as our criteria.

The evaluation results are shown in the first two columns of Table I. We observe that Expansive attention map gets higher  $mS_{rec}$  and  $mS_{IoU}$  score, while Discriminative attention map gets higher  $mS_{prec}$ . This verifies different properties of attention maps: Discriminative attention map is likely to localize partial interior object regions while Expansive attention map highlights the regions of larger expansion and is likely to identify the whole object region.

As described in Sec. III-B, we can merge these two attention maps to improve the quality of generated pseudo-annotations. We merge the two attention maps as Eq. (11) and follow the  $mS_{IoU}$  criteria. The results are shown in the last column of Table I. We observe that  $mS_{IoU}$  of merged attention map is higher than only using single type of attention map, which demonstrates that Expansive attention map and Discriminative attention map are complimentary to each other in localizing the whole object regions. The pseudo-annotations generated from the merged attention map are relatively close to the ground-truth, as is shown in Fig. 5.

TABLE II: Evaluation of different drop-out rates (DR). With the increase of drop-out rates,  $mS_{rec}$  constantly increases while  $mS_{prec}$  constantly decreases. It shows that large drop-out rates enhance the expansion effect of attention maps to some extent.

DR	0	0.3	0.4	0.5	0.6	0.7
$mS_{prec}$	<b>57.3</b>	57.1	56.7	56.2	55.1	51.8
$mS_{rec}$	64.5	65.7	67.6	69.6	71.6	<b>74.1</b>
$mS_{IoU}$	42.6	43.1	43.7	44.0	<b>44.2</b>	42.8

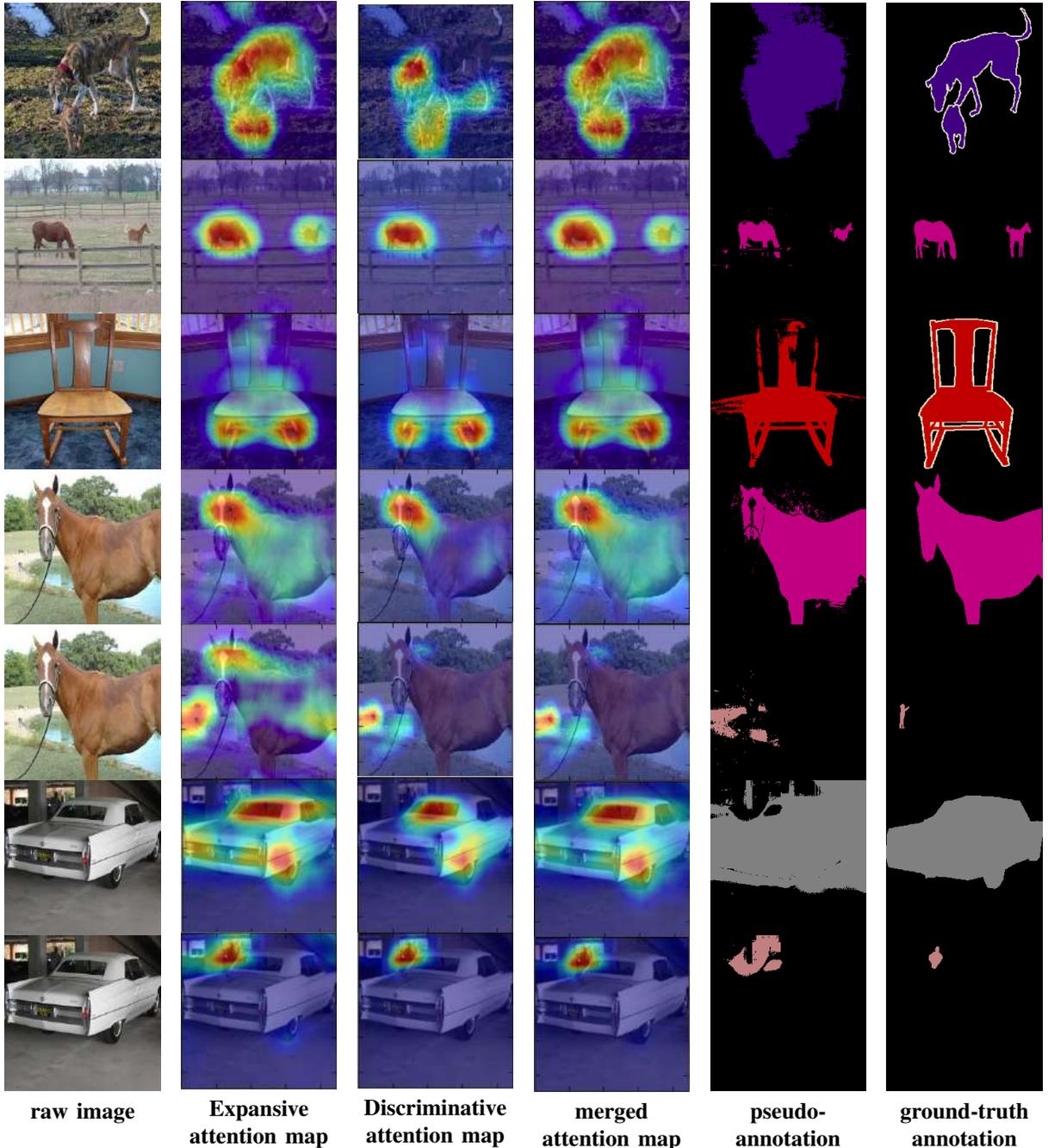


Fig. 5: Visualization of attention maps. The first column shows the input image. The second column shows the Expansive attention maps. The third column shows the Discriminative attention maps. The fourth column shows the merged attention maps. The fifth column shows the pseudo-annotations generated on the merged attention maps. The last column is the real ground-truth annotation for comparison. For the images with multiple classes we show the attention maps separately for each class. We observe that our merged attention maps combine the complimentary information of Expansive attention map and Discriminative attention maps and the generated pseudo-annotations are close to the groundtruth.

2) *Evaluation of dropout layers*: As described in Sec. III-B, we add dropout layers in the Expansive attention detector modular to have expansion effect. In this section we evaluate and discuss the effect of drop-out layers.

In our experiments we use the default drop-out rate of 0.5. We also evaluate other drop-out rates and follow the same criteria as Table I on the merged attention maps. The results

are reported in Table II. It shows that with the increase of drop-out rates,  $mS_{rec}$  constantly increases while  $mS_{prec}$  constantly decreases. This demonstrates that increasing the drop-out rates helps expand the identified region from the interior parts to the entire object scales. We also show the visualization examples of the attention maps generated by different drop-out rates in Fig. 6, which demonstrate the expansion effect results from

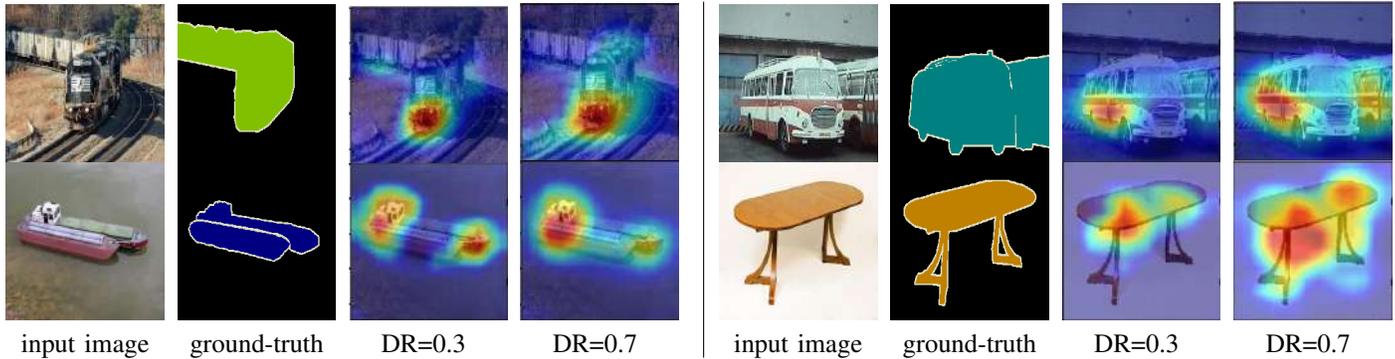


Fig. 6: The visual examples of different drop-out rates (DR). It shows that using larger drop-out rates (DR = 0.7) leads to expansion effect over the small drop-out rates (DR = 0.3).

larger drop-out rates.

The effects of drop-out layers could be explained as follows: Drop-out layers randomly set the feature grids to zeros by a rate which consequently suppress the feature space by noise. Such zero-out process is operated directly on the CNN feature space at random spatial locations and feature dimensions. If the discriminative features are suppressed by drop-out process, the training process will mine other discriminative features for classification purpose. Thus the attention model will adapt to highlight more relevant features instead of the most discriminative ones.

3) *Evaluation of decoupled structure*: In this section, we aim to show that our decoupled attention structure is effective. We implement the case of removing the Expansive attention detector modular and only train the remaining Discriminative attention detector modular. We add one drop-out layer before and after the convolutional layer of Discriminative attention modular respectively. Then we evaluate the Discriminative attention map follow the same criteria as Table I. The result of  $mS_{prec}$ ,  $mS_{rec}$  and  $mS_{IoU}$  are 48.3% , 49.7% and 31.5% respectively. These results are far lower than our decoupled attention in Table I. It verifies that our decoupled attention structure is more effective than the single-stream attention.

### C. Comparisons with Region-Mining Based Methods

Our method is related to region-mining based weakly supervised semantic segmentation approaches. In this section, we compare with two recently proposed region-mining based approaches, *i.e.*, SEC [4] and Adversarial-Erasing (AE) [6]. Both SEC and AE use region-mining methods CAM [7] to generate pseudo-annotations. CAM [7] is a method to mine the discriminate object regions by image-level labels, which is related to our attention based methods.

SEC [4] contains three losses to optimize:  $L_{seed}$  is the segmentation loss based on pseudo-annotations.  $L_{expand}$  is for expanding the localization seed to object scale the and  $L_{constrain}$  is to make the segmentation results agree with the denseCRF [36] result. AE [6] iteratively mine the object regions through multiple iteration steps by erasing the most discriminative regions from the original images and re-train the CAM localization network to mine the next most discriminative regions. The mined regions of all the iterations are

TABLE III: Comparisons with Region-Mining Based Methods. We list the middle-step segmentation results based on the pseudo-annotations generated by different region-mining based methods. We outperform SEC and AE w/o PSL with various settings.

Methods	mIoU
<b>SEC [4]</b>	
$L_{seed}$	45.4
$L_{seed} + L_{constrain}$	50.4
$L_{seed} + L_{constrain} + L_{expand}$	50.7
<b>AE w/o PSL [6]</b>	
AE-step1	44.9
AE-step2	49.5
AE-step3	50.9
AE-step4	48.8
<b>Ours</b>	
Expansive attention map	54.1
Discriminative attention map	52.2
merged attention map	55.4

combined as the final pseudo-annotations. We list segmentation results of the middle steps of SEC and AE in Table III. For SEC, we show the segmentation results with different combination of losses. For AE, we list the segmentation results using different number of erasing steps without prohibitive segmentation learning (PSL).

For our methods, we list the segmentation results using the pseudo-annotations generated by Expansive attention map, Discriminative attention map and merged attention map in Table III. Since merged attention map achieves better performance over the other two attention maps, we will use the segmentation results generated by merged attention map for further experiments by default. We outperform SEC and AE w/o PSL with various settings, which indicates our attention based region mining approach are significantly more effective than the CAM based mining methods. Moreover, our method employed a simple pipeline without complicated iterative training in contrast with AE.

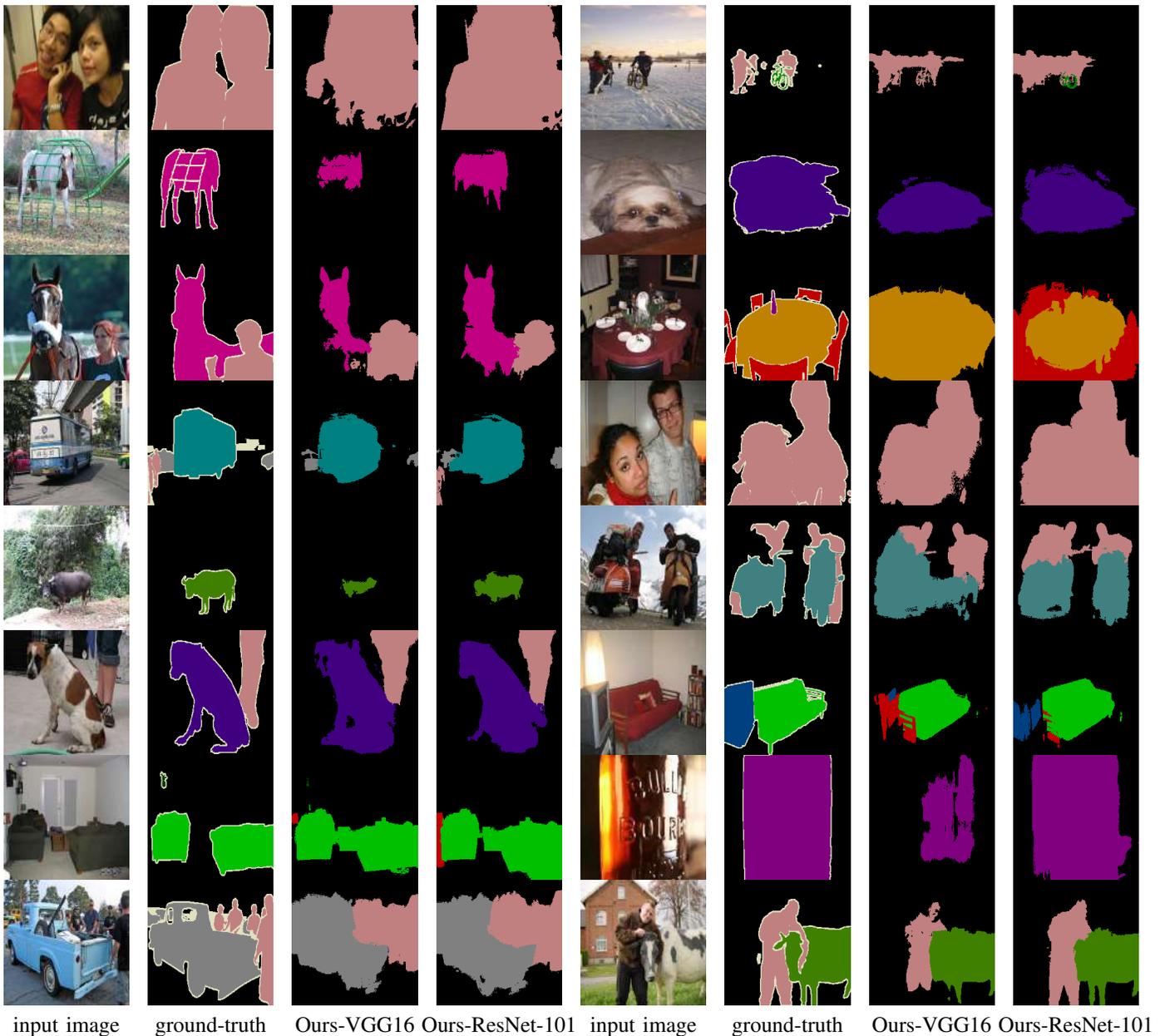


Fig. 7: Qualitative results of PASCAL 2012 val images. We list the qualitative examples generated by Deeplab-largeFOV (Ours-VGG16) segmentation model and Resnet101 based (Ours-Resnet-101) [13] Deeplab-like segmentation model.

#### D. Comparisons with State-of-the-arts

In this section, we compare our segmentation results with other weakly supervised semantic segmentation methods. The comparison results are listed in Table IV and Table V for *val* and *test* images respectively. We divide weakly supervised semantic segmentation into three categories based on different levels of weak supervisions and whether additional information (data/supervision) is implicitly used in their pipeline:

In the first category, the methods utilize interactive input such as scribble [38] or point [39] as supervision, which is relatively more precise indicator of object location and scale. The results in this category are listed in the first block of Table IV and Table V.

In the second category, the methods only use image-level

label as supervision, but introduce information of external sources to improve segmentation results. The segmentation performance in this category are listed in the second block of Table IV and Table V. Some work use additional data to assist training. For example, STC [22] crawled 50K web images for initial training step and Crawled-Video [42] crawled large amount of online videos, which significantly increase the training data amount. Some work implicitly utilize the pixel-wise annotations in their training. For example, TransferNet [43] transfer the pixel-wise annotation from MSCOCO dataset [44] to PASCAL dataset. AF-MCG [45] utilize MCG proposal method which is trained with PASCAL VOC pixel-wise ground-truth. Some works utilize fully-supervised saliency detection methods in localization seed expansion [5] or fore-

TABLE IV: Weakly supervised semantic segmentation result on *val* images. Our methods achieves the state-of-the-art results in the weakly supervised semantic segmentation supervised only with image labels without external sources

Methods	<i>val</i>	Comments
<b>Interactive Input</b>		Annotation
ScribbleSup [38]	63.1	scribble annotation
ScribbleSup(point) [38]	51.6	point annotation
What't the point [39]	46.1	point annotation
BoxSup [40]	62.0	Box annotation
<b>Image-level labels with external source</b>		Additional Info
STC [22]	49.3	Web images
Co-segmentation [24]	56.4	Web images
Webly-supervised [41]	53.4	Web images
Crawled-Video [42]	58.1	Web videos
TransferNet [43]	52.1	MSCOCO [44] pixel-wise label
AF-MCG [45]	54.3	MCG proposal
Joon <i>et al.</i> [5]	55.7	Supervised Saliency
DCSP-VGG16 [46]	58.6	Supervised Saliency
DCSP-ResNet-101 [46]	60.8	Supervised Saliency
Mining-pixels [23]	58.7	Supervised Saliency
AE w/o PSL [6]	50.9	Supervised Saliency
AE-PSL [6]	55.0	Supervised Saliency
<b>Image-level label w/o external source</b>		
MIL-FCN [19]	25.7	
EM-Adapt [21]	38.2	
BFBP [33]	46.6	
DCSM [47]	44.1	
SEC [4]	50.7	
AF-SS [45]	52.6	
Two-phase [25]	53.1	
Anirban <i>et al.</i> [48]	52.8	
<b>Ours-VGG16</b>	<b>55.4</b>	
<b>Ours-ResNet-101</b>	<b>58.2</b>	

ground/background detection [6], [22], which implicitly utilize external saliency ground-truth data to train saliency detector.

In the third category, the methods only use image-level label as supervision without the information of external sources. Our methods belong to this category. The segmentation performance in this category are listed in the third block of Table IV and Table IV.

We mainly focus on comparing with the methods in the third category. We achieve the mIoU score of 55.4 and 56.4 on *val* and *test* images respectively, which significantly outperforms other methods using the same supervision settings. In order to further improve our segmentation results, we use Deeplab-like model in [13] based on Resnet-101 [49] as the segmentation model. We achieve the mIoU score of 58.2 on *val* images and 60.1 on *test* images, which outperforms all the existing methods in weakly supervised semantic segmentation using the same level of supervision setting up to date. We also list our segmentation results for each class in the Table VI for reference. We show the qualitative examples of the segmentation results in Fig. 7.

Moreover, we also want to emphasize that we do not iteratively train the models in multiple steps. The whole pipeline only need training decoupled attention structure and segmentation structure once on the PASCAL VOC *train* data. To our knowledge our approach has the most simple pipeline for weakly-supervised semantic segmentation.

TABLE V: Weakly supervised semantic segmentation result on *test* images. Our methods achieves the state-of-the-art results in the weakly supervised semantic segmentation supervised only with image labels without external sources

Methods	<i>test</i>	Comments
<b>Interactive Input</b>		Annotation
BoxSup [40]	64.2	Box annotation
<b>Image-level labels with external source</b>		Additional Info
STC [22]	51.2	Web images
Co-segmentation [24]	56.9	Web images
Webly-supervised [41]	55.3	Web images
Crawled-Video [42]	58.7	Web videos
TransferNet [43]	51.2	MSCOCO [44] pixel-wise label
AF-MCG [45]	55.5	MCG proposal
Joon <i>et al.</i> [5]	56.7	Supervised Saliency
DCSP-VGG16 [46]	59.2	Supervised Saliency
DCSP-ResNet-101 [46]	61.9	Supervised Saliency
Mining-pixels [23]	59.6	Supervised Saliency
AE w/o PSL [6]		Supervised Saliency
AE-PSL [6]	55.7	Supervised Saliency
<b>Image-level label w/o external source</b>		
MIL-FCN [19]	24.9	
EM-Adapt [21]	39.6	
BFBP [33]	48.0	
DCSM [47]	45.1	
SEC [4]	51.7	
AF-SS [45]	52.7	
Two-phase [25]	53.8	
Anirban <i>et al.</i> [48]	53.7	
<b>Ours-VGG16</b>	<b>56.4</b>	
<b>Ours-ResNet-101</b>	<b>60.1</b>	

## V. CONCLUSION

In this work we have presented a novel decoupled spatial neural attention work to generate pseudo-annotations for weakly-supervised semantic segmentation with image-level labels. This decoupled attention model simultaneously outputs two class-specific attention maps with different properties and are effective for estimating pseudo-annotations. We perform detailed ablation experiments to verify the effectiveness of our decoupled attention structure. Finally we achieve state-of-the-art weakly supervised semantic segmentation results on Pascal VOC 2012 image segmentation benchmark.

## VI. ACKNOWLEDGEMENT

This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the National Research Foundation, Singapore, under its IDM Futures Funding Initiative and administered by the Interactive and Digital Media Programme. We also thank NVIDIA for their donation of GPU.

## REFERENCES

- [1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.
- [2] L. 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.
- [3] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *CVPR*, 2016.

TABLE VI: Our Segmentation results for each class on *val* and *test* images. We list the results of using Deeplab-largeFOV (vgg16) [2] and ResNet-101 based Deeplab-like model (resnet) [13] as segmentation DCNN model.

	background	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dining table	dog	horse	motorbike	person	potted plant	sheep	sofa	train	tv/monitor	mIoU
vgg16- <i>val</i>	84.5	73.3	23.4	66.4	42.7	48.7	76.9	61.0	66.0	21.8	65.2	37.1	70.0	62.5	62.0	58.9	41.1	71.2	32.4	58.7	39.0	55.4
vgg16- <i>test</i>	85.6	71.2	27.1	74.2	35.8	51.0	77.7	63.4	65.0	21.4	60.0	47.7	73.9	60.2	71.0	61.5	39.2	64.0	39.5	54.2	40.0	56.4
resnet- <i>val</i>	84.7	73.5	27.1	68.3	43.8	57.2	80.9	64.7	65.4	22.3	69.7	39.2	70.8	73.0	67.8	60.0	45.8	72.6	35.9	57.7	42.2	58.2
resnet- <i>test</i>	86.3	70.4	29.7	78.8	40.8	53.9	80.6	67.7	67.7	21.0	70.2	46.4	74.8	70.6	74.8	63.7	45.2	76.9	45.0	57.8	39.8	60.1

- [4] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *ECCV*, 2016.
- [5] S. J. Oh, R. Benenson, A. Khoreva, Z. Akata, M. Fritz, and B. Schiele, "Exploiting saliency for object segmentation from image level labels," in *CVPR*, 2017.
- [6] Y. Wei, J. Feng, X. Liang, M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *CVPR*, 2017.
- [7] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *CVPR*, 2016.
- [8] J. Zhang, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," in *ECCV*, 2016.
- [9] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017.
- [10] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *CVPR*, 2016.
- [11] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015.
- [12] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *ICCV*, 2015.
- [13] T. Shen, G. Lin, C. Shen, and I. Reid, "Learning multi-level region consistency with dense multi-label networks for semantic segmentation," in *IJCAI*, 2017.
- [14] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *CVPR*, 2017.
- [15] L. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *CVPR*, 2016.
- [16] 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 *ICCV*, 2015.
- [17] X. Li, Z. Liu, P. Luo, C. C. Loy, and X. Tang, "Not all pixels are equal: difficulty-aware semantic segmentation via deep layer cascade," in *CVPR*, 2017.
- [18] G. Lin, C. Shen, A. Van Den Hengel, and I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in *CVPR*, 2016.
- [19] D. Pathak, E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional multi-class multiple instance learning," in *ICLR Workshop*, 2015.
- [20] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *CVPR*, 2015.
- [21] G. Papandreou, L. Chen, K. Murphy, and A. L. Yuille, "Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *ICCV*, 2015.
- [22] Y. Wei, X. Liang, Y. Chen, X. Shen, M. Cheng, Y. Zhao, and S. Yan, "Stc: A simple to complex framework for weakly-supervised semantic segmentation," *TPAMI*, 2016.
- [23] Q. Hou, P. K. Dokania, D. Massiceti, Y. Wei, M. Cheng, and P. Torr, "Mining pixels: Weakly supervised semantic segmentation using image labels," *arXiv preprint arXiv:1612.02101*, 2016.
- [24] T. Shen, G. Lin, L. Liu, C. Shen, and I. Reid, "Weakly supervised semantic segmentation based on co-segmentation," 2017.
- [25] D. Kim, D. Yoo, I. S. Kweon *et al.*, "Two-phase learning for weakly supervised object localization," in *ICCV*, 2017.
- [26] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *CVPR*, 2016.
- [27] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *CVPR*, 2016.
- [28] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by joint learning to align and translate," in *ICLR*, 2015.
- [29] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang, "Learning spatial regularization with image-level supervisions for multi-label image classification," in *CVPR*, 2017.
- [30] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *CVPR*, 2017.
- [31] J. Kuen, Z. Wang, and G. Wang, "Recurrent attentional networks for saliency detection," in *CVPR*, 2016.
- [32] B. Zhuang, L. Liu, Y. Li, C. Shen, and I. Reid, "Attend in groups: a weakly-supervised deep learning framework for learning from web data," in *CVPR*, 2017.
- [33] F. Saleh, M. S. A. Akbarian, M. Salzmann, L. Petersson, S. Gould, and J. M. Alvarez, "Built-in foreground/background prior for weakly-supervised semantic segmentation," in *ECCV*, 2016.
- [34] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *NIPS*, 2011.
- [35] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *IJCV*, vol. 88, no. 2, 2010.
- [36] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *ICCV*, 2011.
- [37] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters—improve semantic segmentation by global convolutional network," in *CVPR*, 2017.
- [38] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *CVPR*, 2016.
- [39] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "Whats the point: Semantic segmentation with point supervision," in *ECCV*, 2016.
- [40] J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *ICCV*, 2015.
- [41] B. Jin, M. V. Ortiz Segovia, and S. Susstrunk, "Webly supervised semantic segmentation," in *CVPR*, 2017.
- [42] S. Hong, D. Yeo, S. Kwak, H. Lee, and B. Han, "Weakly supervised semantic segmentation using web-crawled videos," 2017.
- [43] S. Hong, J. Oh, B. Han, and H. Lee, "Learning transferable knowledge for semantic segmentation with deep convolutional neural network," in *CVPR*, 2015.
- [44] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.
- [45] X. Qi, Z. Liu, J. Shi, H. Zhao, and J. Jia, "Augmented feedback in semantic segmentation under image level supervision," in *ECCV*, 2016.
- [46] A. Chaudhry, P. K. Dokania, and P. H. Torr, "Discovering class-specific pixels for weakly-supervised semantic segmentation," in *BMVC*, 2017.
- [47] W. Shimoda and K. Yanai, "Distinct class-specific saliency maps for weakly supervised semantic segmentation," in *ECCV*, 2016.
- [48] A. Roy and S. Todorovic, "Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation," in *CVPR*, 2017.
- [49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.