# Two-Phase Learning for Weakly Supervised Object Localization

Dahun Kim
KAIST
mcahny@kaist.ac.kr

Donghyeon Cho
KAIST
cdh12242@gmail.com

Donggeun Yoo*
KAIST
dgyoo@rcv.kaist.ac.kr

In So Kweon
KAIST
iskweon@kaist.ac.kr

## Abstract

*Weakly supervised semantic segmentation and localization have a problem of focusing only on the most important parts of an image since they use only image-level annotations. In this paper, we solve this problem fundamentally via two-phase learning. Our networks are trained in two steps. In the first step, a conventional fully convolutional network (FCN) is trained to find the most discriminative parts of an image. In the second step, the activations on the most salient parts are suppressed by inference conditional feedback, and then the second learning is performed to find the area of the next most important parts. By combining the activations of both phases, the entire portion of the target object can be captured. Our proposed training scheme is novel and can be utilized in well-designed techniques for weakly supervised semantic segmentation, salient region detection, and object location prediction. Detailed experiments demonstrate the effectiveness of our two-phase learning in each task.*

## 1. Introduction

The most fundamental task for image understanding is to localize objects in a scene where each object has different locations and scales. It provides clues to challenging vision problems such as object detection and semantic segmentation. In recent years, deep learning based methods [13, 12, 27, 19, 21, 7, 20, 39] have achieved remarkably improved performance for those tasks by virtue of a large amount of annotated data and GPU parallel processing. However, it is expensive and laborious to obtain huge amounts of annotations such as bounding boxes and pixel-level labels. Therefore, weakly supervised learning [22, 42, 5, 16, 25, 18, 17, 23, 24, 28, 38] using only image-level annotations has begun to attract attention and shown interesting results.

However, there is still a large gap between the object localization power of weakly supervised methods and that of fully supervised methods. One major reason is that the localizability of weakly supervised FCNs is inherently limited to finding the most discriminative parts, rather than estimating the complete extent of objects. This is because image-level annotations simply lack information on the spatial extent of objects. Most existing weakly supervised methods for object localization [22, 42, 41, 10], detection [5, 16, 9, 29, 4, 37], and semantic segmentation [25, 18, 17, 23, 24, 28, 38, 35, 36, 40] suffer from this chronic problem.

In this work, we overcome this problem fundamentally via two-phase learning. Our networks are trained in two phases. During the first phase, a conventional FCN is trained for image-level classification. At this time, pixels belonging to the most important parts in an image are revealed in a heat map, as shown in Fig. 1-(b). During the second phase, another FCN is trained but the activations on highlighted regions in the first stage are suppressed via *inference conditional feedback*. The underlying insight is that when the network is encouraged to discriminate im-

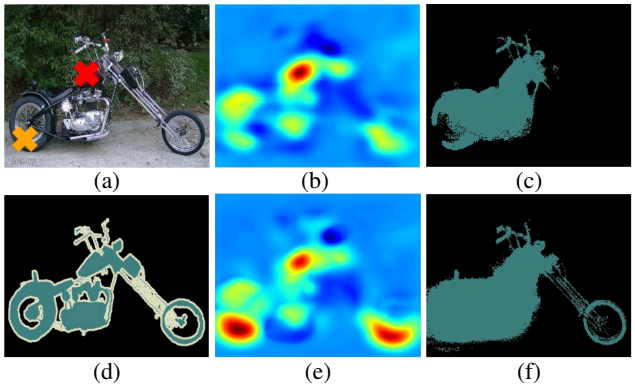


Figure 1: The effects of two-phase learning. (a) An input image, and estimated locations as the most (red) and the next (orange) important parts. (b) The heat map from the first network [42]. (c) The segmentation prediction of our baseline [18]. (d) Ground truth segmentation mask. (e) The heat map from the proposed method. (f) The segmentation prediction using the proposed method.

*This work was done when he was in KAIST. He is currently working in Lunit Inc.

ages into their categories without knowledge of the most distinctive regions, the network will discover the next most discriminative parts of objects. At the inference stage, the entire portion of objects can be captured by combining activations of both phases, as illustrated in Fig. 1-(e). In other words, two-phase learning solves the fundamental problem that heat maps do not contain the entire parts of objects.

Enhanced heat maps are then used to improve the performance of per-class saliency detection and object localization as well as semantic segmentation. We explain in detail how to apply improved heat maps to each task, and discuss the effectiveness of the proposed two-phase learning through various experiments.

In summary, this paper introduces the concept of two-phase learning for weakly supervised object localization. It allows the network to capture the full extent of the objects.

## 2. Related Works

In this section, we review previous studies that have sought to capture the spatial extent or the whole part of objects, not just the location of the most important part. Their goal coincides with that of two-phase learning. These studies can be broadly categorized into two types of approaches.

First, a group of approaches modify score aggregation methods in order to achieve a balance between the two most popular global pooling strategies: global max pooling (GMP) [22] and global average pooling (GAP) [42]. Since each of these pooling methods tends to underestimate or overestimate the extent of objects, respectively, finding a generalized model between these two extremes is essential. Pinheiro and Collobert [25] aggregate activations into image-level scores through the log-sum-exp (LSE) pooling layer. In particular, Sun and Paluri [32] provide a comparison of GMP, GAP, and LSE pooling methods by showing the classification and localization performance of each method. Also, global weighted ranking pooling (GWRP) is proposed by Kolesnikov and Lampert [18] to properly combine properties of GMP and GAP. However, these methods are based on a user-parameter about the object size, which predetermines the portion of an image to be focused on.

The second group of methods employ external algorithms to obtain saliency masks or object proposals. Wei *et al.* [38] construct a new dataset consisting of images with a well-centered single object, and then apply the state-of-the-art saliency detection method proposed by Jiang *et al.* [15] to generate foreground/background masks. Qi *el al.* [26], Pinheiro *el al.* [25], and Bearman *et al.* [3] make use of external region proposal methods to boost their performance. Selective search [34], CMPC [6], BING [8], Objectness [1], and MCG [2] are the popular helpers. One approach with no such dependencies is suggested by Saleh *et al.* [28]. They extract saliency masks from the network itself by fusing feature maps from conv4 and conv5 layers. However, the aforementioned problem of a typical FCN is still inherent, and thus human annotation is further involved to achieve higher performance.

Our proposed method is fundamentally different from the previous approaches. We do not focus on determining aggregation methods but on finding more comprehensive features of objects. Thus, we are able to train the network to collect class-related regions without prior knowledge about the object size. Also, our approach relies on no external module that requires lower-than-image-level annotations.

## 3. Two-Phase Learning

This section describes the dataset and the baseline network architecture used in our approach. We then go into detail about two-phase learning, which consists of the first phase learning, inference conditional feedback, and the second phase learning. We refer to each of the networks trained in the first and second phase learning as the first and the second network, respectively. Finally, we introduce the inference step where the two sets of heat maps obtained from both networks are combined.

### 3.1. Dataset

We train on the Pascal VOC 2012 datasets. In practice, we use *trainaug* part with 10,582 weakly annotated images of Pascal VOC 2012, as configured by [14]. The input images are rescaled to $321 \times 321$, as in [18].

### 3.2. Baseline Architecture

The dominant paradigm of weakly supervised learning for object localization is to use a FCN with global pooling. The network is trained only by image-level supervision and generates heat maps for each class at the last convolutional layer. The global pooling layer then aggregates the heat maps for each class to compare with image-level labels.

Among a number of weakly supervised FCNs, we build on a particular FCN proposed in [42]. It is basically a modified VGG [31] variant, where fc6 and fc7 are converted to conv6 and conv7 and randomly initialized. GAP and a 20-way fully-connected layer are followed, and also pool4 and pool5 are removed.

This network has been imported as a component into one of the state-of-the-art techniques for weakly supervised semantic segmentation. Therefore, it is convenient to manifest the effect of our contribution by simply replacing the component with ours, and testing the segmentation performance of the entire system. Note that, however, our approach is not especially dependent on this very architecture, but can be applied to any types of existing FCNs.

### 3.3. First Phase Learning

In the first phase, a FCN is trained with a multi-label logistic loss for 20 foreground classes. The network is opti-

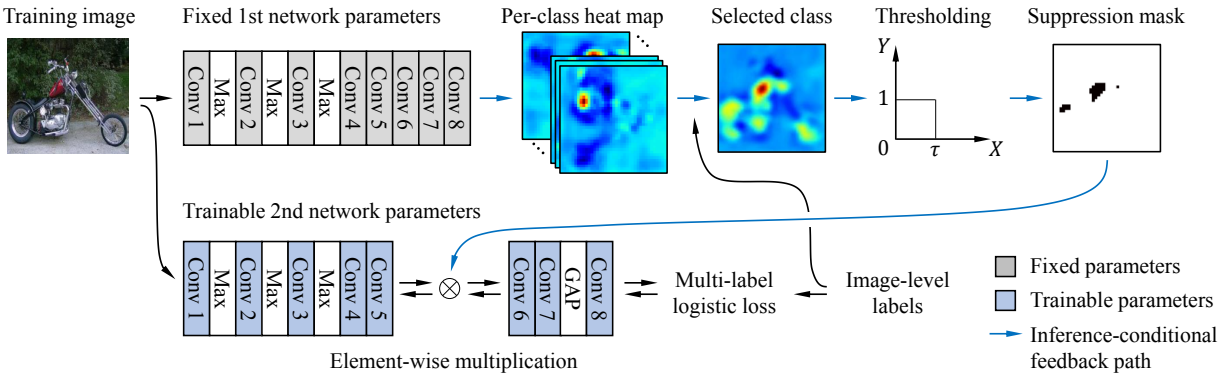


Figure 2: The second phase learning. The overall process of inference conditional feedback is marked as blue arrows: The first network (with fixed layers, colored in gray) takes an input image and outputs heat maps. Only the heat maps corresponding to the classes present in the image labels are selected, and become a suppression mask after applying thresholding. The suppression mask is then element-wise multiplied with the conv5-3 output of the second network (with trainable layers, colored in blue). The forward and backward passes are marked as black arrows.

mized via stochastic gradient descent (SGD) for 8,000 iterations, with a batch size of 15 and a weight decay of 0.0005. The learning rate is initially set to 0.001 and is reduced by a factor of 10 every 2000 iterations. At the inference stage, the network outputs class-specific heat maps; see [42] for details.

### 3.4. Inference Conditional Feedback

The inference conditional feedback suppresses neurons not to fire repeatedly on the locations that had high activations in the first network. In order to realize this, we design a suppression mask to block the first highlighted regions during training. First, out of the 20 heat maps from the first network, we select only the heat maps that are relevant to a given image-level label. We then apply an inverse rectification: for each selected heat map, we apply hard thresholding by 60% of the maximum value. In practice, we assign a value of zero to pixels above the threshold and one otherwise, as

$$M^c_{supp,u} = \begin{cases} 0, & \text{if} \quad H^c_u > 0.6 \cdot max(H^c_*) \\ 1, & \text{otherwise} \end{cases}, \tag{1}$$

where $M^c_{supp,u} \in \mathbb{R}^{41 \times 41}$, $M^c_u \in \mathbb{R}^{41 \times 41}$, $u$ and $c$ denote the binary suppression mask, per-class heat map, pixel position, and the indices of the classes present, respectively.

If there are multiple categories present, and consequently multiple binary suppression masks, they are combined by a logical *AND* operation as

$$M_{supp,u} = \prod_c M^c_{supp,u}. \tag{2}$$

Finally, a resulting binary suppression mask $M_{supp,u}$ is fed to the second network to suppress neurons from being activated at the same locations as in the first network.

### 3.5. Second Phase Learning

During the second phase, the first network with its fixed parameters is considered as a *function* that takes an image as input and produces a binary suppression mask as output. Fig. 2 illustrates how this suppression mask is fed back into the training of the second network. Here, the second network has the same architecture as the first network.

As noted in [13], all the layers up to the conv5-3 layer are regarded as feature extractors, where they learn the class-tuned representations. Based on the insight, we believe that it is semantically most appropriate to apply the feedback just after the conv5-3 layer of the second network. In practice, a suppression mask is multiplied with each channel of the conv5-3 output, element-wise. Thus, the forward pass with suppression mask is given as

$$C'^k_u = M_{supp,u} \cdot C^k_u \quad \forall \quad k, \tag{3}$$

where $C^k_u \in \mathbb{R}^{41 \times 41}$ and $C'^k_u \in \mathbb{R}^{41 \times 41}$ denote activations before and after applying the suppression to the conv5-3 output, and $k$ denotes each channel of the conv5-3 output. Similarly, backward pass is given as

$$\frac{\partial L}{\partial C^k_u} = M_{supp,u} \cdot \frac{\partial L}{\partial C'^k_u} \quad \forall \quad k, \tag{4}$$

where $L$ denotes the output loss. During the forward pass and backward update, the suppressed pixels are ignored. In other words, from the conv5-3 layer, activations on the previously important regions are dropped out by the feedback during the second phase.

The second network is subsequently trained to do image-level classification without the feature information that was most discriminative in the first phase. In this manner, the second network focuses on new features that can still be used to distinguish categories, and thus reveals more regions that were not highlighted in the first phase.

We can further think of the third or more phases using the next inference conditional feedback by lowering the threshold. However, as shown in Table 3, the localization performance gradually decreases as the phase proceeds (the threshold of 40% is used for the third phase). Therefore, only two phases of learning are considered throughout the applications of our approach.

### 3.6. Inference

At the inference stage, the feedback is not defined. The first and second networks produce two sets of heat maps each in a single forward pass. The implementations on how to combine the two sets of heat maps will vary depending on the applications, as we will explain in Sec. 4, Sec. 5, and Sec. 6.

## 4. Semantic Segmentation Experiments

In the task of semantic segmentation, each pixel in the image is classified into one of 21 categories including the background. However, in a weakly supervised setting, the network cannot explicitly learn the information about object boundaries or sizes. Therefore, to successfully perform this task, it is essential to initially retrieve accurate localization cues. Most techniques for weakly supervised segmentation internally train FCNs and obtain localization cues from the heat maps for each category.

The heat maps obtained via two-phase learning can cover not only the most discriminative parts of objects but also the whole parts. Thus, the quality of our localization cues is enhanced, and the performance of semantic segmentation is also increased accordingly. In order to verify this, we apply our two-phase learning algorithm to the SEC model [18], one of the state-of-the-art methods for weakly supervised semantic segmentation.

In this section, we briefly review our baseline segmentation network, SEC, and describe how the localization cues are complemented via two-phase learning. we then experiment on semantic segmentation using the localization cues. Finally, we report and analyze the results.

### 4.1. Review of SEC Architecture

As introduced in [18], SEC stands for *seed*, *expand*, and *constrain*. They are referred to as three important principles in weakly supervised semantic segmentation. First, a *seed* is a module to provide localization cues to the main segmentation network. The segmentation network is implicitly supervised to match the retrieved localization cues. Next, *expand* considers how to aggregate heat maps into image-level scores. It encourages the responses on promising locations to be high and to be consistent with image-level labels. As a new pooling strategy, global weighted rank pooling (GWRP) is proposed in order to recover the spatial information that will be lost in the aggregation process. Lastly, *constrain* is a module that constrains the results of the segmentation networks to follow the boundaries of objects. In practice, fully-connected conditional random fields (dense CRF) [33] are used.

### 4.2. Two-phase Learning for Localization Cues

A set of localization cues, *seed*, is *a cornerstone* for a segmentation network to build on. In the context of the SEC model [18], the localization cues refer to a set of class-specific binary masks that are obtained by a thresholding operation: for each per-class heat map, all pixels with a score larger than 20% of the maximum score are selected.

The localization cues obtained using heat maps from a conventional FCN are considered reliable only for the object positions, so they remain *weak*, as noted in [25, 5, 18]. With our proposed two-phase learning, the heat maps become more comprehensive. As a result, the localization cues for semantic segmentation also become more *powerful*.

In practice, we have two sets of heat maps from the first and second networks. In order to integrate the information on object regions, we merge the two heat maps via *weighted map voting*, which will be described in detail in Sec. 4.3. For the background class, as in [18], we imported the network implementation proposed in [30], which generates class-agnostic saliency detection based on the image gradient. The inferred localization cues are used to supervise semantic segmentation task.

### 4.3. Merging Heat Maps from Both Phases

To effectively combine two heat maps, we consider a simple post-processing technique, *weighted map voting*. We assume that a per-class probability score given by the network represents how confident the network is about the heat map of the same class. That is, if the first network predicts a high probability for a specific class, the information in the corresponding heat map is more confident than that of the second network, which predicts a lower probability for the same class.

Following this insight, *weighted map voting* is integrated into the system by multiplying the per-class heat maps $H^c$ by its class probability scores $p^c$. We then merge the resulting maps by taking the pixel-wise maximal values between the two multiplications, that is:

$$H^c_u = max(p^c_{1st} * H^c_{1st,u}, p^c_{2nd} * H^c_{2nd,u}), \tag{5}$$

| Method | bg | plane | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | motor | person | plant | sheep | sofa | train | tv | mIoU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Semi supervised:** | | | | | | | | | | | | | | | | | | | | | | |
| MIL+seg [25] | 79.6 | 50.2 | 21.6 | 41.6 | 34.9 | 40.5 | 45.9 | 51.5 | 60.6 | 12.6 | 51.2 | 11.6 | 56.8 | 52.9 | 44.8 | 42.7 | 31.2 | 55.4 | 21.5 | 38.8 | 36.9 | 42.0 |
| MIL+bbox [25] | 78.6 | 46.9 | 18.6 | 27.9 | 30.7 | 38.4 | 44.0 | 49.6 | 49.8 | 11.6 | 44.7 | 14.6 | 50.4 | 44.7 | 40.8 | 38.5 | 26.0 | 45.0 | 20.5 | 36.9 | 34.8 | 37.8 |
| STC [38] | 84.5 | 68.0 | 19.5 | 60.5 | 42.5 | 44.8 | 68.4 | 64.0 | 64.8 | 14.5 | 52.0 | 22.8 | 58.0 | 55.3 | 57.8 | 60.5 | 40.6 | 56.7 | 23.0 | 57.1 | 31.2 | 49.8 |
| CheckMask [28] | 86.4 | 70.1 | 21.7 | 53.1 | 52.5 | 50.7 | 70.9 | 66.6 | 63.2 | 16.9 | 45.8 | 39.1 | 61.1 | 50.0 | 56.8 | 56.2 | 40.0 | 51.9 | 29.3 | 63.1 | 35.9 | 51.5 |
| **Weakly supervised:** | | | | | | | | | | | | | | | | | | | | | | |
| EM-Adapt [23] | 67.2 | 29.2 | 17.6 | 28.6 | 22.2 | 29.6 | 47.0 | 44.0 | 44.2 | 14.6 | 35.1 | 24.9 | 41.0 | 34.8 | 41.6 | 32.1 | 24.8 | 37.4 | 24.0 | 38.1 | 31.6 | 33.8 |
| CCNN [24] | 68.5 | 25.5 | 18.0 | 25.4 | 20.2 | 36.3 | 46.8 | 47.1 | 48.0 | 15.8 | 37.9 | 21.0 | 44.5 | 34.5 | 46.2 | 40.7 | 30.4 | 36.3 | 22.2 | 38.8 | 36.9 | 35.3 |
| MIL+sppxl [25] | 77.2 | 37.3 | 18.4 | 25.4 | 28.2 | 31.9 | 41.6 | 48.1 | 50.7 | 12.7 | 45.7 | 14.6 | 50.9 | 44.1 | 39.2 | 37.9 | 28.3 | 44.0 | 19.6 | 37.6 | 35.0 | 36.6 |
| CheckMask-tags [28] | 79.2 | 60.1 | 20.4 | 50.7 | **41.2** | **46.3** | 62.6 | 49.2 | 62.3 | 13.3 | 49.7 | **38.1** | 58.4 | 49.0 | 57.0 | 48.2 | 27.8 | 55.1 | 29.6 | **54.6** | 26.6 | 46.6 |
| SEC (baseline) [18] | 82.4 | **62.9** | **26.4** | 61.6 | 27.6 | 38.1 | 66.6 | 62.7 | 75.2 | **22.1** | 53.5 | 28.3 | 65.8 | 57.8 | **62.5** | 52.5 | 32.5 | 62.6 | 32.1 | 45.4 | 45.3 | 50.7 |
| Ours | **82.8** | 62.2 | 23.1 | **65.8** | 21.1 | 43.1 | **71.1** | **66.2** | **76.1** | 21.3 | **59.6** | 35.1 | **70.2** | **58.8** | 62.3 | **66.1** | **35.8** | **69.9** | **33.4** | 45.9 | **45.6** | **53.1** |

Table 1: Comparison of weakly supervised semantic segmentation methods on VOC 2012 *segmentation, val.* set.

| Method | bg | plane | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | motor | person | plant | sheep | sofa | train | tv | mIoU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Semi supervised:** | | | | | | | | | | | | | | | | | | | | | | |
| MIL+seg [25] | 78.7 | 48.0 | 21.2 | 31.1 | 28.4 | 35.1 | 51.4 | 55.5 | 52.8 | 7.8 | 56.2 | 19.9 | 53.8 | 50.3 | 40.0 | 38.6 | 27.8 | 51.8 | 24.7 | 33.3 | 46.3 | 40.6 |
| MIL+bbox [25] | 76.2 | 42.8 | 20.9 | 29.6 | 25.9 | 38.5 | 40.6 | 51.7 | 49.0 | 9.1 | 43.5 | 16.2 | 50.1 | 46.0 | 35.8 | 38.0 | 22.1 | 44.5 | 22.4 | 30.8 | 43.0 | 37.0 |
| STC [38] | 85.2 | 62.7 | 21.1 | 58.0 | 31.4 | 55.0 | 68.8 | 63.9 | 63.7 | 14.2 | 57.6 | 28.3 | 63.0 | 59.8 | 67.6 | 61.7 | 42.9 | 61.0 | 23.2 | 52.4 | 33.1 | 51.2 |
| CheckMask [28] | 87.4 | 65.7 | 26.0 | 64.2 | 43.7 | 53.2 | 72.6 | 63.6 | 59.5 | 17.1 | 48.0 | 43.7 | 61.2 | 52.0 | 69.3 | 54.8 | 43.0 | 50.3 | 34.6 | 59.2 | 42.0 | 52.9 |
| **Weakly supervised:** | | | | | | | | | | | | | | | | | | | | | | |
| EM-Adapt [23] | 76.3 | 37.1 | 21.9 | 41.6 | 26.1 | 38.5 | 50.8 | 44.9 | 48.9 | 16.7 | 40.8 | 29.4 | 47.1 | 45.8 | 54.8 | 28.2 | 30.0 | 44.0 | 29.2 | 34.3 | 46.0 | 39.6 |
| CCNN [24] | - | 24.2 | 19.9 | 26.3 | 18.6 | 38.1 | 51.7 | 42.9 | 48.2 | 15.6 | 37.2 | 18.3 | 43.0 | 38.2 | 52.2 | 40.0 | 33.8 | 36.0 | 21.6 | 33.4 | 38.3 | 35.6 |
| MIL+sppxl [25] | 74.7 | 38.8 | 19.8 | 27.5 | 21.7 | 32.8 | 40.0 | 50.1 | 47.1 | 7.2 | 44.8 | 15.8 | 49.4 | 47.3 | 36.6 | 36.4 | 24.3 | 44.5 | 21.0 | 31.5 | 41.3 | 35.8 |
| CheckMask-tags [28] | 80.3 | 57.5 | 24.1 | 66.9 | **31.7** | 43.0 | 67.5 | 48.6 | 56.7 | 12.6 | 50.9 | **42.6** | 59.4 | 52.9 | 65.0 | 44.8 | **41.3** | 51.1 | 33.7 | **44.4** | 33.2 | 48.0 |
| SEC (baseline) [18] | **83.5** | 56.4 | **28.5** | 64.1 | 23.6 | 46.5 | **70.6** | 58.5 | 71.3 | **23.2** | 54.0 | 28.0 | 68.1 | **62.1** | 70.0 | 55.0 | 38.4 | 58.0 | 39.9 | 38.4 | **48.3** | 51.7 |
| Ours | 83.4 | **62.2** | 26.4 | **71.8** | 18.2 | **49.5** | 66.5 | **63.8** | **73.4** | 19.0 | **56.6** | 35.7 | **69.3** | 61.3 | **71.7** | **69.2** | 39.1 | **66.3** | **44.8** | 35.9 | 45.5 | **53.8** |

Table 2: Comparison of weakly supervised semantic segmentation methods on VOC 2012 *segmentation, test.* set.

where the subscripts $u$ and $1st$ and $2nd$ denote the pixel position and the first and second networks, respectively.

### 4.4. Improving Segmentation Network

Our baseline segmentation network, SEC [18], performs best when trained with all three losses of *seed*, *expand*, and *constrain*. In its original form, it achieves an average intersection-over-union scores of 50.7%, which is 0.3% higher than the same network trained with only *seed* and *constrain* losses.

Since our two-phase learning enables the localization cues to cover wider object regions in addition to the first predicted locations, it provides the segmentation network with richer information for object localization. In other words, our heat maps are able to perform both *seeding* and *expanding* roles in their former sense. Therefore, we use the only *seed* and *constrain* loss terms to train the segmentation network whose localizing module is replaced by our improved method. At inference, the predicted segmentation masks are rescaled to the size of their original images and refined by dense CRF [33].

### 4.5. Evaluation

To evaluate the contribution of our two-phase learning on the semantic segmentation, we use the metric of intersection-over-union scores, following the protocol of Pascal VOC 2012 semantic segmentation challenge [11]. We evaluated the results on 1,449 images in the *validation* part of the Pascal VOC 2012 segmentation dataset.

### 4.6. Results and Discussion

Table 2 compares the numeric results of our approach with those of previous weakly supervised approaches. For reference, we also provide the results of other methods that utilize additional annotations. They require either additional data from *Flickr* and an external saliency detector pretrained by pixel-level supervision [38] or user clicks [28] or region proposals such as selective search and MCG [25]. Since they are not trained with purely image-level annotations, we refer to them as semi-supervised learning. In this regard, only EM-Adapt [23], CCNN [24], MIL+sppxl [25], CheckMask-tags [28], and SEC [18] would be fair comparisons with ours. Among them, we achieved the best mIoU scores 53.1% on VOC-val and 53.8% on VOC-test, which improve upon the SEC baseline by 2.4% and 2.1% for each set.

Fig. 5, Fig. 6, and Fig. 7 illustrate visual comparisons of the segmentation predicted by the baseline [18] and ours.

**Discovering More Object Regions** A chronic problem of weakly supervised semantic segmentation is that the segmentation covers only parts of objects. This is because their heat maps tend to focus only on the most discriminative parts, e.g. *a person's face*. In particular, when objects are cropped or partially occluded, the object is often totally ignored in the prediction. We observe that our two-phase learning is able to overcome this problem. On this level, Fig. 5 compares qualitative results of the baseline and ours. The segmentation network trained in our method cov-

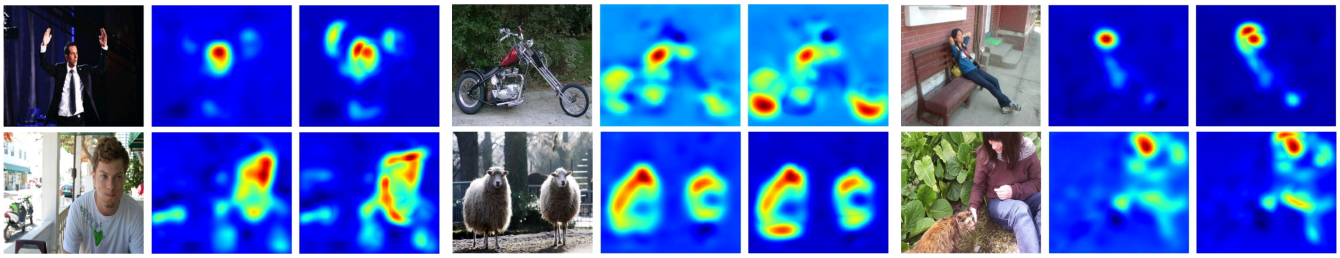

Figure 3: Object saliency detections using the first network (column 2,5, and 8) and proposed method (column 3,6, and 9).

ers more object regions than the baseline [18]. More specifically, it either discovers other parts of objects, e.g. *a torso, arms, and legs of a person*, or reveals new instances that have not been found before.

**Expanding up to Reasonable Extent** As mentioned in Sec. 4.1, various aggregation methods often fail to accurately estimate the extent of objects. This is because they enforce the network to *expand* to a certain degree. This often causes unreasonable expansions, as shown in Fig. 6. However, our approach is immune to this problem. The reason is that our system determines what additional features should be considered important. Therefore, the combination of the heat maps from the first and second networks does not simply widens the segmentation but also restricts it to fall inside the class-related regions. This method of propagation allows our approach to successfully remove the unreasonable expansions that happened in the baseline segmentation.

**Failure Cases** Like typical weakly supervised segmentation techniques, our segmentation also has a problem distinguishing objects that co-occur almost always, e.g. *trains vs. tracks*, as shown in Fig. 7. Another failure case arises rarely, when the newly found regions do not belong to the the predicted class, e.g. *plants but not potted*. We believe this is because the newly highlighted features in the second phase are sometimes not discriminative enough to exclude such confusing regions. This implies that the two-phase learning will have an upper bound on the degree to which the important parts are suppressed, as noted in Sec. 3.5.

**Scope** In order to demonstrate that our method can be applied to other semantic segmentation methods using heat maps, we applied our method to CCNN [24], and confirmed that the benefits of our method are consistent: Our approach achieves an mIoU score of 35.7% on VOC-val, outperforming the CCNN baseline which achieves 34.5% (what we could reproduce) by highlighting the second most important parts that are not found in the baseline. This implies that our two-phase learning is not limited to either the SEC model or the CAM [42] module, but is more generally applicable to other segmentation systems.

## 5. Per-Class Saliency Prediction Experiments

In this section, we demonstrate that the two sets of heat maps obtained via two-phase learning can synergize each other to capture the complete object. Here, we consider the heat maps as per-class saliency maps. Accordingly, we investigate whether those saliency maps are consistent with the ground truth segmentation masks. The two sets of heat maps are combined via *weighted map voting*, as given in Eq. (5).

### 5.1. Evaluation

In order to evaluate the quality of our heat maps, we only consider the heat maps whose corresponding class is present in the images. Similarly, we extract per-class saliency masks only for the classes present, from the ground truth segmentation. We use these as our ground truth saliency masks. In practice, 2,148 pairs of a per-class heat map and the ground truth saliency mask are collected for 1,440 images in Pascal VOC 2012 *val.* set. Each pixel in those heat maps has a response value that we consider as a confidence value, and we generate a precision-recall curve and compute the average precision (AP).

### 5.2. Results

A set of our heat maps combined via *weighted map voting* achieves an AP of 37.7%, which is 5.5% higher than that (32,5%) achieved using only the first heat maps. Fig. 3 illustrates the qualitative results: in our combined heat maps, the regions highlighted by both networks are revealed on object-relevant locations, e.g. *the hands of a person, wheels of a motorcycle*, and *a person's feet*.

## 6. Location Prediction Experiments

The proposed two-phase learning allows the second network to focus on the valuable features that have not been discovered in the first phase learning. In the previous section, we have shown that those newly revealed features can

| Phase | plane | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | motor | person | plant | sheep | sofa | train | tv | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Center | 86.0 | 56.6 | 64.8 | 41.6 | 18.0 | 82.5 | 30.0 | 87.5 | 23.3 | 73.9 | 24.5 | 75.3 | 83.1 | 65.9 | 54.2 | 17.6 | 66.1 | 52.1 | 78.4 | 30.3 | 55.6 |
| First | 98.7 | 94.4 | 93.2 | 88.5 | 67.2 | 93.6 | 81.3 | 99.0 | 65.0 | 94.5 | 67.4 | 96.7 | 98.8 | 95.9 | 92.6 | 72.0 | 98.5 | 88.8 | 92.1 | 83.8 | 88.1 |
| Second | 98.1 | 89.9 | 92.8 | 75.1 | 52.7 | 90.8 | 76.7 | 97.2 | 56.4 | 95.9 | 38.8 | 97.4 | 98.7 | 95.1 | 91.2 | 69.9 | 97.5 | 78.1 | 82.7 | 77.6 | 82.6 |
| Third | 94.6 | 89.3 | 88.5 | 38.0 | 32.8 | 86.0 | 65.2 | 96.4 | 31.7 | 93.9 | 24.8 | 95.1 | 93.2 | 89.1 | 71.2 | 27.2 | 92.1 | 43.4 | 92.3 | 64.8 | 70.5 |

Table 3: Object location prediction for each phase on VOC 2012 *main, val.* set.

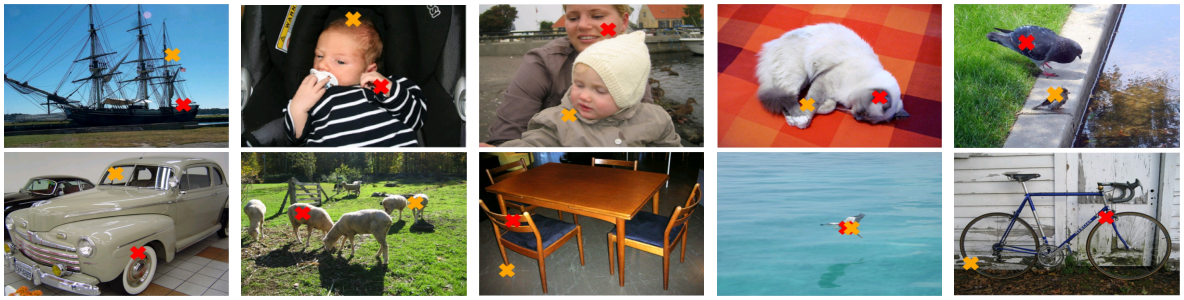

Figure 4: Object location predictions of the first (red) and second (orange) networks.

be combined with the first features to better capture the extent of objects. However, in this section, we also demonstrate that the different features highlighted by each of the first and second networks are semantically consistent with the distinctive parts of objects.

Here, we experiment on 5,823 images and the ground truth bounding boxes of the Pascal VOC 2012 main *val.* set.

### 6.1. Evaluation

In order to pinpoint the locations which the networks focus on, we consider the pixel of the maximal response of a per-class heat map as the predicted object location. For quantitative evaluation, we use the criteria introduced in [22]. First, the heat maps are rescaled to their original image size using bilinear interpolation. With 18-pixel tolerance, the predicted location within any ground truth bounding boxes of the target category is counted as correct and false negative otherwise, see [22] for details. For each image, for each class, the maximal response is considered as the confidence for the prediction, and this is then used to compute AP. Note that the heat maps from each network are not combined here but investigated separately because only the maximal value locations are considered.

Moreover, in order to confirm that the features that are considered important in both networks do not overlap, we measured the Euclidean pixel distance between the predicted locations of the first and second networks.

### 6.2. Results and Discussion

The per-class precisions of the location prediction for Pascal VOC are summarized in Table 3. To show the difficulty of the location prediction task, we report the performance of our naive baseline, *center*, which predicts the center of the image as the object location.

As it has been widely noted in the literature [42, 22, 30, 18] that weakly supervised FCNs reliably predict approximate positions of objects, our first network also successfully captures object locations, achieving an mAP of 88.1%. However, our second network is at a great disadvantage in predicting object locations because the most discriminative parts of objects have not been shown during training. Nevertheless, the second network was able to highlight the next most important parts with a small performance reduction of 5.5%, achieving an mAP of 82.6%.

Likewise, as shown in the previous experiments, the second network tends to highlight either different important parts of objects, e.g. *sails of a boat, pillars of a car*, or other instances even of small sizes, e.g. *a bird in front*. Also, even when the object region is small, it maintains the ability to predict the location, e.g. *a small bird flying*, implying that the second learned features are also representative of the object. Fig. 4 visualizes some pairs of predictions.

In most cases, two networks focus on different parts of images. The average Euclidean distance of the predictions of the two networks appeared to be 69 pixels. Considering that the average size of the images in the Pascal VOC 2012 dataset is $390 \times 470$, it is shown that the second network found fairly distant objects from those detected by the first network. Consequently, we demonstrate that different features highlighted in both networks can complement each other to localize objects.

## 7. Conclusion

Weakly supervised object localization has an inherent weakness that it often fails to capture the extent of objects because the network focuses only on the most distinc-

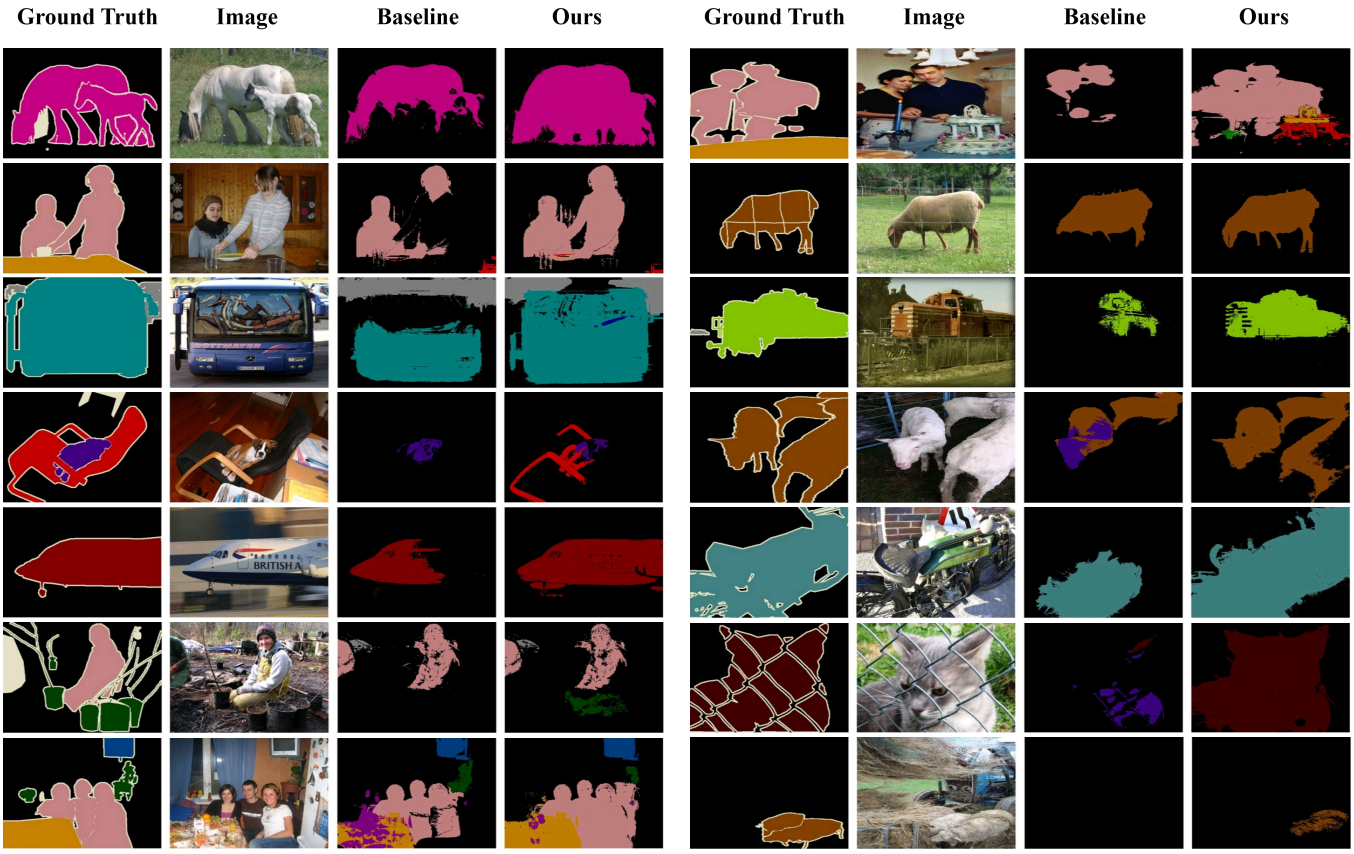


Figure 5: Qualitative segmentation results. Discovering more object regions (on VOC 2012 *segmentation, val.* set).

Figure 6: Qualitative segmentation results. Expanding up to reasonable extent (on VOC 2012 *segmentation, val.* set).

Figure 7: Qualitative segmentation results. Some failure cases (on VOC 2012 *segmentation, val.* set).

tive parts of the objects. In this paper, we propose a two-phase learning algorithm that can fundamentally mitigate this problem. We have been motivated by the insight that if we retrain the network while covering the most discriminative parts of the objects, it will highlight feature regions that are different from the first, while those features still fall inside the range of the objects. We propose inference conditional feedback in order to train an additional network in this manner. Finally, the heat maps of the first and second networks are combined to enhance object localization. Experiments on semantic segmentation, object saliency detection, and object location prediction tasks have shown the effectiveness of our two-phase learning on the challenging Pascal VOC 2012 dataset.

**Acknowledgements** This work was supported by DMC R&D Center of Samsung Electronics Co.

## References


[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 34(11):2189–2202, 2012. 2

[2] P. A. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marqués, and J. Malik. Multiscale combinatorial grouping. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[3] A. L. Bearman, O. Russakovsky, V. Ferrari, and F. Li. What's the point: Semantic segmentation with point supervision. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2016. 2

[4] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with convex clustering. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[5] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 4

[6] J. Carreira and C. Sminchisescu. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 34(7):1312–1328, 2012. 2

[7] 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 *Proc. of Int'l Conf. on Learning Representations (ICLR)*, 2015. 1

[8] M. Cheng, Z. Zhang, W. Lin, and P. H. S. Torr. BING: binarized normed gradients for objectness estimation at 300fps. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[9] H. Cholakkal, J. Johnson, and D. Rajan. Backtracking scspm image classifier for weakly supervised top-down saliency. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[10] R. G. Cinbis, J. J. Verbeek, and C. Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 39(1):189–203, 2017. 1

[11] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *Int'l Journal of Computer Vision (IJCV)*, 88(2):303–338, 2010. 5

[12] R. Girshick. Fast r-cnn. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, 2015. 1

[13] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2014. 1, 3

[14] B. Hariharan, P. Arbelaez, L. D. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, 2011. 2

[15] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li. Salient object detection: A discriminative regional feature integration approach. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2013. 2

[16] V. Kantorov, M. Oquab, M. Cho, and I. Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2016. 1

[17] A. Kolesnikov and C. H. Lampert. Improving weakly-supervised object localization by micro-annotation. *Proc. of British Machine Vision Conference (BMVC)*, 2016. 1

[18] A. Kolesnikov and C. H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2016. 1, 2, 4, 5, 6, 7

[19] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[20] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[21] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, 2015. 1

[22] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free? weakly-supervised learning with convolutional neural networks. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2, 7

[23] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille. Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, 2015. 1, 5

[24] D. Pathak, P. Krähenbühl, and T. Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *International Conference on Computer Vision (ICCV)*, 2015. 1, 5, 6

[25] P. O. Pinheiro and R. Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2, 4, 5

[26] X. Qi, Z. Liu, J. Shi, H. Zhao, and J. Jia. Augmented feedback in semantic segmentation under image level supervision. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2016. 2

[27] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proc. of Neural Information Processing Systems (NIPS)*, 2015. 1

[28] 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 *Proc. of European Conf. on Computer Vision (ECCV)*, 2016. 1, 2, 5

[29] M. Shi and V. Ferrari. Weakly supervised object localization using size estimates. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2016. 1

[30] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *Proc. of Int'l Conf. on Learning Representations (ICLR)*, 2014. 4, 7

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *Proc. of Int'l Conf. on Learning Representations (ICLR)*, 2014. 2

[32] C. Sun, M. Paluri, R. Collobert, R. Nevatia, and L. D. Bourdev. Pronet: Learning to propose object-specific boxes for cascaded neural networks. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[33] T. Toyoda and O. Hasegawa. Random field model for integration of local information and global information. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 30(8):1483–1489, 2008. 4, 5

[34] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *Int'l Journal of Computer Vision (IJCV)*, 104(2):154–171, 2013. 2

[35] M. Vasconcelos, N. Vasconcelos, and G. Carneiro. Weakly supervised top-down image segmentation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2006. 1

[36] A. Vezhnevets and J. M. Buhmann. Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2010. 1

[37] C. Wang, K. Huang, W. Ren, J. Zhang, and S. J. Maybank. Large-scale weakly supervised object localization via latent category learning. *IEEE Trans. Image Processing (TIP)*, 24(4):1371–1385, 2015. 1

[38] Y. Wei, X. Liang, Y. Chen, X. Shen, M. M. Cheng, J. Feng, Y. Zhao, and S. Yan. STC: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, PP(99):1–1, 2015. 1, 2, 5

[39] W. Xia, C. Domokos, J. Dong, L. Cheong, and S. Yan. Semantic segmentation without annotating segments. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, 2013. 1

[40] J. Xu, A. G. Schwing, and R. Urtasun. Tell me what you see and I will show you where it is. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2014. 1

[41] J. Zhang, Z. L. Lin, J. Brandt, X. Shen, and S. Sclaroff. Top-down neural attention by excitation backprop. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2016. 1

[42] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 3, 6, 7