Weakly-Supervised Semantic Segmentation by Iteratively Mining Common Object Features

Xiang Wang[1] Shaodi You[2,3] Xi Li[1] Huimin Ma[1]*
[1] Department of Electronic Engineering, Tsinghua University
[2] DATA61-CSIRO [3] Australian National University
{wangxiang14@mails., lixi16@mails., mhmpub@}tsinghua.edu.cn, shaodi.you@data61.csiro.au

Abstract

Weakly-supervised semantic segmentation under image tags supervision is a challenging task as it directly associates high-level semantic to low-level appearance. To bridge this gap, in this paper, we propose an iterative bottom-up and top-down framework which alternatively expands object regions and optimizes segmentation network. We start from initial localization produced by classification networks. While classification networks are only responsive to small and coarse discriminative object regions, we argue that, these regions contain significant common features about objects. So in the bottom-up step, we mine common object features from the initial localization and expand object regions with the mined features. To supplement non-discriminative regions, saliency maps are then considered under Bayesian framework to refine the object regions. Then in the top-down step, the refined object regions are used as supervision to train the segmentation network and to predict object masks. These object masks provide more accurate localization and contain more regions of object. Further, we take these object masks as initial localization and mine common object features from them. These processes are conducted iteratively to progressively produce fine object masks and optimize segmentation networks. Experimental results on Pascal VOC 2012 dataset demonstrate that the proposed method outperforms previous state-of-the-art methods by a large margin.

1. Introduction

Weakly-supervised semantic segmentation under image tags supervision is to perform a pixel-wise segmentation of an image, providing only the labels of existing semantic objects in the image. Because it relies on very slight human labeling, it benefits a number of computer vision tasks, such as object detection [8] and autonomous driving [3].

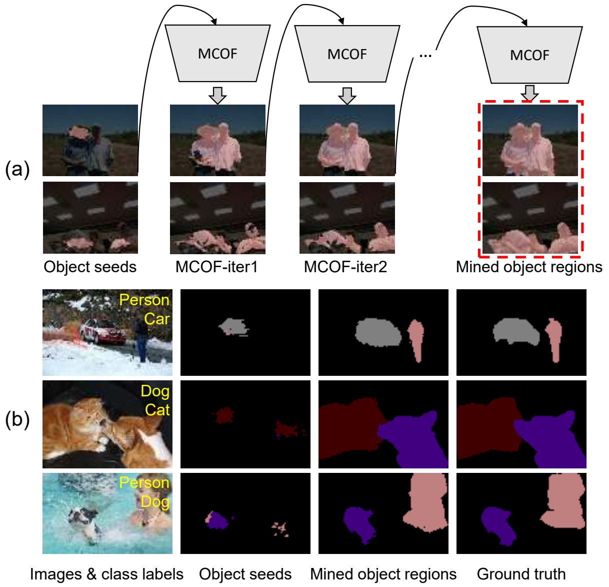

Figure 1. (a) Illustration of the proposed *MCOF* framework. Our framework iteratively mines common object features and expands object regions. (b) Examples of initial object seeds and our mined object regions. Our method can tolerate inaccurate initial localization and produce quite satisfactory results.

Weakly-supervised semantic segmentation is, however, very challenging as it directly associates high-level semantic to low-level appearance. Since only image tags are available, most previous works rely on classification networks to localize objects. However, while no pixel-wise annotation is available, classification networks can only produce inaccurate and coarse discriminative object regions, which can not meet the requirement of pixel-wise semantic segmentation, and thus harms the performance.

To address this issue, in this paper, we propose an iterative bottom-up and top-down framework, which tolerates inaccurate initial localization by Mining Common Object Features (*MCOF*) from initial localization to progressively expand object regions. Our motivation is, though the initial localization produced by classification network is coarse, it

*corresponding author

gives certain discriminative regions of objects, these regions contain important knowledge about objects, *i.e.* common object features. For example, as shown in Figure 1 (a), some images may locate hands of person, while other images may locate heads. Given a set of training images, we can learn common object features from them to predict regions of whole object. So in the bottom-up step, we take the initial object localization as object seeds and mine common object features from them to expand object regions. Then in the top-down step, we train segmentation network using the mined object regions as supervision to predict fine object masks. The predicted object masks contain more regions of objects, which are more accurate and provide more training samples of objects, so we can further mine common object features from them. And the processes above are conducted iteratively to progressively produce fine object regions and optimize segmentation networks. With iterations, inaccurate regions in the initial localization are progressively corrected, so our method is robust and can tolerate inaccurate initial localization. Figure 1 (b) shows some examples in which the initial localization is very coarse and inaccurate, while our method can still produce satisfactory results.

Concretely, we first train an image classification network and localize discriminative regions of object using Classification Activation Maps (CAM) [34]. Images are then segmented into superpixel regions and are assigned with class labels using CAM, these regions are called initial object seeds. The initial object seeds contain certain key parts of objects, so in bottom-up step, we mine common object features from them and then expand object regions. We achieve this by training a region classification network and use the trained network to predict object regions. While these regions may still only focus on key part regions of objects, to supplement non-discriminative regions, saliency-guided refinement method is proposed which considers both the expanded object regions and saliency maps under Bayesian framework. Then in top-down step, we train segmentation network using the refined object regions as supervision to predict segmentation masks. With the aforementioned procedure, we can get segmentation masks which contain more complete object regions and are much more accurate than the initial object seeds. We further take the segmentation masks as object seeds, and conduct the processes iteratively. With iterations, the proposed *MCOF* framework progressively produces more accurate object regions and enhances the performance of the segmentation network. The final trained segmentation network is applied for inference.

The main contributions of our work are three-fold:

- We propose an iterative bottom-up and top-down framework which tolerates inaccurate initial localization by iteratively mining common object features to progressively produce accurate object masks and optimize segmentation network.
- Saliency-guided refinement method is proposed to supplement non-discriminative regions which are ignored in initial localization.
- Experiments on PASCAL VOC 2012 segmentation dataset demonstrate that our method outperforms previous methods and achieves state-of-the-art performance.

2. Related Work

In this section, we introduce both fully-supervised and weakly-supervised semantic segmentation networks which are related to our work.

2.1. Fully-Supervised Semantic Segmentation

Fully-supervised methods acquire a large number of pixel-wise annotations, according to the process mode, they can be categorized as region-based and pixel-based networks.

Region-based networks take images as a set of regions and extract features of them to predict their labels. Mostajabi *et al*. [17] proposed zoom-out features which combines features of local, proximal, distant neighboring superpixels and the entire scene to classify each superpixel.

Pixel-based networks take the entire image as input and predict pixel-wise labels end-to-end with fully convolutional layers. Long *et al*. [16] proposed fully convolutional network (FCN) and skip architecture to produce accurate and detailed semantic segmentation. Chen *et al*. [2] proposed DeepLab which introduces “hole algorithm” to enlarge the receptive field with lower stride to produce denser segmentation. A large number of works [1, 18, 32] have been proposed based on FCN and DeepLab.

Pixel-based networks have been proved to be more powerful than Region-based networks for semantic segmentation. However, in this paper, we take advantages of both kinds of networks. We show that region-based networks are powerful in learning common features of objects and thus can produce fine object regions as supervision to train pixel-based networks.

2.2. Weakly-Supervised Semantic Segmentation

While fully-supervised methods require a large number of pixel-wise annotation which is very expensive, recent advances have exploited semantic segmentation with weak supervision, including bounding box [4, 19, 12], scribble [15] and image-level labels [21, 22, 25, 19, 31, 13, 23, 30]. In this paper, we only focus on the weakest supervision, *i.e.*, image-level supervision.

In image-level weakly-supervised semantic segmentation, since only image tags are available, most methods are based on classification methods, and these methods can be coarsely classified into two categories: *MIL-based* methods, which directly predict segmentation masks with classi-

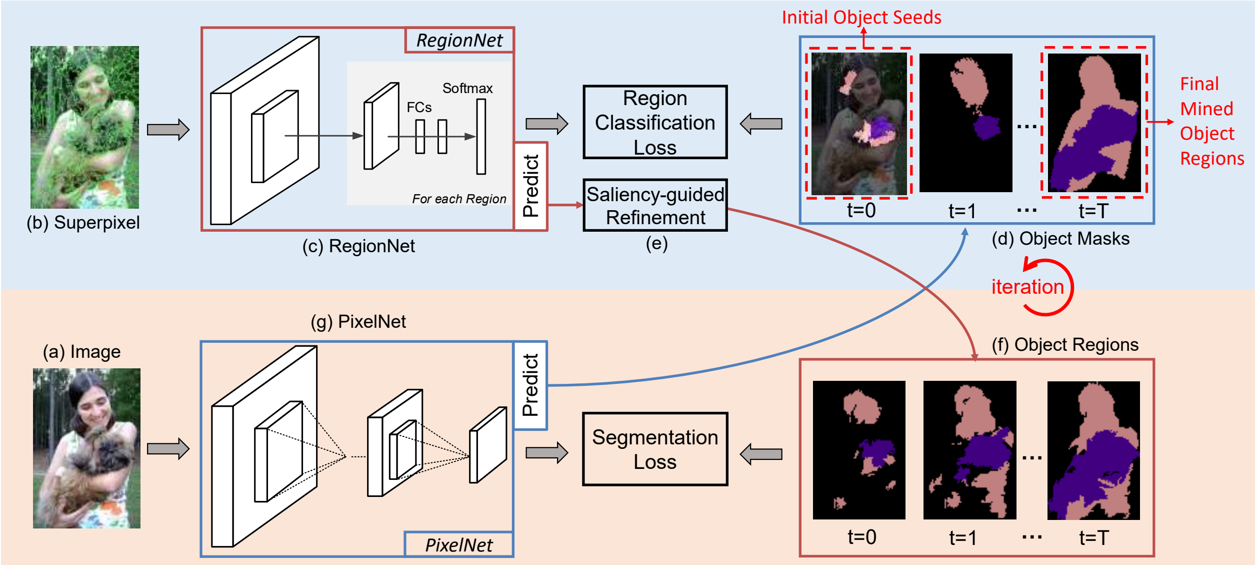

Figure 2. Pipeline of the proposed *MCOF* framework. At first (t=0), we mine common object features from initial object seeds. We segment (a) image into (b) superpixel regions and train the (c) region classification network *RegionNet* with the (d) initial object seeds. We then re-predict the training images regions with the trained *RegionNet* to get object regions. While the object regions may still only focus on discriminative regions of object, we address this by (e) saliency-guided refinement to get (f) refined object regions. The refined object regions are then used to train the (g) *PixelNet*. With the trained *PixelNet*, we re-predict the (d) segmentation masks of training images, are then used them as supervision to train the *RegionNet*, and the processes above are conducted iteratively. With the iterations, we can mine finer object regions and the *PixelNet* trained in the last iteration is used for inference.

fication networks; and *localization-based* methods, which utilize classification networks to produce initial localization and use them to supervise segmentation networks.

Multi-instance learning (MIL) based methods [21, 22, 13, 25, 5] formulate weakly-supervised learning as a MIL framework in which each image is known to have at least one pixel belonging to a certain class, and the task is to find these pixels. Pinheiro *et al*. [22] proposed Log-Sum-Exp (LSE) to pool the output feature maps into image-level labels, so that the network can be trained end-to-end as a classification task. Kolesnikov *et al*. [13] proposed global weighted rank pooling (GWRP) method which gives more weights to promising location in the last pooling layer. However, while MIL-based methods can locate discriminative object regions, they suffer from coarse object boundaries and thus the performance is not satisfactory.

Localization-based methods [19, 31, 13, 23, 30] aim to generate initial object localization from weak labels and then use it as supervision to train segmentation networks. Kolesnikov *et al*. [13] used localization cues generated from classification networks as a kind of supervision, they also proposed classification loss and boundary-aware loss to consider class and boundary constrain. Wei *et al*. [30] proposed adversarial erasing method to progressively mine object region with classification network. While Wei *et al*. [30] also aims to expand object regions from the initial localization. They rely on the classification network to sequentially produce the most discriminative regions in erased images. It will cause error accumulation and the mined object regions will have coarse object boundary. The proposed *MCOF* method mines common object features from coarse object seeds to predict finer segmentation masks, and then iteratively mines features from the predicted masks. Our method progressively expands object regions and corrects inaccurate regions, which is robust to noise and thus can tolerate inaccurate initial localization. By taking advantages of superpixel, the mined object regions will have clear boundary.

3. Architecture of the Proposed MCOF

Classification networks can only produce coarse and inaccurate discriminative object localization, which are far from the requirement of pixel-wise semantic segmentation. To address this issue, in this paper, we argue that, though the initial object localization is coarse, it contains important features about objects. So we propose to mine common object features from initial object seeds to progressively correct inaccurate regions and produce fine object regions to supervise segmentation network.

As shown in Figure 2, our framework consists of two iterative steps: bottom-up step and top-down step. The bottom-up step mines common object features from object seeds to produce fine object regions, and the top-down step uses the produced object regions to train weakly-supervised segmentation network. The predicted segmentation masks contain more complete object regions than initial. We then

Algorithm 1 Framework of the proposed *MCOF*

```
Input: Training images $\mathcal{I}$ and Superpixel regions $\mathcal{R}$
Initialize: Generate initial object seeds $\mathcal{S}$, $t = 0$.
while iteration is effective do
    Train the RegionNet with $\mathcal{R}$ and $\mathcal{S}$
    Predict with the trained RegionNet to get object regions $\mathcal{O}$.
    if $t == 0$ then
        Refine object regions $\mathcal{O}$ with saliency maps to get refined object regions $\mathcal{O}^R$
    else
        $\mathcal{O}^R \leftarrow \mathcal{O}$
    end if
    Train the PixelNet with $\mathcal{I}$ and $\mathcal{O}^R$
    Predict with the trained PixelNet to get object masks $\mathcal{M}$
    Update $\mathcal{S} \leftarrow \mathcal{M}, t \leftarrow t + 1$.
end while
Output: Mined object masks $\mathcal{M}$ and the trained PixelNet
```

take them as object seeds to mine common object features and the processes are conducted iteratively to progressively correct inaccurate regions and produce fine object regions.

Note that, in the first iteration, the initial object seeds only contain discriminative regions, after mining common object features, some non-discriminative regions are still missing. To address this, we propose to incorporate saliency maps with the mined object regions. After the first iteration, the segmented masks contain much more object regions and are more accurate, while the accuracy of saliency maps are also limited, so in the later iterations, the saliency maps are not used to prevent introducing additional noise. The overall procedure is summarized as Algorithm 1.

It is worth noting that the iterative processes are only applied in the training stage, for inference, only the segmentation network of the last iteration is utilized, so the inference is efficient.

4. Mining Common Object Features

4.1. Initial Object Seeds

To get initial object localization, we train a classification network and use CAM method [34] to produce heatmap of each object. As shown in Figure 3, the heatmap is very coarse, to localize discriminative regions of objects, first, we segment images into superpixel regions using graph-based segmentation method [7] and average the heatmap within each region. We observe that the CAM map usually has several center regions with low-confidence regions surrounding them, and the center regions are mostly the key part of objects. So for each heatmap, we select its local maximum region as initial seeds. However, this may miss lots of regions, so regions with heatmap larger than a threshold are also selected as initial seeds. Some examples are shown in Figure 3.

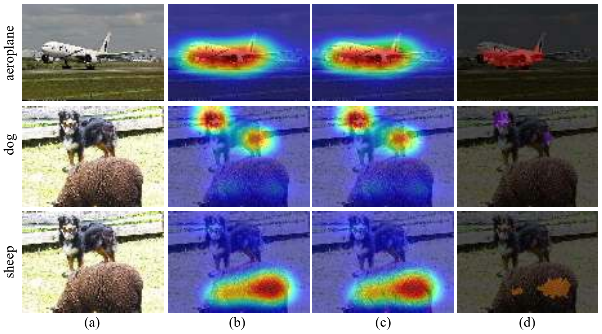

Figure 3. Examples of generating initial object seeds from object heatmaps. (a) Images, (b) object heatmaps of CAM [34], (c) object heatmaps averaged in each superpixel, (d) initial object seeds.

4.2. Mining Common Object Features from Initial Object Seeds

The initial object seeds are too coarse to meet the requirement of semantic segmentation, however, they contain discriminative regions of objects. For example, as shown in Figure 4, one image may locate hands of a person, while another may give the location of face. We argue that, regions of same class have some shared attributions, namely, common object features. So given a set of training images with seed regions, we can learn common object features from them and predict the whole regions of object, thus to expand object regions and suppress noisy regions. We achieve this by training a region classification network, named *RegionNet*, using the object seeds as training data.

Formally, given N training images $\mathcal{I} = \{I_i\}_{i=1}^{N}$, we first segment them into superpixel regions $\mathcal{R} = \{R_{i,j}\}_{i=1,j=1}^{N,n_i}$ using graph-based segmentation method [7], where n_i is the number of superpixel regions of the image I_i. In Sec 4.1, we have got initial object seeds, with them, we can give labels for superpixel regions $\mathcal{R}$ and denote them as $\mathcal{S} = \{S_{i,j}\}_{i=1,j=1}^{N,n_i}$, where $S_{i,j}$ is one-hot encoding with $S_{i,j}(c) = 1$ and others as 0 if $R_{i,j}$ belongs to class c. Based on training data $\mathcal{D} = \{(R_{i,j}, S_{i,j})\}_{i=1,j=1}^{N,n_i}$, our goal is to train a region classification network $f^r(R; \theta_r)$ parameterized by θ_r to model the probability of region $R_{i,j}$ being class label c , namely, $f_c^r(R_{i,j}|\theta_r) = p(y = c|R_{i,j})$.

We achieve this with the efficient mask-based Fast R-CNN framework [9, 28, 29]. In this framework, we take *external rectangle* of each region as the *RoI* of the original Fast R-CNN framework. In the *RoI* pooling layer, features inside superpixel regions are pooled while features inside the external rectangle but outside the superpixel regions are pooled as zero. To train this network, we minimize the

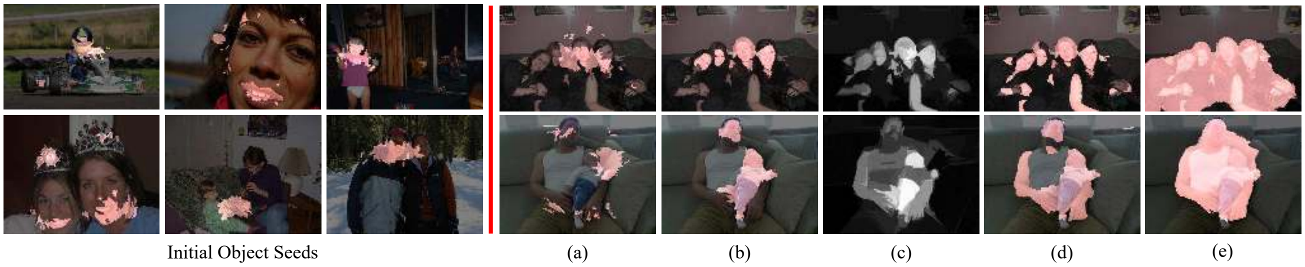

Figure 4. Left: examples of object seeds. They give us features of objects of different locations. However, they mainly focus on key parts which are helpful for recognition. Right: (a) initial object seeds, (b) object masks predicted by *RegionNet*, (c) saliency map, (d) refined object regions via Bayesian framework, (e) segmentation results of *PixelNet*.

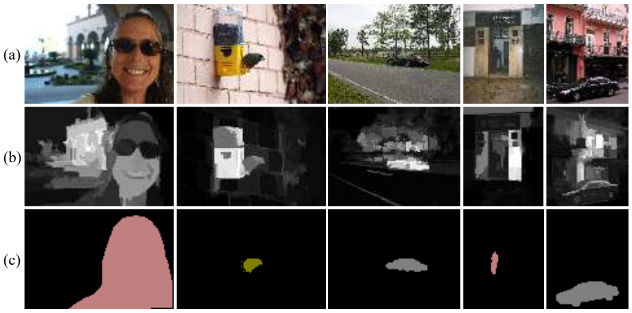

Figure 5. For images with single object class, salient object regions may not be consistent with semantic segmentation. In addition, they may be inaccurate and may locate other objects which are not included in semantic segmentation datasets. (a) Images, (b) saliency map of DRFI [11], (c) semantic segmentation.

cross-entropy loss function:

$$\mathcal{L}_r = -\sum_{i,j,c} S_{i,j}(c) log(f_c^r(R_{i,j}|\theta_r)). \tag{1}$$

By training the *RegionNet*, common object features can be mined from the initial object seeds. We then use the trained network to predict the label of each region of the training images. In the prediction, some incorrect regions and regions initially labeled as background can be classified correctly, thus to expand object regions. Some examples are shown in Figure 4 (a) and (b), we can see that object regions predicted by *RegionNet* contain more regions of objects and some noisy regions in initial object seeds are corrected. In this paper, we call these regions as object regions and denote them as $\mathcal{O} = \{O_i\}_{i=1}^N$.

Note that since we have the class labels of training images, we can remove wrong predictions and label them as background. This will guarantee that the produced object regions do not contain any non-existent class, which is important for training the following segmentation network.

4.3. Saliency-Guided Object Region Supplement

Note that the *RegionNet* is learned from the initial seed regions which mainly contain key regions of objects. With the *RegionNet*, the object regions can be expanded while there still exists some regions that are ignored. For example, the initial seed regions mainly focus on heads and hands of a person, while other regions, such as the body, are often ignored. After expanding by *RegionNet*, some regions of the body are still missing (Figure 4 (b)).

To address this issue, we propose to supplement object regions by incorporating saliency maps for images with single object class. Note that we do not directly use saliency map as initial localization as previous works [31], since in some cases, salient object may not be the object class we need in semantic segmentation, and the saliency map itself also contains noisy regions which will affect the localization accuracy. Some examples are shown in Figure 5.

We address this by proposing saliency-guided object region supplement method which considers both the mined object regions and saliency maps under Bayesian framework. In Sec 4.2, we have mined object regions which contains key parts of objects. Based on these key parts, we aim to supplement object regions with saliency maps. Our idea is, for a region with high saliency value, if it's similar with the mined object objects, then it is more likely to be part of that object. We can formulate the above hypothesis under Bayesian optimization [33, 27] as:

$$p(obj|\boldsymbol{v}) = \frac{p(obj)p(\boldsymbol{v}|obj)}{p(obj)p(\boldsymbol{v}|obj) + p(bg)p(\boldsymbol{v}|bg)}, \tag{2}$$

where $p(obj)$ is the saliency map, and $p(bg) = 1 - p(obj)$, $p(\boldsymbol{v}|obj)$ and $p(\boldsymbol{v}|bg)$ are the feature distribution at object regions and background regions, $\boldsymbol{v}$ is the feature vector, $p(obj|\boldsymbol{v})$ is the refined object map which represents the probability of region with feature $\boldsymbol{v}$ being object. By binarizing the refined object map $p(obj|\boldsymbol{v})$ with a CRF [14], we can get refined object regions which incorporate saliency maps to supplement the original object regions. In our work, we use saliency map of the DRFI method [11] as in [31].

Some examples are shown in Figure 4, by incorporating saliency maps, more object regions are included. In this paper, we call these regions as refined object regions and denote them as $\mathcal{O}^R = \{O_i^R\}_{i=1}^N$.

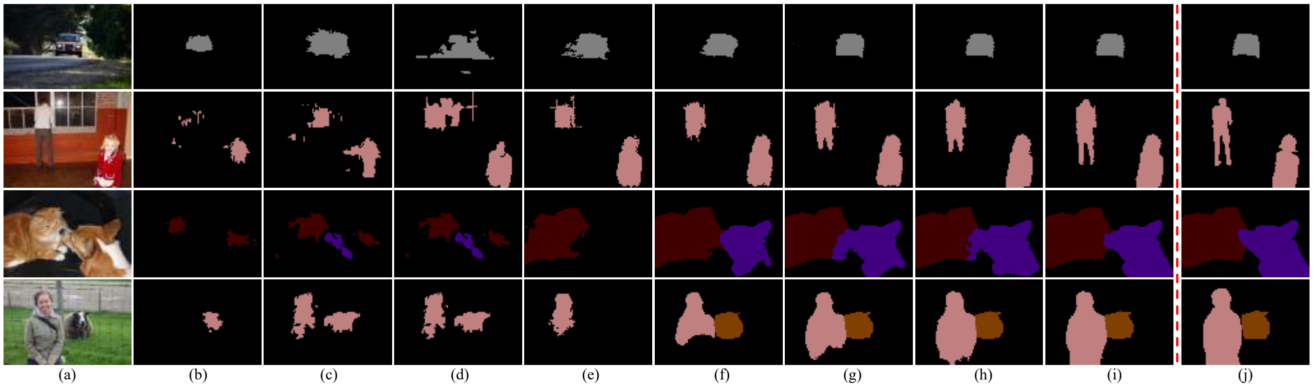

Figure 6. Intermediate results of the proposed framework. (a) Image, (b) initial object seeds, (c) expanded object regions predicted by *RegionNet*, (d) saliency-guided refined object regions. Note that, the saliency-guided refinement is only applied to images with single class, for images with multiple classes (3rd and 4th rows), the object regions remain unchanged. Segmentation results of *PixelNet* in (e) 1st, (f) 2nd, (g) 3rd, (h) 4th and (i) 5th iteration, (j) ground truth.

5. Iterative Learning Framework

The refined object regions give us some reliable localization of object, we can use them as supervision to train the weakly-supervised semantic segmentation network. While previous works [13, 30, 5] rely on both localization cues and class labels to design and train segmentation network, in our work, we have removed wrong class regions in the previous *RegionNet*, thus the refined object regions do not contain any wrong class. So we can only use the localization cues as supervision, this is completely compatible with fully-supervised framework, and thus we can benefit from existing fully-supervised architecture. In this paper, we utilize the popular DeepLab-LargeFOV model [2] as the basic network of our segmentation network, named *PixelNet*.

Formally, given the training images $\mathcal{I} = \{I_i\}_{i=1}^N$ and corresponding refined object regions $\mathcal{O}^R = \{O_i^R\}_{i=1}^N$, our goal is to train the segmentation network $f^s(I;\theta_s)$ parameterized by θ_s to model the probability that location u being the class label c, namely, $f^s_{u,c}(I|\theta_s) = p(y_u = c|I)$. The loss function is the cross-entropy loss which encourages the predictions to match our refined object regions:

$$\mathcal{L}_s = -\frac{1}{\sum_{c=1}^{C}|S_c|}\sum_{c=1}^{C}\sum_{u\in S_c} log(f^s_{u,c}(I|\theta_s)), \qquad (3)$$

where C is the number of classes and S_c is a set of locations that are labelled with class c in the supervision.

The supervision cues, namely, the object regions, is produced by the region classification network, it only considers features inside each region. While in the *PixelNet*, the whole image is considered and thus the context information is utilized. Using the trained *PixelNet* to predict the segmentation masks of the training images, the segmentation masks will further include more object regions. Some examples are shown in Figure 4, we can see that the predicted segmentation masks locate more regions of objects and suppress the noisy regions in the previous steps.

Further, we take the predicted segmentation masks as object seeds and conduct the processes above iteratively. With iterations, more robust common object features can be mined thus to produce finer object regions, and the segmentation network is progressively optimized with better supervision. Figure 6 shows the results with iterations. With iterations, the object regions are expanded and the inaccurate regions are corrected, so the segmentation results become more and more accurate. Finally, we use the trained *PixelNet* of the last iteration for inference and evaluate it in the experiment section.

6. Experiments

6.1. Setup

We evaluate the proposed *MCOF* framework on the PASCAL VOC 2012 image segmentation benchmark [6] *. The dataset contains 20 object classes and 1 background class. For the segmentation task, it contains 1464 training, 1449 validation and 1456 test images. Following previous works [13, 23, 30], we use the augmentation data [10] which contains 10,582 images as training set. We evaluate our method and compare with other methods on validation and test sets for segmentation task in terms of intersection-over-union averaged on all 21 classes (mIoU).

6.2. Comparison with State-of-the-art Methods

We compare our method with previous state-of-the-art image-level weakly-supervised semantic segmentation methods: CCNN [20], EM-Adapt [19], MIL-sppxl [22], STC [31], DCSM [26], BFBP [25], AF-SS [23], SEC [13], CBTS [24] and AE-PSL [30]. As we mentioned above, our *PixelNet* is completely compatible with fully-supervised framework and thus we can benefit from existing fully-supervised architecture. In this paper, we utilize DeepLab-LargeFOV [2] built on top of both VGG16 and ResNet101

*The models and results are available at https://wangxiang10.github.io/

Method	bkg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	**mIoU**
CCNN (ICCV'15) [20]	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
EM-Adapt (ICCV'15) [19]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	38.2
MIL-sppxl (CVPR'15) [22]	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
STC (PAMI'16) [31]	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
DCSM (ECCV'16) [26]	76.7	45.1	24.6	40.8	23.0	34.8	61.0	51.9	52.4	15.5	45.9	32.7	54.9	48.6	57.4	51.8	38.2	55.4	32.2	42.6	39.6	44.1
BFBP (ECCV'16) [25]	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
AF-SS (ECCV'16) [23]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	52.6
SEC (ECCV'16) [13]	82.2	61.7	26.0	60.4	25.6	45.6	70.9	63.2	72.2	20.9	52.9	30.6	62.8	56.8	63.5	57.1	32.2	60.6	32.3	44.8	42.3	50.7
CBTS (CVPR'17) [24]	85.8	65.2	29.4	63.8	31.2	37.2	69.6	64.3	76.2	21.4	56.3	29.8	68.2	60.6	66.2	55.8	30.8	66.1	34.9	48.8	47.1	52.8
AE-PSL (CVPR'17) [30]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	55.0
Ours:																						
MCOF-VGG16	85.8	74.1	23.6	66.4	36.6	62.0	75.5	68.5	78.2	18.8	64.6	29.6	72.5	61.6	63.1	55.5	37.7	65.8	32.4	68.4	39.9	**56.2**
MCOF-ResNet101	87.0	78.4	29.4	68.0	44.0	67.3	80.3	74.1	82.2	21.1	70.7	28.2	73.2	71.5	67.2	53.0	47.7	74.5	32.4	71.0	45.8	**60.3**

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

Method	bkg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	**mIoU**
CCNN (ICCV'15) [20]	70.1	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
EM-Adapt (ICCV'15) [19]	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
MIL-sppxl (CVPR'15) [22]	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
STC (PAMI'16) [31]	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
DCSM (ECCV'16) [26]	78.1	43.8	26.3	49.8	19.5	40.3	61.6	53.9	52.7	13.7	47.3	34.8	50.3	48.9	69.0	49.7	38.4	57.1	34.0	38.0	40.0	45.1
BFBP (ECCV'16) [25]	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
AF-SS (ECCV'16) [23]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	52.7
SEC (ECCV'16) [13]	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
CBTS (CVPR'17) [24]	85.7	58.8	30.5	67.6	24.7	44.7	74.8	61.8	73.7	22.9	57.4	27.5	71.3	64.8	72.4	57.3	37.0	60.4	42.8	42.2	50.6	53.7
AE-PSL (CVPR'17) [30]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	55.7
Ours:																						
MCOF-VGG16	86.8	73.4	26.6	60.6	31.8	56.3	76.0	68.9	79.4	18.8	62.0	36.9	74.5	66.9	74.9	58.1	44.6	68.3	36.2	64.2	44.0	**57.6**
MCOF-ResNet101	88.2	80.8	31.4	70.9	34.9	65.7	83.5	75.1	79.0	22.0	70.3	31.7	77.7	72.9	77.1	56.9	41.8	74.9	36.6	71.2	42.6	**61.2**

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

as *PixelNet*. Table 1 and Table 2 show the comparison on mIoU on PASCAL VOC 2012 validation and test sets, respectively. We can see that our method outperforms previous methods by a large margin and achieves new state-of-the-art. When using VGG16 as basic network (MCOF-VGG16), our method outperforms the second best method, AE-PSL [30] by 1.2% and 1.9% on *val* and *test* sets, respectively. And when using the more powerful ResNet101 (MCOF-ResNet101), the improvement can reach 5.3% and 5.5%, respectively. For the training samples, MIL-sppxl [22] used 700K images and STC [31] used 50K images, our method and other methods use 10K images. We also show some qualitative segmentation results of the proposed framework in Figure 7, we can see that our weakly-supervised method can produce quite satisfactory segmentation, even in complex images.

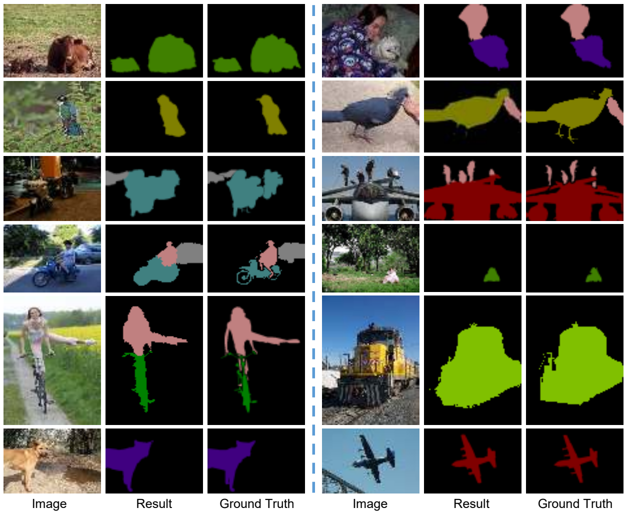

Figure 7. Qualitative segmentation results of the proposed framework on PASCAL VOC 2012 *val* set.

6.3. Ablation Studies

6.3.1 Progressive Common Object Features Mining and Network Training Framework

To evaluate the effectiveness of the proposed progressive common object features mining and network training framework, we evaluate the *RegionNet* and *PixelNet* of each iteration on training and validation set. In the ablation studies, we use VGG16 as base network for *PixelNet*. The results are shown in Table 3. We can see that the initial object seeds are very coarse (14.27% mIoU on *train* set), by applying the *RegionNet* to learn the common features

		train	*val*
	Initial Object Seeds	14.27	-
iter1	*RegionNet*	29.1	-
	Saliency-guided refinement	34.8	-
	PixelNet	48.4	44.4
iter2	*RegionNet*	53.8	-
	PixelNet	57.9	51.6
iter3	*RegionNet*	58.2	-
	PixelNet	60.9	53.3
iter4	*RegionNet*	61.1	-
	PixelNet	63.1	55.5
iter5	*RegionNet*	62.5	-
	PixelNet	63.2	56.2

Table 3. Results of the iteration process. We evaluate the *RegionNet* and *PixelNet* of each iteration on training and validation sets of PASCAL 2012 dataset.

of objects, the performance achieves 29.1%, by introducing saliency-guided refinement, it achieves 34.8%, and after learning with the *PixelNet*, it achieves 48.4%. And in the later iterations, the performance improves gradually, which demonstrates that our method is effective.

6.3.2 Comparison with Direct Iterative Training

We extensively conduct experiments to verify effectiveness of the proposed progressive common object features mining and network training framework by comparing with direct iterative training method. For the direct iterative training method, we start from the segmentation results of our first iteration, and then in later iterations, use the segmentation masks of the previous iteration to train the segmentation network.

Figure 8 shows the comparison. With the iterations, the performance of the direct iterative method increases slowly and only reaches a low accuracy, while in the proposed *M-COF*, the performance increases rapidly and achieves much higher accuracy. This result demonstrates that our *MCOF* framework is effective. The *MCOF* progressively mines common object features from previous object masks and then to expand more reliable object regions to optimize the semantic segmentation network, thus the accuracy can increase rapidly to a quite satisfactory results.

6.3.3 Effectiveness of Saliency-Guided Refinement

The initial object seeds only locate discriminative regions of objects, for example, heads and hands of a person, while other regions, such as the body, are often ignored. To supplement other object regions, saliency maps are incorporated with initial object seeds. This is very important for mining the whole regions of objects. To evaluate the effectiveness, we conduct experiment on framework without saliency-guided refinement, and compare the performance of the *PixelNet* of each iteration. The result is shown in Table 4. Without incorporating saliency maps, some object regions will be missing and thus the performance will be limited and can not reach satisfactory accuracy.

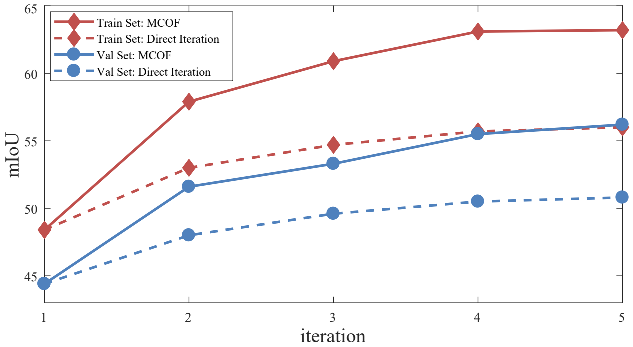

Figure 8. Comparison with direct iterative training method. Our performance improves rapidly while performance of the direct iterative training method increases slowly and only reaches a low accuracy.

iterations	1	2	3	4	5
w/o saliency refinement	41.8	46.2	47.7	51.5	52.1
w/ saliency refinement	**44.4**	**51.6**	**53.3**	**55.5**	**56.2**

Table 4. Evaluate the effectiveness of saliency-guided refinement. We show the mIoU of the *PixelNet* of each iteration on Pascal VOC 2012 val set. Without saliency-guided refinement, the performance will be limited and can not reach satisfactory accuracy.

7. Conclusion

In this paper, we propose *MCOF*, an iterative bottom-up and top-down framework which tolerates inaccurate initial localization by iteratively mining common object features from object seeds. Our method progressively expands object regions and optimizes segmentation network. In bottom-up step, starting from coarse but discriminative object seeds, we mine common object features from them to expand object regions. To supplement non-discriminative object regions, saliency-guided refinement method is proposed. Then in top-down step, these regions are used as supervision to train the segmentation network and predict segmentation masks. The predicted segmentation masks contain more complete object regions than initial, so we can further mine common object features from them. And the processes are conducted iteratively to progressively correct inaccurate initial localization and produce more accurate object regions for semantic segmentation. Our bottom-up and top-down framework bridges the gap between high-level semantic and low-level appearance in weakly-supervised semantic segmentation, and achieves new state-of-the-art performance.

Acknowledgement. This work is supported by National Key Basic Research Program of China (No. 2016YF-B0100900) and National Natural Science Foundation of China (No. 61171113 and No.61773231).

References

[1] L.-C. Chen, J. T. Barron, G. Papandreou, K. Murphy, and A. L. Yuille. Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform. In *IEEE CVPR*, 2016. 2

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *ICLR*, 2015. 2, 6

[3] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE CVPR*, 2016. 1

[4] J. Dai, K. He, and J. Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *IEEE ICCV*, 2015. 2

[5] T. Durand, T. Mordan, N. Thome, and M. Cord. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In *IEEE CVPR*, 2017. 3, 6

[6] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 2010. 6

[7] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 2004. 4

[8] S. Gidaris and N. Komodakis. Object detection via a multi-region and semantic segmentation-aware cnn model. In *IEEE ICCV*, 2015. 1

[9] R. Girshick. Fast R-CNN. In *IEEE ICCV*, 2015. 4

[10] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *IEEE ICCV*, 2011. 6

[11] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li. Salient object detection: A discriminative regional feature integration approach. In *IEEE CVPR*, 2013. 5

[12] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *IEEE CVPR*, 2017. 2

[13] A. Kolesnikov and C. H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*, 2016. 2, 3, 6, 7

[14] V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011. 5

[15] D. Lin, J. Dai, J. Jia, K. He, and J. Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *IEEE CVPR*, 2016. 2

[16] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *IEEE CVPR*, 2015. 2

[17] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich. Feedforward semantic segmentation with zoom-out features. In *IEEE CVPR*, 2015. 2

[18] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *IEEE ICCV*, 2015. 2

[19] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille. Weakly-and semi-supervised learning of a dcnn for semantic image segmentation. *IEEE ICCV*, 2015. 2, 3, 6, 7

[20] D. Pathak, P. Krahenbuhl, and T. Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *IEEE ICCV*, 2015. 6, 7

[21] D. Pathak, E. Shelhamer, J. Long, and T. Darrell. Fully convolutional multi-class multiple instance learning. In *ICLR Workshop*, 2015. 2, 3

[22] P. O. Pinheiro and R. Collobert. From image-level to pixel-level labeling with convolutional networks. In *IEEE CVPR*, 2015. 2, 3, 6, 7

[23] X. Qi, Z. Liu, J. Shi, H. Zhao, and J. Jia. Augmented feedback in semantic segmentation under image level supervision. In *ECCV*, 2016. 2, 3, 6, 7

[24] A. Roy and S. Todorovic. Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation. In *IEEE CVPR*, 2017. 6, 7

[25] 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. 2, 3, 6, 7

[26] W. Shimoda and K. Yanai. Distinct class-specific saliency maps for weakly supervised semantic segmentation. In *ECCV*, 2016. 6, 7

[27] X. Wang, H. Ma, and X. Chen. Geodesic weighted Bayesian model for salient object detection. In *IEEE ICIP*, 2015. 5

[28] X. Wang, H. Ma, and X. Chen. Salient object detection via fast r-cnn and low-level cues. In *IEEE ICIP*, 2016. 4

[29] X. Wang, H. Ma, X. Chen, and S. You. Edge preserving and multi-scale contextual neural network for salient object detection. *IEEE Transactions on Image Processing*, 2018. 4

[30] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *IEEE CVPR*, 2017. 2, 3, 6, 7

[31] 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 Transactions on Pattern Analysis and Machine Intelligence*, 2016. 2, 3, 5, 6, 7

[32] F. Xia, P. Wang, L.-C. Chen, and A. L. Yuille. Zoom better to see clearer: Human part seg-mentation with auto zoom net. *ECCV*, 2016. 2

[33] Y. Xie, H. Lu, and M.-H. Yang. Bayesian saliency via low and mid level cues. *IEEE Transactions on Image Processing*, 2013. 5

[34] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *IEEE CVPR*, 2016. 2, 4