Mixed-dual-head Meets Box Priors: A Robust Framework for Semi-supervised Segmentation

Chenshu Chen Tao Liu Wenming Tan Shiliang Pu
Hikvision Research Institute
{chenchenshu, liutao46, tanwenming, pushiliang.hri}@hikvision.com

Abstract

As it is costly to densely annotate large scale datasets for supervised semantic segmentation, extensive semi-supervised methods have been proposed. However, the accuracy, stability and flexibility of existing methods are still far from satisfactory. In this paper, we propose an effective and flexible framework for semi-supervised semantic segmentation using a small set of fully labeled images and a set of weakly labeled images with bounding box labels. In our framework, position and class priors are designed to guide the annotation network to predict accurate pseudo masks for weakly labeled images, which are used to train the segmentation network. We also propose a mixed-dual-head training method to reduce the interference of label noise while enabling the training process more stable. Experiments on PASCAL VOC 2012 show that our method achieves state-of-the-art performance and can achieve competitive results even with very few fully labeled images. Furthermore, the performance can be further boosted with extra weakly labeled images from COCO dataset.

1. Introduction

Deep neural networks have achieved remarkable success in many computer vision tasks. Their performance highly depends on the amount of labeled data and the quality of annotation. Semantic segmentation labeling is one of the most costly tasks, which requires manual annotation for each pixel in an image.

Aiming to reduce the annotation cost of semantic segmentation, numerous weakly supervised methods have been proposed to leverage weak labels such as image-level labels [1, 10, 15, 16, 21, 24, 25, 43–45, 47, 49], bounding boxes [6, 19, 23, 26, 32, 35], scribbles [27, 38, 39, 41, 42], and points [2]. Although these methods can largely reduce the annotation cost by utilizing weak labels, especially image-level labels, their performance fails to meet the needs of real-world applications. In comparison, semi-supervised methods can achieve promising results with a small set of fully labeled images and a relatively large set of unlabeled images [3, 11, 17, 20, 30, 31, 36, 51]. However, they still underperform their fully supervised counterparts. Some semi-supervised methods seek to realize better results by utilizing both fully labeled and weakly labeled images [18, 29], but the way to utilize multiple types of supervision remains to be further explored.

In this paper, we train semantic segmentation models in a semi-supervised manner using a small set of fully labeled images and a set of weakly labeled images with bounding box labels. The proposed framework contains two semantic segmentation networks: an annotation network (or AnnNet for short) and a segmentation network (or SegNet for short). The AnnNet is used to estimate pseudo masks for weakly labeled images, and the SegNet is the resulting model for deployment. We first train the AnnNet using fully labeled images with ground-truth masks and weakly labeled images with proposal masks. The proposal masks of weakly labeled images are generated using GrabCut [34] and MCG [33]. Then pseudo masks for weakly labeled images are predicted by the trained AnnNet. Finally, we train the SegNet using fully labeled images with ground-truth masks and weakly labeled images with pseudo masks.

We focus on improving the performance of semi-supervised semantic segmentation in two aspects: generating high quality pseudo masks and improving the training stability. To generate accurate pseudo masks, we extract the position and class priors from bounding boxes to guide the training of AnnNet. The RGB image and the extracted priors are concatenated together as the input for AnnNet. Inspired by [29], we propose the mixed-dual-head training method to improve the training stability, which applies parallel strong and weak prediction heads to both AnnNet and SegNet during training. The strong head is trained with a mixture of fully labeled and weakly labeled images, and the weak head is trained with only weakly labeled images. This can reduce the interference of label noise in proposal masks and pseudo masks while maximizing the utilization of weakly labeled images.

The proposed framework is simple and flexible. There are no hyperparameters that need to be carefully tuned, so it is easy to implement. We can choose a large model with high performance as AnnNet, but a small model as SegNet for efficient deployment. Moreover, the proposed framework can benefit from the large number of weakly labeled images and iterative training. The contributions of this paper are summarized as follows:

- We propose an effective and flexible framework for semi-supervised semantic segmentation using a small set of fully labeled images and a set of weakly labeled images with bounding box labels.
- We extract the position and class priors from bounding boxes to guide the training of AnnNet, which can effectively improve the accuracy of generated pseudo masks.
- We propose the mixed-dual-head training method, which can reduce the interference of label noise in proposal masks and pseudo masks while maximizing the utilization of weakly labeled images.
- Extensive experiments on PASCAL VOC 2012 dataset show that the proposed framework is effective and achieves state-of-the-art performance. With additional weakly labeled images from COCO dataset, the performance can be further boosted.

2. Related work

2.1. Weakly-supervised semantic segmentation

To alleviate the burden of manual annotation, numerous methods have been proposed to address the semantic segmentation task based on weak labels, including image-level labels [1,10,15,16,21,24,25,43–45,47,49], bounding boxes [6,19,23,26,32,35], scribbles [27,38,39,41,42], and points [2]. Among them, image-level labels and bounding boxes are most commonly used. Since the image-level label requires the least annotation effort, it is the most investigated weak label. In comparison, bounding boxes provide more information about objects, so higher performance can be achieved. Weakly supervised methods based on bounding boxes are more related to this paper. BoxSup [6] generates a set of candidate segmentation masks using MCG, and then iterates between selecting one candidate mask for each bounding box and training the segmentation network. Similarly, SDI [19] generates segment proposals by combining GrabCut and MCG, and enhances the recursive training with a denoising procedure. BCM [35] uses dense-CRF [22] to generate segment proposals from bounding boxes and calculates the mean filling rates of each class to guide the model training. In Box2Seg [23], bounding boxes are considered as noisy labels for foreground objects. They predict a per-class attention map to guide the loss to focus on foreground pixels and learn pixel embeddings to encourage high intra-class feature affinity.

Some weakly supervised methods also present results under semi-supervised settings by combining the generated pseudo masks with a few ground-truth masks [6,19,32,35]. However, they just use the pseudo masks to improve the semi-supervised performance without further exploring the proper utilization of multiple types of supervision.

2.2. Semi-supervised semantic segmentation

Semi-supervised methods can reduce the annotation cost by using a small set of fully labeled images while keeping competitive performance. Existing semi-supervised methods for semantic segmentation mainly resort to self-training [3, 18, 51], consistency regularization [11, 20, 31], and adversarial learning [17, 30, 36]. Self-training methods usually first train a model using fully labeled images to predict pseudo masks for unlabeled or weakly labeled images, and then train another model using both ground-truth and pseudo masks. Consistency regularization methods improve the generalization performance of the model by adding a consistency loss term. They usually achieve perturbations with data augmentation methods such as Cutout [8] and CutMix [48], and resort to consistency training paradigms such as Mean Teacher [40]. Adversarial learning methods usually make the generated pseudo masks close to the real ones with the help of GAN [12]. A comprehensive literature survey about these methods is beyond the scope of this work, and thus we focus on comparing our framework with the most related ones.

Besides fully labeled images, most of semi-supervised methods utilize unlabeled images and a few methods utilize weakly labeled images [18,29]. Ibrahim et al. [18] propose a framework containing a primary segmentation model, an auxiliary segmentation model and a self-correction module to perform semantic segmentation using fully labeled images and images with bounding boxes. However, the auxiliary segmentation model and self-correction module in their framework are only trained with fully labeled images and thus cannot benefit from the large number of weakly labeled images. In addition, the training procedure of their framework is quite complicated. Luo et al. [29] propose to impose separate treatments of strong and weak annotations via a strong-weak dual-branch network. The strong and weak branches share the backbone network and are trained with fully labeled images and weakly labeled images (with image-level labels), respectively. However, the strong branch is trained using only fully labeled images. Hence, the weakly labeled images cannot be fully utilized. In comparison, our proposed mixed-dual-head training method can maximize the utilization of weakly labeled images while reducing the interference of label noise, which can improve the training stability.

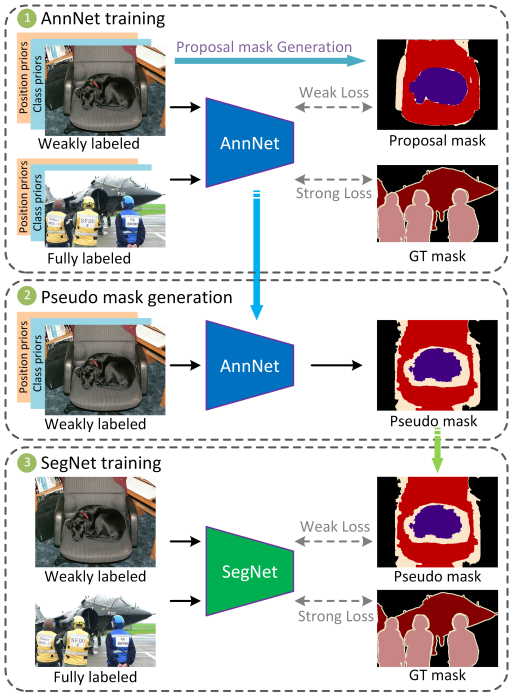


Figure 1. The training procedure of proposed semi-supervised semantic segmentation framework. We first train the AnnNet using fully labeled images and weakly labeled images with proposal masks generated by GrabCut and MCG. Then we predict pseudo masks for weakly labeled images using the trained AnnNet. Finally, we train the SegNet using fully labeled images and weakly labeled images with pseudo masks.

3. Method

The training procedure of our proposed semi-supervised semantic segmentation framework is shown in Figure 1. There are two semantic segmentation networks in our framework: an annotation network (AnnNet) and a segmentation network (SegNet). The training procedure consists of three stages: (1) AnnNet training, (2) pseudo masks generation, and (3) SegNet training. In stage (1), we first generate proposal masks for weakly labeled images using GrabCut and MCG. Then the AnnNet is trained using fully labeled images with ground-truth masks and weakly labeled images with proposal masks. The role of AnnNet is to generate pseudo masks, so in stage (2), we obtain pseudo masks of weakly labeled images based on the predictions of AnnNet. In the final stage, we train the SegNet, the resulting model for deployment, using fully labeled images with ground-truth masks and weakly labeled images with pseudo masks.

In the following parts of this section, we first introduce the generation of proposal masks for weakly labeled images in Section 3.1. Next, we introduce the position and class priors used to form the 5-channel input for AnnNet in Section 3.2. Then, we introduce the mixed-dual-head training method applied to both AnnNet and SegNet in Section 3.3.

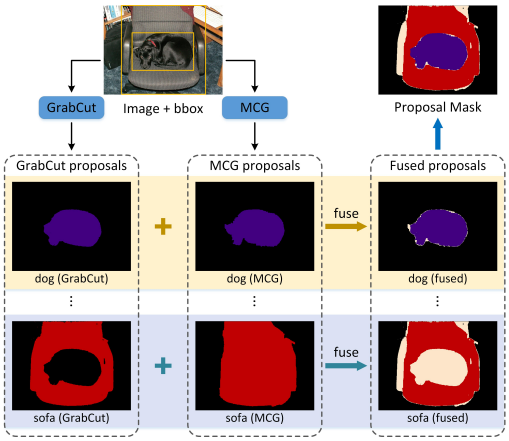


Figure 2. Generation of proposal masks. We generate the proposal masks using GrabCut [34] and MCG [33]. We first generate the GrabCut proposal and MCG proposal for each object. Then we fuse the GrabCut and MCG proposals to obtain the proposal masks. Only pixels with consistent GrabCut and MCG segmentation results are retained and others are ignored.

Finally, we present the training procedure of our framework in detail in Section 3.4.

3.1. Proposal mask generation

To train the AnnNet using weakly labeled images, it is necessary to obtain pixel-level labels from bounding box labels. There are several successful methods for this purpose, of which dense CRF, GrabCut and MCG are the mostly used ones. Similar to previous works [19, 26], we use GrabCut and MCG to obtain the segmentation results of images with bounding box labels, which we called proposal masks.

The proposal masks are generated in the way shown in Figure 2. GrabCut can estimate an object proposal from its bounding box. So GrabCut proposals can be obtained by performing GrabCut on each object in an image. Different from GrabCut, MCG is a region proposal method that can yield lots of proposals for an image. Be aware that the final stage of MCG uses a random forest model trained on PASCAL VOC 2012 dataset with ground-truth masks to rank all the proposals. To avoid introducing extra pixel-level supervision, we do not use this ranking stage, but select one proposal for each object that has the highest Intersection over Union (IoU) with its bounding box.

The resulting proposal mask for an image is obtained by fusing GrabCut and MCG proposals in two steps. First, we fuse the GrabCut proposal and MCG proposal of each object. Only pixels with consistent GrabCut and MCG segmentation results are retained, and the rest are ignored. After getting the fused proposal of each object, we put them back to the correct position in the original image according to the foreground area in descending order (a small object has a higher priority to cover the pixels of large ones). Note

(a) (b)

Figure 3. Visualization of the distance map. (a) Image. (b) Distance map.

that the pixels marked as ignored have the lowest priority, which can be covered by foreground pixels from any other objects.

The proposal masks obtained in this way contain fewer mislabeled pixels, which we believe is extremely crucial for the training of AnnNet. When applying the proposal masks as supervision for AnnNet, the pixels marked as ignored are not involved in the loss calculation.

3.2. Position and class priors

The proposal masks obtained by GrabCut and MCG based on bounding boxes are much coarser than ground-truth masks. Moreover, we ignore some of the pixels that may be mislabeled when generating proposal masks, which also discards some useful information. It is not possible to train a high-performance AnnNet using such coarse and incomplete proposal masks. To compensate for these limitations of proposal masks, we hope to extract useful position and class priors from bounding boxes to guide the training of AnnNet. The extracted position and class priors are in the form of 2-D maps with the same size as the input image. We concatenate the RGB image with the position priors map and class priors map to form the 5-channel input for AnnNet. In the following, we detail the construction of the position and class priors maps.

Position priors. The bounding box can only provide the approximate position of the object. We assume that the closer to the center of the bounding box the more likely the object exists. Based on this assumption, we use the distance from each pixel to the bounding box to construct the position priors map [46]. Given a bounding box of an object, p_i is the coordinate of the i-th pixel in an image. S_e, S_i, and S_o are the sets of pixel coordinates located on, inside, and outside the bounding box, respectively. The distance map $\boldsymbol{D}$ of an object can be calculated as follows:

$$\boldsymbol{D}\left(p_i\right)=\begin{cases}128-\min\limits_{p_j\in S_e}\left\|p_i-p_j\right\| & \text{if } p_i\in S_i\\ 128 & \text{if } p_i\in S_e\\ 128+\min\limits_{p_j\in S_e}\left\|p_i-p_j\right\| & \text{if } p_i\in S_o\end{cases} \quad (1)$$

where $\|\cdot\|$ denotes the Euclidean distance, $\boldsymbol{D}\left(p_i\right)$ denotes the distance value at coordinate p_i. The distance map $\boldsymbol{D}$ has the same size as the input image.

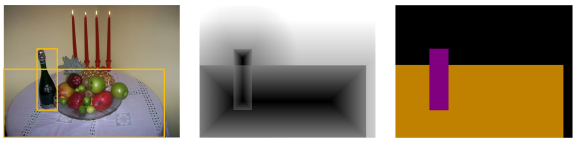
For the convenience of storage and computation, we truncate the distance values in the range of [0, 255]. To ensure the distance value at the center of the bounding box is always zero, we then rescale the distance values for the pixels inside the bounding box:

$$\boldsymbol{D}\left(p_i\right)=\begin{cases}128\times\frac{\boldsymbol{D}(p_i)-\boldsymbol{D}_{\min}}{128-\boldsymbol{D}_{\min}} & \text{if } p_i\in S_i\\ \boldsymbol{D}\left(p_i\right) & \text{otherwise}\end{cases} \quad (2)$$

where $\boldsymbol{D}_{min}$ denotes the minimum value in $\boldsymbol{D}$. A visualization result of the distance map is shown in Figure 3.

The position priors map $\boldsymbol{G}$ is obtained by fusing the distance maps of all objects in an image. We refer to pixels within at least one bounding box as foreground pixels, otherwise as background pixels. Assume there are M objects in an image, their corresponding distance maps are $\mathcal{D}=\{\boldsymbol{D}_i\}_{i=1}^{M}$. For a foreground pixel p, if it is inside only one bounding box whose distance map is $\boldsymbol{D}_i$, $\boldsymbol{G}\left(p\right)$ directly takes the corresponding distance value in $\boldsymbol{D}_i$, i.e., $\boldsymbol{G}\left(p\right)=\boldsymbol{D}_i\left(p\right)$; if it is inside more than one bounding boxes whose distance map are $\mathcal{D}'=\{\boldsymbol{D}_i\}_{i=1}^{m}$, $1<m\leq M$, $\boldsymbol{G}\left(p\right)$ is obtained by averaging the corresponding distance values in $\mathcal{D}'$, i.e., $\boldsymbol{G}\left(p\right)=[\boldsymbol{D}_1\left(p\right)+\cdots+\boldsymbol{D}_m\left(p\right)]/m$. For a background pixel p, $\boldsymbol{G}\left(p\right)$ is obtained by averaging the corresponding distance values of all distance maps in $\mathcal{D}$.

The visualization result of the position priors map is shown in Figure 4(b). The position priors map can be regarded as a heatmap, where smaller values (darker in Figure 4(b)) indicate a higher potential for the existence of an object.



(a) (b) (c)

Figure 4. Visualization of the position priors map and class priors map. (a) Image. (b) Position priors map. (c) Class priors map.

Class priors. The pixels enclosed by the bounding box form a superset of the object pixels. Therefore, pixels inside the bounding box may belong to the bounding box class, and pixels outside all bounding boxes must belong to the background class. Based on this property, we use the bounding box class to construct the class priors map. We refer to pixels within at least one bounding box as foreground pixels, otherwise as background pixels. For a foreground pixel, if it is inside only one bounding box, its value in class priors map is the index of the bounding box class; if it is inside more than one bounding box, its value in class priors map depends on the bounding box with smaller area. For a background pixel, its value in class priors map is the index of the background class.

The visualization result of the class priors map is shown in Figure 4(c). The class priors map encodes the possible classes of foreground pixels, which can guide the AnnNet to distinguish between different classes of objects.

3.3. Mixed-dual-head

If we want to train the semantic segmentation network using weakly labeled images in a supervised manner, we need to estimate pixel-level labels based on weak labels somehow. Despite the fact that we already apply appropriate methods to filter out the potential mislabeled pixels when generating proposal masks and pseudo masks, there are still mislabeled pixels. Compared with strong supervision (ground-truth masks), the estimated proposal masks and pseudo masks can be considered as weak supervision that contains label noise. Aiming to reduce the interference of label noise while ensuring the training stability, we seek to a proper way to utilize both strong and weak supervision.

We can split the semantic segmentation network into the feature extractor and the prediction head. The feature extractor is responsible for extracting feature representation from the image. The prediction head predicts the class of pixels using the extract features. As shown in Figure 5, there are three ways to utilize both strong and weak supervision.

Figure 5(a) shows the commonly used structure with one single head. Fully labeled images and weakly labeled images go through the same prediction head to calculate the loss. In this case, strong and weak supervision are treated equally.

Figure 5(b) shows the dual-head training method. Fully labeled images and weakly labeled images go through the same feature extractor but different prediction heads to calculate the loss separately. After training, the weak head is discarded. We believe that the prediction head is more vulnerable to the interference from label noise than the feature extractor. The design of such separate strong and weak prediction heads can reduce the interference of label noise from weak supervision on the strong head. However, since the strong head is supervised with only strong supervision, a major drawback of this method is that the strong head cannot be well trained when there are few fully labeled images.

Figure 5(c) shows the proposed mixed-dual-head training method. The network structure of mixed-dual-head is the same as that of dual-head, and only strong head is used for inference as well. In order to address the above-mentioned drawback of dual-head, we train the strong head with a mixture of fully labeled and weakly labeled images. In this way, the strong head can benefit from the large number of weakly labeled images while reducing the interference of label noise. We use the hyperparameter r (7:1 by default) to control the mixing ratio between fully labeled and weakly labeled images in strong head. We apply the mixed-dual-head training method to both AnnNet and SegNet. We will show the effectiveness of the mixed-dual-head training method and its superiority over single-head and dual-head in Section 4.4.

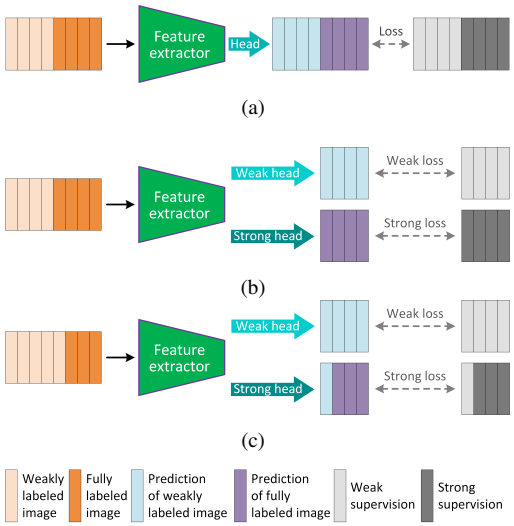


Figure 5. Three ways to utilize both strong and weak supervision. (a) Single-head simply supervises the prediction head with a mixture of strong and weak supervision. (b) Dual-head supervises the strong and weak head using strong and weak supervision separately. (c) Mixed-dual-head supervises the strong head with a mixture of strong and weak supervision, which can improve the training robustness while reducing the interference of label noise.

3.4. Training

As mentioned before, our framework contains three training stages: (1) AnnNet training, (2) pseudo mask generation, and (3) SegNet training. We call performing stage (1), stage (2), and stage (3) once as a training round. Details about one training round of our framework can be found in Section A of the supplemental material.

The training set consists of two parts: the fully labeled dataset $\mathcal{F} = \{(x_i, y_i, b_i)\}_{i=1}^{F}$ containing F fully labeled images and the weakly labeled dataset $\mathcal{W} = \{(x_i, b_i)\}_{i=1}^{W}$ containing W weakly labeled images, where x_i denotes i-th image, y_i and b_i denote its ground-truth mask and bounding box label.

We train AnnNet and SegNet with the proposed mixed-dual-head method in stage (1) and (3), respectively. The differences are: 1) the proposal masks of weakly labeled images for AnnNet are generated by GrabCut and MCG, whereas the pseudo masks for SegNet are generated by AnnNet (the visualization of some proposal and pseudo masks are shown in Section E of the supplementary material); 2) only AnnNet is trained with position and class priors.

In stage (2), we use a threshold τ to obtain pseudo masks from the prediction results of AnnNet. For each pixel p, its

pseudo label is generated as follows:

$$\tilde{y}^c(p) = \begin{cases} 1 & \text{if } c = \arg\max_j \hat{y}^j(p) \text{ and } \hat{y}^c(p) > \tau \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where $\hat{y}(p)$ denotes the predicted probability of pixel p and $\hat{y}^j(p)$ is the prediced probability of class j. Similarly, $\tilde{y}(p)$ denotes the pseudo label of pixel p which is a one-hot vector and $\tilde{y}^c(p)$ is the c-th element. Note that $\tilde{y}(p)$ can be a zero vector, meaning the pseudo label of this pixel is ignored when used to train the SegNet.

4. Experiment

4.1. Datasets and evaluation metrics

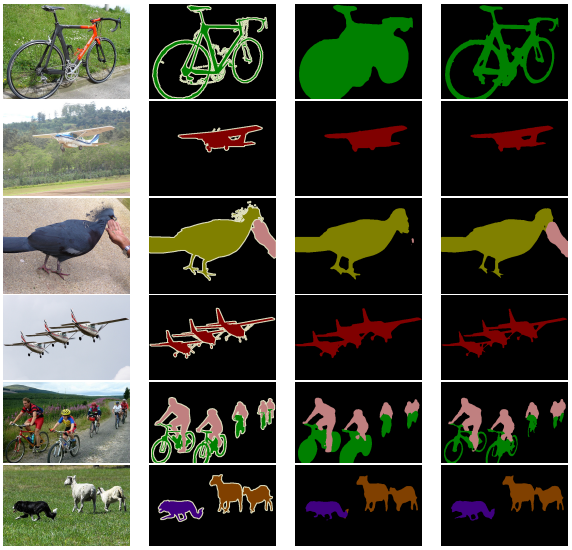
We evaluate our framework on PASCAL VOC 2012 dataset [9], which contains 21 classes including background. In addition, we utilize the extra annotated images from SBD [13] dataset. In total, there are 10582 images for training, 1449 images for validation. We use the standard mean Intersection over Union (mIoU) as the evaluation metric for all experiments.

4.2. Implementation Details

We adopt HRNetV2-W48 [37] as AnnNet (additional experiments about the selection of AnnNet can be found in Section C of the supplemental material). Since SegNet is used for deployment, we use the commonly used DeepLabv3+ [4] as SegNet for fair comparison. We use ResNet-101 [14] and ResNeSt-101 [50] as the backbone network of SegNet to compare with other methods, and mainly use ResNet-50 [14] in ablation experiments. Following the common practice, we use the weights pretrained on ImageNet [7] for all backbone networks. The threshold τ required for generating the pseudo masks is set to 0.9. The mixing ratio r between fully labeled and weakly labeled images in strong head is set to 7:1 by default.

We train both AnnNet and SegNet for 60 epochs. We adopt SGD as the optimizer with a momentum of 0.9 and a batch size of 32. The initial learning rates of AnnNet and SegNet are set to 0.0008 and 0.004, respectively. We employ a poly learning rate policy with power of 0.9. We apply common data augmentation, including random horizontal flipping, random scaling in the range of [0.5, 2.0] and random cropping with a size of 512×512. Without special statements, we use single-scale input during inference.

We modify the first convolutional layer of HRNetV2-W48, which is used as our AnnNet, to allow it to accept 5-channel data as input. We also remove the last convolutional layer of HRNetV2-W48 and DeepLabv3+ used for prediction and then add two parallel prediction heads to implement the dual-head or mixed-dual-head training method. The two added heads have the same structure, including a 1×1 convolutional layer followed by a batch normalization layer and ReLU activation, and a 1×1 convolutional layer for prediction.



(a) (b) (c) (d)

Figure 6. Qualitative results on PASCAL VOC 2012 val set. Our framework can achieve better segmentation results with fewer ground-truth masks than the fully supervised baseline. (a) Image. (b) Ground Truth. (c) Fully Supervised. (d) Ours.

4.3. Quantitative and qualitative results

We train the DeepLabv3+ model using 10582 images with ground-truth masks as our fully supervised baseline. Following the common practice, we train our framework using 1464 images with ground-truth masks from PASCAL VOC 2012 training set and extra 9118 images with only bounding box labels from SBD. We use the same DeepLabv3+ framework as our SegNet, except that an additional prediction head is added during training phase. But the network structure of SegNet is exactly the same as the baseline model at inference time.

The results on PASCAL VOC 2012 validation set are shown in Table 1. "Fully Supervised" is our implementation of baseline DeepLabv3+. Our reproduction has a mIoU score of 80.69% with ResNet-101, which is slightly better than 80.22% reported in [4]. Therefore, our reproduced fully supervised baseline is convincing. It can be seen that our framework outperforms the fully supervised baseline by a large margin with even fewer ground-truth masks. We attribute this result to two aspects. First, we observe that the extra ground-truth masks from SBD are not as fine-grained as those from PASCAL VOC 2012. Possibly the label noise from SBD affects the performance of the baseline model. More importantly, our framework extract position and class priors from bounding boxes to ensure the quality of pseudo

Method	Backbone	F	W	MS	mIoU
Fully Supervised	ResNet-50	10582	-		78.55
Ours	ResNet-50	1464	9118		**81.31**
Fully Supervised	ResNet-50	10582	-	✓	79.39
Ours	ResNet-50	1464	9118	✓	**82.26**
Fully Supervised	ResNet-101	10582	-	✓	80.69
Ours	ResNet-101	1464	9118	✓	**83.78**
Fully Supervised	ResNeSt-101	10582	-	✓	82.69
Ours	ResNeSt-101	1464	9118	✓	**85.05**

Table 1. Results on PASCAL VOC 2012 val set. F and W are the numbers of fully labeled images and weakly labeled images, respectively. "MS" denotes using multi-scale and left-right flipped inputs at inference time.

	Backbone	F	W	mIoU
w/o priors	ResNet-50	1464	9118	79.30
w/ priors	ResNet-50	1464	9118	**81.31**

Table 2. Ablation study of our framework on PASCAL VOC 2012 val set with or without priors information.

masks and use mixed-dual-head training method to reduce the negative effect of label noise. Some qualitative results are presented in Figure 6.

4.4. Ablation studies

The effectiveness of priors. To verify the effect of priors added to the input of AnnNet, we evaluate the performance of our framework with or without priors. Without priors, the input of AnnNet is the 3-channel RGB image. The experiment results are shown in Table 2. With priors added to the input of AnnNet, the performance of SegNet is significantly improved by 2.01%, implying that priors can make AnnNet produce better pseudo masks which bring gains to SegNet. Additional experiments about adding priors or not are shown in supplemental material Section B.

Choice of mixing ratio r. We use r to control the mixing ratio between fully labeled and weakly labeled images in strong head. While keeping more fully labeled images than weakly labeled ones in strong head, we vary the mixing ratio r and show the results in Table 3. A relatively high performance is obtained when $r = 7:1$, which is the default value for subsequent experiments. In addition, performance of our framework still outperforms the fully supervised baseline (78.55%) when $r = 5:3$ or $6:2$.

r	Backbone	F	W	mIoU
5:3	ResNet-50	1464	9118	80.22
6:2	ResNet-50	1464	9118	80.64
7:1	ResNet-50	1464	9118	**81.31**

Table 3. Ablation study of our framework on PASCAL VOC 2012 val set for different mixing ratio r.

The effectiveness of mixed-dual-head. To verify the effectiveness of proposed mixed-dual-head training method, we conduct experiments with three kinds of training methods shown in Figure 5. The results are shown in Table 4. It can be observed that both dual-head and mixed-dual-head are significantly superior to single-head. This indicates that the separate strong and weak prediction heads can improve the performance by reducing the interference of label noise. In addition, the performance of mixed-dual-head is slightly higher than that of dual-head.

	Backbone	F	W	mIoU
single-head	ResNet-50	1464	9118	78.96
dual-head	ResNet-50	1464	9118	81.03
mixed-dual-head	ResNet-50	1464	9118	**81.31**

Table 4. Ablation study of our framework on PASCAL VOC 2012 val set for different prediction head.

	Backbone	F	W	mIoU
dual-head	ResNet-50	374	10235	**78.95**
mixed-dual-head	ResNet-50	374	10235	77.99
dual-head	ResNet-50	166	10416	73.32
mixed-dual-head	ResNet-50	166	10416	**76.88**
dual-head	ResNet-50	83	10499	67.07
mixed-dual-head	ResNet-50	83	10499	**76.62**

Table 5. Ablation study of our framework on PASCAL VOC 2012 val set for varying numbers of fully labeled images.

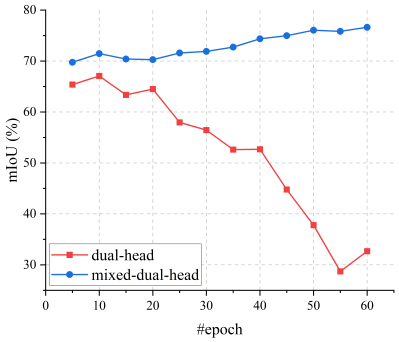


Figure 7. The performance of dual-head and mixed-dual-head on PASCAL VOC 2012 val set during training when F=83.

To further verify the superiority of mixed-dual-head, we conduct experiments with fewer fully labeled images. As shown in Table 5, when the number of fully labeled images decreases, the performance of dual-head degrades dramatically (78.95% → 73.32% → 67.07%). In comparison, the performance drop of mixed-dual-head is much smaller (77.99% → 76.88% → 76.62%). We also observe that the performance of dual-head becomes unstable during training when $F = 83$, as shown in Figure 7. It shows that the proposed mixed-dual-head training method can improve training stability, especially when there are few fully labeled images. A better training stability means a better robustness of

our framework, which can avoid the dramatic performance degradation due to a decreasing number of fully labeled images.

Iterative training. Our framework can benefit from iterative training. To verify this, we perform multiple rounds of iterative training for AnnNet. At the end of each training round, the AnnNet predicts pseudo masks for weakly labeled images, and the pseudo masks are then used as supervision for AnnNet in the next training round. We record the performance of SegNet in each round, and the results are shown in Table 6. The experiments are conducted when $F = 83$. It can be observed that the performance of our framework is further improved by iterative training. Surprisingly, the performance of our framework after four rounds of iterative training with only 83 fully labeled images even exceeds the fully supervised baseline (78.55%). The visualization of some pseudo masks can be found in supplemental material Section D.

Round	Backbone	F	W	mIoU
1	ResNet-50	83	10499	76.62
2	ResNet-50	83	10499	78.53
3	ResNet-50	83	10499	77.88
4	ResNet-50	83	10499	**79.49**

Table 6. Ablation study of our framework on PASCAL VOC 2012 val set for iterative training.

Backbone	F	W	MS	mIoU
ResNet-50	1464	9118	✓	82.26
ResNet-50	1464	9118+94k	✓	**83.94**
ResNet-101	1464	9118	✓	83.78
ResNet-101	1464	9118+94k	✓	**85.16**
ResNeSt-101	1464	9118	✓	85.05
ResNeSt-101	1464	9118+94k	✓	**86.10**

Table 7. Ablation study of our framework on PASCAL VOC 2012 val set for fine-tuning on COCO dataset. "+94k" denotes using extra images from COCO 2017. "MS" denotes using multi-scale and left-right flipped inputs at inference time.

Fine-tuning with COCO dataset. To investigate the performance gain of our framework when more weakly labeled images are available, we perform fine-tuning experiments with COCO [28] dataset. Specifically, we first follow the same procedure as before to train the AnnNet and SegNet for 60 epochs using 1464 fully labeled images and 9118 weakly labeled images from PASCAL VOC 2012 dataset. We then use the trained AnnNet to predict pseudo masks for around 94k images in COCO 2017 training set that contain objects belonging to the 20 classes in PASCAL VOC 2012 dataset. The bounding box labels of these 94k images are used to form the 5-channel input for AnnNet. The visualization of some pseudo masks of COCO dataset can be found in Section F of the supplemental material. We can see from the visualization results that the quality of pseudo masks is quite impressive even in some complex scenarios, which implies the effectiveness of position and class priors introduced in AnnNet. Finally, we fine-tune the trained SegNet with 9118+94k images for 30 epochs. When fine-tuning, the learning rate of SegNet is set to 0.001. As shown in Table 7, the performance of our framework is further improved with more weakly labeled images.

Method	Backbone	F	W	MS	mIoU	FS%
Self-correcting [18]	Xception-65	1464	9118	✓	82.33	101.38
Ours	ResNet-50	1464	9118	✓	82.26	103.62
Ours	ResNet-101	1464	9118	✓	83.78	**103.83**
Ours	ResNeSt-101	1464	9118	✓	**85.05**	102.80
BCM [34]	ResNet-101	1464	9118		71.60	96.11
BoxSup [6]	VGG-16	1464	9118		63.50	99.53
SDI [19]	VGG-16	1464	9118		65.80	95.22
WSSL [31]	VGG-16	1464	9118		65.10	96.30

Table 8. Comparison with other methods on PASCAL VOC 2012 val set. "MS" denotes using multi-scale and left-right flipped inputs at inference time.

4.5. Comparison with state-of-the-art

We compare our framework with state-of-the-art methods on PASCAL VOC 2012 dataset, as shown in Table 8. We mainly compare our framework with Self-correcting [18], which also uses fully labeled images and weakly labeled images with bounding box labels to perform semi-supervised semantic segmentation. They use the same DeepLabv3+ model as we do, but with Xception [5] as backbone. Note that the performance of DeepLabv3+ with Xception (81.21%) is better than that with ResNet-101 (80.22%), as reported in [4]. In the lower half of Table 8, we record the performance of semi-supervised semantic segmentation from some weakly supervised methods which also use bounding boxes as weak labels. FS% denotes the performance relative to the fully supervised counterpart. It can be seen that the performance of both Self-correcting and our framework exceeds the respective fully supervised counterparts. Moreover, our framework achieves a new state-of-the-art performance.

5. Conclution

We believe that improving the quality of pseudo masks and reducing the interference of label noise are critical for semi-supervised semantic segmentation. Based on this, we design the position and class priors and propose mixed-dual-head training method. Despite it is simple, experiment results show that our proposed framework is effective and stable. Exploration about how to further weaken the effect of label noise while maximizing the utilization of weakly labeled images, and extending our framework to instance segmentation will be studied in our future work.

References

[1] Jiwoon Ahn and Suha Kwak. Learning Pixel-Level Semantic Affinity With Image-Level Supervision for Weakly Supervised Semantic Segmentation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4981–4990. IEEE Computer Society, 2018.

[2] Amy L. Bearman, Olga Russakovsky, Vittorio Ferrari, and Fei-Fei Li. What's the Point: Semantic Segmentation with Point Supervision. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII*, volume 9911 of *Lecture Notes in Computer Science*, pages 549–565. Springer, 2016.

[3] Miriam Bellver, Amaia Salvador, Jordi Torres, and Xavier Giró-i Nieto. Budget-aware Semi-Supervised Semantic and Instance Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 93–102. Computer Vision Foundation / IEEE, 2019.

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, volume 11211 of *Lecture Notes in Computer Science*, pages 833–851. Springer, 2018.

[5] François Chollet. Xception: Deep Learning with Depthwise Separable Convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1800–1807. IEEE Computer Society, 2017.

[6] Jifeng Dai, Kaiming He, and Jian Sun. BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1635–1643. IEEE Computer Society, 2015.

[7] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. ISSN: 1063-6919.

[8] Terrance DeVries and Graham W. Taylor. Improved Regularization of Convolutional Neural Networks with Cutout. *arXiv:1708.04552 [cs]*, Nov. 2017. arXiv: 1708.04552.

[9] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[10] Junsong Fan, Zhaoxiang Zhang, Tieniu Tan, Chunfeng Song, and Jun Xiao. CIAN: Cross-Image Affinity Net for Weakly Supervised Semantic Segmentation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 10762–10769. AAAI Press, 2020.

[11] Geoffrey French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham D. Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020.

[12] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative Adversarial Nets. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014.

[13] Bharath Hariharan, Pablo Arbelaez, Lubomir D. Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In Dimitris N. Metaxas, Long Quan, Alberto Sanfeliu, and Luc Van Gool, editors, *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, pages 991–998. IEEE Computer Society, 2011.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[15] Qibin Hou, Peng-Tao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-Erasing Network for Integral Object Attention. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 547–557, 2018.

[16] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-Supervised Semantic Segmentation Network With Deep Seeded Region Growing. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7014–7023. IEEE Computer Society, 2018.

[17] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial Learning for Semi-supervised Semantic Segmentation. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 65. BMVA Press, 2018.

[18] Mostafa S. Ibrahim, Arash Vahdat, Mani Ranjbar, and William G. Macready. Semi-Supervised Semantic Image Segmentation With Self-Correcting Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 12712–12722. IEEE, 2020.

[19] Anna Khoreva, Rodrigo Benenson, Jan Hendrik Hosang, Matthias Hein, and Bernt Schiele. Simple Does It: Weakly Supervised Instance and Semantic Segmentation. In *2017*

IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pages 1665–1674. IEEE Computer Society, 2017.

[20] Jongmok Kim, Jooyoung Jang, and Hyunwoo Park. Structured Consistency Loss for semi-supervised semantic segmentation. *arXiv:2001.04647 [cs]*, Jan. 2020. arXiv: 2001.04647.

[21] Alexander Kolesnikov and Christoph H. Lampert. Seed, Expand and Constrain: Three Principles for Weakly-Supervised Image Segmentation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, volume 9908 of *Lecture Notes in Computer Science*, pages 695–711. Springer, 2016.

[22] Philipp Krähenbühl and Vladlen Koltun. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 109–117, 2011.

[23] Viveka Kulharia, Siddhartha Chandra, Amit Agrawal, Philip H. S. Torr, and Ambrish Tyagi. Box2Seg: Attention Weighted Loss and Discriminative Feature Learning for Weakly Supervised Segmentation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXVII*, volume 12372 of *Lecture Notes in Computer Science*, pages 290–308. Springer, 2020.

[24] Suha Kwak, Seunghoon Hong, and Bohyung Han. Weakly Supervised Semantic Segmentation Using Superpixel Pooling Network. In Satinder P. Singh and Shaul Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4111–4117. AAAI Press, 2017.

[25] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. FickleNet: Weakly and Semi-Supervised Semantic Image Segmentation Using Stochastic Inference. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5267–5276. Computer Vision Foundation / IEEE, 2019.

[26] Qizhu Li, Anurag Arnab, and Philip H. S. Torr. Weakly- and Semi-supervised Panoptic Segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV*, volume 11219 of *Lecture Notes in Computer Science*, pages 106–124. Springer, 2018.

[27] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. ScribbleSup: Scribble-Supervised Convolutional Networks for Semantic Segmentation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 3159–3167. IEEE Computer Society, 2016.

[28] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014.

[29] Wenfeng Luo and Meng Yang. Semi-supervised semantic segmentation via strong-weak dual-branch network. In *European Conference on Computer Vision*, pages 784–800. Springer, 2020.

[30] Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox. Semi-Supervised Semantic Segmentation With High- and Low-Level Consistency. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(4):1369–1379, 2021.

[31] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-Supervised Semantic Segmentation With Cross-Consistency Training. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 12671–12681. IEEE, 2020.

[32] George Papandreou, Liang-Chieh Chen, Kevin P. Murphy, and Alan L. Yuille. Weakly-and Semi-Supervised Learning of a Deep Convolutional Network for Semantic Image Segmentation. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1742–1750. IEEE Computer Society, 2015.

[33] J. Pont-Tuset, P. Arbeláez, J. T. Barron, F. Marques, and J. Malik. Multiscale Combinatorial Grouping for Image Segmentation and Object Proposal Generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(1):128–140, 2017.

[34] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. GrabCut: Interactive Foreground Extraction Using Iterated Graph Cuts. *ACM Trans. Graph.*, 23(3):309–314, 2004. Place: New York, NY, USA Publisher: Association for Computing Machinery.

[35] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Box-Driven Class-Wise Region Masking and Filling Rate Guided Loss for Weakly Supervised Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3136–3145. Computer Vision Foundation / IEEE, 2019.

[36] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi Supervised Semantic Segmentation Using Generative Adversarial Network. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5689–5697. IEEE Computer Society, 2017.

[37] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-Resolution Representations for Labeling Pixels and Regions. *arXiv:1904.04514 [cs]*, Apr. 2019. arXiv: 1904.04514.

[38] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized Cut Loss for Weakly-Supervised CNN Segmentation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*

2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 1818–1827. IEEE Computer Society, 2018.

[39] Meng Tang, Federico Perazzi, Abdelaziz Djelouah, Ismail Ben Ayed, Christopher Schroers, and Yuri Boykov. On Regularized Losses for Weakly-supervised CNN Segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI*, volume 11220 of *Lecture Notes in Computer Science*, pages 524–540. Springer, 2018.

[40] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, volume 30, pages 1195–1204. Curran Associates, Inc., 2017.

[41] Paul Vernaza and Manmohan Chandraker. Learning Random-Walk Label Propagation for Weakly-Supervised Semantic Segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2953–2961. IEEE Computer Society, 2017.

[42] Bin Wang, Guojun Qi, Sheng Tang, Tianzhu Zhang, Yunchao Wei, Linghui Li, and Yongdong Zhang. Boundary Perception Guidance: A Scribble-Supervised Semantic Segmentation Approach. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 3663–3669. ijcai.org, 2019.

[43] Xiang Wang, Sifei Liu, Huimin Ma, and Ming-Hsuan Yang. Weakly-Supervised Semantic Segmentation by Iterative Affinity Learning. *International Journal of Computer Vision*, 128(6):1736–1749, 2020.

[44] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-Supervised Semantic Segmentation by Iteratively Mining Common Object Features. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1354–1362. IEEE Computer Society, 2018.

[45] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object Region Mining with Adversarial Erasing: A Simple Classification to Semantic Segmentation Approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6488–6496. IEEE Computer Society, 2017.

[46] Ning Xu, Brian L. Price, Scott Cohen, Jimei Yang, and Thomas S. Huang. Deep grabcut for object selection. In *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*. BMVA Press, 2017.

[47] Zeng Yu, Yun-Zhi Zhuge, Huchuan Lu, and Lihe Zhang. Joint Learning of Saliency Detection and Weakly Supervised Semantic Segmentation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7222–7232. IEEE, 2019.

[48] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6022–6031. IEEE, 2019.

[49] Bingfeng Zhang, Jimin Xiao, Yunchao Wei, Mingjie Sun, and Kaizhu Huang. Reliability Does Matter: An End-to-End Weakly Supervised Semantic Segmentation Approach. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 12765–12772. AAAI Press, 2020.

[50] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, and Alexander Smola. ResNeSt: Split-Attention Networks. *arXiv:2004.08955 [cs]*, Apr. 2020. arXiv: 2004.08955.

[51] Yi Zhu, Zhongyue Zhang, Chongruo Wu, Zhi Zhang, Tong He, Hang Zhang, R. Manmatha, Mu Li, and Alexander Smola. Improving Semantic Segmentation via Self-Training. *arXiv:2004.14960 [cs]*, May 2020. arXiv: 2004.14960.