Self-Supervised Difference Detection for Weakly-Supervised Semantic Segmentation

Wataru Shimoda and Keiji Yanai

Artificial Intelligence eXploration Research Center, The University of Electro Communications, Tokyo
1-5-1 Chofugaoka, Chofu, Tokyo 182-8585 JAPAN

{shimoda-k,yanai}@mm.inf.uec.ac.jp

Abstract

To minimize the annotation costs associated with the training of semantic segmentation models, researchers have extensively investigated weakly-supervised segmentation approaches. In the current weakly-supervised segmentation methods, the most widely adopted approach is based on visualization. However, the visualization results are not generally equal to semantic segmentation. Therefore, to perform accurate semantic segmentation under the weakly supervised condition, it is necessary to consider the mapping functions that convert the visualization results into semantic segmentation. For such mapping functions, the conditional random field and iterative re-training using the outputs of a segmentation model are usually used. However, these methods do not always guarantee improvements in accuracy; therefore, if we apply these mapping functions iteratively multiple times, eventually the accuracy will not improve or will decrease.

In this paper, to make the most of such mapping functions, we assume that the results of the mapping function include noise, and we improve the accuracy by removing noise. To achieve our aim, we propose the self-supervised difference detection module, which estimates noise from the results of the mapping functions by predicting the difference between the segmentation masks before and after the mapping. We verified the effectiveness of the proposed method by performing experiments on the PASCAL Visual Object Classes 2012 dataset, and we achieved 64.9% in the val set and 65.5% in the test set. Both of the results become new state-of-the-art under the same setting of weakly supervised semantic segmentation.

1. Introduction

Semantic segmentation is a promising image recognition technology that enables the detailed analysis of images for various practical applications. However, semantic segmentation methods require training data with pixel-level annotation, which is costly to create. On the other hand, image-level annotation is much easier to obtain than pixel-level annotation. In recent years, various weakly-supervised semantic segmentation (hereinafter WSS) methods that required only image-level annotation have been proposed to resolve the annotation problems. However, there is still a large performance gap between fully-supervised and weakly-supervised methods.

In weakly-supervised segmentation methods, visualization-based approaches [39, 33, 41] have been widely adopted. The visualization results highlight the regions that contributed to the classification, and we can roughly estimate the regions of the target objects by visualization. Class Activation Map (CAM) [41] is a standard method to visualize the classification results. However, the visualization results do not always match actual segmentation results; therefore, it is usually necessary to consider the mapping from the visualization results to the semantic segmentation in weakly-supervised segmentation. Conditional Random Field (CRF) [17] is widely used as a mapping function. CRF is a method for optimizing the probability distribution to be fitted to the edge of regions by using color and position information as features. The iterative approach for the learning segmentation models proposed by Wei et al. [37] is a versatile approach for improving weakly supervised segmentation results. In this method, we generate pseudo pixel-level labels under weakly supervised conditions, and we train a segmentation model with the pseudo labels. Subsequently, we generate pseudo pixel-level labels from the outputs of the trained segmentation model, and we re-train a new segmentation model using the generated pseudo labels. Wei et al. [37] showed that repeating this process absorbed outliers and gradually improved the accuracy. These methods can be regarded as mapping functions that bring inputs closer to the segmentation. However, the mapping functions of these methods [17, 37] do not guarantee any improvement in the accuracy of the semantic segmentation; therefore, the mapping results contain noise. In this paper, the mapping functions that make the above inputs close to the segmentation are treated as supervision containing noise, and we propose a robust learning method for such noise.

In this paper, we denote the information used as the inputs of the mapping functions as *knowledge*, and we consider the supervision containing the noise as *advice*. The supervision for fully supervised learning that allows one-to-

one mapping is *teacher*. We assume that the *advice* provides supervision, which includes some correct and incorrect information. To make effective use of the information obtained from this *advice*, it is necessary to select useful information. In this paper, we regard the regions where opinions differ between *knowledge* and *advice* as *difference*. Since *difference* in the two segmentation masks can be obtained by simple processing without annotation, it is a kind of self-supervised learning to train a model, which predicts *difference*. Self-supervised learning is a pretext task as a form of indirect supervision. For example, as notable works, colorization [4] and predicting the patch ordering [5] have been proposed.

Inferring *difference* in *knowledge* and *advice* from *knowledge* leads to predicting the advisor's *advice* in advance. In predicting *advice*, there are predictable *advice* and unpredictable *advice*. Certain *advice* can be easily inferred because many similar samples are included during training. Here, we assumed that `advice` contains a sufficient number of good information, and predictable information can be considered to be useful information. Based on this idea, we propose a method for selecting information by finding the true information in `advice` that can be predicted from the inference results of difference detection. Fig.1 shows the concept of the proposed approach.

In this paper, we demonstrate that the proposed Self-Supervised Difference Detection (SSDD) module can be used in both the seed generation stage and the training stage of fully supervised segmentation. In the seed generation stage, we refine the CRF results for pixel-level semantic affinity (PSA) [1] by using the SSDD module. In the training stage, we introduce two SSDD modules inside the training loop of a fully supervised segmentation network. In the experiments, we demonstrate the effectiveness of the SSDD modules in both stages. In particular, the SSDD modules greatly boosted the performance of the WSS on the PASCAL visual object classes (VOC) 2012 dataset, and achieved new state-of-the-art. To summarize it, our contributions are as follows:

- We propose an SSDD module, which estimates the noise of the mapping functions of the weakly supervised segmentation and select useful information.
- We show that the SSDD modules can be effectively applied to both the seed generation stage and the training stage of a fully supervised segmentation model.
- We obtained the best results on the PASCAL VOC 2012 dataset with 64.9% mean IoU on the *val set* and 65.5% on the *test set*.

2. Related Works

In this section, we review related research on CNN-based WSS methods by classifying them into several types.

Visualization In the early works of CNN-based WSS, visualization-based methods were studied. The pixels that contributed to the classification were correlated to the regions of the target objects; therefore, the visualization methods can be used as segmentation methods under weakly supervised settings. Zeiler et al. [40] showed that the derivatives obtained by back-propagation from the CNN models trained for classification tasks highlight the region of a target object in an image. Simonyan et al. [33] used derivatives such as the GrabCut seeds and extended the visualization method to the WSS method. They also demonstrated that the regions of multi-class objects could also be captured by the difference in class-specific derivatives [13, 32]. Oquab et al.[21] visualized the attention region by the forwarding process using activation and trained a classification model with large input images by using global max pooling. After this approach, several derived methods employing global pooling were also proposed [25, 41, 16]. In particular, CAM [41] has been widely adopted in recent weakly supervised segmentation methods.

Region refinement for WSS results using CRF In general, the segmentation results based on fully convolutional neural network (FCN) [19] tend to output ambiguous outlines. CRF [17] can refine the ambiguous outlines using low-level features such as the pixel colors. Chen et al. [22] and Pathak et al. [23] adopted CRF as a post-processing method for region refinement and demonstrated the effectiveness of the CRF for WSS. Kolesnikov et al. [16] proposed the use of CRF during the training of a semantic segmentation model. Ahn et al. [1] proposed a method to learn pixel-level similarity from the CRF results, and apply a random walk-based region refinement, which achieved the best results on the PASCAL VOC 2012 dataset. CRF plays an important role to improve the accuracy of weakly supervised segmentation. Furthermore, various researches employed the CRF for refining the coarse segmentation masks [32, 29, 28, 15, 37, 36, 10, 31]. However, CRF does not guarantee any improvement in the mean intersection over union (IoU) score, and it often degrades the segmentation masks and the scores. Therefore, we focus on preventing a segmentation mask from being degraded by applying CRF. We estimate the confidence maps of both the initial mask and the mask after CRF post-processing, and we integrate both masks based on the estimated confidence maps.

Training fully supervised segmentation model under weakly supervised setting Certain researchers trained a fully supervised semantic segmentation (hereinafter FSS) model under a weakly supervised setting. First, Papandreou et al. [24] proposed MIL-FCN, which trained a fully supervised semantic segmentation model with a global max-pooling loss using only image-level labels. Wei et al. [37] proposed a novel approach to train an FSS model using pixel-level labels obtained by saliency maps [12]. This method is simple, and the obtained results are impressive. Wei et al. [37] also demonstrated that the outputs of the trained semantic segmentation model could be used as a new pixel-level annotation for re-training, and the re-trained FSS model achieved better results than the original model.

Generating pixel-level labels during training of an FSS model Constrained convolutional neural network

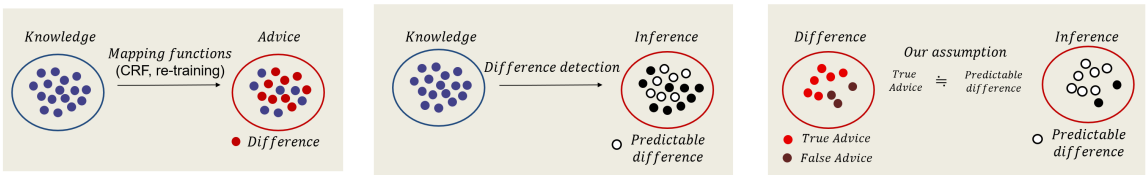

Figure 1. The concept of the proposed approach. (a) We denote the inputs of the mapping functions as *knowledge* and the outputs as *advice*. (b) The proposed difference detection network (DD-Net) estimates the *difference* between *knowledge* and *advice*. (c) In *difference*, the *advice* is divided into true *advice* and false *advice*. We assume that if the amount of true *advice* is larger than the amount of false *advice*, that is, if a set of false *advice* are outliers, then the predictable *advice* has a strong correlation with the true *advice*.

(CCNN) [23] and EM-adopt [22] generated pixel-level labels during training using class labels and outputs of the segmentation model. In both the studies similar constraints were made for generating pixel-level labels to obtain better results. They set the ratios of the foreground and the background in an image and generated pixel-level labels within the ratio. Wei et al. [36] proposed an online prohibitive segmentation learning (PSL). They generated pixel-level seed labels of training samples before the first training of an FSS model and re-generated pixel-level labels using the outputs of the segmentation model and the classification results. The semantic segmentation model was trained by both the pixel-level labels, and they achieved good performance without costly manual pixel-level annotation. We expected that the pixel-level seed labels would play the role of the constraint. Huang et al. [11] proposed deep seeded region growing (DSRG), which is a method to expand the seed region during training. Before training, the authors prepared pixel-level seed labels that had unlabeled regions for unconsidered pixels. In this research, we proposed new constraints for generating pixel-level labels during the training of the FSS model. We trained an FSS model and the difference detection model in an end-to-end manner. Then, we interpolated a few pixel-level seed labels, that had different regions in the newly generated pixel-level labels and these labels could also be predicted by the difference detection model.

WSS methods using additional information A few recent weakly supervised approaches achieved high accuracy by using additional annotations for image-level labels. Researchers have proposed the bounding box annotation for WSS [22], and they showed that the bounding box annotation substantially boosted performance. As weaker additional annotation, point annotation and scribble annotation were also proposed [2]. Saleh et al. [29] proposed an approach to check the generated initial masks by minimal additional supervision by human visions. Motion segmentation of videos as additional training information for weakly supervised segmentation has also been proposed [34, 9]. There are also reports that web images were helpful for improving the weakly supervised segmentation accuracy [25, 37, 14, 31]. Recently, fully supervised saliency methods are being widely used for detecting the background regions, and certain researchers have reported that this approach could substantially boost performance [30, 36, 38, 11, 10, 35, 3]. Region proposal methods trained with fully supervised foreground masks such as MCG [26] have also been used in [25, 27]. Hu et al. [6] used instance-level saliency maps for WSS. The concept of saliency can be used and helpful in various situation; however, the fully supervised saliency model was affected by its training data domain, which may cause negative effects on applications. WSS methods without saliency maps are also beneficial. In this paper, we do not use any additional information, and we use only PASCAL VOC images with image-level labels and CNN models pre-trained with ImageNet images and their image-level labels.

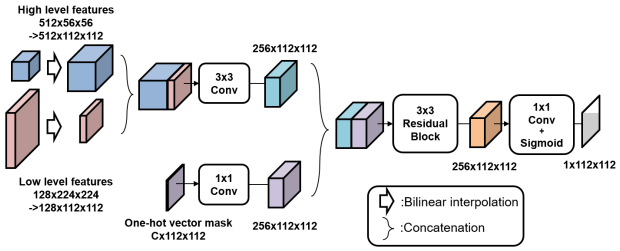

Figure 2. Difference Detection Network (DD-Net).

3. Method

There was no supervision for the mapping functions of segmentation in the weakly supervised setting; therefore, it was necessary to consider a mapping for bringing the input close to the better segmentation results by using a method that incorporated human knowledge. In this paper, we propose a method for selecting useful information from the results of the mapping functions by treating the results as supervision containing noise. We define the inputs of the mapping functions as *knowledge*, and the mapped results as *advice*. We predict the regions of *differences* between *knowledge* and *advice*, and we call this as the difference detection task. Using the inference results, we select the information of the *advice*.

3.1. Difference detection network

In this section, we formulate the difference detection task. In the proposed method, we predict the *difference* between *knowledge* and *advice*. Here, we define the segmen-

tation mask of *knowledge* as m^K, the segmentation mask of *advice* as m^A, and their *difference* as $M^{K,A} \in \mathbb{R}^{H \times W}$.

$$M_u^{K,A} = \begin{cases} 1 & \text{if } (m_u^K = m_u^A) \\ 0 & \text{if } (m_u^K \neq m_u^A) \end{cases}, \tag{1}$$

where $u \in \{1, 2, .., n\}$ indicates a location of pixels, and n is the number of pixels. Next, we define a network of difference detection for deducing the *difference*. We use feature maps extracted from a trained CNN to assist the difference detection. In particular, we use high-level features $e^h(x; \theta_e)$ and low-level features $e^l(x; \theta_e)$ extracted from a backbone network, such as ResNet. Here, x is an input image, and e is an embedding function parameterized by θ_e. As shown in Fig.3, the confidence map of the input mask d is generated by difference detection network (DD-Net), $\mathrm{DDnet}(e^h(x; \theta_e), e^l(x; \theta_e), \hat{m}; \theta_d), d \in \mathbb{R}^{H \times W}$, where $\hat{m}$ is a one-hot vector mask with the same number of channels to the target class number, θ_d is the parameter of the DD-Net, and $e(x) = (e^l(x), e^h(x))$. The architecture of DD-Net is shown in Fig.2; it consists of three convolutional layers and one Residual block with three inputs and one output. DD-Net takes either a raw mask or a processed mask as an input, and outputs the difference mask. This network performs learning using the following losses:

$$\begin{aligned} \mathcal{L}_{diff} = \frac{1}{|S|} \sum_{u \in S} (&J(M^{K,A}, d^K, u; \theta_d) \\ &+ J(M^{K,A}, d^A, u; \theta_d)), \end{aligned} \tag{2}$$

where S is a set of pixels of the input spaces, and $J()$ is assumed to be a function that returns a loss for the binary cross entropy.

$$J(M, d, u) = M_u \log d_u + (1 - M_u) \log(1 - d_u).$$

Note that the parameters of the embedding function θ_e are independent of the optimization of θ_d. The training of DD-Net is self-supervised; therefore, neither special annotation nor additional data are needed.

3.2. Self-supervised difference detection module

In this section, we describe the details of the SSDD module shown in Fig.3, which integrates two masks adaptively according to the confidence maps. We denote a set of *advice* that are true in *difference* as $S^{A,T}$, and a set of *advice* that are false as $S^{A,F}$. The purpose of the method is to extract as many samples of $S^{A,T}$ as possible from the entire set of *advice* S^A. Let d^K be the inference results of *advice* from the given *knowledge*. The inference results are the probability distributions from 0 to 1, and the values have variations. The variations are caused by the difference in the difficulty of inference. The presence of similar patterns during training can have a strong influence on the difference in the difficulty of inference. Here, if there are a sufficient number of *advice* that are true values rather than false values, that is, if $|S^{A,T}| > |S^{A,F}|$, the larger values indicate that their *advice* most likely belong to $S^{A,T}$. However, for the values

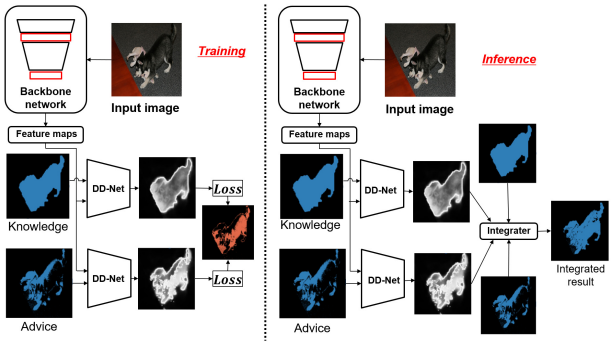

Figure 3. Overview of the DD-Net. The figure on the left shows the training of the DD-Net, and the right figure shows the processing of the integration using the results of difference detection.

of d^K at a boundary, it is not clear whether *advice* belongs to $S^{A,T}$ or not; this should probably be different from sample to sample. Therefore, it is difficult to deduce a good *advice* directly from the size of the value of d^K. To alleviate the problem, we use the inference results about the state of *knowledge* for each *advice*. Although *advice*s have large variations in their distribution, these variations are less than the variations in the distribution of *knowledge* in general. Therefore, using *advice* to infer *knowledge* is assumed to be easier than using *knowledge* to *advice* inference. In this paper, we consider the results of the inference of *knowledge* to *advice* for evaluating the difficulty of inference in each sample; we use the inferences for the thresholds for each sample. Specifically, we calculate the confidence scores of *advice* from the viewpoint of how close the values of d^K to d^A. The confidence score $w_u \in \mathbb{R}$ is defined by the following expression:

$$w_u = d_u^K - d_u^A + bias_u \tag{3}$$

Here, $bias$ is a hyper parameter for a threshold of the selection obtained by the difference detection, and it is also an enhanced value for the categories in the presence labels of the input image. The refined masks m^D obtained from m^K and m^A are defined by the following expression:

$$m_u^D = \begin{cases} m_u^A & \text{if } (w_u \geq 0) \\ m_u^K & \text{if } (w_u < 0) \end{cases} \tag{4}$$

We denote this processing flow for generating new segmentation mask as an SSDD module in the after notation.

$$m^D = SSDD(e(x), m^K, m^A; \theta_d) \tag{5}$$

4. Introducing SSDD modules into the processing flow of WSS

In this section, we explain how to use SSDD modules in the processing flow of WSS. The proposed method can be adapted to various cases by applying inputs of the mapping function as *knowledge* and the results of the mapping

function as *advice*. The processing flow that we adopted in this paper consists of two stages: the seed generation stage with static region refinement and the training stage of a segmentation model with dynamic region refinement. In the first stage, we adapted the proposed method by applying the results of PSA as *knowledge* and its CRF results as *advice* (Sec.4.1). In the second stage, we adapted the proposed method by applying the results of the first stage (Sec.4.1) as *knowledge*, and the outputs of the segmentation models trained by the masks were applied as *advice* (Sec.4.2).

4.1. Seed mask generation stage with static region refinement

PSA [1] is a method to propagate label responses to nearby areas that belong to the same semantic entity. Though PSA employs CRF for the refinement of the segmentation mask, CRF often fails to improve the segmentation masks; in fact, it degrades the masks. In this section, we refine the outputs of CRF in PSA by using the proposed SSDD module. We illustrate the processing flow of the first seed generation stage in Fig.4. Note that we omitted the input of the given image to an SSDD module for the sake of simplifying in the figure.

We denote an input image as x; the probability maps obtained by PSA are denoted as $p^{K0} = PSA(x; \theta_{psa})$, and its CRF results are denoted as p^{A0}. We obtain the segmentation masks (m^{K0}, m^{A0}) from the probability maps (p^{K0}, p^{A0}) by taking the argument of the maximum of the presence labels including a background category. We computed the loss of the DD-Net as follows:

$$\mathcal{L}_{diff0} = \frac{1}{|S|}\sum_{u\in S}(J(M^{K0,A0}, d^{K0}, u; \theta_{d0}) + J(M^{K0,A0}, d^{A0}, u; \theta_{d0})), \quad (6)$$

The proposed method is not effective when either of the segmentation masks or both of them do not have the correct labels. These cases are not only meaningless for the proposed refinement approach, but they may also harm the training of the DD-Net. We define the bad training samples by simple processing based on the difference in the number of the class-specific pixels, and we exclude them from the training.

In this work, we also train the embedding function by training a segmentation network with m^{K0} to obtain good representation for the inputs of high-level features and low-level features:

$$\mathcal{L}^{base} = \mathcal{L}^{seg}(x, m^{K0}; \theta_{e0}, \theta_{base}), \quad (7)$$

$$\mathcal{L}^{seg}(x, m; \theta) = -\frac{1}{\sum_{k\in K}|S_k^m|}\sum_{k\in K}\sum_{u\in|S_k^m|}\log(h_u^k(\theta)), \quad (8)$$

where S_k^m is a set of locations that belong to the class k on the mask m; h_u^k is the conditional probability of observing any label k at any location $u \in \{1, 2, ..., n\}$; and $\mathcal{C}$ is a set of class labels. θ_{e0} are parameters of embedding functions

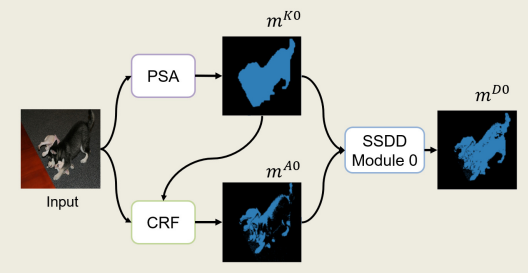

Figure 4. Processing flow at the seed mask generation stage with static region refinement.

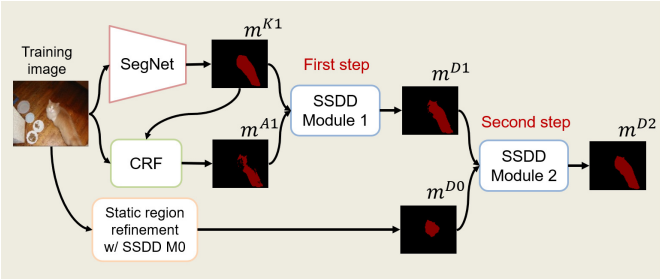

Figure 5. Illustration of the processing flow for the dynamic region refinement. ("SegNet" does not represent any specific network but represents any kind of network for fully supervised semantic segmentation.)

and θ_{base} are parameters for the segmentation branch. The training of θ_{e0} is independent of θ_{d0}.

The final loss function for the static region refinement using the difference detection is as follows:

$$\mathcal{L}_{static} = \mathcal{L}_{base} + \mathcal{L}_{diff0}. \quad (9)$$

After training, we integrate the masks (m^{K0}, m^{A0}) and obtain the integrated masks m^{D0} using the SSDD module with the trained parameter θ_{d0} as follows:

$$m^{D0} = SSDD(e(x), m^{K0}, m^{A0}; \theta_{d0}). \quad (10)$$

4.2. Training stage of a fully supervised segmentation model with a dynamic region refinement

When we train a fully supervised semantic segmentation model with pixel-level seed labels, the accuracy of the seed labels directly effects the performance of the segmentation. The performance gain is expected by replacing the seed labels to better the pixel-level labels during training. In this study, we propose a novel approach to constrain the interpolation of the seed labels during the training of a segmentation model. The idea of the constraint is to limit the interpolation of seed labels only to predictable regions of difference detection between newly generated pixel-level labels and seed labels.

In practice, we interpolate the pixel-level seed labels in two steps of each iteration as shown in Fig.5. Note

that "SegNet" in the figure does not represent a specific segmentation network; it represents any fully supervised segmentation network. In the first step, for an input image x, we obtain the outputs of the segmentation model $p^{K1} = Seg(e(x); \theta_{main})$ and its CRF outputs p^{A1}. We obtain the segmentation masks (m^{K1}, m^{A1}) from the probability maps (p^{K1}, p^{A1}) by taking the argument of the maximum of the presence labels including a background category. Then, we obtain the refined pixel-level labels m^{D1} by applying the proposed refinement method as follows: $m^{D1} = SSDD(e(x), m^{K1}, m^{A1}; \theta_{d1})$. In the second step, we apply the proposed method to the seed labels m^{D0} and to the mask m^{D1} obtained in the first step. The further refined mask m^{D2} is obtained by $m^{D2} = SSDD(e(x), m^{D0}, m^{D1}; \theta_{d2})$. We generate the mask m^{D2} in each iteration and train the segmentation model using the generated mask m^{D2}. We train the semantic segmentation model with the generated mask m^{D2} as follows:

$$\mathcal{L}_{main} = \mathcal{L}_{seg}(x, m^{D2}; \theta_{e1}, \theta_{main}), \tag{11}$$

The loss of DD-Net for m^{A1} and m^{K1} is as follows:

$$\begin{aligned}\mathcal{L}_{diff1} = \frac{1}{|S|}\sum_{u \in S}(&J(M^{K1,A1}, d^{K1}, u; \theta_{d1}) \\ &+J(M^{K1,A1}, d^{A1}, u; \theta_{d1})),\end{aligned} \tag{12}$$

In the second stage, we also exclude the bad samples (as done in Sec.static) based on the change ratio of pixels because the proposed method is not effective if the input segmentation masks do not have correct regions.

We explain how to train the DD-Net for (m^{D0},m^{D1}). The masks (m^{K1}, m^{A1}, m^{D1}) depend on the outputs of the segmentation model $Seg(e(x), \theta_{main})$. Therefore, if the learning of the segmentation model falls into a local minimum, the masks will become meaningless; all the pixels become background pixels or single foreground pixels. In this case, the inference results of the difference detection is also always constant, that is, $(D^K = 1, d^A = 1, d^A = d^K)$, and Eq.(3) becomes $w = bias$. To escape from this local minimum, we create a new branch of a segmentation model and use it for learning the difference detection between m^{D0} and m^{D1}. Assume that the mask m^{sub} was obtained from outputs of the branch of the new segmentation model $p^{sub} = Seg(e(x); \theta_{sub})$. In the training of difference detection, we trained the network to learn the differences among (m^{D0}, m^{sub}) and (m^{sub}, m^{D1}) as follows:

$$\begin{aligned}\mathcal{L}_{diff2} = \frac{1}{|S|}\sum_{u \in S}(&J(M^{D0,sub}, d^{D0}, u; \theta_{d2}) \\ &+J(M^{sub,D1}, d^{D1}, u; \theta_{d2})),\end{aligned} \tag{13}$$

If m^{sub} is the output, which is halfway between m^{D0} and m^{D1}, the replacement of the training samples will let the segmentation model exit from the situation $(d^K = 1, d^A = 1, d^A = d^K)$, and the inference results of the difference detection will predict the regions that correlate with the *difference* between m^{D0} and m^{D1}. We train the parameters θ_{sub} from the following loss to achieve the outputs that are halfway between m^{D0} and m^{D1}.

$$\mathcal{L}_{sub} = \alpha\mathcal{L}_{seg}(x, m^{D0}; \theta_{e1}, \theta_{sub}) + (1-\alpha)\mathcal{L}_{seg}(x, m^{D1}; \theta_{e1}, \theta_{sub}), \tag{14}$$

where α is a hyper parameter of the mixing ratio of m^{D0} and m^{D1} .

The final loss function of the proposed dynamic region refinement method is calculated as follows:

$$\mathcal{L}_{dynamic} = \mathcal{L}_{main} + \mathcal{L}_{sub} + \mathcal{L}_{diff1} + \mathcal{L}_{diff2} \tag{15}$$

5. Experiments

We evaluated the proposed methods using the PASCAL VOC 2012 data. The PASCAL VOC 2012 segmentation dataset has 1464 training images, 1449 validation images, and 1456 test images including 20 class pixel-level labels and image-level labels. Similar to the methodology followed by [25, 22, 16], we used the augmented PASCAL VOC training data provided by [8] as well, wherein the training image number was 10,582. For evaluation, we used an IoU metric, which is the official evaluation metric in the PASCAL VOC segmentation task. For calculating the mean IoU on the val and test sets, we used the official evaluation server. We compared the best performance of our method with the state-of-the-art methods on both the val and test sets.

5.1. Implementation details

Our experiments are heavily based on the previous research [1]. For the generating results of PSA results, we used implementations and trained parameters provided by the authors that are publicly available. We followed the methodology of [1] and set hyperparameters that gave the best performance. For the CRF parameters, we used the default settings provided by [17]. For the semantic segmentation model, we used a ResNet-38 model, which had almost the same architecture as that in [1]. The only difference was in the last upsampling rate; in the paper on PSA, the authors set the upsampling rate to 8, while we set the rate to 2 for reducing the computational cost of CRF. The input image size was 448 for training, and the test images and the output feature map size before the upsampling was 56. In the DD-Net, we used features obtained from the segmentation model before the last layer as the high-level features e^h and the features obtained before the second pooling layer as the low level features e^l. These feature map sizes were adjusted to 112 by 112 using the simple linear interpolation approach. We initialized the parameters of the segmentation models by using parameters trained with the PASCAL VOC images and their image-level labels with a pre-trained model using ImageNet, which was also provided in [1]. The codes provided by [1] did not include the training and test code for the segmentation models; therefore, we implemented our own codes. In the original paper on PSA, though the authors optimized the segmentation models by Adam; however, the performance was unstable in

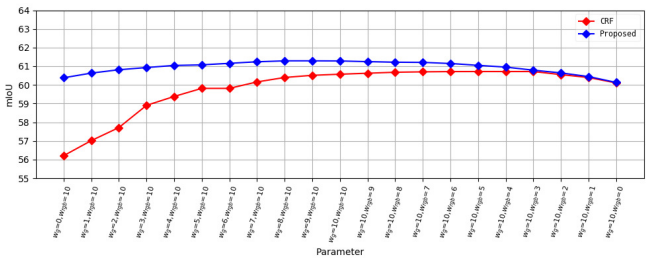

Figure 6. mIoU of the seed masks of the training images with different params values with only CRF and with SSDD and CRF.

our re-implementation, and there were several unclear settings. Therefore, we used SGD for training the entire networks. We set an initial learning rate to 1e-3 (1e-2 for initialization without the pre-trained model), and we decreased learning rate with cosine LR ramp down [20]. For the static region refinement, we trained the network with batch sizes of 16 and 10 epochs. For the dynamic region refinement, we trained the network with batch sizes of 8 and 30 epochs. For the data augmentation and inference technique, we carefully followed the methodology used in [1]. We implemented the proposed method using PyTorch. All the networks are trained using four NVIDIA Titan X PASCAL. We will open the results of the proposed method and training codes.

5.2. Analysis of static region refinement

In the proposed method, we used fully connected CRF [17] with the same parameter settings as those for PSA [1], ($w_g = 3$, $w_{rgb} = 10$,$\theta_\alpha = 80$, $\theta_\beta = 13$, $\theta_\gamma = 3$) in the following kernel potentials: $k(f_i, f_j) = w_g exp\left(-\frac{|p_i - p_j|}{2\theta_\alpha^2} - \frac{|I_i - I_j|}{2\theta_\beta^2}\right) + w_{rbg} exp\left(-\frac{|p_i - pj|^2}{2\theta_\gamma^2}\right)$. To examine the relationship between the CRF params and results, we changed the values of (w_g, w_{rgb}) and evaluated the accuracy. Fig.6 shows a comparison of the proposed static region refinement with the PSA [1] and its CRF results on the training set. The weakening of w_{rgb} decreases the difference only between the CRF and the SSDD+CRF results; therefore the effectiveness of the proposed method reduces. However, the proposed method always indicates a high accuracy. The optimal weights are different for each image, and it is expected to be difficult to search them for each image. We consider that the proposed method realized the improvement of CRF by correcting the partial failure of CRF.

Fig.7 shows the difference detection results and their refined segmentation masks. In the fourth and fifth rows of Fig.7, we show the typical failure cases of the proposed method. The regions of small objects tend to vanish in the CRF, and the DD-Net also learns such tendencies, which causes the failure of the proposed re-refinement method. In the fifth row, both of the input segmentation masks fail to provide segmentation. In such cases, the proposed method is also not effective.

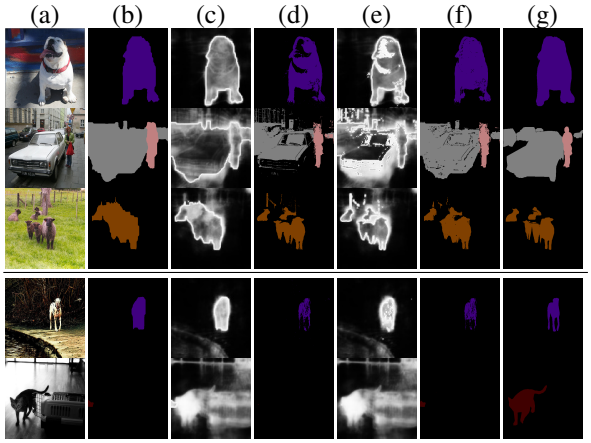

Figure 7. Each row shows (a) input images, (b) raw PSA segmentation masks, (c) difference detection maps of (b), (d) CRF masks of (b), (e) difference detection maps of (d), (f) refined segmentation masks by the proposed method, and (g) ground truth masks.

5.3. Analysis of the whole proposed method

We denote the dynamic region refinement as "SSDD" in all the tables. The score of the SSDD is with the CRF with parameters ($w_g = 3$, $w_{rgb} = 10$) that are default values from the author's public implementation. We also used the parameters for the CRF during training.

Comparison with PSA Table 1 shows the comparison of the dynamic region refinement method with the PSA. We observe that the proposed method outperforms PSA by more than 3.2 point margins. This clearly proves the effectiveness of the interpolation for the seed labels with the novel constraint by difference detection. The accuracy is greatly improved as compared with the results of the static region refinement because of the increase in the number of good *advice* by end-to-end learning of the segmentation model, that is, $|S^{A1,T}| > |S^{A0,T}|$.

In Table 1, we also show the gains between the proposed method and PSA for detailed analysis. We obtain over 10% gain on the cat, cow, horse, and sheep classes. Interestingly, all the classes that gave the large gain belonged to the animal category. However, in the potted plant, airplane, and person class objects, it was hard to improve the segmentation mask by using the proposed method. In the proposed method, we considered the precondition that *advise*, which is a true value, was larger than the value that was not a true value($|S^{A,T}| > |S^{A,F}|$). When this precondition was satisfied, the accuracy of the classes improved. If the precondition was not satisfied, the accuracy did not improve or the accuracy decreased.

Fig.8 shows the examples of the results of re-implementation of PSA, the static region refinement, and the dynamic region refinement. Dynamic region refinement shows more accurate predictions on object location and boundary. The results of the static region refinement are outputs of a segmentation model re-trained with the masks in case of ($w_g = 3$, $w_{rgb} = 10$) in Fig.6. Note that we show the results of before the CRF for detailed comparisons.

Table 1. Results on PASCAL VOC 2012 *val set*.

Methods	Bg	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Motor	Person	Plant	Sheep	Sofa	Train	Tv	**mIoU**
PSA [1]	88.2	68.2	30.6	81.1	49.6	61.0	77.8	66.1	75.1	29.0	66.0	40.2	80.4	62.0	70.4	73.7	42.5	70.7	42.6	68.1	51.6	61.7
SSDD	89.0	62.5	28.9	83.7	52.9	59.5	77.6	73.7	87.0	34.0	83.7	47.6	84.1	77.0	73.9	69.6	29.8	84.0	43.2	68.0	53.4	64.9
Gain	*+0.8*	*-5.7*	*-1.7*	*+2.6*	*+3.3*	*-1.5*	*-0.2*	*+7.6*	*+11.9*	*+5.0*	*+17.7*	*+7.4*	*+3.7*	*+15.0*	*+3.5*	*-4.1*	*-12.7*	*+13.3*	*+0.6*	*-0.1*	*+1.8*	*+3.2*

Table 2. Comparison with the WSS methods without additional supervision.

Method	Val	Test
FCN-MIL [24]ICLR2015	25.7	24.9
CCNN [23]ICCV2015	35.3	35.6
EM-Adapt [22]ICCV2015	38.2	39.6
DCSM [32]ECCV2016	44.1	45.1
BFBP [29]ECCV2016	46.6	48.0
SEC [16]ECCV2016	50.7	51.7
CBTS [28]CVPR2017	52.8	53.7
TPL [15]ICCV2017	53.1	53.8
MEFF [7]CVPR2018	-	55.6
PSA [1]CVPR2018	61.7	63.7
SSDD	**64.9**	**65.5**

Table 3. Comparison of the WSS methods with additional supervision.

Method	Additional supervision	Val	Test
MIL-seg [25]CVPR2015	Saliency mask + Imagenet images	42.0	40.6
MCNN [34]ICCV2015	Web videos	38.1	39.8
AFF [27]ECCV2016	Saliency mask	54.3	55.5
STC [37]PAMI2017	Saliency mask + Web images	49.8	51.2
Oh et al. [30]CVPR2017	Saliency mask	55.7	56.7
AE-PSL [36]CVPR2017	Saliency mask	55.0	55.7
Hong et al. [9]CVPR2017	Web videos	58.1	58.7
WebS-i2 [14]CVPR2017	Web images	53.4	55.3
DCSP [3]BMVC2017	Saliency mask	60.8	61.9
GAIN [18]CVPR2018	Saliency mask	55.3	56.8
MDC [38]CVPR2018	Saliency mask	60.4	60.8
MCOF [35]CVPR2018	Saliency mask	60.3	61.2
DSRG [11]CVPR2018	Saliency mask	61.4	63.2
Shen et al. [31]CVPR2018	Web images	63.0	63.9
SeeNet [10]NIPS2018	Saliency mask	63.1	62.8
AISI [6]ECCV2018	Instance saliency mask	63.6	64.5
SSDD	-	**64.9**	**65.5**

Comparison with the state-of-the-art methods Table 2 shows the results of the proposed method and the recent weakly supervised segmentation methods that do not use additional supervisions on the PASCAL VOC 2012 validation data and PASCAL VOC 2012 test data. We observed that our method achieves the highest score as compared with all the existing methods, which use the same types of supervision [23, 22, 32, 29, 16, 15, 28, 7, 1]. The proposed method outperforms the recent previous works on MEFF and TPL by large margins. As discussed earlier, the proposed method also outperforms the current state-of-the-art methods [1]. This result clearly indicates the effectiveness of the proposed method.

Table 3 shows the comparison of the proposed method with a few weakly supervised segmentation methods that employ relatively cheap additional information. Surprisingly, the proposed method also outperforms all the listed weakly supervised segmentation methods. The proposed methods outperformed the following methods: SeeNet [29], DSRG [37], MDC [16], GAIN [18], and MCOF [35] that employed fully supervised saliency methods. In addition, the score of the proposed method was also better than the results of AISC [6], which used instance-level saliency map methods. Note that AISC achieved 64.5% on the val set and 65.6% on the test set using an additional 24,000 ImageNet images for training. The score of the proposed method was also higher than the score of Shen et al. [31], which used 76.7k web images for training. It is not possible to have a completely fair comparison for them because of the difference of the network model, the augmentation technique, the number of iteration epochs, and so on. However, the proposed method demonstrates comparable performance or better performance without any additional training information.

6. Conclusions

In this paper, we proposed a novel method to refine a segmentation mask from a pair of segmentation masks before and after the refinement process such as the CRF by using the proposed SSDD module. We demonstrated that the proposed method could be used effectively in two stages: the static region refinement in the seed generation stage and the dynamic region refinement in the training stage. In the first stage, we refined the CRF results of PSA [1] by using the SSDD module. In the second stage, we refined the generated semantic segmentation masks by using a fully supervised segmentation model and CRF during the training. We demonstrated that three SSDD modules could greatly boost the performance of WSS and achieve the best results on the PASCAL VOC 2012 dataset over all the weakly supervised methods with and without additional supervision.

Acknowledgements This work was supported by JSPS KAKENHI Grant Number 17J10261, 15H05915, 17H01745, 17H06100 and 19H04929.

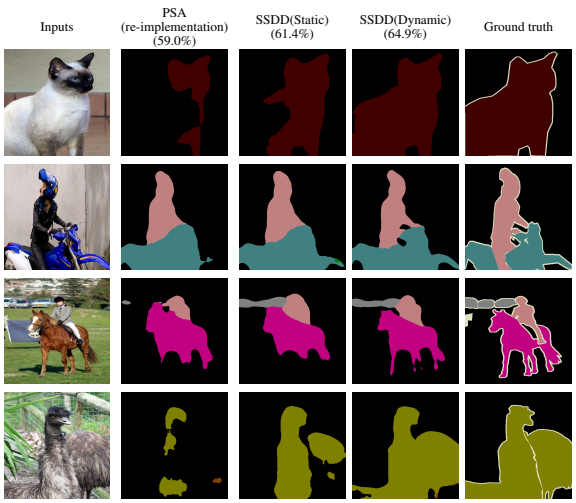

Figure 8. Segmentation examples of results on PASCAL VOC 2012.

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *CVPR*, 2018. 2, 5, 6, 7, 8, 12

[2] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *ECCV*, 2016. 3

[3] Arslan Chaudhry, K. Puneet Dokania, and H.S. Philip Torr. Discovering class-specific pixels for weakly-supervised semantic segmentation. In *Proc. of British Machine Vision Conference*, 2017. 3, 8, 12

[4] Zezhou Cheng, Qingxiong Yang, and Bin Sheng. Deep colorization. In *ICCV*, 2015. 2

[5] Carl Doersch, Abhinav Gupta, and A. Alexei Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015. 2

[6] Ruochen Fan, Qibin Hou, Ming-Ming Cheng, Gang Yu, R. Ralph Martin, and Shi-Min Hu. Associating inter-image salient instances for weakly supervised semantic segmentation. In *ECCV*, 2018. 3, 8, 12

[7] Weifeng Ge, Sibei Yang, and Yizhou Yu. Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning. In *CVPR*, 2018. 8, 12

[8] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *ECCV*, 2014. 6

[9] Seunghoon Hong, Donghun Yeo, Suha Kwak, Honglak Lee, and Bohyung Han. Weakly supervised semantic segmentation using web-crawled videos. In *CVPR*, 2017. 3, 8, 12

[10] Qibin Hou, Peng-Tao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In *NIPS*, 2018. 2, 3, 8, 12

[11] Zilong Huang, Wang Xinggang, Wang Jiasi, Wenyu Liu, and Wang Jingdong. Weakly-supervised semantic segmentation network with deep seeded region growing. In *CVPR*, 2018. 3, 8, 12

[12] Huaizu Jiang, Zejian Yuan, Ming-Ming Cheng, Yihong Gong, Nanning Zheng, and Jingdong Wang. Salient object detection: A discriminative regional feature integration approach. In *CVPR*, 2013. 2

[13] Zhang Jianming, Lin Zhe, Brandt Jonathan, Shen Xiaouhui, and Stan Sclaroff. Top-down neural attention by excitation backprop. In *ECCV*, 2016. 2

[14] Bin Jin, Maria V. Ortiz Segovia, and Sabine Susstrunk. Webly supervised semantic segmentation. In *CVPR*, 2018. 3, 8, 12

[15] Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon. Two-phase learning for weakly supervised object localization. In *ICCV*, 2017. 2, 8, 12

[16] Alexander Kolesnikov and Christoph H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*, 2016. 2, 6, 8, 12

[17] Philipp Krahenbuhl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011. 1, 2, 6, 7

[18] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernest, and Yun Fu. Tell me where to look: Guided attention inference network. In *CVPR*, 2018. 8, 12

[19] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2

[20] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 7

[21] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014. 2

[22] George Papandreou, Liang-Chieh Chen, Kevin Murphy, and Alan L. Yuille. Weakly-and semi-supervised learning of a dcnn for semantic image segmentation. In *ICCV*, 2015. 2, 3, 6, 8, 12

[23] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*, 2015. 2, 3, 8, 12

[24] Deepak Pathak, Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional multi-class multiple instance learning. In *ICLR*, 2015. 2, 8, 12

[25] Pedro O. Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, 2015. 2, 3, 6, 8, 12

[26] Jordi Pont-Tuset, Pablo Arbelaez, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping. In *CVPR*, 2014. 3

[27] Xiaojuan Qi, Zhengzhe Liu, Jianping Shi, Hengshuang Zhao, and Jiaya Jia. Augmented feedback in semantic segmentation under image level supervision. In *ECCV*, 2016. 3, 8, 12

[28] Anirban Roy and Sinisa Todorovic. Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation. In *CVPR*, 2017. 2, 8, 12

[29] Fatemehsadat Saleh, Mohammad Sadegh Ali Akbarian, Mathieu Salzmann, Lars Petersson, Stephen Gould, and Jose M. Alvares. Built-in foreground/background prior for weakly-supervised semantic segmentation. In *ECCV*, 2016. 2, 3, 8, 12

[30] Joon Oh Seong, Benenson Rodrigo, Khoreva Anna, Akata Zeynep, and Fritz Mario. Exploiting saliency for object segmentation from image level labels. In *CVPR*, 2017. 3, 8, 12

[31] Tong Shen, Guosheng Lin, Chunhua Shen, and Reid Ian. Bootstrapping the performance of webly supervised semantic segmentation. In *CVPR*, 2018. 2, 3, 8, 12

[32] Wataru Shimoda and Keiji Yanai. Distinct class saliency maps for weakly supervised semantic segmentation. In *ECCV*, 2016. 2, 8, 12

[33] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR WS*, 2014. 1, 2

[34] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Weakly-supervised semantic segmentation using motion cues. In *ECCV*, 2016. 3, 8, 12

[35] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *CVPR*, 2018. 3, 8, 12

[36] 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 *CVPR*, 2017. 2, 3, 8, 12

[37] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. STC: A simple to complex framework for weakly-supervised semantic segmentation. In *IEEE Trans. on PAMI*, 2017. 1, 2, 3, 8, 12

[38] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly- and semisupervised semantic segmentation. In *CVPR*, 2018. 3, 8, 12

[39] Matthew D. Zeiler and Rob Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *ICCV*, 2011. 1

[40] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 2

[41] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 1, 2