IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 12, 2022, accepted January 31, 2022, date of publication February 7, 2022, date of current version February 23, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3149587

Pseudo-Label-Free Weakly Supervised Semantic Segmentation Using Image Masking

SANGTAE KIM, LUONG TRUNG NGUYEN, KYUHONG SHIM, JUNHAN KIM, AND BYONGHYO SHIM, (Senior Member, IEEE)

Department of Electrical and Computer Engineering, Seoul National University, Gwanak-gu, Seoul 08826, South Korea

Corresponding author: Byonghyo Shim (bshim@islab.snu.ac.kr)

This work was supported in part by the Research Grant of Samsung Electronics Company Ltd., under Grant IO210201-08353-01; and in part by the National Research Foundation of Korea (NRF) Grant by the Korean Government through MSIT under Grant 2020R1A2C2102198.

ABSTRACT Weakly-supervised semantic segmentation (WSSS) aims to train a semantic segmentation network using weak labels. Recent approaches generate the pseudo-label from the image-level label and then exploit it as a pixel-level supervision in the segmentation network training. A potential drawback of the conventional WSSS approaches is that the pseudo-label cannot accurately express the object regions and their classes, causing a degradation of the segmentation performance. In this paper, we propose a new WSSS technique that trains the segmentation network without relying on the pseudo-label. Key idea of the proposed approach is to train the segmentation network such that the object erased by the segmentation map is not detected by the classification network. From extensive experiments on the PASCAL VOC 2012 benchmark dataset, we demonstrate that our approach is effective in WSSS.

INDEX TERMS Weakly-supervised semantic segmentation, image classification, visual attention, saliency map.

I. INTRODUCTION

Image semantic segmentation, a task to classify each pixel among the interested classes, is an important problem with a wide range of applications such as autonomous driving, medical diagnosis, industrial automation, and aerial imaging [1], [2]. Recently, deep neural networks (DNN)-based semantic segmentation has received special attention due to its excellent segmentation performance [3], [4]. A potential drawback of the DNN-based approach is that a large number of fully-annotated data are needed to train the networks. Since the generation of fully-annotated dataset is laborious, alternative approaches such as unlabeled or weakly-labeled learning have been suggested in recent years [5]–[7]. There are various forms of weak labels such as image-level labels [8], points [9], scribbles [10], and bounding boxes [11]. Among these, image-level label is popularly used for its simplicity [12]–[14]. In essence, image-level label indicates whether the foreground objects appear in an image or not (e.g., bird is in an image and cat is not). We henceforth refer to the DNN-based semantic segmentation using the image-level labels as weakly-supervised semantic segmentation (WSSS).

A central challenge of WSSS is that the image-level labels do not provide information on object regions required to train the semantic segmentation networks. A simple way to localize object regions is to use class activation mapping [15]. Basically, this approach figures out what regions in the image are relevant to the semantic classes. The localization map obtained from this technique, called *class activation map* (CAM), indicates the discriminative object regions. In recent WSSS approaches, CAM is used to generate a pseudo-label for the training of semantic segmentation network [5], [16]. While the pseudo-label can well express the object region of interest, it might cause some potential problems hindering the accurate image segmentation. First, the object extent in the non-discriminative region is not accurately expressed (see Fig. 1-(a)). This is because the classification network focuses only on the existence of the objects so that the network tends to ignore the non-discriminative regions which are also parts of the objects. Second, the class assigned in each pixel of the pseudo-label might not be correct when an image contains multiple objects with distinct classes (see Fig. 1-(b)) since the CAMs are spread to unwanted regions outside the

The associate editor coordinating the review of this manuscript and approving it for publication was Turgay Celik.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

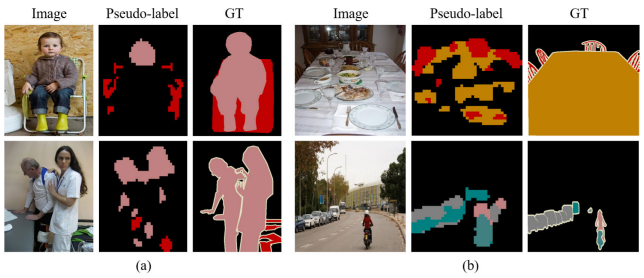


FIGURE 1. Problems of pseudo-labels obtained from CAM. (a): The labeled regions are focused on the most discriminative object regions (e.g., face of person). (b): The labeled regions are misclassified.

foreground objects. For these reasons, an approach that trains the semantic segmentation network using the pseudo-label might not achieve the satisfactory performance in many practical scenarios.

An aim of this paper is to propose a novel WSSS technique that can train the semantic segmentation network without relying on the pseudo-label. Basically, our approach is inspired by the visual attention mechanism of the human visual system (HVS) [17]. When HVS perceives the visual information, HVS focuses on the desired object without being interfered by other objects. In order to mimic the human behavior and thereby reduce the interference from irrelevant regions, we mask an input image using an attention map that guides which pixels to attend or ignore. Specifically, a segmentation network generates the segmentation map describing the discovered object regions. Then, the attention map is generated by collecting the discovered regions in the segmentation maps of interesting classes. We exploit the attention map in erasing the discovered regions and therefore focus on the remaining regions in the masked image.

In order to check whether the objects are erased properly, we employ a classification network trained for multi-class multi-label classification. In summary, in the training process, the segmentation network tries to generate the segmentation map covering the object regions. Then, for the image masked by the segmentation map, the classification network tries to find out the interesting objects. For example, when an image contains bird and car, a segmentation network is guided to generate an accurate map of bird or car. If the generated segmentation map contains the bird, then the bird is erased in the masked image, helping the detection of a car in the classification network.

To train the segmentation network in the absence of the pseudo-label, we adopt a novel combination of two complementary loss functions: *attention loss* and *saliency loss*. The attention loss is used to penalize the segmentation network if the segmentation map does not completely cover the objects of a target class. The saliency loss is used to encourage the segmentation network to learn the accurate object extent which cannot be identified by the classification network. By learning the object classes using the attention loss and object extent using the saliency loss, the segmentation network can segment the image without obtaining the class-specific knowledge from pixel-level supervision.

As a means to enhance the segmentation performance, we propose a training strategy for the classification network and a refining technique for the saliency map. First, for the training of the classification network, we exploit the dilated convolutional blocks (see Fig. 3). The dilated convolutional blocks are used to find out the object regions outside the most discriminative regions. The regions discovered by dilated convolutional blocks are then used as an additional supervision for the classification network in finding out complete object regions. Second, we refine the saliency map using the CAM obtained by the classification network (see Fig. 4). Note that the value in each pixel of the CAM indicates the probability of an object being contained in that pixel. Using these values, we can find out the missing objects and also remove the unwanted objects in the saliency map.

The main contributions of this paper are as follows:

- We propose a novel segmentation technique for weakly-supervised semantic segmentation. In our work, instead of learning the class-specific knowledge from the pseudo-label, the segmentation network learns the class-specific knowledge directly from the classification network by exploiting the image masking technique.
- We propose a training strategy for the semantic segmentation network. In the proposed approach, the segmentation network is trained using the combination of the attention loss and the saliency loss to accomplish the semantic segmentation task (see Section III-C).
- From numerical experiments on *val* and *test* of the PASCAL VOC 2012 semantic segmentation benchmark [18], we show that our approach achieves mean-intersection-over union 66.5% and 66.9% using VGG16-based network and 69.0% and 69.2% using ResNet101-based network, respectively, which are competitive with the state-of-the-arts.

II. RELATED WORK

A. WEAKLY-SUPERVISED SEMANTIC SEGMENTATION

Image-level label has been used in many WSSS approaches due to its simplicity. Early works include multiple-instance learning [12], constrained optimization [19], and expectation-maximization techniques [20]. Recently, the class activation mapping technique that finds out the most discriminative object regions has been used to generate a pixel-level pseudo-label from the image-level label [15]. The generated pseudo-label depicting the reliable object regions is used as a supervision for the semantic segmentation network. The segmentation performance of this approach depends strongly on the accuracy of the generated pseudo-labels. Hence, it is of importance to find out accurate object regions for the proper training of the semantic segmentation network.

In order to obtain a reliable pseudo-label, various segmentation techniques have been proposed. In [5], fully-connected conditional random field (CRF) is applied to the predicted segmentation maps to refine the object boundaries. In [14],

seeded region growing technique is used to assign classes to unlabeled pixels. Recently, approaches generating the reliable pseudo-labels without relying on the segmentation algorithms have been proposed. In [21], for example, a large number of localization maps are generated and then aggregated into a single localization map. In [22], the localization maps are accumulated through the training process to collect the discriminative regions of different parts in the objects. In [23], multiple dilated convolutional blocks are used to enlarge the receptive fields and transfer the discriminative information to the non-discriminative regions. In [24], adversarial manipulation technique is used to expand the discriminative object regions.

In a nutshell, the proposed approach is a bit similar to the CAM-based approach in the sense that we find out the object regions from CAM. However, key distinctive point of the proposed approach is that the segmentation network learns the classes of pixels by directly utilizing the classification network meaning that we can train the segmentation network without relying on the pseudo-label.

B. VISUAL ATTENTION

Visual attention, an approach to select the search regions and analyze their effects, has been applied to various computer vision tasks such as image classification [25], object detection [26], and image caption generation [27]. In the semantic segmentation, visual attention is often implemented using the image masking, a technique to erase part of an image. In many approaches, discovered object regions are erased to help the discovery of new object regions [28]–[31]. For example, in [28], [29], discovered object regions are repetitively erased to find out new object regions. In [30], two-phase learning strategy has been proposed to get a complete region of the foreground objects from the attention maps of two networks. The drawback of these approaches is that it is difficult to figure out whether the masked image still contains part of foreground objects or not. As a consequence, one might simultaneously find out unwanted background objects and the main foreground objects (e.g., water with boat, rail with train). In [31], discriminative object regions are erased to guide the network to find out new object regions. In [32], discriminative object regions are suppressed to spread the attention of the network to adjacent non-discriminative object regions.

In [16], [33], visual attention mechanism is applied to the adversarial learning. In these approaches, an attention map obtained from the main network is used to mask an input image and then the masked image is delivered to the adversarial network. By training the network using the adversarial loss function, the main network is encouraged to generate an attention map which makes the adversarial network output consistent with the image-level label. In [33], an adversarial network is used to discriminate whether the input map is ground truth or generated from the segmentation network. In [16], an input image is masked by the self-attention map. The masked image is passed to the adversarial network to check if the attention map covers regions contributing to the classification output.

C. SALIENCY DETECTION

The main goal of the salient object detection is to identify the visually distinctive objects (or regions) in an image and then segment them out from the background. Since the image-level label does not contain any information on the background regions in WSSS systems, one cannot directly find out the confident background regions using the classification network. To overcome this limitation, the saliency map has been widely used in many WSSS approaches [5], [14], [16], [21], [22]. Key idea of these schemes is to identify the background regions using the pixels with low salient probabilities.

In [34]–[38], the saliency map is directly used in the training process of the segmentation networks. For example, in [34], the segmentation network is trained using the saliency maps of simple images to generate the pseudo-labels for complex images. In [35], saliency maps are used to supplement non-discriminative object regions. In [36], saliency maps are exploited to guide the seeded region growing method. In [37], saliency-guided self-attention module is used to capture rich contextual information for discovering the integral extent of objects and retrieving high-quality pseudo-label. In [38], an approach that trains the network using pixel-level feedback from combination of saliency maps and image-level labels has been proposed.

III. PROPOSED WEAKLY-SUPERVISED SEMANTIC SEGMENTATION NETWORK

In this section, we discuss the proposed WSSS framework. We first discuss the classification network training using dilated convolutional blocks and then discuss the refinement of the saliency maps using CAMs obtained from the classification network. We also explain how to train the semantic segmentation network using the image masking technique. The overall network architecture is illustrated in Fig. 2

A. TRAINING OF CLASSIFICATION NETWORK

A classification network is a key ingredient in our approach. Basically, the classification network is trained using the multi-class multi-label classification loss. One well-known problem in the conventional classification network is that the network cannot detect the non-discriminative object regions. To address this issue, we use an extra-supervision on the non-discriminative object regions in the training of the classification network. To find out the non-discriminative object regions, we use a dilated convolution which enlarges the receptive field without changing the computational cost [39]. With the increased receptive field, the information in the discriminative object regions can be transferred to distant regions, helping the detection of the non-discriminative object regions.

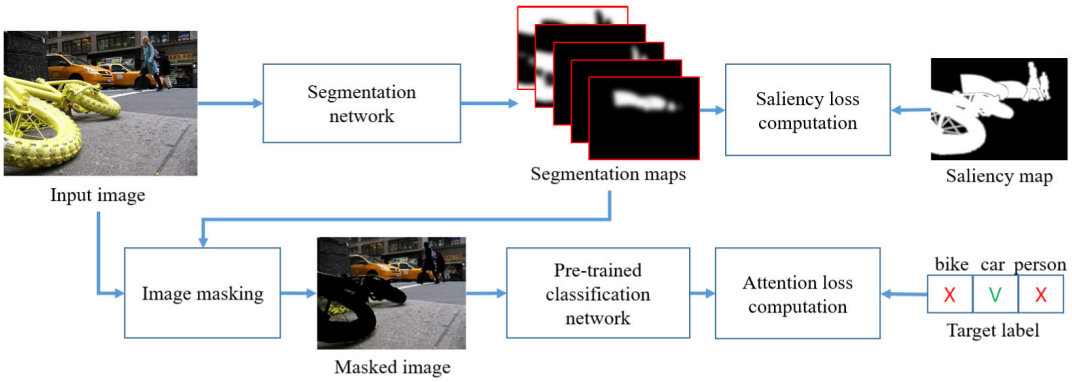


FIGURE 2. The overview of the proposed method to train weakly-supervised semantic segmentation network.

In the training of the classification network, we append dilated convolutional blocks to the classification network (see Fig. 3). The dilated convolutional blocks are similar to the standard convolutional block except that their first convolutional layers have unique dilation rates d. Let M^0 be the CAM obtained from the standard convolutional block and $M^1, \cdots, M^D$ be the CAMs obtained from D dilated convolutional blocks. Then, the object regions found by multiple dilated convolutional blocks are added to M^0 using the max-fusion to supplement the non-discriminative object region. A dense CAM, denoted as M, covering the discriminative and non-discriminative object regions is obtained as $M = \max(M^0, \frac{1}{D}\sum_{i=1}^{D}(M^i))$.

The classification network is trained using the multi-class multi-label classification loss and the CAM loss. First, the multi-class multi-label classification loss ℓ_{sig} is

$$\ell_{sig} = \frac{1}{C}\sum_{i\in\mathcal{D}}\sum_{c\in\mathcal{C}}\Big(-t_c\log(\sigma(\widehat{z}_{ic})) - (1-t_c)\log(1-\sigma(\widehat{z}_{ic}))\Big), \quad (1)$$

where C is the number of foreground classes, $\mathcal{D}$ is the set of indices of convolutional blocks, $\mathcal{C}$ is the set of indices of foreground classes, t_c is the image-level label for class c, $\widehat{z}_{ic} = \text{GAP}(M_c^i)$ is the predicted class score for class c (GAP is the global average pooling operation), and $\sigma(x) = 1/(1+e^{-x})$ is the sigmoid function. Second, the CAM loss ℓ_{cam}, used to match the CAM M^0 to the dense CAM M, is the mean square error (MSE) between M and M^0:

$$\ell_{cam} = \frac{1}{C|\mathcal{S}|}\sum_{u\in\mathcal{S}}\sum_{c\in\mathcal{C}^p}(\phi(M_{u,c}) - \phi(M^0_{u,c}))^2 \quad (2)$$

where $\mathcal{S}$ is the set of all positions, $\mathcal{C}^p$ is the set of indices of the present classes, and $\phi(x) = \max(0, x)$ is the ReLU activation function. Also, $M_{u,c}$ is the class score of class c at position u of class activation map M.

The overall loss ℓ_{cls} for training the classification network is

$$\ell_{cls} = \ell_{sig} + \lambda_{cls}\ell_{cam} \quad (3)$$

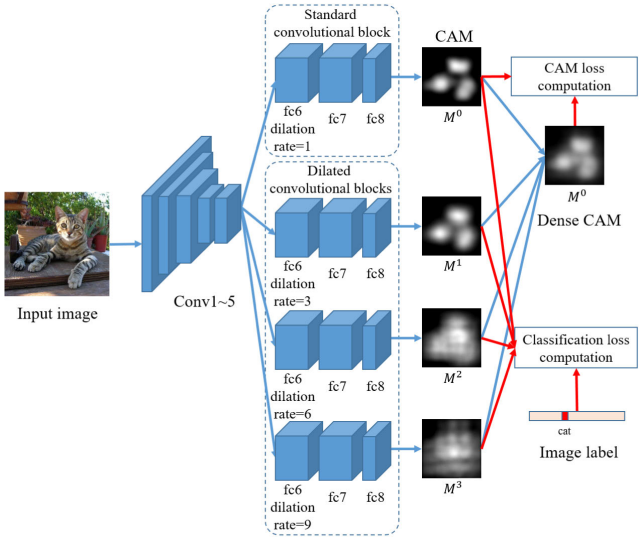


FIGURE 3. The architecture for the training of the classification network using dilated convolutional blocks. In this example, three dilated convolutional blocks with dilation rates {3,6,9} are appended to the classification network.

where λ_{cls} is the weighting factor for balancing two losses.

B. SALIENCY MAP REFINEMENT

In the segmentation network training, the saliency map is used to learn which pixels belong to either background or foreground regions. While the saliency detector (SD) can find out the detailed shape of the objects, it might also find out unwanted background objects or miss interesting foreground objects since SD is trained without the semantic classes. To overcome this potential drawback, we correct the pixels in the saliency map based on the CAM score. The score in each pixel indicates the probability of an object being contained in that pixel. Since the object detection in the classification network is fairly accurate, we can readily find out the missing foreground regions from the high-scored pixels in the CAM. Note that this does not necessarily mean that the low-scored pixels belong to the background regions since these pixels

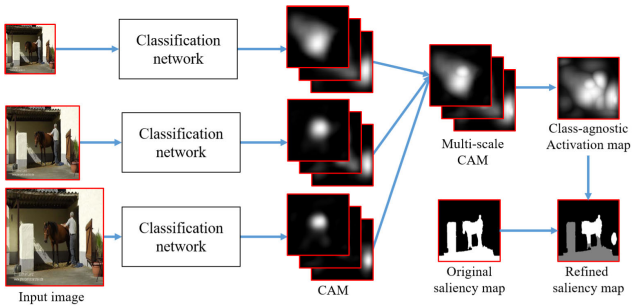


FIGURE 4. Overall procedure to refine the saliency map using the CAM.

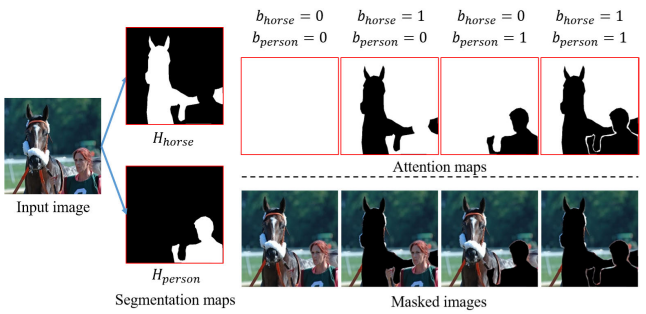


FIGURE 5. Illustration for examples of the attention maps and masked images.

might belong to the non-discriminative object regions. From our extensive experiments, we observe that correcting these pixels to the background pixels causes a degradation of the segmentation performances. In our work, we set pixels with low scores to unlabeled pixels.

In Fig. 4, we illustrate the overall procedure of refining the saliency map. We first obtain the CAM of an input image from the classification network. To improve the reliability of the CAM, we merge the CAMs of multiple scaled input images. Let $M^0(s_i)$ be the CAM of an input image scaled by a factor s_i ($s_i \in \{s_0, \cdots, s_n\}$), then the reliable CAM M^* is obtained as

$$M^*_{u,c} = \max_i scale(\phi(M^0_{u,c}(s_i))) \tag{4}$$

where *scale* is the scaling operator that changes the size of map to the size of the input image. To obtain a map expressing the foreground object regions, we merge the CAMs of present classes, generating a class-agnostic activation map B whose pixels indicate the probabilities of an object being contained in that pixel:

$$B_u = \max_{c \in \mathcal{C}^p} \frac{M^*_{u,c}}{\max_u M^*_{u,c}}. \tag{5}$$

If B_u is larger than the pre-defined threshold τ_1 and the pixel u belongs to the background regions in the saliency map O, we consider this pixel as a foreground pixel. On the other hand, if B_u is smaller than the pre-defined threshold τ_2 and the pixel u belongs to the foreground regions, we consider this pixel as an unlabeled pixel. That is, the refined saliency map R is obtained as

$$R_u = \begin{cases} 1, & \text{if } B_u > \tau_1 \text{ and } O_u = 0 \\ \text{unlabeled}, & \text{if } B_u < \tau_2 \text{ and } O_u = 1 \\ O_u, & \text{else.} \end{cases} \tag{6}$$

C. TRAINING OF SEGMENTATION NETWORK

For the training of the segmentation network, we use the saliency loss that encourages the segmentation network to learn the object regions from the saliency map. While the segmentation map has $C + 1$ classes, the saliency map has only two classes (background and foreground). To connect the segmentation map to the saliency map, we design the background map H^b and foreground map H^f from the segmentation map.

Let H_c be the segmentation map of class c ($c = 0$ denotes the background class). Then, the background map is $H^b = H_0$ and the foreground map is $H^f = \sum_{c \in \mathcal{C}^p} H_c$. Let $\mathcal{S}^b$ and $\mathcal{S}^f$ be the set of positions of background and foreground pixels in saliency map, respectively. Then, the saliency loss is defined as a weighted cross-entropy with only two classes (background and foreground):

$$\ell_{sal} = -\sum_{u \in \mathcal{S}^b} \frac{1}{|\mathcal{S}^b|} \log H^b_u - \sum_{u \in \mathcal{S}^f} \frac{1}{|\mathcal{S}^f|} \log H^f_u, \tag{7}$$

where u is the position of the pixels. The weights for background and foreground pixels are set to $\frac{1}{|\mathcal{S}^b|}$ and $\frac{1}{|\mathcal{S}^f|}$, respectively, to balance the losses for background and foreground pixels. The first and second terms in (7) correspond to the loss for background and foreground classes, respectively. Note that the losses on unlabeled pixels in the saliency map are not computed during the training process. To improve reliability of the network in various scales, we feed the multiple scaled input images to the network and compute the losses individually. Thus, the resulting saliency loss is the sum of cross-entropy losses for $|S|$ scaled outputs (S is the set of input scales).

One potential weakness using the saliency loss is that the segmentation network might predict the class of pixel incorrectly since the class of each pixel is unspecified in the saliency map. In order to make sure that the segmentation network predicts the correct class for each pixel, we exploit the image masking technique in the training of the segmentation network. During the training process, an input image is masked using an attention map F that designates which regions are erased. The attention map is obtained from the predicted regions in the segmentation map:

$$F = 1 - \sum_{c \in \mathcal{C}^p} H_c b_c \tag{8}$$

where b_c is the binary random number that decides whether the segmentation map H_c is erased in the attention map or not. Using the F, the masked image I' can be expressed as

the product of the input image I and the attention map F:

$$I' = F \odot (I - \mu) \tag{9}$$

where $\odot$ is the element-wise multiplication and μ is the RGB mean of the training images. For a given class c, when $b_c = 1$, we expect that the objects of class c are erased in I'. Whereas, when $b_c = 0$, we expect that the objects of class c remain in I'. Hence, it is natural to choose $t' = t(1 - b)$ as the modified label corresponding to I'. We illustrate the attention maps and masked images corresponding to b in Fig. 5.

The segmentation network is trained to predict the correct object regions so that the class score corresponding to I' matches t'. An associated loss, the attention loss ℓ_{attn} is defined as the cross-entropy between the class score $\widehat{z'}$ of I' and the target label t':

$$\ell_{attn} = \frac{1}{|\mathcal{C}^p|} \sum_{c \in \mathcal{C}^p} \Big(- t'_c \log(\sigma(\widehat{z'}_c)) - (1 - t'_c) \log(1 - \sigma(\widehat{z'}_c)) \Big). \tag{10}$$

In contrast to the classification loss ℓ_{sig}, the attention loss only considers the present classes. By generating multiple masked images using different attention maps, we can investigate the effects of different combinations of the segmentation maps. As illustrated in Fig. 6, we can add additional classification paths for other masked images generated using different attention maps. In each path, the same classification network is employed to compute the class score and the attention loss individually. The total attention loss is computed as the average of the attention losses:

$$\ell_{total_attn} = \frac{1}{|\mathcal{N}|} \sum_{n \in \mathcal{N}} \ell_{attn}^{(n)} \tag{11}$$

where N is the number of classification paths.

In summary, an overall loss for training the semantic segmentation network is

$$\ell_{seg} = \ell_{sal} + \lambda_{seg} \ell_{total_attn} \tag{12}$$

where λ_{seg} is the weighting factor for balancing two losses in the segmentation loss. Note that during the training process of the segmentation network, we fix the parameters of the classification network to keep its learned knowledge.

IV. EXPERIMENTS

A. DATASET AND EXPERIMENT SETTINGS

We evaluate the proposed approach on the PASCAL VOC 2012 segmentation benchmark dataset [18] which has 20 foreground classes and one background class. This dataset has 1,464 training images, 1,449 validation images, and 1,456 test images. As in many practices [4], [28], we use augmented training dataset consisting of 10,582 images [40]. In our experiments, we only utilize image-level annotations for the network training. We employ the saliency detector [41] to obtain saliency map that expresses class-agnostic pixel-wise object scores. As a performance measure, we use mean intersection-over-union (mIOU), average of IOUs over 21 categories. We obtain the result on the test set by submitting the predicted results to the official PASCAL VOC evaluation server.

For the classification network, we employ VGG16 [42] pre-trained on ImageNet classification dataset [43]. As illustrated in Fig. 3, we replace the last three fully-connected (fc) layers in VGG16 with a standard convolutional block consisting of three convolutional layers. The convolutional blocks consist of two 3×3 convolutional layers (fc6 and fc7 both 1024 outputs) and one 1×1 convolutional layer (fc8). We append three dilated convolutional blocks to the classification network (see Fig. 3). The dilation rates in three dilated convolutional blocks are set to $d = \{3, 6, 9, 12, 15, 18, 21, 24\}$. The parameters of the standard and dilated convolutional blocks are initialized from the normal distribution. We apply the GAP layer after fc8 for the training of the classification network.

For the segmentation network, we employ DeepLab-ASPP [4] whose backbone architecture is either VGG16 [42] or ResNet101 [44]. We initialize the parameters of VGG16- and ResNet101-based DeepLab using the convolutionalized VGG16 and ResNet101 pre-trained on MS-COCO [45], respectively. For the last layer, the parameters are initialized from the normal distribution. When training the segmentation network, we use the classification network only with standard convolutional block (i.e., the dilated convolutional blocks are removed). In the training of the ResNet101-based DeepLab, we only update parameters of convolutional layers while fixing the parameters of batch normalization layers. The softmax output of the segmentation network is post-processed by CRF with default parameters [46].

To improve the robustness of the classification network and the segmentation network, we apply data augmentation techniques. We randomly flip and scale (from 0.5 to 1.5) input images. The resulting images are cropped to 321×321 at random location. We also apply color augmentation techniques by randomly changing brightness, contrast, saturation, and hue. We use multi-scale inputs with scales, $S = \{1, 0.75, 0.5\}$ in both training and test phases [4], [47]. We use stochastic gradient descent optimizer with the momentum 0.9. We set the weight decay to 0.0005 and the batch size to 20. We employ polynomial learning rate policy [48] with initial learning rate 10^{-3} and power 0.9, i.e., learning rate $= 10^{-3} \times (1 - \frac{iter}{maxiter})^{0.9}$. The learning rate for the last layers are multiplied by 10. We set the two thresholds τ_1 and τ_2 in (6) used to refine the saliency map to 0.8 and 0.3, respectively, which are found by grid search. The weighting factors λ_{cls} in (3) and λ_{seg} in (12) are set to 0.1 and 2, respectively. The entries of the binary random vectors b are drawn uniformly. We train the classification network and the segmentation network for 50 and 30 epochs, respectively. Our approach is implemented based on Tensorflow [49]. The classification network and the segmentation network are trained on a single NVIDIA GeForce Titan Xp.

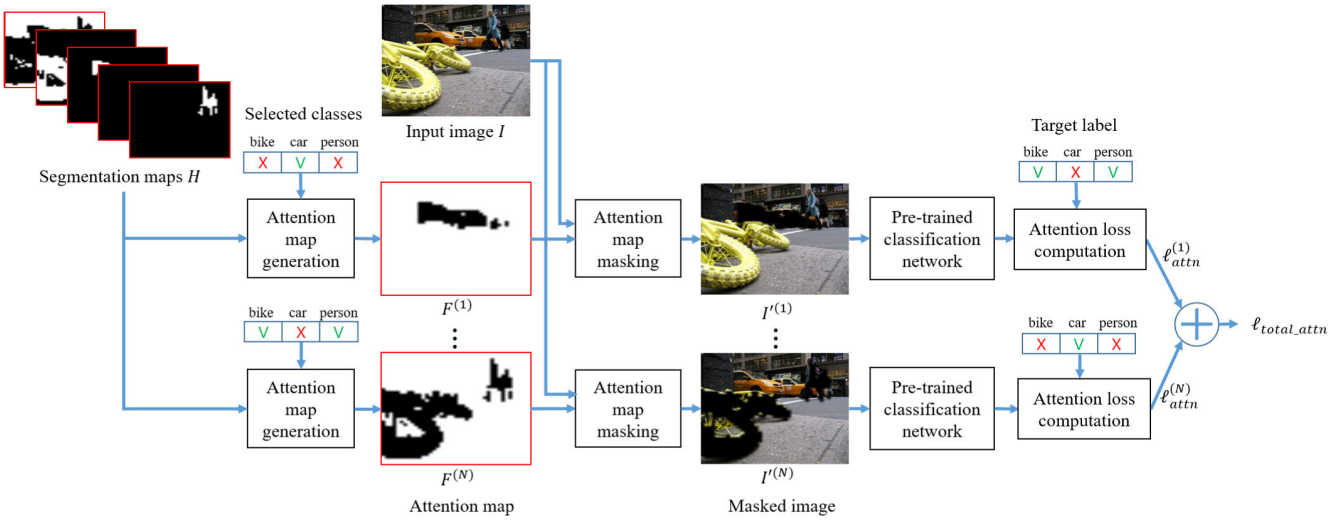


FIGURE 6. Illustration for the attention loss computation.

TABLE 1. Comparison of VGG16-based weakly-supervised semantic segmentation methods' mean IOUs on PASCAL VOC 2012 *val* and *test* set.

Method	Use saliency map	Need pseudo-label?	Val	Test
SEC [5]	✓	✓	50.7	51.1
TPL [30]	✓	✓	53.1	53.8
AE-PSL [28]	✓	✓	55.0	55.7
DCSP [29]	✓	✓	58.6	59.2
GAIN [16]	✓	✓	55.3	56.8
MCOF [35]	✓	✓	56.2	57.6
AffinityNet [50]		✓	58.4	60.5
DSRG [14]	✓	✓	59.0	60.4
MDC [23]	✓	✓	60.4	60.8
FickleNet [21]	✓	✓	61.2	61.9
OAA [22]	✓	✓	63.1	62.8
RRM [51]		✓	60.7	61.0
SGAN [37]	✓	✓	64.2	65.0
SAFN [52]	✓	✓	61.9	62.3
ICD [53]	✓	✓	64.0	63.9
DRS [32]	✓	✓	63.5	64.5
GSM [54]	✓	✓	63.3	63.6
NSR [55]	✓	✓	65.5	65.3
ESP [38]	✓	✓	67.0	67.3
ECS-Net [31]		✓	62.1	63.4
Ours	✓	✗	**66.5**	**66.9**

TABLE 2. Comparison of ResNet-based weakly-supervised semantic segmentation methods' mean IOUs on PASCAL VOC 2012 *val* and *test* set.

Method	Use saliency map	Need pseudo-label?	Val	Test
DCSP [29]	✓	✓	60.8	61.9
MCOF [35]	✓	✓	60.3	61.2
AffinityNet [50]		✓	61.7	63.7
DSRG [14]	✓	✓	61.4	63.2
FickleNet [21]	✓	✓	64.9	65.3
OAA [22]	✓	✓	65.2	66.4
RRM [51]		✓	66.3	66.5
SGAN [37]	✓	✓	67.1	67.2
SAFN [52]	✓	✓	61.9	62.3
ICD [53]	✓	✓	67.8	68.0
DRS [32]	✓	✓	71.2	71.4
GSM [54]	✓	✓	68.2	68.5
LIID [56]	✓	✓	66.5	67.5
NSR [55]	✓	✓	70.4	70.2
ESP [38]	✓	✓	71.0	71.8
advCAM [24]		✓	68.1	68.0
GCN [57]		✓	68.7	69.3
AuxSegNet [58]	✓	✓	69.0	68.6
ECS-Net [31]		✓	66.6	67.6
Ours	✓	✗	**69.0**	**69.2**

B. COMPARISONS WITH STATE-OF-THE-ARTS

We compare the performance of the proposed method with that of state-of-the-art WSSS methods. In Tables 1 and 2, we summarize the mIOU obtained by VGG16- and ResNet-based WSSS approaches. From the results, we observe that our approach performs competitive with the conventional WSSS approaches. Specifically, our approach achieves mIOU of 66.5% and 69.0% for *val* set of PASCAL VOC 2012 segmentation dataset with VGG16- and ResNet101-based DeepLab-ASPP, respectively. Using our approach, we can train the segmentation network such that it learns the class-specific knowledge directly from the classification network. Our results clearly demonstrate that the generation of pseudo-labels is unnecessary for WSSS.

We compare the proposed approach with a few notable WSSS approaches. In GAIN [16], since the main network and the adversarial network are sharing the parameters and also trained simultaneously, the network might be confused when the object regions are poorly discovered. Our approach can avoid this by training the adversarial network in advance and fixing the parameters in the network. MDC uses the classification network having multiple convolutional blocks to generate the pseudo-label [23]. In our approach, the classification network trained to predict the dense CAM is used for the training of the segmentation network directly. Similarly to the proposed approach, MCOF trains the

TABLE 3. Comparison of performances on *val* set with different settings of our approach. GT indicates the ground truth saliency map.

Model	Segmentation network backbone	Saliency map refinement	Classification network	The number of classification paths	val w/o crf	val w/ crf
A1	VGG16	✗	none		62.0	62.2
A2	VGG16	✓	none		61.4	63.8
A3	VGG16	✗	standard	1	62.4	64.0
A4	VGG16	✗	dilation	1	62.9	64.7
A5	VGG16	✓	dilation	1	62.4	66.4
A6	VGG16	✓	dilation	2	62.0	**66.5**
A7	VGG16	✓	dilation	3	62.2	66.4
B1	ResNet101	✗	none		64.2	65.1
B2	ResNet101	✓	none		65.7	66.8
B3	ResNet101	✓	dilation	1	66.2	68.8
B4	ResNet101	✓	dilation	2	66.1	68.7
B5	ResNet101	✓	dilation	3	66.1	**69.0**

segmentation network using the classification network [35]. In MCOF, the classification network is used to classify the superpixels of an input image. Whereas, in our approach, the pre-trained classification network is used to classify the regions of input image after applying image masking technique.

C. ABLATION STUDIES

In order to show the effectiveness of each component, we conduct ablation experiments with different settings of the proposed work. In Table 3, we summarize the segmentation performance of the proposed approach in different settings. When we say 'standard' classification network, it means that the network trained only using the multi-class multi-label classification loss function. The 'dilation' classification network means the network trained using the classification loss and the CAM loss described in Section III-A.

Our baselines are the VGG16- and ResNet101-based segmentation networks trained only using the saliency loss associated with the original saliency map obtained by SD [41] (see A1 and B1). From the results, we observe that the segmentation performance can be improved by refining the saliency map. Specifically, the models using the refined saliency map (A2 and B2) achieve about 3% improvement in mIOU over the baseline models. By comparing the performance of A1 and A3, we also observe that the segmentation performance can be improved by exploiting the classification network in the training of the segmentation network. Moreover, we can observe that the segmentation performance can be further improved by exploiting the dilated convolution-based classification network (see A3 and A4). We also observe that the performances can be enhanced by employing multiple classification paths (see A5 to A7 and B3 to B5).

To investigate the efficacy of refining saliency map, we conduct experiments using different saliency maps: 1) original saliency map obtained from saliency detector, 2) refined saliency map in which low-scored foreground pixels are corrected to background pixels, and 3) refined saliency map in which low-scored foreground pixels are considered as unlabeled pixels. From the results in Table 4, we observe that the segmentation performance is degraded when the low-scored foreground pixels are corrected to background pixels. We also observe that the segmentation performance is significantly improved when the low-scored foreground pixels are considered as unlabeled pixels.

TABLE 4. Comparison of the segmentation performance using different saliency maps.

	Saliency map	mIOU on val
1	original saliency map	60.7
2	refined saliency map without unlabeled pixels	59.9
3	refined saliency map with unlabeled pixels	65.5

TABLE 5. Segmentation performances with respect to the combination of input scales for refining saliency map.

Configuration	Input scales	mIOU on val
S1	$\{0.5, 0.75, 1, 1.25, 1.5\}$	66.0
S2	$\{0.5, 0.75, 1, 1.25\}$	66.0
S3	$\{0.75, 1, 1.25, 1.5\}$	66.1
S4	$\{0.5, 0.75, 1\}$	66.5
S5	$\{0.75, 1, 1.25\}$	65.7
S6	$\{1, 1.25, 1.5\}$	64.8
S7	$\{0.5, 1, 1.5\}$	66.5

To observe the effect of combination of input scale used for refining saliency maps, we conduct experiments by varying the number of the input scales. The input scales are chosen among the scales used in data augmentation $\{0.5, 0.75, 1, 1.25, 1.5\}$. From the results, we see that the best segmentation performance is obtained when three input scales are used (see S4 and S7 in Table 5).

We also test the performances for various number of dilated convolutional blocks D. From the results shown in Table 6, we observe that the segmentation performance slightly improves with the number of dilated convolutional blocks at the expense of the additional computations and training time.

D. QUALITATIVE RESULTS

In Fig. 7, we provide qualitative results obtained from ResNet101-based DeepLab-ASPP. From the results, we can

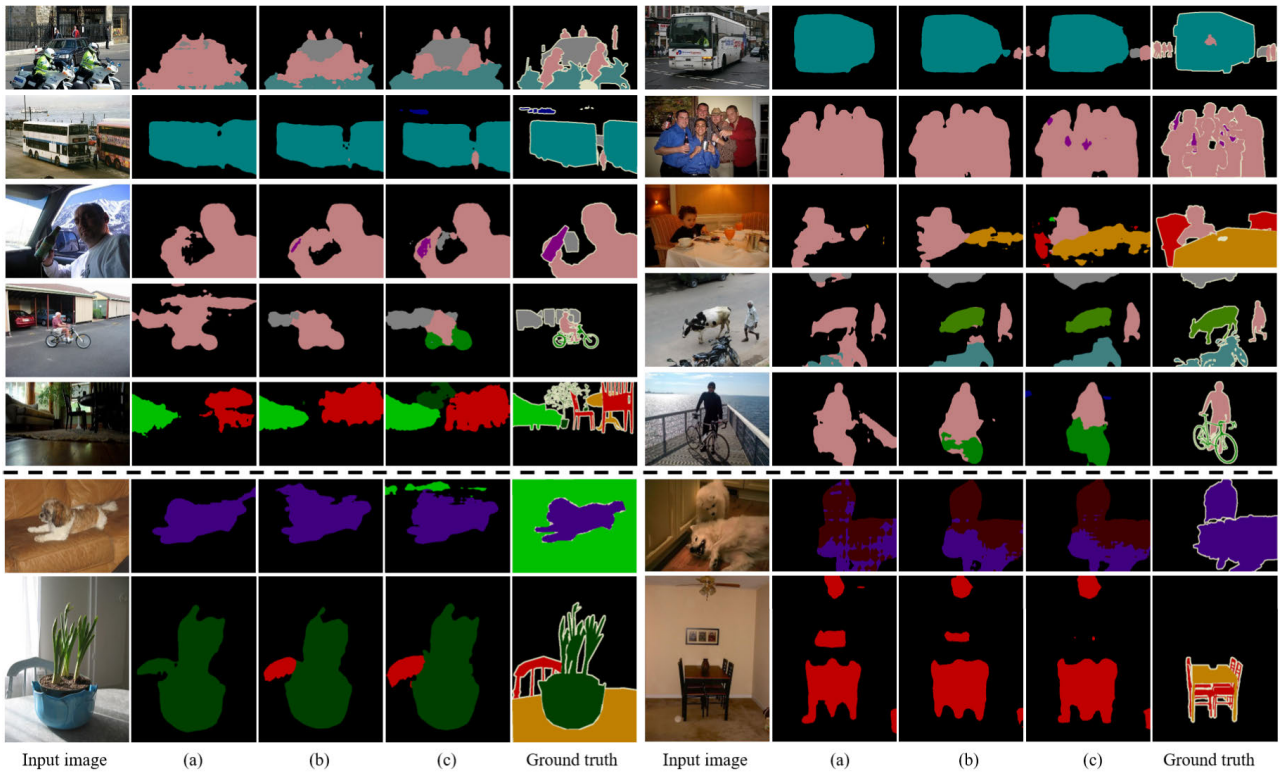


FIGURE 7. Qualitative results obtained from ResNet101-based DeepLab-ASPP. (a): the baseline network trained only using the saliency loss with the original saliency map, (b): the network trained only using the saliency loss with the refined saliency map, and (c): the network trained using the attention loss in addition to the saliency loss with the refined saliency map. The bottom two rows show some failure cases.

TABLE 6. Segmentation performances with respect to different number of dilated convolutional blocks.

Model	# of dilated conv. blocks	Dilation rates	mIOU on val
D1	$D = 8$	$\{3, 6, 9, 12, 15, 18, 21, 24\}$	65.8
D2	$D = 4$	$\{6, 12, 18, 24\}$	65.6
D3	$D = 3$	$\{3, 6, 9\}$	65.5
D4	$D = 2$	$\{3, 6\}$	65.2

observe that our saliency map refining strategy can find out the objects which might not be detected by SD and removing the falsely activated background objects. Also, we can observe that our image masking-based training strategy can help the segmentation network to learn the object classes precisely even when the objects are very small. Also, we would like to mention some failure cases. One of the most frequent failure scenarios is that there is an object which covers a large portion of the image. For example, sofa or table can be confused as background.

In Fig. 8, we provide qualitative results for the proposed approach and conventional approaches. From the results, we observe that the proposed approach predicts the detailed object region (see the first three columns in Fig. 8) while the conventional approaches make false activations (see the last two columns in Fig. 8).

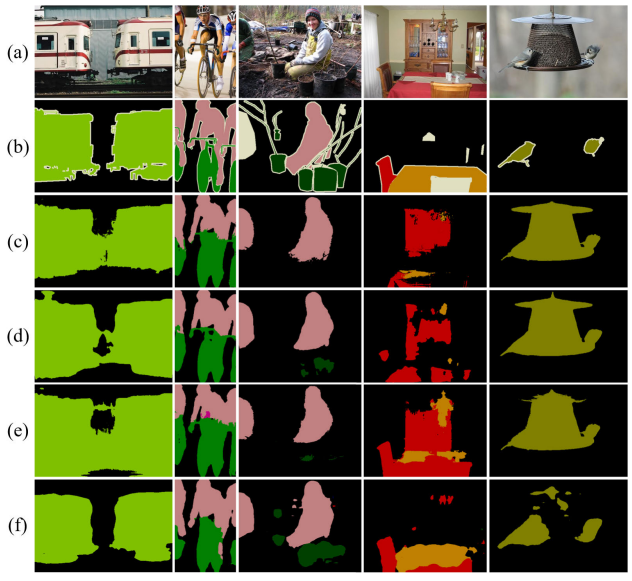


FIGURE 8. Qualitative results obtained from various WSSS approaches. (a): input image, (b): ground truth, (c): DRS [32], (d): GSM [54], (e): NSR [55], (f): ours.

V. CONCLUSION

In this paper, we proposed a new WSSS technique that can train the segmentation network without pixel-level pseudo-labels. To prevent the performance degradation caused by

inaccurate pseudo-label in conventional WSSS approaches, we have exploited the image masking technique in the training of the segmentation network. We also introduced an approach to refine the saliency map, which significantly improves the segmentation performance. Extensive experiments demonstrate that our approach is effective in solving the problem of WSSS.

REFERENCES

[1] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[5] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 695–711.

[6] P. Tang, X. Wang, X. Bai, and W. Liu, "Multiple instance detection network with online instance classifier refinement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2843–2851.

[7] J. Kim, K. Shim, and B. Shim, "Semantic feature extraction for generalized zero-shot learning," 2021, *arXiv:2112.14478*. [Online]. Available: https://arxiv.org/abs/2112.14478

[8] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1742–1750.

[9] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 549–565.

[10] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3159–3167.

[11] J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1635–1643.

[12] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1713–1721.

[13] S. J. Oh, R. Benenson, A. Khoreva, Z. Akata, M. Fritz, and B. Schiele, "Exploiting saliency for object segmentation from image level labels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5038–5047.

[14] Z. Huang, X. Wang, J. Wang, W. Liu, and J. Wang, "Weakly-supervised semantic segmentation network with deep seeded region growing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2018, pp. 7014–7023.

[15] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[16] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu, "Tell me where to look: Guided attention inference network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9215–9223.

[17] M. M. Chun and J. M. Wolfe, "Visual attention," in *Blackwell Handbook of Sensation and Perception*. Malden, MA, USA: Blackwell, 2005, pp. 272–310.

[18] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, pp. 98–136, Jun. 2014.

[19] D. Pathak, P. Krahenbuhl, and T. Darrell, "Constrained convolutional neural networks for weakly supervised segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1796–1804.

[20] G. Papandreou, L.-C. Chen, K. Murphy, and A. Yuille, "Weakly- and semi-supervised learning of a DCNN for semantic image segmentation," 2015, *arXiv:1502.02734*.

[21] J. Lee, E. Kim, S. Lee, J. Lee, and S. Yoon, "FickleNet: Weakly and semi-supervised semantic image segmentation using stochastic inference," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5267–5276.

[22] P.-T. Jiang, Q. Hou, Y. Cao, M.-M. Cheng, Y. Wei, and H. Xiong, "Integral object mining via online attention accumulation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2070–2079.

[23] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang, "Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7268–7277.

[24] J. Lee, E. Kim, and S. Yoon, "Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 4071–4080.

[25] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.

[26] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. ICCV*, 2017, pp. 2961–2969.

[27] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[28] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1568–1576.

[29] A. Chaudhry, P. K. Dokania, and P. H. Torr, "Discovering class-specific pixels for weakly-supervised semantic segmentation," 2017, *arXiv:1707.05821*.

[30] D. Kim, D. Cho, D. Yoo, and I. S. Kweon, "Two-phase learning for weakly supervised object localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3534–3543.

[31] K. Sun, H. Shi, Z. Zhang, and Y. Huang, "Ecs-Net: Improving weakly supervised semantic segmentation by using connections between class activation maps," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 7283–7292.

[32] B. Kim, S. Han, and J. Kim, "Discriminative region suppression for weakly-supervised semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 2, 2021, pp. 1754–1761.

[33] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic segmentation using adversarial networks," 2016, *arXiv:1611.08408*.

[34] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, and S. Yan, "STC: A simple to complex framework for weakly-supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2314–2320, Nov. 2016.

[35] X. Wang, S. You, X. Li, and H. Ma, "Weakly-supervised semantic segmentation by iteratively mining common object features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1354–1362.

[36] F. Sun and W. Li, "Saliency guided deep network for weakly-supervised image segmentation," *Pattern Recognit. Lett.*, vol. 120, pp. 62–68, Apr. 2019.

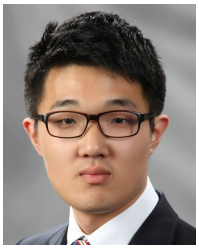
[37] Q. Yao and X. Gong, "Saliency guided self-attention network for weakly and semi-supervised semantic segmentation," *IEEE Access*, vol. 8, pp. 14413–14423, 2020.

[38] S. Lee, M. Lee, J. Lee, and H. Shim, "Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 5495–5505.

[39] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*.

[40] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 991–998.

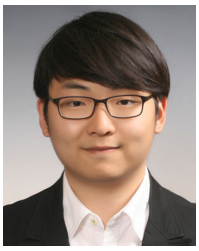
[41] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3917–3926.
[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[43] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[45] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.
[46] Y. Zhang and T. Chen, "Efficient inference for fully-connected CRFs with stationarity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 109–117.
[47] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3640–3649.
[48] W. Liu, A. Rabinovich, and A. C. Berg, "ParseNet: Looking wider to see better," 2015, *arXiv:1506.04579*.
[49] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, and M. Kudlur, "Tensorflow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Operating Syst. Design Implement. (OSDI)*, 2016, pp. 265–283.
[50] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4981–4990.
[51] B. Zhang, J. Xiao, Y. Wei, M. Sun, and K. Huang, "Reliability does matter: An end-to-end weakly supervised semantic segmentation approach," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, pp. 12765–12772, Apr. 2020.
[52] L. Xu, M. Bennamoun, F. Boussaid, and F. Sohel, "Scale-aware feature network for weakly supervised semantic segmentation," *IEEE Access*, vol. 8, pp. 75957–75967, 2020.
[53] J. Fan, Z. Zhang, C. Song, and T. Tan, "Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 4283–4292.
[54] X. Li, T. Zhou, J. Li, Y. Zhou, and Z. Zhang, "Group-wise semantic mining for weakly supervised semantic segmentation," 2020, *arXiv:2012.05007*.
[55] Y. Yao, T. Chen, G.-S. Xie, C. Zhang, F. Shen, Q. Wu, Z. Tang, and J. Zhang, "Non-salient region object mining for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 2623–2632.
[56] Y. Liu, Y.-H. Wu, P. Wen, Y. Shi, Y. Qiu, and M.-M. Cheng, "Leveraging instance-, image- and dataset-level information for weakly supervised instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1415–1428, Mar. 2022.
[57] S.-Y. Pan, C.-Y. Lu, S.-P. Lee, and W.-H. Peng, "Weakly-supervised image semantic segmentation using graph convolutional networks," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2021, pp. 1–6.
[58] L. Xu, W. Ouyang, M. Bennamoun, F. Boussaid, F. Sohel, and D. Xu, "Leveraging auxiliary tasks with affinity learning for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 6984–6993.



SANGTAE KIM received the B.S. degree in information and communication engineering from Korea University, South Korea, in 2014. He is currently pursuing the Ph.D. degree with Seoul National University. His research interests include big data analysis, deep learning, image classification, and image segmentation.



LUONG TRUNG NGUYEN received the B.S. and M.S. degrees from the Ho Chi Minh City University of Technology, Vietnam, in 2010 and 2012, respectively, and the Ph.D. degree in electrical and computer engineering from Seoul National University, South Korea, in 2020. He is currently a Postdoctoral Researcher at the Institute of New Media and Communications, Seoul National University. His research interests include matrix completion, federated learning, and machine learning.



KYUHONG SHIM received the B.S. degree in electrical and computer engineering from Seoul National University, South Korea, in 2015, where he is currently pursuing the Ph.D. degree. His research interests include speech and language processing, efficient deep learning algorithms and implementations, neural network compression, compressed sensing, low-rank matrix completion, big data analysis, and machine learning.



JUNHAN KIM received the B.S. degree in electrical and computer engineering from Seoul National University, South Korea, in 2017, where he is currently pursuing the Ph.D. degree. His research interests include compressed sensing, low-rank matrix completion, big data analysis, and machine learning.



BYONGHYO SHIM (Senior Member, IEEE) received the B.S. and M.S. degrees in control and instrumentation engineering from Seoul National University (SNU), South Korea, in 1995 and 1997, respectively, and the M.S. degree in mathematics and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC), USA, in 2004 and 2005, respectively.

From 1997 to 2000, he was with the Department of Electronics Engineering, Korean Air Force Academy, as an Officer (the First Lieutenant) and an Academic Fulltime Instructor. From 2005 to 2007, he was with Qualcomm Inc., San Diego, CA, USA, as a Staff Engineer. From 2007 to 2014, he was with the School of Information and Communication, Korea University, Seoul, as an Associate Professor. Since 2014, he has been with SNU, where he is currently a Professor with the Department of Electrical and Computer Engineering. His research interests include wireless communications, statistical signal processing, compressed sensing, and machine learning.

Prof. Shim is an Elected Member of the Signal Processing for Communications and Networking (SPCOM) Technical Committee of the IEEE Signal Processing Society. He was a recipient of the M. E. Van Valkenburg Research Award from the ECE Department, University of Illinois, in 2005; the Hadong Young Engineer Award from IEIE, in 2010; the Irwin Jacobs Award from Qualcomm and KICS, in 2016; and the Shinyang Research Award from the Engineering College of SNU. He has served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE WIRELESS COMMUNICATIONS LETTERS, and *Journal of Communications and Networks*. He has served as a Guest Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS.