ECS-Net: Improving Weakly Supervised Semantic Segmentation by Using Connections Between Class Activation Maps

Kunyang Sun[1,2], Haoqing Shi[1,2], Zhengming Zhang[1,2], Yongming Huang[1,2,*]
[1] National Mobile Communications Research Laboratory, Southeast University
[2] Pervasive Communication Research Center, Purple Mountain Laboratories
[1] Nanjing 210096, China
[2] Nanjing 211111, China
Sunky@seu.edu.cn, shihaoqing619@seu.edu.cn, zmzhang@seu.edu.cn, huangym@seu.edu.cn

Abstract

Image-level weakly supervised semantic segmentation is a challenging task. As classification networks tend to capture notable object features and are insensitive to over-activation, class activation map (CAM) is too sparse and rough to guide segmentation network training. Inspired by the fact that erasing distinguishing features force networks to collect new ones from non-discriminative object regions, we using relationships between CAMs to propose a novel weakly supervised method. In this work, we apply these features, learned from erased images, as segmentation supervision, driving network to study robust representation. In specifically, object regions obtained by CAM techniques are erased on images firstly. To provide other regions with segmentation supervision, Erased CAM Supervision Net (ECS-Net) generates pixel-level labels by predicting segmentation results of those processed images. We also design the rule of suppressing noise to select reliable labels. Our experiments on PASCAL VOC 2012 dataset show that without data annotations except for ground truth image-level labels, our ECS-Net achieves 67.6% *mIoU on test set and* 66.6% *mIoU on val set, outperforming previous state-of-the-art methods.*

1. Introduction

Semantic segmentation, with the goal of assigning a category label to each pixel on the image, is one of the fundamental computer vision tasks. Due to the rapid development of Fully Convolutional Network (FCN), fully supervised semantic segmentation (FSSS) methods [5, 6], widely applied to applications like assisted driving and medical imaging, have achieved excellent performance over a short period. However, as training those algorithms requires datasets equipped with pixel-level annotations, existing datasets suffer from the severe time consuming and numerous workforce investment. Aware of these difficulties, a line of research takes a weakly supervised strategy. Those methods usually follow a two-stage paradigm: (1) Generating reliable pseudo labels by low degree supervision like bounding boxes [16, 20], scribbles [19, 28], points as well as image-level annotations [17, 2]. (2) Training existing fully-supervised semantic segmentation approaches by treating these pseudo labels as ground truth annotations. Obviously, generating more accurate pseudo labels can enhance the performance of algorithms in (2). Compared with other weak supervisions, image-level annotation, which can even be gotten directly from almost all existing datasets enjoys lowest time and labour costs. Therefore, we choose

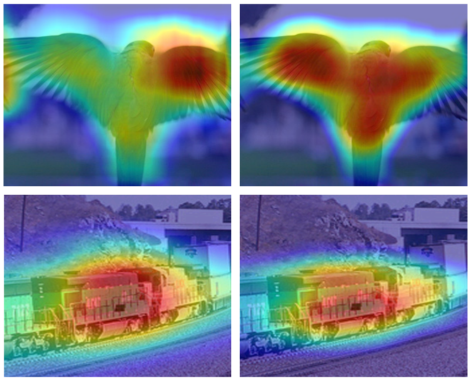


Figure 1. The deficiency of spatial dimension annotations leads the network to focus too much on salient object regions and reduce the sensitivity to the edges of objects (left). Our method makes the network not omit another valuable region, and can better capture the edges (right).

This work was supported by the National Key R&D Program of China under Grant 2018YFB1800801
[*]Corresponding author

image-level labels as weak annotations in this work.

Image-level weakly supervised semantic segmentation [23, 13, 21, 22] is a challenging task as classification labels fail to retain the object localization information, which is essential in segmentation task. In order to study this issue, Class Activation Map (CAM) [38] is widely applied in weakly supervised semantic segmentation (WSSS) methods to introduce additional location supervisions. Albeit being simple in structure, CAM has two main obstacles that prevent it from being directly used as pseudo labels: (1) CAM trends to give the high response to part instead of the whole region of an object. (2) Rough positioning activation introduces noise like over-activation. Many researchers solved these two issues by designing different object region mining techniques varying from random activation [18] to dilated convolutions [34]. By sampling from various areas of target objects, networks shift attention from discriminate features to their surrounding regions. However, these approaches only use image-level labels as supervision, making the second issue challenging to resolve. We noticed that the big performance gap between weakly and fully supervised methods mainly comes from differences between classification and segmentation tasks. Specifically, as all pixels are correctly annotated in fully supervised frameworks, segmentation tasks can be settled directly. On the contrary, image-level weakly supervised approaches trend to solve segmentation task by turning it into a classification problem. In original task, the whole object regions as well as boundaries need to be finely segmented while rough activation on part of objects is enough to deal with the converted task, preventing these methods from achieving comparable performance as fully supervised methods.

In this work, we consider narrowing the performance gap. It is natural to ask a question: *Can we introduce additional supervision that guides CAMs to study segment information?* We recognize a fact: in classification task, erasing highlight regions in images can guide the network to explore and activate new object regions. Previous work like adversarial erasing (AE) [32] also takes advantage of this phenomenon. AE erases images iteratively until the network fails to converge. It generates final result by combining all CAMs together. However, with iterations increasing, these overly simplified erasing designs may fail to avoid over-activation. Besides, simple assembly does not make good use of connections between different CAMs. We notice that CAM predicted by an erased image contains object segment information which is lack in the original CAM. In other words, the CAM of the erased image may provide the original CAM with additional segment supervision.

We focus on investigating ways to provide additional segmentation supervisions by utilizing predictions of erased images. Particularly, we firstly erase high response regions from images and generate new CAMs of those erased images. Then, we sample reliable pixels from new CAMs and apply their segmentation predictions as semantic labels to train corresponding original CAMs. Instead of erasing multiple times, our method only needs to erase once, avoiding introducing excessive noise. We carry out extensive ablation studies to discover the optimal hyperparameters like the sampling threshold. Concretely, we can achieve the followings:

- We propose a simple, efficient, and novel framework: Erased CAM Supervision Net (ECS-Net), to solve the problems in weakly supervised semantic segmentation. Taking advantage of object region mining techniques and relation of twice CAMs, our method supplies additional segmentation cues. Shown in experiments, CAMs predicted by ECS-Net learn better segment information such as boundaries and shapes of objects.
- As noise like over-activation seriously hurt segmentation performance, our ECS-Net devise the sampling rule to suppress those brought from erased images' CAMs. We confirm that the proposed method is helpful for weeding out unreliable samples and speeding up network convergence.
- As shown in experiments on the test set of PASCAL VOC 2012 datasets, our framework achieves mIoU of 63.4% with the VGG-16 backbone and 67.6% mIoU with ResNet-38 backbone outperforming the previous state-of-the-art methods.

2. Related Work

2.1. Segmentation with Low-degree Supervision

Recent advances in weakly semantic segmentation unveil the possibilities of using weak labels instead of pixel-level annotations, largely reducing the cost of hand-operated labeling. Various types of weak labels are learned to solve the issue, e.g., bounding boxes [16, 20, 9], scribbles [28, 19] and points [3]. However, when faced with massive volumes of unlabeled data, those coarse annotations still suffer from a great deal of manual labeling pressure.

2.2. Segmentation with Image-level Supervisions

Image-level annotations can be obtained directly from almost all existing large datasets without human investment. Therefore, many researchers attempt to apply classification labels to give guides to network training. Pinheiro et al. [22] establish the framework to learn semantic segmentation from image-level labels by multi-instance learning. Lack of object localization cues, its performance is far behind contemporaneous FSSS algorithms. This problem is partly solved by introducing rough object localization supervision through discriminative localization techniques like saliency detection [36] and class activation map (CAM) [38]. However, both of those localization cues respond to part of objects instead of the whole object regions, failing to be used as pseudo labels directly.

The most dominant approaches refine and expand the class activation map to extend over the whole object. SEC [17] introduces three loss functions, seeding, expansion, and constrain loss to guide network training. However, static seed cues, which are too few and sparse, limit the segmentation performance. In order to improve the recognition ability of low response object regions, AE [32] erases high response features from input images iteratively, forcing the network to study new highlight features from low response areas. However, iterative learning is time-consuming. MDC [34] employs dilated convolutions with high dilated rates to sample and study features from whole object regions. Due to fixed sampling positions, MDC has difficulty in capturing object boundaries flexibly. This issue is studied by FickleNet [18], which attempts to utilize the Dropout method with different drop rates to select and combine features randomly. FickleNet generates multiple position maps on a single image, obtaining regions with different shapes. Due to the large randomness of dropout, FickleNet can not avoid introducing noise. PSA [2] generates an affinity matrix to study the similarity between pixels and applies a random walk to predict the final results.

We notice that methods like [32, 34, 18] force network to learn from low response areas to expand object regions. However, most of them have no way to suppress over-activation introduced by sampling since a small number of background pixels are incorrectly classified, the classification loss may not be affected. To break through these limitations, we propose ECS-Net. To our best knowledge, our method maybe is the first algorithm that introduces reliable pseudo segmentation supervision during the exploration phase.

Using Connections between CAMs: Many superior image-level weakly supervised methods consider sample splice different CAMs to predict final segmentation results. MDC [34] sums CAMs predicted by different dilated convolutions. Similarly, RRM [37] calculates the average of CAM with different scales. AE [32] crops highlight pieces from every CAMs and paste them together according to corresponding positions. We think that overly simplified assembling designs can not develop the power of different CAMs.

Recently, SEAM [31] produces corresponding CAMs by resizing an image into two scales. Further, it exploits equivariant regularization to narrow the difference between those two CAMs. Through this self-supervised learning method, SEAM generates more robust CAMs for the segmentation task. To some extent, CAM of a small-scale image activates more objects' parts but aggravates over-activation. Conversely, CAM of a large-scale image has less activation regions including less over-activation. These two CAMs supervise each other, providing a good balance between expanding object regions and over-activation. However, the same prediction errors in both CAMs are difficult to be corrected in SEAM.

3. Our Approach

This section describes our approach in detail. First of all, we elaborate on the exhaustive process of applying CAMs to produce segment supervisions. We also introduce the method of suppressing noisy label and further discuss the implementation of our framework, including loss function, network structure, some other improvements like scaling and multi dilated overlay module. The whole framework is illustrated in Figure 2. Finally, we give an exhaustive explanation of how the algorithm works.

3.1. Segment Labels Generation

The main idea of our proposed ECS-Net is building connections between CAMs by erasing. First, we use CAMs of erased images to generate segment labels. Then, those pseudo labels are considered as supervisions to refine CAMs of original images. In particular, by applying class activation map techniques, an image $\mathbf{I}$ with classification label $\mathbf{L}$ is firstly fed into the network $\mathcal{F}$ to predict heatmap $\mathbf{H} \in \mathbb{R}_{>0}^{C\times H\times W}$, where C is the object categories and $H \times W$ is the size of the raw image $\mathbf{M}$ ($e.g.$ 448×448). Then, we normalize $\mathbf{H}$ to produce original CAM $\mathbf{a}$ and apply classification weights $W = \{w_c \mid w_c = 1, \text{ if } c \in \mathbf{L}, \text{ else } w_c = 0. \ \forall \text{c} \in \{1 \dots C\}\}$ to prohibit non-existent category from being activated:

$$\mathbf{a}_c(x,y) = w_c \cdot \frac{\mathbf{H}_c(x,y) - min_{x,y}\mathbf{H}_c(x,y)}{max_{x,y}\mathbf{H}_c(x,y) - min_{x,y}\mathbf{H}_c(x,y)}, \tag{1}$$

where (x, y) is the location on $\mathbf{H}$. We generate score map $\mathbf{s}$ as follows

$$\mathbf{s}(x,y) = max_c \mathbf{a}_c(x,y). \tag{2}$$

A higher score means more distinct classified features, we set a threshold $\delta = 0.6$ to select erased regions $\mathbf{R}$ from $\mathbf{s}$. Further, we erase those features on $\mathbf{M}$ by applying Gaussian Blur to pixels in selected regions. Then, ECS-Net sends the processed image $\mathbf{I}'$ into the network $\mathcal{F}'$ which shared weights with $\mathcal{F}$ and outputs the heatmap $\mathbf{H}'$. Following Eq.(1), we get the CAM of erased image $\mathbf{a}'$. Then, $\mathbf{a}'$ is processed by argmax function to get rough segment labels $\mathbf{L}'$.

$$\mathbf{L}'(x,y) = \text{argmax}_c \mathbf{a}'(x,y). \tag{3}$$

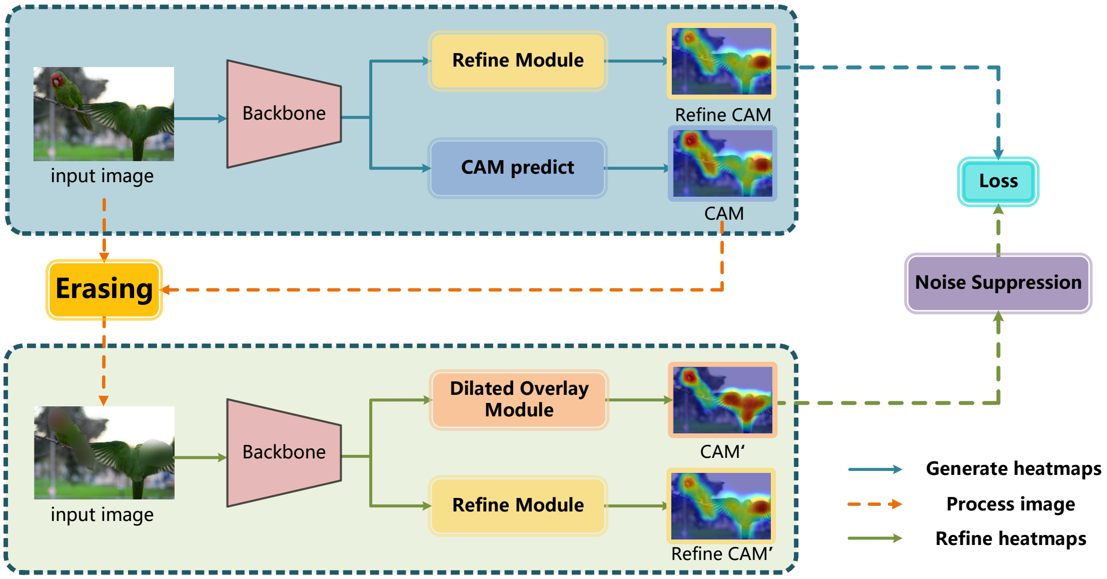


Figure 2. The whole framework of our proposed ECS-Net. High respond regions of CAMs in $\mathcal{F}$ are erased on images. CAMs in $\mathcal{F}'$ are employed as additional segment labels by suppressing noise. $\mathcal{F}$ and $\mathcal{F}'$ share weights.

Figure 2 shows that by erasing high response features(red regions), our network shifts its attention to other low response object regions. Until now, our produced rough labels contain massive prediction errors, being far below requirements.

Noise Suppression: We follow a rule to select reliable segment labels from $\mathbf{L}'$. Firstly, we ignore labels from erased regions. There are two reasons: (1) Those regions, considered as easy examples, have no contribution. (2) Caused by erasing, those regions miss features and lead to unreliable predictions. We further ignore background labels. Finally, the reliable labels are obtained by applying a threshold θ on score map s'.

3.2. Implementation of ECS-Net

Network structure: We follow the work of [38] to calculate CAM predictions. In our ECS-Net, a classification convolutional layer $\mathbf{B}$ with 3×3 kernels followed by a global average pooling (GAP) is added behind the last layer of the backbone. Considering to avoid mutual interference between segmentation task and classification task, we add spatial attention to refine CAM results $\mathbf{E} \in \mathbb{R}^{C\times H\times W}_{>0}$. As shown in Figure 3, even though sharing weights, our two networks $\mathcal{F}$ and $\mathcal{F}'$ are different in structure. We give a detailed introduction in the section of improvements 3.2.

Loss function: In our work, as both classification labels and produced pseudo semantic labels are used for supervision, our loss function $\mathcal{L}$ consists of two-part: classification loss and segmentation loss. For the image classification task, we follow the work of CAM technique [38] to define our classification loss as follows:

$$l_{cls}(\mathbf{o},\mathbf{L}) = \frac{1}{C-1}\sum_{c=1}^{C-1}\left[l_c log\left(\frac{1}{1+f(o_c)}\right)\right.$$
$$\left. +(1-l_c)log\left(\frac{f(o_c)}{1+f(o_c)}\right)\right], \qquad (4)$$

where $f(x) = exp(-x)$ and $\mathbf{o}$ is a vector with length C which predicted by the GAP layer. Our l_{cls} ignores background category, i.e. $c = 0$. We define $\mathbf{o}_{cam}$ and $\mathbf{o}_{pro}$ as GAP results of $\mathbf{H}$ and refine CAM $\mathbf{E}$ in $\mathcal{F}$ respectively. Similarly, $\mathbf{o}'_{cam}$ and $\mathbf{o}'_{pro}$ denote GAP results of those in $\mathcal{F}'$. Then the final classification loss is formulated as:

$$\mathcal{L}_{cls} = \frac{1}{2}\left(l_{cls}(\mathbf{o}_{cam},\mathbf{L}) + l_{cls}(\mathbf{o}'_{cam},\mathbf{L})\right)$$
$$+\frac{1}{2}\left(l_{cls}(\mathbf{o}_{pro},\mathbf{L}) + l_{cls}(\mathbf{o}'_{pro},\mathbf{L})\right). \qquad (5)$$

For semantic segmentation task, we adopt cross entropy loss which is defined as:

$$\mathcal{L}_{ce}(\mathbf{P},\mathbf{Q}') = \sum_{i\in\Phi,c\in C} \mathbf{Q}'(i,c)log\left(\mathbf{P}(i,c)\right), \qquad (6)$$

where $\mathbf{Q}'$ denotes the onehot results of pseudo segmentation labels $\mathbf{L}'$, Φ is defined as the location set of reliable labels and $\mathbf{P}$ is the refine CAM $\mathbf{E}$ in $\mathcal{F}$ followed by a softmax operation. Therefore, the finally Loss function in our ECS-Net is defined as

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{ce}. \qquad (7)$$

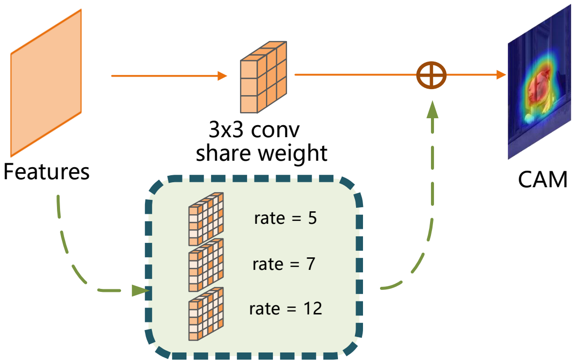


Figure 3. Multi dilated overlay module in our method. The solid line represents the process of predicting CAM in $\mathcal{F}$. The dotted line represents the process of generating $\mathbf{H}'$ from the erased image in $\mathcal{F}'$. The 3×3 convolution layer shares weight with each dilated convolution layer in the multi dilated overlay module.

Other improvements: We also discuss other improvements that increase prediction performance. Firstly, before sending images into $\mathcal{F}$, we rescale the raw image $\mathbf{M}$ with scale factor $\beta \in [0, 1]$. It means that the first input image $\mathbf{I}$ is smaller than $\mathbf{M}$. More specifically, the second input image $\mathbf{I}'$ has the same shape with $\mathbf{M}$. Besides, motivated by the fact that dilated convolution layers are able to expand the receptive field [8], we add K dilated convolution layers with different rates paralleling with layer $\mathbf{B}$ (shown in Figure 3). It is worth mentioning that these additional layers share weights with $\mathbf{B}$ and only are applied in $\mathcal{F}'$ in the training stage. Therefore, our network may capture more robust features. The heatmap $\mathbf{H}'$ is computed as followed

$$\mathbf{H}' = \frac{1}{2}\mathbf{H}'_0 + \frac{1}{2K}\left(\sum_{k=1}^{K}\mathbf{H}'_k\right), \tag{8}$$

where $\mathbf{H}'_0$ is defined as the output of $\mathbf{B}$ while $\mathbf{H}'_k$ is the output of kth dilated convolution layer.

4. Experiments

4.1. Dataset

Our approach is trained and evaluated on the PASCAL VOC 2012 segmentation benchmark [11]. This dataset has been annotated by 21 class pixel-level labels, including one background and 20 different object categories. It is noticed that ground-truth image-level labels instead of pixel-level labels can be obtained in our ECS-Net. The original subsets of PASCAL VOC 2012 consist of 1464 training images, 1449 validation images and 1456 test images. We train our approach on the augmented training set with 10582 images provided by SBD dataset [12]. The results of segmentation are evaluated by the official evaluation metric the mean Intersection-over-Union (mIoU) ratio after submitting the predicted results to the official PASCAL VOC evaluation server.

4.2. Implementation Details

Training: We choose the convolutional layers of ResNet-38 [35] as our backbone and initialize the parameters of it by a pre-trained module on ImageNet [10]. The training images are firstly randomly rescaled with the longest edge being in the range of [448, 768]. Then those rescaled images are randomly cropped into 448 × 448 patches. We take a batch size of 8 patches and train the network for 8 epochs on 4 GPUs. We warm up the network with learning rate 0.01 for 100 steps. Then, the initial learning rate is set as 0.05 and decayed following the poly policy $lr_{step} = lr_{step}(1 - \frac{step}{max_step})^{\gamma}$ where $\gamma = 0.9$. The learning rate of the additional classification layer and convolution layer in the refine module is 10 times that of the backbone.

Configurations and baselines: We consider following configurations hyperparameter for ECS-Net:

- β, rescale factor of the first input image,
- θ, the threshold for selecting reliable segment labels,
- heatmaps for producing pesduo segment labels, it can either be $\mathbf{H}'$ or the refine heatmap, output of refine module, in $\mathcal{F}'$,

We represent our models with the abbreviation β_θ. For example, 0.5_0.8 represents that the resolution of first input images is 244 × 244 (0.5 factor of 448), we ignore labels with a score smaller than 0.8 on $\mathbf{s}'$. By default, we use $\mathbf{H}'$ to predict segment labels and do not add multi dilated overlay module in our baseline. Besides, we choose 1_0.8 as our baseline in the rest of the paper.

4.3. Ablation Experiments

We investigate the effectiveness of our ECS-Net by carrying out ablation experiments on the configurable hyperparameters mentioned above. We evaluate our methods with semantic segmentation metric (mIoU).

Rescale factor of the first input image: We measure the performances of our model with different ratio of the twice input size. Table 1 demonstrates that an appropriate rescale factor between 0.5 to 0.7 can enhance the performance. Using a smaller input can provide more object localization in

β	1	0.7	0.5
mIoU	55.1	55.7	**56.1**

Table 1. **Rescale factor of the first input image:** Performances of β_0.8 models on PASCAL VOC 2012 train set. After data augmentation, input images are cropped into $H \times W$ patches. The input size is $\beta H \times \beta W$ in $\mathcal{F}$ while $H \times W$ in $\mathcal{F}'$.

$\mathbf{H}$. Moreover, studying different scales of the original image leads to a stable prediction. We set $\beta = 0.5$ for our final model.

θ	0.6	0.7	0.8	0.9
mIoU	53.8	54.1	**55.1**	54.9

Table 2. **Threshold for selecting segmentation labels:** Performance with twice same size inputs.

Threshold for selecting segment labels: We validate the influence of different threshold for selecting segmentation labels. We change θ from 0.6 to 0.9. Table 2 reports the results. A smaller threshold means that more labels are selected to introduce segment supervision, while a larger threshold that fewer reliable pixels are labeled. There is a trade-off that too many label regions introduce noise since incorrect labels are selected, yet too few segmentation labels are not enough for network training. If not specified, we use $\theta = 0.8$ in later experiments.

Features	CAM heatmap $\mathbf{H}'$	Refine heatmap $\mathbf{E}'$
mIoU	**55.1**	54.2

Table 3. **Segmentation feature locations:** Performance of 1_0.8 model on train set. Both $\mathbf{H}'$ and $\mathbf{E}'$ are outputs of $\mathcal{F}'$. $\mathbf{H}'$ comes from CAM branch while $\mathbf{E}'$ is refine module outputs.

Segmentation features locations: CAM *vs.* refine module: We compare our segment labels generation locations. In Table 3, by using CAM heatmap $\mathbf{H}'$, we can improve the performance. We think that $\mathbf{E}'$ are highly coupled with $\mathbf{E}$ as they go through the same refine module, making the same prediction errors between them. However, $\mathbf{H}'$ has a low degree of coupling with $\mathbf{E}$, reducing the occurrence of this phenomenon. In the rest experiments, if not specified, we apply CAM heatmap $\mathbf{H}'$ to generate segment labels.

Effectiveness of each part: Table 4 illustrates the effectiveness of each part in our ECS-Net. It is observed that the localization maps generated by ECS-Net outperform baseline. If we do not using connections between CAMs, erasing just brings a slight improvement on segmentation performance. After applying the segment supervision, segmentation performance improves a lot from 48.5 to 55.1. The localization maps with CRF trained on the 0.5_0.8 model with a multi dilated overlay module are shown in Figure 4.

Baseline	Erasing	CELoss	Other Improvements	CRF	mIoU
✓					47.4
✓	✓				48.5
✓	✓	✓			55.1
✓	✓	✓	✓		56.6
✓	✓	✓	✓	✓	58.6

Table 4. The ablation study for each part of ECS-Net. We report segmentation performance on the train set. The erasing threshold is 0.6 and selecting reliable labels with a threshold 0.8. Baseline: original CAMs. Erasing: erasing discriminative object regions. CELoss: using connections between CAMs. CRF: conditional random field. Other Improvements: using 0.5_0.8 model and adding multi dilated overlay module with dilation rate (5, 7, 12).

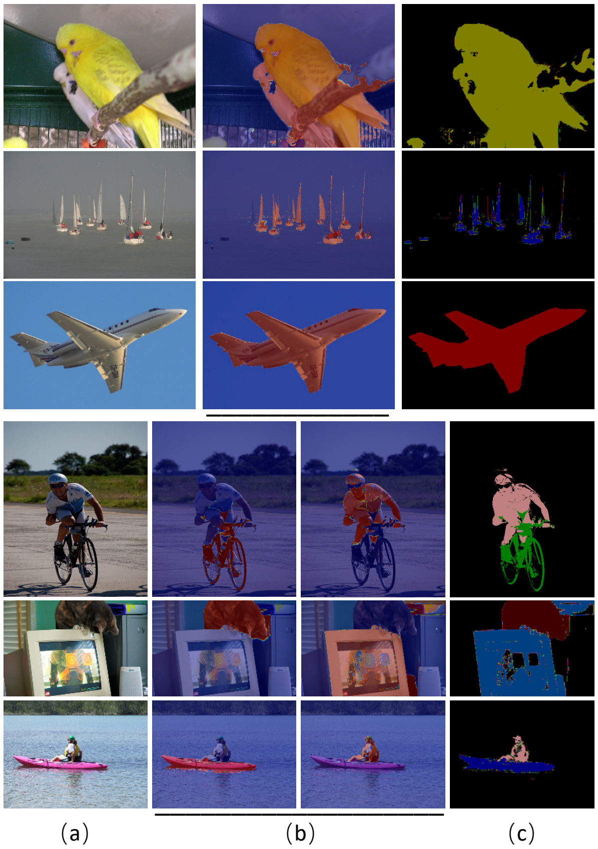
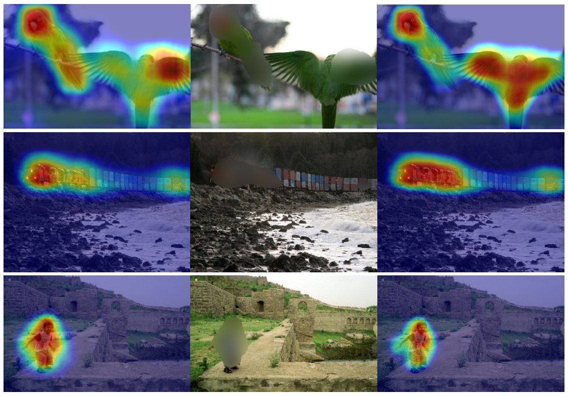


Figure 4. The localization maps with CRF obtained by ECS-Net on the PASCAL VOC 2012 validation set. Two sets of images are listed in rows. (a) Images. (b) CRF results on single category. (c) Segmentation results of localization maps with CRF.

Method	mI0U
CAM [38]	47.4
GradCAM++ [4]	47.4
CAM+SEAM [31]	55.4
CAM+ECS-Net	**56.6**

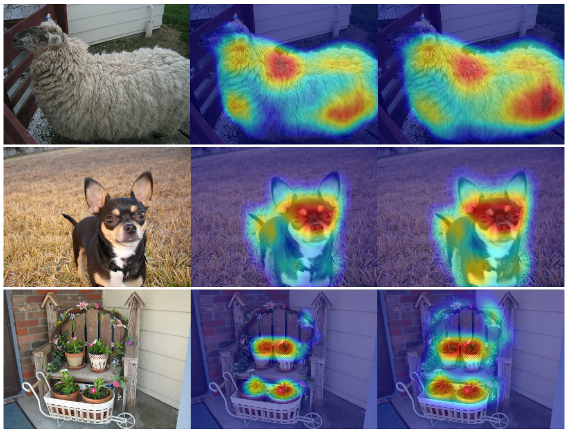
Table 5. **Comparison of different weakly supervised localization methods**: We evaluate those methods on PASCAL VOC 2012 train set.

Comparision with other localization methods: Similar to our method, CAM [38], GradCAM++ [4], and SEAM [31] provide localization information for generating pseudo segment labels. As shown in Table 5, our ECS-Net surpasses the other methods on semantic segmentation task. Since taking the advantage of erased CAMs, the proposed method learns more refined segmentation representation, being more matchable with the segmentation task. We will give a further explanation that our ECS-Net provides



(a) (b) (c)

Figure 5. (a) CAM results of orginal images. (b) Processed images with erased. (c) CAM results of erased images. We find that lacking of discriminative features, ECS-Net focuses on new discriminative features as well as object boundaries.



(a) Input (b) CAM1 (c) CAM2

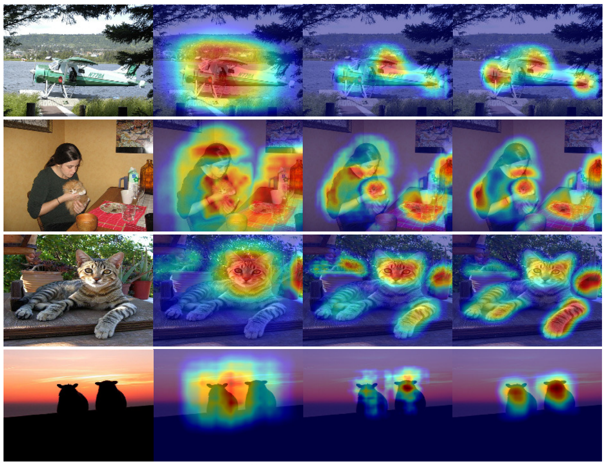
Figure 6. (a) Input images. (b) CAM results with 5k iterations. (c) CAM results with 10k iterations. Our ECS-Net can expand the object regions during training.

more precise localization information.

4.4. Discussions

Object region mining in ECS-Net: As shown in the first row of Figure 5, the network firstly gives a high response to the left bird and the right wing of the right bird. When blurring those features, the right bird's left wing is detected as new high response regions. Figure 6 illustrates that during training, our ECS-Net detects more and more object regions, benefiting from erasing operation. Different from the AE [32] approach, whose erasing is for object mining, our erasing is mainly for generating new supervision.

Improved segment information in ECS-Net: We agree that localization methods like CAM [38], which is generated by just applying image-level labels as supervision, are not suitable for segmentation work. More specifically, the classification task is insensitive to incorrect classification between foreground and background pixels. However, these mistakes hurt the performance of semantic segmentation. As shown in Figure 5, our heatmaps with fewer over-activation provide more detailed boundary information for network training. We also compare ECS-Net with baseline [38] as well as SEAM [31] in Figure 7. Compared with other methods, our results have sharper boundaries and fewer over-activation pixels. Those make our results look more like segment masks. We believe that our ECS-Net can narrow the gap between goals of semantic segmentation and classification.



(a) images (b) Baseline (c) SEAM CAM (d) Our CAM

Figure 7. Multi-scale test results. CAM generated by our method have shaper boundries and less over-activation background pixels.

Methods	Backbone	Training Set	val	test
DCSM [25]	VGG16	10K	44.1	45.1
BFBP [24]	VGG16	10K	46.6	48.0
SEC [17]	VGG16	10K	50.7	51.1
STC [33]	VGG16	50K	49.8	51.2
AE_PSL [32]	VGG16	10K	55.0	55.7
MDC [34]	VGG16	10K	60.4	60.8
MCOF [30]	ResNet-101	10K	60.3	61.2
DSRG [14]	ResNet-101	10K	61.4	63.2
IRNet [1]	ResNet-50	10K	63.5	64.8
FickleNet [18]	ResNet-101	10K	64.9	65.3
SSDD [26]	ResNet-38	10K	64.9	65.5
WSIAL [29]	ResNet-38	10K	64.3	65.4
SEAM [31]	ResNet-38	10K	64.5	65.7
OOA [15]	ResNet-101	10K	65.2	66.4
BES [7]	ResNet-101	10K	65.7	66.6
Ours	VGG16	10K	62.1	63.4
Ours	ResNet-38	10K	**66.6**	**67.6**

Table 6. Evaluation results on the PASCAL VOC 2012 dataset. We compare our ECS-Net with previous state-of-the-art Image-level WSSS approaches.

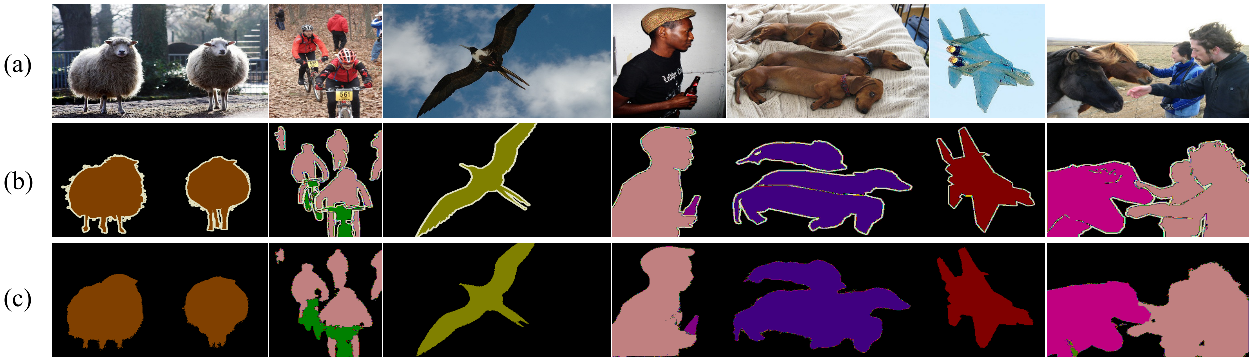


Figure 8. Qualitative results on the PASCAL VOC 2012 validation set. (a) Original images. (b) Ground truth labels. (c) Our results obtained by retain DeepLab-resnet38 network on our pseudo labels.

Methods	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbk	person	plant	sheep	sofa	train	tv	mean
SEC [17]	82.4	62.9	26.4	61.6	27.6	38.1	66.6	62.7	75.2	22.1	53.5	28.3	65.8	57.8	62.3	52.5	32.5	62.6	32.1	45.4	45.3	50.7
PSA [2]	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
SEAM [31]	88.8	**68.5**	33.3	**85.7**	40.4	67.3	78.9	76.3	81.9	29.1	75.5	48.1	79.9	73.8	71.4	75.2	48.9	79.8	40.9	58.2	53.0	64.5
SSDD [26]	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
BES [7]	88.9	74.1	29.8	81.3	53.3	69.9	**89.4**	**79.8**	84.2	27.9	76.9	46.6	78.8	75.9	72.2	70.4	**50.8**	79.4	39.9	65.3	44.8	65.7
Ours	**89.8**	68.4	**33.4**	85.6	48.6	**72.2**	87.4	78.1	86.8	33.0	77.5	41.6	81.7	76.9	**75.4**	**75.6**	46.2	80.7	**43.9**	59.8	**56.3**	**66.6**

Table 7. Performance on the PASCAL VOC 2012 *val* dataset.

4.5. Comparisons with State-of-the-arts

We follow the work of [2] to train an Affinity Net based on our refined CAM. Then, we generate pseudo labels by random walk operation. The final pseudo labels achieve 67.82% mIoU on the train set, surpassing SEAM [31] by 4.2 mIoU. We also train the fully supervised semantic segmentation model DeepLabv1 [5] with our pseudo labels. We apply both VGG16 [27] and ResNet-38 [35] as the network backbone. For ResNet-38 [35], we replace the three fully connected layers with the dilated convolution at the end of the backbone. The data augmentation operations include randomly scale, random cropped, flip and color jittering. After that, we resize the input images to 321 × 321. The initial learning rate is 0.001, following the poly policy proposed in DeepLabV2 [6]. We train the networks with batch size 10 for 20 epochs. We choose the SGD as the optimizer. The segmentation networks are implemented on Pytorch framework and performed on 4 NVIDIA Tesla-v100 GPUs.

Table 7 shows the final results on PASCAL VOC 2012 val set. Compared with results trained with pseudo labels that produced based on original CAM, our ECS-Net has good performances on all categories. We make extensive comparisons with previous state-of-the-art weakly-supervised semantic segmentation solutions with image-level annotations. As shown in Table 6 our ECS-Net surpasses other methods without any other auxiliary algorithms like Saliency Detection. Qualitative results are illustrated in Figure 8.

5. Conclusion

In this paper, we propose a powerful method (ECS-Net) to narrow the performance gap between image-level supervised methods and fully supervised methods. We introduce segmentation supervision by utilizing relationships between original and erased CAMs to generating part reliable pixel-level labels for WSSS methods. Besides, we design the rule of selecting segmentation labels to suppress noise. We also introduce other improvements to further increase prediction performance. Compared with other weakly supervised localization approaches, our ECS-Net refines CAMs with a more similar shape of objects. We train a fully supervised semantic segmentation model by our produced pseudo labels on PASCAL VOC 2012 dataset. Results show that our ECS-Net achieves state-of-the-art performance.

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2209–2218. Computer Vision Foundation / IEEE, 2019. 7

[2] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4981–4990. IEEE Computer Society, 2018. 1, 3, 8

[3] Amy L. Bearman, Olga Russakovsky, Vittorio Ferrari, and Fei-Fei Li. What's the point: Semantic segmentation with point supervision. In Bastian Leibe, Jiri Matas, Nicu Sebe,

and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII*, volume 9911 of *Lecture Notes in Computer Science*, pages 549–565. Springer, 2016. 2

[4] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, Lake Tahoe, NV, USA, March 12-15, 2018*, pages 839–847. IEEE Computer Society, 2018. 6

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 1, 8

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2018. 1, 8

[7] Liyi Chen, Weiwei Wu, Chenchen Fu, Xiao Han, and Yuntao Zhang. Weakly supervised semantic segmentation with boundary exploration. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXVI*, volume 12371 of *Lecture Notes in Computer Science*, pages 347–362. Springer, 2020. 7, 8

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, volume 11211 of *Lecture Notes in Computer Science*, pages 833–851. Springer, 2018. 5

[9] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1635–1643. IEEE Computer Society, 2015. 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009. 5

[11] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.*, 88(2):303–338, 2010. 5

[12] Bharath Hariharan, Pablo Arbelaez, Lubomir D. Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In Dimitris N. Metaxas, Long Quan, Alberto Sanfeliu, and Luc Van Gool, editors, *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, pages 991–998. IEEE Computer Society, 2011. 5

[13] Seunghoon Hong, Donghun Yeo, Suha Kwak, Honglak Lee, and Bohyung Han. Weakly supervised semantic segmentation using web-crawled videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2224–2232. IEEE Computer Society, 2017. 2

[14] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7014–7023. IEEE Computer Society, 2018. 7

[15] Peng-Tao Jiang, Qibin Hou, Yang Cao, Ming-Ming Cheng, Yunchao Wei, and Hongkai Xiong. Integral object mining via online attention accumulation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2070–2079. IEEE, 2019. 7

[16] Anna Khoreva, Rodrigo Benenson, Jan Hendrik Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1665–1674. IEEE Computer Society, 2017. 1, 2

[17] Alexander Kolesnikov and Christoph H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, volume 9908 of *Lecture Notes in Computer Science*, pages 695–711. Springer, 2016. 1, 3, 7, 8

[18] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5267–5276. Computer Vision Foundation / IEEE, 2019. 2, 3, 7

[19] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 3159–3167. IEEE Computer Society, 2016. 1, 2

[20] George Papandreou, Liang-Chieh Chen, Kevin Murphy, and Alan L. Yuille. Weakly- and semi-supervised learning of a DCNN for semantic image segmentation. *CoRR*, abs/1502.02734, 2015. 1, 2

[21] Deepak Pathak, Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional multi-class multiple instance learning. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR*

2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings, 2015. 2

[22] Pedro H. O. Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1713–1721. IEEE Computer Society, 2015. 2

[23] Xiaojuan Qi, Zhengzhe Liu, Jianping Shi, Hengshuang Zhao, and Jiaya Jia. Augmented feedback in semantic segmentation under image level supervision. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, volume 9912 of *Lecture Notes in Computer Science*, pages 90–105. Springer, 2016. 2

[24] Fatemehsadat Saleh, Mohammad Sadegh Ali Akbarian, Mathieu Salzmann, Lars Petersson, Stephen Gould, and Jose M. Alvarez. Built-in foreground/background prior for weakly-supervised semantic segmentation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, volume 9912 of *Lecture Notes in Computer Science*, pages 413–432. Springer, 2016. 7

[25] Wataru Shimoda and Keiji Yanai. Distinct class-specific saliency maps for weakly supervised semantic segmentation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, volume 9908 of *Lecture Notes in Computer Science*, pages 218–234. Springer, 2016. 7

[26] Wataru Shimoda and Keiji Yanai. Self-supervised difference detection for weakly-supervised semantic segmentation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 5207–5216. IEEE, 2019. 7, 8

[27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 8

[28] Meng Tang, Federico Perazzi, Abdelaziz Djelouah, Ismail Ben Ayed, Christopher Schroers, and Yuri Boykov. On regularized losses for weakly-supervised CNN segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI*, volume 11220 of *Lecture Notes in Computer Science*, pages 524–540. Springer, 2018. 1, 2

[29] Xiang Wang, Sifei Liu, Huimin Ma, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation by iterative affinity learning. *Int. J. Comput. Vis.*, 128(6):1736–1749, 2020. 7

[30] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1354–1362. IEEE Computer Society, 2018. 7

[31] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 12272–12281. IEEE, 2020. 3, 6, 7, 8

[32] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6488–6496. IEEE Computer Society, 2017. 2, 3, 7

[33] 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. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(11):2314–2320, 2017. 7

[34] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S. Huang. Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7268–7277. IEEE Computer Society, 2018. 2, 3, 7

[35] Zifeng Wu, Chunhua Shen, and Anton van den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognit.*, 90:119–133, 2019. 5, 8

[36] Huaxin Xiao, Jiashi Feng, Yunchao Wei, and Maojun Zhang. Self-explanatory deep salient object detection. *CoRR*, abs/1708.05595, 2017. 3

[37] Bingfeng Zhang, Jimin Xiao, Yunchao Wei, Mingjie Sun, and Kaizhu Huang. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020,*, pages 12765–12772. AAAI Press, 2020. 3

[38] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2921–2929. IEEE Computer Society, 2016. 2, 3, 4, 6, 7