Unlocking the Potential of Ordinary Classifier: Class-specific Adversarial Erasing Framework for Weakly Supervised Semantic Segmentation

Hyeokjun Kweon*, Sung-Hoon Yoon*, Hyeonseong Kim, Daehee Park, and Kuk-Jin Yoon
Visual Intelligence Lab., KAIST, Korea
{0327june,yoon307,brain617,bag2824,kjyoon}@kaist.ac.kr

Abstract

Weakly supervised semantic segmentation (WSSS) using image-level classification labels usually utilizes the Class Activation Maps (CAMs) to localize objects of interest in images. While pointing out that CAMs only highlight the most discriminative regions of the classes of interest, adversarial erasing (AE) methods have been proposed to further explore the less discriminative regions. In this paper, we review the potential of the pre-trained classifier which is trained on the raw images. We experimentally verify that the ordinary classifier[1] already has the capability to activate the less discriminative regions if the most discriminative regions are erased to some extent. Based on that, we propose a class-specific AE-based framework that fully exploits the potential of an ordinary classifier. Our framework (1) adopts the ordinary classifier to notify the regions to be erased and (2) generates a class-specific mask for erasing by randomly sampling a single specific class to be erased (target class) among the existing classes on the image for obtaining more precise CAMs. Specifically, with the guidance of the ordinary classifier, the proposed CAMs Generation Network (CGNet) is enforced to generate a CAM of the target class while constraining the CAM not to intrude the object regions of the other classes. Along with the pseudo-labels refined from our CAMs, we achieve the state-of-the-art WSSS performance on both PASCAL VOC 2012 and MS-COCO dataset only with image-level supervision. The code is available at https://github.com/KAIST-vilab/OC-CSE.

1. Introduction

Deep learning has been spotlighted for its effectiveness and evolved to achieve a higher level of performance than conventional techniques. In semantic segmentation [6, 7, 30, 41, 42], it has also achieved significant performance improvement. However, unlike other tasks such as object detection and classification, semantic segmentation requires dense pixel-level annotated labels that are time-consuming and costly to acquire. Accordingly, many attempts have been made for weakly-supervised semantic segmentation (WSSS) that only uses image-level classification labels [1–3, 10, 22, 33, 34, 37], scribbles [24, 31], and bounding boxes [8, 16, 26]. Among them, the most widely used approach is to utilize only image-level classification labels that can be easily obtained on massive amounts of data. In order to localize the object regions with the image-level labels, most existing approaches [1–3, 5, 10–12, 23, 28, 29, 33, 34, 37] utilize Class Activation Maps (CAMs) [40], represent the importance of image regions for the class prediction. To the best of our knowledge, most of the existing WSSS researches have pointed out that the CAMs high-

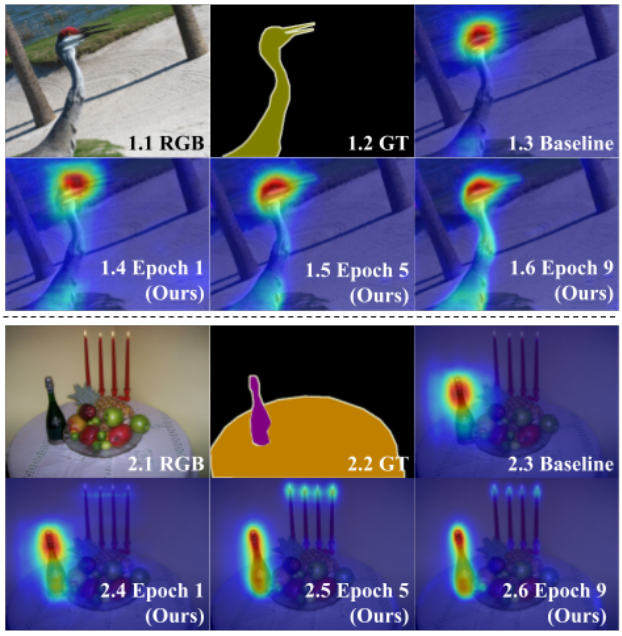

Figure 1: Qualitative comparison between the CAMs of baseline (ordinary classifier [2]) and ours on the PASCAL VOC 2012. From 1 to 6: original images, ground truth segmentations, baseline CAMs, our CAMs at epoch 1, 5, 9.

*The first two authors contributed equally. In alphabetical order.
[1]Throughout this paper, we will refer to a classifier pre-trained on raw images as a term 'ordinary classifier'.

light only the most discriminative regions rather than the whole object regions (*e.g.* 1.3 and 2.3 in Fig. 1). To dispel this under-activation issue, Adversarial Erasing (AE) methods [13, 22, 34, 39] have been widely used. They mask out the most highlighted parts of the CAMs from the image, and then a new classifier is trained on the masked images to seek the less highlighted regions.

In this paper, with a simple experiment inspired by the AE methods in Fig. 2 (which will be explained in Sec. 3), we review the potential of the ordinary classifier. We find that the ordinary classifier already has sufficient capability to identify the less discriminative regions without additional training. So, in our view, it is redundant to train a new classifier for the masked images as in existing AE methods. We experimentally verify that aggregating such regions using an ordinary classifier can be beneficial to generate the pseudo-labels for WSSS.

To fully exploit the potential of the ordinary classifier, we propose a class-specific AE-based framework that aggregates the regions from the most discriminative to the less discriminative. Our framework is composed of two networks: a CAMs Generating Network (CGNet) and the ordinary classifier used for guidance. First, we randomly sample a single class to be erased (target class) among the existing classes on the image. Then, the CAM of the target class is picked up among the CAMs generated by CGNet for masking the input image in a back-propagable manner. Finally, from the masked image, the ordinary classifier makes a prediction score of each class. We train the CGNet to lower the score of the erased target class, while the scores of the other existing classes are kept high.

The main advantage of the proposed class-specific erasing method is that it enables the CGNet to generate more precise CAMs. When all existing classes are simultaneously erased from the image in a class-agnostic manner [22], a confusion of the CGNet at the object boundaries between different classes cannot be resolved. Our class-specific erasing method can reduce such confusion by penalizing the intrusion of the CAMs at the object boundaries.

Figure 1 is a qualitative comparison between the CAMs of the baseline [2] (ordinary classifier) and the CGNet in the proposed framework. It shows that the localization ability of our CAMs gets better as the training proceeds, which supports the effectiveness of the proposed framework in a qualitative manner. We also conduct extensive ablation studies in Sec. 5.3 and experimentally verify that the proposed framework achieves additional performance gain in mean Intersection over Union (mIoU).

The contributions of our work are four-fold:

- We experimentally verify that an ordinary classifier has sufficient capability to segment the whole object region.
- To exploit the potential of the ordinary classifier, we propose an adversarial erasing-based framework.
- We design a class-specific erasing method that fully utilizes multi-class images which yields CAMs with more accurate boundaries.
- We achieve new *state-of-the-art* performance both on the PASCAL VOC 2012 *val/test* set and MS-COCO *val* set in the WSSS task with only the image-level classification labels.

2. Related Works

Utilizing only the image-level classification labels for semantic segmentation requires much less labeling costs among the various WSSS approaches, so we adopt this approach.

Earlier works in WSSS Most WSSS methods have employed CAMs to localize the object by only using the image-level classification labels. The CAMs, however, have been criticized for that they tend to focus on the most discriminative region, which can be an important classification cue, rather than the whole object regions. A group of studies attempted to expand and refine the sparse CAMs with seed growing methods [14, 18] or pixel-level affinity-based methods [2, 12, 28] to make dense pixel-level pseudo-labels for semantic segmentation. The aforementioned seed-based approaches, however, are highly dependent on the quality of the initial CAMs. Accordingly, numerous studies have been conducted to improve the quality of CAMs. Multiple dilated convolution blocks [35] and self-equivariant regularization [33] have been proposed to make the classifier for the CAMs robust under the scale variation. Also, much research has also been conducted to improve the localization ability with stochastic feature selection [21], an accumulation at different training phases [15], and cross-image approaches based on sub-category classification [3] or class-wise co-attention constraints [23, 29].

Adversarial erasing Adversarial Erasing (AE) method [13, 22, 34, 39] is one of the most commonly used method in WSSS. By explicitly erasing specific regions from an image, this method forces the network to explore the complete object region rather than to be biased on the most discriminative region. Wei *et al*. [34] firstly proposed a recursive find-and-erase scheme while training multiple classification networks. This scheme is repeated until newly adopted classification network fails to find meaningful object regions. Zhang *et al*. [39] improved the recursive scheme as an end-to-end framework composed of two branches with feature-level masking. In these works, however, even if the initial classifier succeeded to erase the object perfectly, the complementary network would not notice that fact and suffer from an over-erasing problem. SeeNet [13] attempted to reduce the over-erasing effect by replacing binary thresholding in [39] with the ternary thresholding that includes potential regions during the mask generation process. However,

this strategy requires the additional aid of a saliency detection module. Recently, Li *et al*. [22] proposed a soft mask generation network that can be jointly trained by a standard classification loss and an attention mining loss. The attention mining loss provides a self-guidance with the weight-shared networks to erase all the objects from the image by minimizing the overall class prediction scores of masked objects. The network obtains better localization capability while finding and masking out the objects from the image in a simultaneous manner. However, since the self-guidance is from the mask generation network itself, it is difficult to self-correct the overly activated regions judged already.

3. Potential of Ordinary Classifier

Aforementioned, it has been commonly regarded that the CAMs from the ordinary classifier usually highlight only the most discriminative parts of the objects rather than the entire object region. However, in this paper, we find that the ordinary classifier already has sufficient capability to activate the entire regions of the objects.

To reveal the potential of the ordinary classifier, we conduct a simple experiment with a recursive erase-and-infer process as visualized in Fig. 2. With an ordinary pre-trained classifier, we get initial CAMs from an input image. Then the image is masked by thresholding the highlighted regions of the CAMs. Interestingly, we can see that, even without the additional training step, when we re-infer the secondary CAMs from the masked image, the ordinary classifier activates the object-relevant regions which were originally suppressed on the initial CAMs. Note that the classifier stays fixed throughout the process, unlike conventional AE schemes which train a complementary classifier with masked images in each phase [34] or branch [13, 39]. With this simple erase-and-infer scheme, the aggregated CAMs achieve 51.3% mIoU on PASCAL VOC 2012 *train* set, which is significantly higher than the performance of baseline CAMs (47.8%).

It is true that the less discriminative regions are less activated on initial CAMs. The experimental results, however, suggest that such regions are not conspicuous due to highly-discriminative regions rather than being simply ignored by the classifier. In our perspective, the main limitation of CAMs is not in their sparsity and incompleteness but in the imbalance between the activation. Therefore, to generate more precise pseudo-labels for WSSS, it is possible to aggregate the less-activated regions with an ordinary classifier if it is well exploited.

This scheme, however, simply processes the image and aggregates the activated regions from the ordinary classifier in a sequential manner. So, during the process, there is no chance to recognize and learn the innate patterns of such regions which can be helpful to generate more complete CAMs. Moreover, it is extremely difficult to optimize

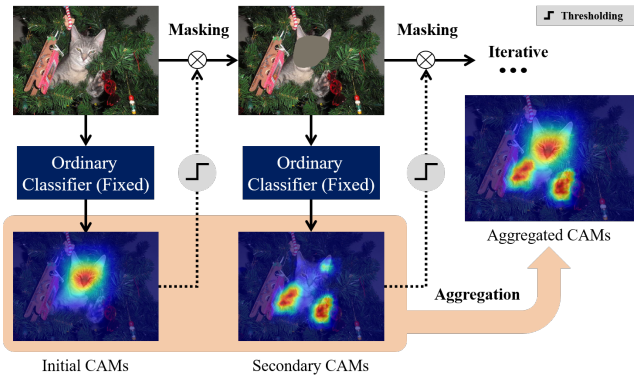

Figure 2: Diagram of a recursive erase-and-infer scheme. With a fixed ordinary classifier, initial CAMs are inferred from an input image. Then the image is masked by thresholding the highlighted regions on the CAMs and the CAMs are re-inferred in a recursive manner. *Note that the classifier stays fixed throughout the whole process in this scheme.*

the masking threshold for each image without the ground truth semantic segmentation labels. Therefore, to unlock the potential of the ordinary classifier while handling these issues, we propose a learning-based AE framework that harnesses the aforementioned scheme in an adaptive/recursive manner.

4. Proposed Method

4.1. CAMs Generation

We follow the approach of [40] to compute CAMs from an ordinary classification network with a small modification. Unlike the final layer of the classification network in [40], which is designed as a Global Average Pooling (GAP) followed by a fully-connected layer, we use a 1×1 convolution layer which has the number of classes (n_c) output channels followed by GAP as in [39]. Thereby the CAM of a class c_k is represented as $A^{c_k}(x,y) = f^{cam}_{c_k}(x,y)$, where $f^{cam}_{c_k}(x,y)$ denotes the feature vector at a location (x,y) on the feature map of the last convolution layer with the class c_k. A class prediction result of the network p for an image I can be defined as follows:

$$p = \sigma(GAP(f^{cam})), \tag{1}$$

where σ denotes the sigmoid activation function.

In order to utilize the CAM A^{c_k} as a back-propagable mask for erasing, we further take the Rectified Linear Unit (ReLU) on it and divide it by its max value so that the the feature maps be normalized between 0 and 1. Bilinear upsampling is applied to match the resolution of the image. The aforementioned process is shown as follows:

$$A^{c_k} = \frac{ReLU(A^{c_k})}{max(ReLU(A^{c_k}))}. \tag{2}$$

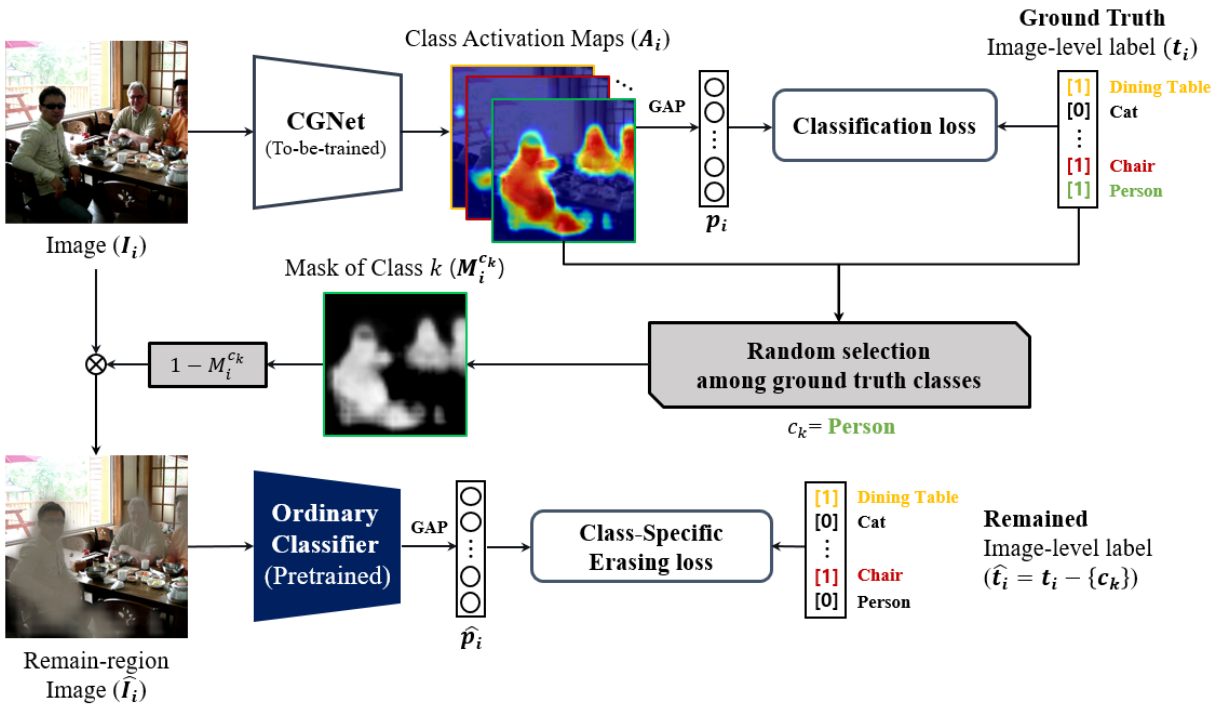

Figure 3: Overview of our AE-based framework with the proposed CSE method. For a given image I_i, class activation maps A_is are generated from the CGNet. Among the class labels of an input image, one target class, c_k, is randomly selected. Then corresponding mask $M_i^{c_k}$ is generated and used for masking the image. After that the remain-region image $\hat{I}_i$ is fed into the ordinary classifier, and the gradients back-propagated from the class-specific erasing loss guide the CGNet to generate better CAMs and a mask.

4.2. Proposed Framework

We denote the training data for a multi-label problem as $D = \{(I_i, t_i)\}_i$, where the label $t_i = \{c_1, c_2, \cdots, c_{n_i}\}$. As shown in Fig. 3, our framework is composed of two networks: CGNet and the ordinary classifier. The CGNet gets an image I_i as an input, then generates the class activation maps A_i and class prediction p_i. In order to mask the image, $M_i^{c_k}$ is selected among A_i, where c_k is a single mask-class label randomly sampled from the ground truth classes t_i. Throughout the following explanation, we name the class c_k as a "target class" and the other classes as "remaining classes". Then the masked image $\hat{I}_i$ is computed as follows:

$$\hat{I}_i = (1 - M_i^{c_k}) \odot I_i, \tag{3}$$

where $\odot$ denotes the element-wise multiplication.

After that, the fixed ordinary classifier gets the masked image $\hat{I}_i$ and makes a class prediction $\hat{p}_i$. Our framework enforces the CGNet to make the masked image not contain the target class anymore, but still include the remaining classes. The CGNet is trained with a combination of the two classification loss functions as:

$$\mathcal{L}_{ours} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{cse} = \ell_{bce}(p_i, t_i) + \lambda \ell_{bce}(\hat{p}_i, \hat{t}_i), \tag{4}$$

where $\mathcal{L}_{cls}$ and $\mathcal{L}_{cse}$ correspond to $\ell_{bce}(p_i, t_i)$ and $\ell_{bce}(\hat{p}_i, \hat{t}_i)$, respectively. Here, ℓ_{bce} denotes the binary cross entropy loss and λ is a weighting parameter which balances between both terms. Note that $\hat{t}_i$ stands for the remaining class labels defined as $\hat{t}_i = t_i - \{c_k\}$.

As discussed in Sec. 3, once the most discriminative regions are masked from the image, the ordinary classifier focuses on secondary discriminative regions and uses them as cues for classification. Therefore, if the CGNet fails to erase some portions of the object of the target class, the ordinary classifier could detect such under-erasing and notifies the CGNet to erase them. Also, with extensive ablation studies in Sec. 5.3, we verify that receiving guidance from the fixed ordinary classifier is more reliable than from the weight-shared networks (self-guidance) as in [22].

4.3. Class-specific Erasing Method (CSE)

The proposed CSE method has a superiority over the Class-Agnostic Erasing (CAE) method [22] in the perspective of generating the precise CAMs. In the CAE method, a mask is generated by applying a pixel-wise max function on CAMs. Since the goal of the CAE method is to erase all existing classes, the loss function is as follows:

$$\mathcal{L}_{agno} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{cae} = \ell_{bce}(p_i, t_i) + \lambda \ell_{bce}(\hat{p}_i, \varnothing) \tag{5}$$

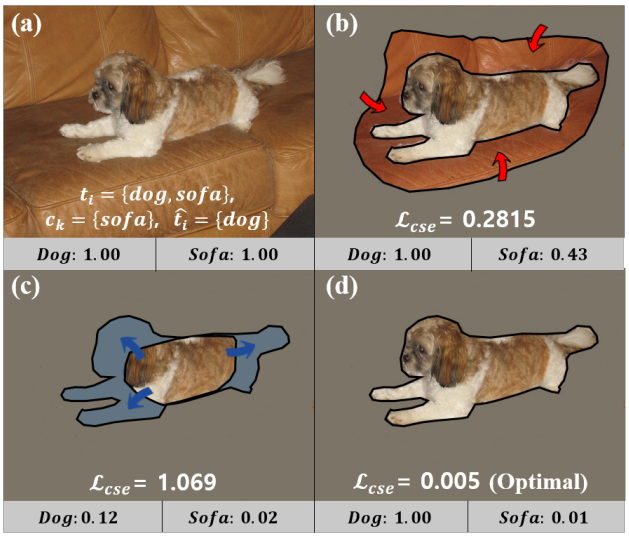

Figure 4: The illustration showing how our CSE method works. The confidence score from the ordinary classifier is listed for each image. This figure is an example of masking a *sofa* class, and notation follows the description of the framework. Class-specific erasing loss ($\mathcal{L}_{cse}$) values are also shown.

where $\mathcal{L}_{cls}$ and $\mathcal{L}_{cae}$ correspond to $\ell_{bce}(p_i, t_i)$ and $\ell_{bce}(\hat{p}_i, \varnothing)$, respectively. Here, $\mathcal{L}_{cls}$ is a standard classification loss and $\mathcal{L}_{cae}$ is a loss between the class prediction of masked image at the ordinary classifier and $\varnothing$, which denotes that the label is an empty set.

In this case, the loss function for CAE in Eq. 5 only checks if there is any object in the masked image. So, even if the generated mask is imprecise at the boundaries between the objects of different categories, such unwanted intrusion can not be penalized since there is no difference on the loss function. With the class-specific erasing method, on the other hand, $\mathcal{L}_{cse}$ in Eq. 4 enables CGNet to learn localization from the ordinary classifier. Since the remaining classes should be predicted from the masked image on the ordinary classifier, the mask $M_i^{c_k}$ is constrained not to intrude the regions of the remaining classes. The loss $\mathcal{L}_{cse}$, therefore, induces the CGNet to generate mask that fits along the object boundaries.

In Fig. 4, we visualize how the proposed CSE method works in more detail. Suppose an image containing two classes *dog, sofa* is given, where the target class c_k is *sofa*, and the remaining class $\hat{t}_i$ is *dog*. If the generated activation map of the target class (*sofa*) is under-activated as shown in Fig. 4-(b), the ordinary classifier would predict 0.43 of confidence score on the *sofa* class since the ordinary classifier has sufficient capability to find the remained regions of *sofa*. Since the *dog* is the remaining class ($\hat{t}_i$) for the CSE loss ($\mathcal{L}_{cse}$), the CGNet is trained to decrease the confidence score on the *sofa* class by expanding the *sofa* activation map.

Conversely, when the activation map of the *sofa* class is over-activated as shown in Fig. 4-(c) and intrudes the region of the other class (*dog*), the confidence score of the *dog* predicted by the ordinary classifier would be decreased. In this case, since the remained image-level label ($\hat{t}_i$) still demands the *dog* class to be left, $\mathcal{L}_{cse}$ will successfully notice this intrusion and punish CGNet to reduce the activation map of the *sofa* class. Through the aforementioned process, we can expect that the CGNet ultimately be optimized to the optimal solution (Fig. 4-(d)) while balancing between the over-activation and under-activation with the help of penalization from $\mathcal{L}_{cse}$.

As the conventional adversarial erasing approaches have faced, the classification loss of the image itself cannot spatially constrain the activation maps to be along the object boundary. However, if we adopt the class-specific erasing method on our framework, then the CGNet could learn the boundary information between the objects on the multi-class image while reducing and expanding its activation regions under reliable guidance from the ordinary classifier. This is a simple but effective way to spatially constrain the activation maps by only using mere classification loss from image-level labels.

Even though the ordinary classifier might not be able to notice when the regions of the remaining classes that intruded by the over-activated CAM of the target class are less-discriminative, the CSE method enables the CGNet to learn the concept of "object boundary" from cases like in the Fig 4. Furthermore, qualitative and quantitative results in Sec 5.3 support that the proposed CSE method properly works according to the design intent.

4.4. CAM Refinement

To refine the CAMs generated from the CGNet for more accurate pseudo pixel-level labels, we follow the work of [2]. The foreground and background labels that are required to train AffinityNet in [2] are obtained by applying crf [19] to our refined CAMs. In order to apply CRF to our CAMs, we compute a background activation map as:

$$A_{bg}(x, y) = \left\{1 - \max_{c \in t} A(x, y)\right\}^{\alpha}. \quad (6)$$

After training AffinityNet, we use those pseudo labels to train Deeplab [4] to accomplish the goal of WSSS.

5. Experiments

5.1. Dataset and Evaluation Metric

We evaluate the proposed method on the PASCAL VOC 2012 dataset [9] and MS-COCO dataset [25]. PASCAL VOC 2012 dataset contains 20 foregrounds and one background categories. As conventional approaches, augmented training set (10,528) with image-level class labels is used

Table 1: The ablation study of proposed framework. $\mathcal{W}.\mathcal{S}.$: Weight Sharing as in [22] with CGNet. $\mathcal{O}.\mathcal{C}.$: the Ordinary Classifier. **CAE**: Class-Agnostic Erasing method. **CSE**: Class-Specific Erasing method. **crf**: Conditional random fields. For the implementation of the GAIN [22], we employ the same backbone as Ours. The performance is evaluated on the PASCAL VOC *train* set.

	Erasing method		Guidance		mIoU(%)	mIoU(%) w/ crf
	CAE	CSE	$\mathcal{O}.\mathcal{C}.$	$\mathcal{W}.\mathcal{S}.$		
Baseline					47.8	53.7
GAIN [22]	✓			✓	48.3	53.7
Ours (w/o CSE)	✓		✓		53.3	59.7
Ours (w/o $\mathcal{O}.\mathcal{C}.$)		✓		✓	47.1	52.5
Ours		✓	✓		**56.0**	**62.8**

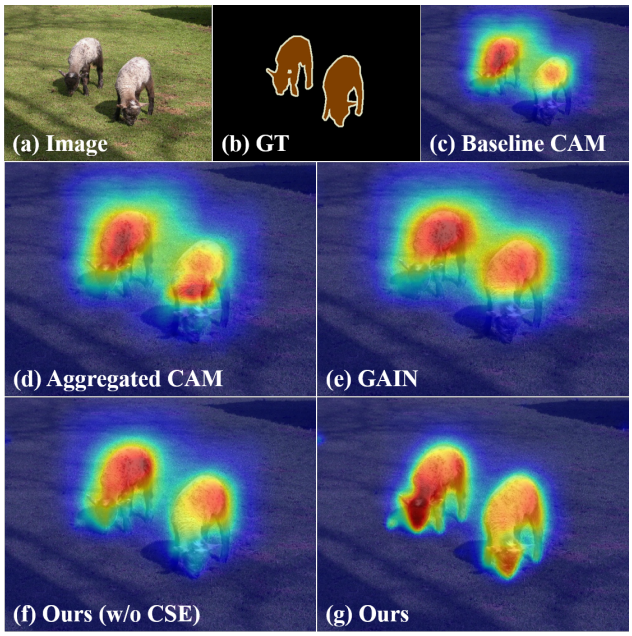

Figure 5: Qualitative comparison of CAMs among several methods on the PASCAL VOC 2012. From (a) to (e): Image, Ground truth segmentation, Baseline CAM, Aggregated CAM, GAIN [22], Ours (w/o CSE), and Ours. Aggregated CAM is generated as the method shown in Fig. 2.

for training. We use validation (1,464) and test sets (1,456) to evaluate our results and to compare with other methods. The other dataset, MS-COCO [25], contains 81 classes including the background class with 80k *train* and 40k *val* images, which is more general and difficult in the perspective of WSSS. As an evaluation metric, we employ the mean Intersection over Union (mIoU) which is a common standard for semantic segmentation task.

5.2. Implementation Details

Framework The proposed network is implemented with PyTorch. In our experiments, ResNet38 [36] is employed as a backbone network for both the CGNet and the ordinary classifier. As our framework is guided by the ordinary classifier, the dependency on it will be discussed in the *Supplementary material* with additional experiments. Both networks are initialized with ImageNet [27] weights. Before training our full framework, the ordinary classifier is pre-trained by a standard classification loss with PASCAL VOC 2012 *train* dataset. Likewise, for experiment using MS-COCO, we pre-trained the classifier in the same manner. For data augmentation, random resizing, horizontal flipping, color jittering [20] and random cropping are applied to the input images. The model is trained on 4 TITAN-RTX GPUs with batch size 8 for 15 epochs. We use a poly learning rate which multiplies $(1-\frac{iter}{max\ iter})^{power}$ to an initial learning rate as in [6]. We set the initial learning rate as 0.01 and the power is set to 0.9.

Figure 6: Qualitative comparison between CAMs on PASCAL VOC 2012. From left to right: Images, Ground-truth segmentations, CAMs of Baseline [2], CAMs of Ours.

AffinityNet and Deeplab To refine the pseudo label, we design both AffinityNet and Deeplab with ResNet38 backbone as in [2]. We use 3/24 for α in Eq. 6 to get the confident foreground/background regions to train the AffinityNet. Learning rate for training Deeplab is set to 0.001.

5.3. Ablation Studies

The ablation study of our method on PASCAL VOC 2012 dataset is shown in Table 1. While adjusting the erasing method (Class-agnostic or Class-specific) and the type of guidance (Ordinary Classifier or Weight Sharing), performance of each method is evaluated. For the implementation of weight-shared guidance, we follow the work of GAIN [22] which receives the guidance from the weight-shared network and uses CAE method. Comparing with the baseline, the GAIN achieves slight performance increase of 0.5%. By utilizing the ordinary classifier as the guidance, the performance is greatly increased to 53.3%. It implies that the penalization from the ordinary classifier is more beneficial comparing with that from the weight-sharing method. We also conduct an experiment to verify the effectiveness of our CSE method. While the CSE method can be applied independently to the ordinary classifier, combining it with the weight-sharing guidance lowers the performance than the baseline. With the ordinary classi-

Table 2: Performance (mIoU,%) comparison with other state-of-the-art WSSS methods on the PASCAL VOC 2012 *val* and *test* set. $\mathcal{I}$ and $\mathcal{S}$ represent image-level labels and the external saliency module used for supervision, respectively. **Bold** numbers represent the best results, while underlined numbers are the second best ones.

Methods	*Backbone*	*Sup.*	*Pub.*	*Val*	*Test*
AdvErasing [34]	VGG16	$\mathcal{I}$	CVPR17	55.0	55.7
GAIN [22]	VGG16	$\mathcal{I}$	CVPR18	55.3	56.8
AffinityNet [2]	ResNet38	$\mathcal{I}$	CVPR18	61.7	63.7
ICD [10]	ResNet101	$\mathcal{I}$	CVPR20	64.1	64.3
IRNet [1]	ResNet50	$\mathcal{I}$	CVPR19	63.5	64.8
SSDD [28]	ResNet38	$\mathcal{I}$	ICCV19	64.9	65.5
SEAM [33]	ResNet38	$\mathcal{I}$	CVPR20	64.5	65.7
Sub-category [3]	ResNet101	$\mathcal{I}$	CVPR20	66.1	65.9
RRM [37]	ResNet101	$\mathcal{I}$	AAAI20	66.3	66.5
BES [5]	ResNet101	$\mathcal{I}$	ECCV20	65.7	66.6
Ours	ResNet38	$\mathcal{I}$	-	**68.4**	68.2
MCOF [32]	ResNet101	$\mathcal{I}$+$\mathcal{S}$	CVPR18	60.3	61.2
SeeNet [13]	ResNet101	$\mathcal{I}$+$\mathcal{S}$	NIPS18	63.1	62.8
DSRG [14]	ResNet101	$\mathcal{I}$+$\mathcal{S}$	CVPR18	61.4	63.2
FickleNet [21]	ResNet101	$\mathcal{I}$+$\mathcal{S}$	CVPR19	64.9	65.3
CIAN [12]	ResNet101	$\mathcal{I}$+$\mathcal{S}$	AAAI20	64.3	65.4
OAA+ [15]	ResNet101	$\mathcal{I}$+$\mathcal{S}$	ICCV19	65.2	66.4
EME [11]	ResNet101	$\mathcal{I}$+$\mathcal{S}$	ECCV20	67.2	66.7
MCIS [29]	ResNet38	$\mathcal{I}$+$\mathcal{S}$	ECCV20	66.2	66.9
ICD [10]	ResNet101	$\mathcal{I}$+$\mathcal{S}$	CVPR20	67.8	68.0
Group-WSSS [23]	ResNet101	$\mathcal{I}$+$\mathcal{S}$	AAAI21	68.2	**68.5**

fier, on the other hand, the performance is further increased to 56.0%, which is a significant gain (8.2% in mIoU) compared to the baseline. In our view, it is difficult to self-correct the miss-activated regions with the self-guidance from the weight-shared network itself, while our scheme can handle such error with the ordinary classifier. We also observe that simultaneously updating the ordinary classifier with the CGNet leads to slightly worse result. In the perspective of the optimization, keeping the guidance network fixed seems to be beneficial since it can provide a more stable gradient to the CGNet.

When the image has only one object class, it is true that the proposed CSE method is the same as the CAE method. But as shown in Table 1, the CSE method achieves significant performance gain over CAE method on the PASCAL VOC 2012 dataset, even though only 40% of the *train* set is multi-class. It implies that the CSE method can effectively exploit the rich information of multi-class images, which is clearly superior to the CAE method. Also, in the perspective of the segmentation tasks, it is much more general and practical to handle a dataset with multi-class images such as MS COCO. As shown in Table 3, the proposed framework also achieves *state-of-the-art* on the MS COCO dataset.

When applying crf to our framework, the mIoU of CAMs is drastically increased to 62.8%. Considering the performance gap between the baseline and baseline with crf is 5.9%, our framework could even more benefit from the crf.

Table 3: Quantitative comparison of the proposed frameworks with other *state-of-the-art* method on MS-COCO [25]. The results of [1, 33](*) are re-implemented by [38].

Method	Publication	Backbone	*val* (mIoU)
SEC [18]	ECCV16	VGG16	22.4
DSRG [14]	CVPR18	VGG16	26.0
Group-WSSS [23]	AAAI21	VGG16	28.4
SEAM* [33]	CVPR20	ResNet38	31.9
IRNet* [1]	CVPR19	ResNet50	32.6
SEAM+CONTA [38]	NeurIPS20	ResNet38	32.8
IRNet+CONTA [38]	NeurIPS20	ResNet50	33.4
Ours	-	ResNet38	**36.4**

We interpret this performance gain comes from the capability of CGNet to generate more precise CAMs that match along object boundaries. As the CAMs less invade across the object boundaries, the crf less confuses while refining the CAMs.

Figure 5 shows the qualitative comparison among five different methods. Comparing (d) the Aggregated CAM and (e) GAIN with (c) baseline CAM, highly-discriminative regions become wider. Since the proposed framework without CSE method has much less risk for an over-erasing with reliable guidance from the ordinary classifier, localization ability of (f) Ours (w/o CSE) is improved comparing with the CAM of (c)-(e). With the class-specific erasing method, as shown in (g) Ours, the CGNet generates much more precise CAM. Qualitative results shown in Fig. 6 also support the effectiveness of our framework.

5.4. Comparison with State-of-the-arts

To improve the quality of pixel-level pseudo labels, we follow the work of [2] as in [3, 33]. After training the AffinityNet with the CAMs from the proposed framework and applying the crf, the synthesized pseudo labels achieve 66.9% mIoU on PASCAL VOC 2012 *train* set. The pseudo pixel-level labels are employed to train the Deeplab-LargeFOV [4] with a ResNet38 backbone network. As shown in Table 2, we achieve 68.4% and 68.2% mIoU on PASCAL VOC 2012 *val* and *test* sets, respectively, achieving new state-of-the-arts. Table 4 gives class-wise IoU comparison with previous methods on *val* set. As shown in Fig. 7, our framework more accurately and precisely segments the objects comparing with [2].

To show the superiority of the proposed framework more clearly, we also conduct experiments on the MS-COCO [25] dataset. Since there are more multi-class images in MS-COCO than PASCAL VOC 2012, the benefit of the proposed CSE method would be more evident. We just apply the crf on the CAMs generated by the CGNet to acquire pseudo-labels, which achieves 37.2% mIoU on the *train* set. Note that we skip the phase for training the Affinity network since it consumes too many resources, and the performance could be even higher with the Affinity phase. As shown in

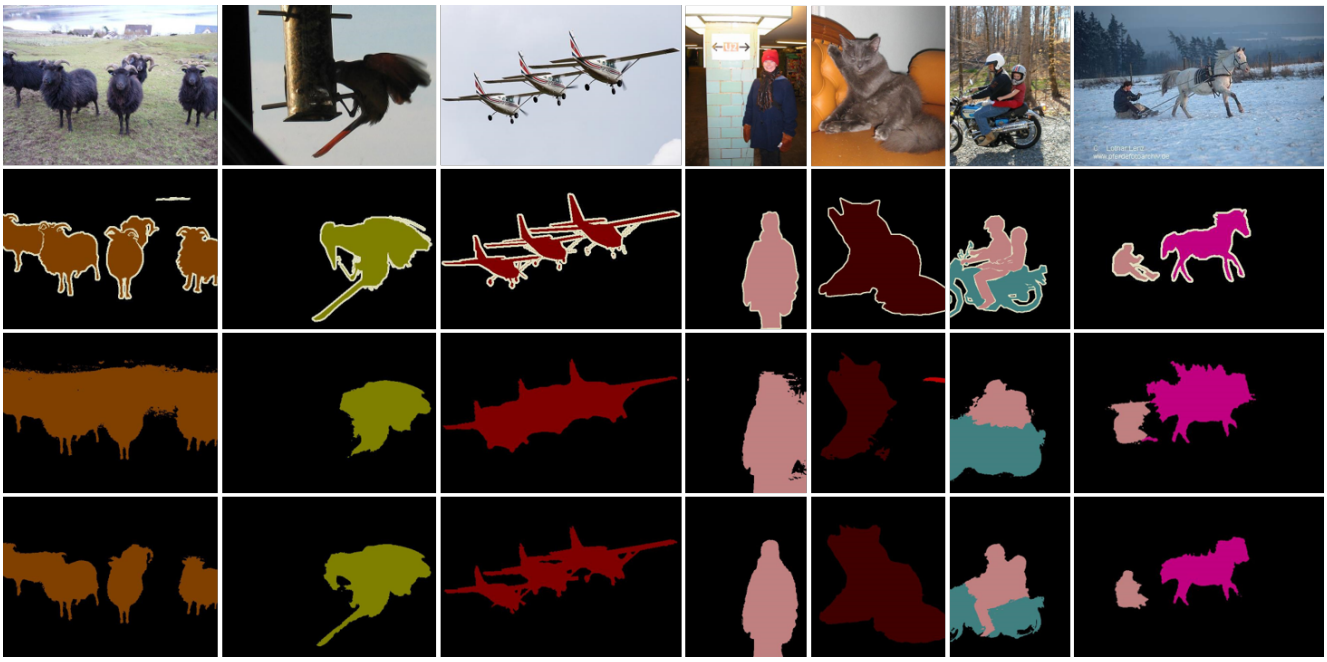

Figure 7: Qualitative results of the segmentation networks trained with pseudo pixel-level labels. Note that those pseudo labels are generated using only image-level labels. From top to bottom: Image, Ground truth, Segmentation results of a baseline [2], Segmentation results of ours.

Table 4: Class-wise IoU comparison on PASCAL VOC 2012 *val* set with only image-level supervision.

Method	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbk	person	plant	sheep	sofa	train	tv	mIoU
TPL [17]	82.8	62.2	23.1	65.8	21.1	43.1	71.1	66.2	76.1	21.3	59.6	35.1	70.2	58.8	62.3	66.1	35.8	69.9	33.4	45.9	45.6	53.1
AdvErasing [34]	83.4	71.1	30.5	72.9	41.6	55.9	63.1	60.2	74.0	18.0	66.5	32.4	71.7	56.3	64.8	52.4	37.4	69.1	31.4	58.9	43.9	55.0
AffinityNet [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 [33]	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 [28]	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 [5]	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	**90.2**	**82.9**	**35.1**	**86.8**	**59.4**	**70.6**	82.5	78.1	**87.4**	30.1	79.4	45.9	83.1	**83.4**	**75.7**	73.4	48.1	**89.3**	42.7	60.4	52.3	**68.4**

Table 3, we experimentally verify the effectiveness of the proposed framework. Our framework achieves **36.4%** on the MS-COCO *val* set, which is a new *state-of-the-art* that surpasses the previous best method by 3.0%.

6. Conclusion and Future Works

In this paper, we proposed a class-specific adversarial erasing-based framework while fully exploiting the potential of ordinary classifier. Motivated by that the ordinary classifier already has enough capability to identify less discriminative regions, we designed the CGNet to extract the full potential from the ordinary classifier. Furthermore, the proposed class-specific erasing (CSE) methods guide the CGNet to generate more precise CAMs by learning boundary information between the objects on multi-class images. Extensive qualitative and quantitative experimental results support the effectiveness of the proposed framework. Along with the pseudo pixel-level labels refined from our CAMs, we achieved the state-of-the-art WSSS performance on both PASCAL VOC 2012 *val* / *test* set and MS COCO *val* set only with image-level supervision.

The proposed framework succeeds to unlock the valuable potential of the ordinary classifier, however, by nature of using a pre-trained classifier for guidance, the proposed framework has a limitation. We observe that the miss-classification results of the ordinary classifier sometimes lead to failure cases and therefore limit the upper bounds of the performance of our framework. In this sense, future studies could investigate the training scheme to handle such miss-classification, or replacing the ordinary classifier with another classifier specialized in WSSS to further enhance the performance of the proposed framework.

Acknowledgements This work was supported by Institute of Information and Communications Technology Planning & Evaluation(IITP) Grant funded by Korea Government (MSIT) (No. 2020-0-00440, Development of Artificial Intelligence Technology that Continuously Improves Itself as the Situation Changes in the Real World)

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2209–2218, 2019. 1, 7

[2] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4981–4990, 2018. 1, 2, 5, 6, 7, 8

[3] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8991–9000, 2020. 1, 2, 7

[4] 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. 5, 7

[5] Liyi Chen, Weiwei Wu, Chenchen Fu, Xiao Han, and Yuntao Zhang. Weakly supervised semantic segmentation with boundary exploration. In *European Conference on Computer Vision*, pages 347–362. Springer, 2020. 1, 7, 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 transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 1, 6

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1

[8] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1635–1643, 2015. 1

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 5

[10] Junsong Fan, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4283–4292, 2020. 1, 7

[11] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Employing multi-estimations for weakly-supervised semantic segmentation. Springer, 2020. 1, 7

[12] Junsong Fan, Zhaoxiang Zhang, Tieniu Tan, Chunfeng Song, and Jun Xiao. Cian: Cross-image affinity net for weakly supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10762–10769, 2020. 1, 2, 7

[13] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In *Advances in Neural Information Processing Systems*, pages 549–559, 2018. 2, 3, 7

[14] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7014–7023, 2018. 2, 7

[15] Peng-Tao Jiang, Qibin Hou, Yang Cao, Ming-Ming Cheng, Yunchao Wei, and Hong-Kai Xiong. Integral object mining via online attention accumulation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2070–2079, 2019. 2, 7

[16] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 876–885, 2017. 1

[17] Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon. Two-phase learning for weakly supervised object localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3534–3543, 2017. 8

[18] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European conference on computer vision*, pages 695–711. Springer, 2016. 2, 7

[19] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011. 5

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 6

[21] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5267–5276, 2019. 2, 7

[22] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9215–9223, 2018. 1, 2, 3, 4, 6, 7

[23] Xueyi Li, Tianfei Zhou, Jianwu Li, Yi Zhou, and Zhaoxiang Zhang. Group-wise semantic mining for weakly supervised semantic segmentation. *arXiv preprint arXiv:2012.05007*, 2020. 1, 2, 7

[24] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3159–3167, 2016. 1

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence

Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5, 6, 7

[26] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1742–1750, 2015. 1

[27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 6

[28] Wataru Shimoda and Keiji Yanai. Self-supervised difference detection for weakly-supervised semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5208–5217, 2019. 1, 2, 7, 8

[29] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. *arXiv preprint arXiv:2007.01947*, 2020. 1, 2, 7

[30] Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5229–5238, 2019. 1

[31] Paul Vernaza and Manmohan Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7158–7166, 2017. 1

[32] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1354–1362, 2018. 7

[33] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020. 1, 2, 7, 8

[34] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1568–1576, 2017. 1, 2, 3, 7, 8

[35] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7268–7277, 2018. 2

[36] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133, 2019. 6

[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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12765–12772, 2020. 1, 7

[38] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xiansheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 2020. 7

[39] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1325–1334, 2018. 2, 3

[40] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 1, 3

[41] Yi Zhu, Karan Sapra, Fitsum A Reda, Kevin J Shih, Shawn Newsam, Andrew Tao, and Bryan Catanzaro. Improving semantic segmentation via video propagation and label relaxation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8856–8865, 2019. 1

[42] Yueqing Zhuang, Fan Yang, Li Tao, Cong Ma, Ziwei Zhang, Yuan Li, Huizhu Jia, Xiaodong Xie, and Wen Gao. Dense relation network: Learning consistent and context-aware representation for semantic image segmentation. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3698–3702. IEEE, 2018. 1