Anti-Adversarially Manipulated Attributions for Weakly and Semi-Supervised Semantic Segmentation

Jungbeom Lee[1] Eunji Kim[1] Sungroh Yoon[1,2,*]
[1] Department of Electrical and Computer Engineering, Seoul National University, Seoul, South Korea
[2] ASRI, INMC, ISRC, and Institute of Engineering Research, Seoul National University
{jbeom.lee93, kce407, sryoon}@snu.ac.kr

Abstract

Weakly supervised semantic segmentation produces a pixel-level localization from a classifier, but it is likely to restrict its focus to a small discriminative region of the target object. AdvCAM is an attribution map of an image that is manipulated to increase the classification score. This manipulation is realized in an anti-adversarial manner, which perturbs the images along pixel gradients in the opposite direction from those used in an adversarial attack. It forces regions initially considered not to be discriminative to become involved in subsequent classifications, and produces attribution maps that successively identify more regions of the target object. In addition, we introduce a new regularization procedure that inhibits the incorrect attribution of regions unrelated to the target object and limits the attributions of the regions that already have high scores. On PASCAL VOC 2012 test images, we achieve mIoUs of 68.0 and 76.9 for weakly and semi-supervised semantic segmentation respectively, which represent a new state-of-the-art. The code is available at: https://github.com/jbeomlee93/AdvCAM.

1. Introduction

Semantic segmentation involves the allocation of a semantic label to each pixel of an image. It is an essential task in image recognition and scene understanding. Deep neural networks (DNNs) have facilitated tremendous progress in semantic segmentation [8, 22]; but they require a large number of training images annotated with pixel-level labels. Preparing such a training dataset is very expensive: pixel-level annotation of images containing an average of 2.8 objects takes about 4 minutes [4] per image, and a single large (2048×1024) image depicting a complicated scene requires more than 90 minutes for pixel-level annotation [9].

The need for pixel-level annotation is addressed by weakly supervised learning, in which a segmentation network is trained on images with less comprehensive annotations that are cheaper to obtain than pixel-level labels. Weakly supervised methods can use scribbles [54], points [4], bounding boxes [26, 51], and class labels [2, 6, 29, 48] as annotations. Labeling an image with class labels takes about 20 seconds [4], making class labels the cheapest option. In addition, many public datasets are already annotated with class labels [10, 12], and automated web searches can also provide images with class labels [20, 30, 47] although the accuracy of such labels may be low. These considerations make class labels the most popular form of weak supervision.

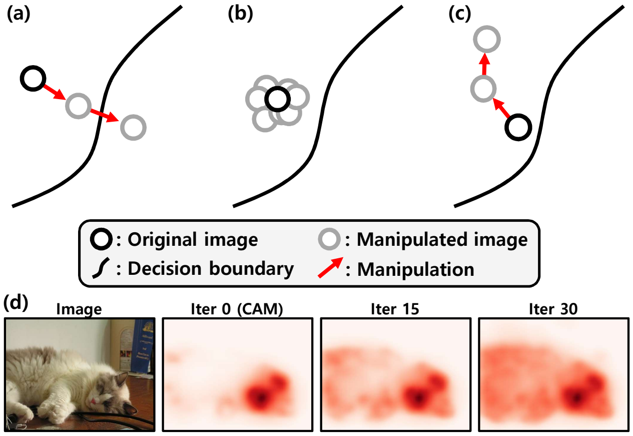

Figure 1: Conceptual description of image manipulation methods for weakly supervised semantic segmentation: (a) erasure [21, 57, 63]; (b) FickleNet [29]; and (c) AdvCAM. (d) Examples of successive attribution maps obtained from iteratively manipulated images.

Most weakly supervised segmentation methods that use class labels depend on attribution maps obtained from a trained classifier [46, 64]. Such a map identifies the image regions on which the classifier concentrated. However, these important, or discriminative, regions are relatively small, and most attribution maps do not represent the whole region occupied by a target object, which makes those attribution maps unsuitable for training a semantic segmentation network. Therefore, many researchers have tried to extend

[*] Correspondence to: Sungroh Yoon <sryoon@snu.ac.kr>.

regions to cover more of a target object, by manipulating images [33, 50, 57] or feature maps [21, 29, 63].

One popular method for manipulation is erasure: the classifier is forced to find new regions of the target object from which discriminative regions previously located have been removed. Erasure is effective, but it requires modification of the network, often by adding additional layers [21, 63], or additional training steps [57]. Another difficulty is the provision of a reliable termination condition for the iterative erasure; the erasure of discriminative region of an image can cause the DNN to misclassify that image. If the image from which the discriminative region has been erased crosses the decision boundary as shown in Figure 1(a), an erroneous attribution map may be generated. An alternative method for manipulation is a stochastic perturbation shown in Figure 1(b). FickleNet [29] diversifies attribution maps from an image by applying random dropout to the feature maps of a DNN and aggregates them into a unified map.

We propose a new manipulation method for extending the discriminative regions of a target object. Our method is based on adversarial attack [16, 28], but with a benign purpose. Adversarial attack finds a small perturbation of an image that pushes it across the decision boundary to change the classification result. By contrast, our method operates in an anti-adversarial manner , which is the reversal of adversarial attack. It aims to find a perturbation that pushes the manipulated image away from the decision boundary, as shown in Figure 1(c). This manipulation is realized by adversarial climbing, in which an image is perturbed along pixel gradients which increase the classification score of the target class. The result is that non-discriminative regions, which are nevertheless relevant to that class, gradually become involved in the classification, so that the CAM of the manipulated image identifies more regions of the object. Figure 1(d) shows examples of CAMs obtained by applying this manipulation technique iteratively.

Ascending the gradient ensures that classification score increases, but the repetitive ascending may cause irrelevant areas, such as parts of the backgrounds or regions of other objects, to be activated together or the attribution scores of some part of the target object to be increased dramatically. We can address these problems by introducing regularization terms that suppress the scores of other classes and limit the attribution scores of the regions that already have high scores. The attribution maps obtained from images that have been iteratively manipulated in this way can be used as pseudo ground-truth masks to train a semantic segmentation network in a weakly and semi-supervised manner.

Our method is a post-hoc analysis of the trained classifier, and can be used to improve the performance of existing methods without modification, resulting in new state-of-the-art performance on the PASCAL VOC 2012 benchmark in both weakly and semi-supervised semantic segmentation.

The main contributions of this paper are three-fold:

- We propose AdvCAM, an attribution map of an image that is manipulated to increase the classification score, allowing it to identify more regions of an object.
- We empirically demonstrate that our method improves the performance of several methods of weakly supervised semantic segmentation without modification or re-training of their networks.
- Our technique produces significantly better performance on the Pascal VOC 2012 benchmark than existing methods, in both weakly and semi-supervised semantic segmentation.

2. Related Work

2.1. Weakly Supervised Learning

Existing weakly supervised semantic segmentation methods aim to find the whole region occupied by a target object by obtaining an improved initial seed which contains a good approximation of the region occupied by the object, and growing that region so that more of the object is identified.

Obtaining a High Quality Seed: Several methods have been proposed to improve the quality of the initial seeds obtained from classifiers. Wang *et al.* [56] use equivariance regularization during the training of their classifier so that the attribution maps obtained from differently transformed images are equivariant to those transformations. Chang *et al.* [6] improve feature learning by using latent semantic classes that are sub-categories of annotated parent classes, which can be pseudo-labeled by clustering image features. Fan *et al.* [13] and Sun *et al.* [53] capture information shared between several images by considering cross-image semantic similarities and differences. Wei *et al.* [58] and Lee *et al.* [32] consider the target object in several contexts by combining multiple attribution maps from differently dilated convolutions or from different layers of a DNN.

Growing the Object Region: Some researchers expand an initial CAM [64] seed using a method analogous to region growing by examining the neighborhood of each pixel. Semantic labels are propagated from regions which can confidently be associated with the target object to regions which were initially ambiguous. SEC [27] and DSRG [23] start with a initial CAM seed containing ambiguous regions, and allocates pseudo labels to those ambiguous region during the training of the segmentation network. PSA [2] and IRN [1] extend the object region to semantically similar areas by a random walk. BEM [7] synthesizes a pseudo boundary from a CAM and then uses a similar propagation with PSA [2].

2.2. Semi-Supervised Learning

In semi-supervised learning, a segmentation network is trained using a small number of images with pixel-level annotations, together with a much larger number of images with weak annotations or none at all. Cross-consistency training (CCT) [42] involves the training of a segmentation network with unlabeled, or weakly labeled, images by enforcing an invariance of the predictions over different perturbations, such as injecting random noise. Souly *et al.* [52] improve feature learning by using images synthesized by generative adversarial network [15]. Hung *et al.* [24] adopt adversarial training scheme that reduces the distribution gap between predicted segmentation maps and ground-truth maps.

2.3. Adversarial Attack

Methods of adversarial attack attempt to fool a DNN by presenting it with manipulated input with the intent to deceive. Adversarial attack can be applied to classification [16, 40], semantic segmentation [3], and object detection [60]. Deceptive attribution maps can also be produced by adversarial image manipulation [11] or model parameter manipulation [19]. The aim of such attacks is to replace an attribution map with a spurious map, which highlights another location in the same image, without significantly changing the output of the DNN. Those methods are interested in manipulating the image to cause the neural network's unintended behavior. By contrast, we are interested in finding the proper manipulation of the input image, so the resulting attribution map can cover the target object better.

3. Proposed Method

We look more closely at adversarial attack methods and class activation map in Section 3.1. In Sections 3.2 and 3.3, we introduce AdvCAM and explain how we generate pseudo ground truth for weakly supervised semantic segmentation. Finally, we show how to train a semantic segmentation network with generated pseudo ground-truth in Section 3.4.

3.1. Preliminaries

Adversarial Attack in more detail: An adversarial attack aims to find a small pixel-level perturbation that can change the output from a DNN. In other words, given an input x, it finds the perturbation n satisfying $\texttt{NN}(x) \neq \texttt{NN}(x+n)$, where $\texttt{NN}(\cdot)$ is the output of the neural network. A representative method [16] of constructing n is to consider the normal vector to the decision boundary of $\texttt{NN}$, which can be realized by finding the gradients of $\texttt{NN}$ with respect to x. A manipulated image x' can then be obtained as follows:

$$x' = x - \xi \nabla_x \texttt{NN}(x), \tag{1}$$

where ξ determines the extent of the change to the image. This process can be understood as gradient descent.

Class Activation Map (CAM): It identifies the region of an image which a classifier has used. A CAM is computed from the class-specific contribution of each channel of the feature map to the classification score. It is based on a convolutional neural network that has global average pooling (GAP) before the last classification layer. A class activation map $\texttt{CAM}(x)$ from an image x can be computed as follows:

$$\texttt{CAM}(x) = \mathbf{w}_c^{\intercal} f(x), \tag{2}$$

where $\mathbf{w}_c$ is the weights of the final classification layer for class c, and $f(x)$ is the feature map of x prior to GAP.

A CAM bridges the gap between image-level and pixel-level annotations. However, the regions obtained by a CAM are usually much smaller than the full extent of the target object, since the small discriminative regions provide sufficient information for classification.

3.2. AdvCAM

3.2.1 Adversarial Climbing

AdvCAM is an attribution map obtained through adversarial climbing, which is an anti-adversarial technique that manipulates the image so as to increase the classification score of that image, with the result that the classifier identifies more regions of objects. This is the reverse of an adversarial attack based on Eq. 1, which manipulates the image to reduce the classification score. Inspired by PGD [28], iterative adversarial climbing of the initial image x^0 can be performed using the following relation:

$$x^t = x^{t-1} + \xi \nabla_{x^{t-1}} y_c^{t-1}, \tag{3}$$

where t ($1 \leq t \leq T$) is the adversarial step index, x^t is the manipulated image at the $t-$th step, and y_c^{t-1} is the classification logit of x^{t-1} for class c.

This process makes the previously non-discriminative yet relevant features become more involved in the classification. Thus, the CAMs obtained from successive images manipulated by the iteration can be expected to identify an increasing amount of the region of the target object. We produce a localization map $\mathcal{A}$ which encapsulates the results of the iteration by aggregating the CAMs obtained from the manipulated images at each iteration t, as follows:

$$\mathcal{A} = \frac{\sum_{t=0}^{T} \texttt{CAM}(x^t)}{\max \sum_{t=0}^{T} \texttt{CAM}(x^t)}. \tag{4}$$

3.2.2 How can Adversarial Climbing Improve CAMs?

The connection between a classification logit y_c and a CAM, *i.e.* $y_c = \text{GAP}(\texttt{CAM})$ [63], infers that adversarial climbing increases y_c, and thus the CAM. In this process, features involved in classification are enhanced. To provide a better understanding how adversarial climbing generates a denser CAM, we consider two questions: ① Can non-discriminative features be enhanced? ② Are those enhanced features class-relevant from a human point of view?

① Can non-discriminative features be enhanced?: One might think that changing a pixel with a large gradient primarily enhances discriminative features. This pixel

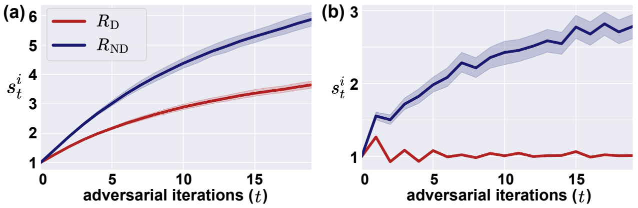

Figure 2: Distributions of the pixel amplification ratio s_t^i for $i \in R_{\mathrm{D}}$ and $i \in R_{\mathrm{ND}}$ for 100 images, (a) without regularization and (b) with regularization.

change affects many features due to the receptive field. However, not all the affected features are necessarily discriminative. We support this analysis empirically. We define the discriminative region $R_{\mathrm{D}} = \{i | \texttt{CAM}(x^0)_i \geq 0.5\}$ and the non-discriminative region $R_{\mathrm{ND}} = \{i | 0.1 < \texttt{CAM}(x^0)_i < 0.5\}$, where i is the location index. The pixel amplification ratio s_t^i is $\texttt{CAM}(x^t)_i / \texttt{CAM}(x^0)_i$ at location i and step t. Figure 2(a) shows that adversarial climbing makes both $s_t^{i \in R_{\mathrm{D}}}$ and $s_t^{i \in R_{\mathrm{ND}}}$ grow, but enhances non-discriminative features more than discriminative ones, resulting in a denser CAM.

② **Are those enhanced features class-relevant from a human point of view?** We now consider whether the highlighted non-discriminative features are class-relevant from a human point of view. Moosavi *et al.* [41] argued that a loss landscape that is sharply curved with respect to input makes a NN vulnerable to adversarial attack. Researchers have subsequently shown that a flattened loss landscape, obtained by reducing the curvature of the loss surface [41] or encouraging the loss to behave linearly [44], can improve the robustness of a NN. Systems which are robust in this sense have been shown to produce features that align better with human perception and operate in a easier way to understand [25, 45, 55].

By the same token, we can expect that images manipulated by adversarial climbing will produce features that align with human perception well because the curvature of loss surface affected by adversarial climbing is small. To support this, we visualize the loss landscape of our trained classifier, following Moosavi *et al.* [41]: we obtain a manipulation vector $\vec{n}$ and a random vector $\vec{r}$ from the classification loss ℓ computed from an image. We determine the surfaces of classification loss values computed from images, manipulated by a vector which is interpolated between $\vec{n}$ and $\vec{r}$ using a range of interpolation ratios. The loss landscape obtained by adversarial climbing (Figure 3(a)) is much more flatten than that obtained by adversarial attacking (Figure 3(b)). Therefore, we can legitimately expect it to increase the attribution of features relevant to the class from a human point of view, resulting in a better CAM.

3.3. Regularization

Even if the loss surface obtained by adversarial climbing is reasonably flat, too much repetitive adversarial manipulation may cause regions corresponding to objects in the wrong class to be activated, or increase the attribution scores of the regions that already have high scores. We address this by (i) suppressing the logit values associated with other classes and (ii) restricting high attributions on discriminative regions of the target object.

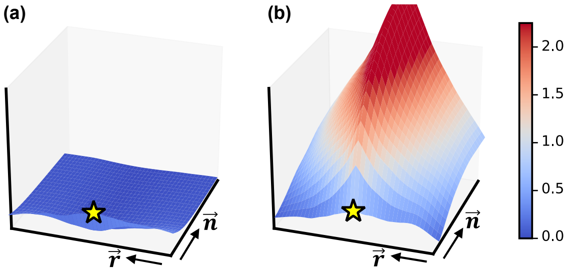

Figure 3: Loss landscapes by manipulating images with weighted sums of the normal vector $\vec{n}$ and a random vector $\vec{r}$ for (a) adversarial climbing and (b) adversarial attack. The yellow star corresponds to the original image.

Suppressing Other Classes: In an image, objects of different classes can mutually increase logit values. For example, since a chair and a dining table mainly occur together in an image, a NN may infer an increased logit value for the chair from the region of the table. We thus add regularization that reduces logit values for all classes except c.

Restricting High Attributions: As mentioned in Section 3.2.2, adversarial climbing increases the attribution scores for both discriminative and non-discriminative regions in the feature map. However, the growth of attribution scores for discriminative regions is problematic for two reasons: 1) it prevents new regions from being additionally attributed to the classification score, and 2) if the maximum value of the attribution score increases during adversarial climbing, the normalized scores of the remaining area may decrease. Please see the blue boxes in Figure 4(b).

Therefore we limit the attribution scores in regions that already have high scores during adversarial climbing, so the attribution scores of those regions remain similar to that of x^0. We realize this scheme by introducing a restricting mask $\mathcal{M}$ that contains the regions whose attribution scores of $\texttt{CAM}(x^{t-1})$ are higher than the threshold τ. More specifically, $\mathcal{M}$ can be represented as follows:

$$\mathcal{M} = \mathbb{1}(\texttt{CAM}(x^{t-1}) > \tau), \tag{5}$$

where $\mathbb{1}(\cdot)$ is an indicator function. An example mask $\mathcal{M}$ is shown in Figure 4(a).

We add the regularization term so that the values of the CAM corresponding to the regions of $\mathcal{M}$ are forced to equal to that of $\texttt{CAM}(x^0)$. With this regularization, $s_t^{i \in R_{\mathrm{D}}}$ remains fairly constant but $s_t^{i \in R_{\mathrm{ND}}}$ still grows during adversarial climbing (Figure 2(b)). Figure 2 shows that, adversarial climbing enhances non-discriminative features more than

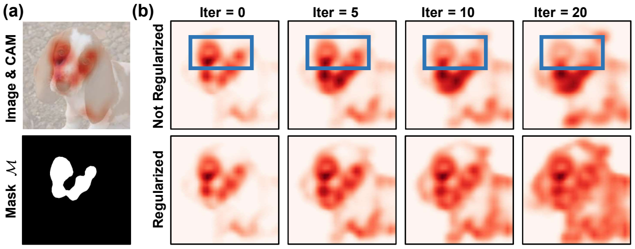

Figure 4: (a) An example image with its CAM and restricting mask $\mathcal{M}$. (b) The initial CAM, and CAMs after 5, 10 and 20 steps of adversarial climbing, with and without regularization.

discriminative features (< 2×), and regularization makes this difference even larger (> 2.5×). Thus, new regions of the target object are found more effectively, resulting in a denser CAM (Figure 4(b)).

To apply regularization, we modify Eq. 3 as follows:

$$x^t = x^{t-1} + \xi \nabla_{x^{t-1}} \mathcal{L}, \quad \text{where} \tag{6}$$

$$\begin{aligned} \mathcal{L} = y_c^{t-1} - \sum_{k \in \mathcal{C} \setminus c} y_k^{t-1} \\ - \lambda \left\| \mathcal{M} \odot \left| \texttt{CAM}(x^{t-1}) - \texttt{CAM}(x^0) \right| \right\|_1 . \end{aligned} \tag{7}$$

$\mathcal{C}$ is the set of all classes, λ is a hyper-parameter that controls the influence of masking regularization, and $\odot$ is element-wise multiplication.

3.4. Training Segmentation Networks

Since CAM is obtained from down-sampled intermediate features produced by the classifier, it localizes the target object coarsely and cannot represent its exact boundary. Many methods of generating an initial seed for weakly supervised semantic segmentation construct a pseudo ground-truth by modifying their initial seeds using existing seed refinement methods [1, 2, 23]. For example, SEAM [56] and Chang *et al.* [6] use PSA [2]; and MBMNet [37] and CONTA [61] use IRN [1]. We also apply the seed refinement method to the coarse map $\mathcal{A}$. For weakly supervised learning, we use the resulting profiles as pseudo ground-truth for training DeepLab-v2, pre-trained on the ImageNet dataset [10]. For semi-supervised learning, we employ CCT [42], which uses IRN [1] to generate pseudo-ground truth masks; we replace these with our masks, constructed as just described.

4. Experiments

4.1. Experimental Setup

Dataset: We conducted experiments on the PASCAL VOC 2012 [12] dataset. The images in this dataset come with masks for fully supervised semantic segmentation, but we only used them for evaluation. In a weakly supervised setting, we trained our network on 10,582 training images provided by Hariharan *et al.* [17], which have image-level annotations. In a semi-supervised setting, we used 1,464 training images with pixel-level annotations and 9,118 training images with class labels, following previous works [29, 38, 42, 58]. We evaluated our results by calculating mean intersection-over-union (mIoU) values for 1,449 validation images and 1,456 test images. Since the labels for test images are not publicly available, the results for those images were obtained from the official PASCAL VOC evaluation server.

Reproducibility: We performed iterative adversarial climbing with $T = 27$ and $\xi = 0.008$. We set λ to 7 and τ to 0.5. To generate the initial seed, we followed the procedure of Ahn *et al.* [1], including the use of ResNet-50 [18]. For final segmentation, we used DeepLab-v2-ResNet101 [8] as the backbone network. We followed the default settings of [8] for training, which included cropping the images to 321×321 pixels. In a semi-supervised setting we used the same settings as Ouali *et al.* [42].

4.2. Experimental Results

Quality of the Mask: Table 1 compares the initial seed and pseudo ground truth masks obtained by our method and by other recent techniques. Both seeds and masks were generated from training images of the PASCAL VOC dataset. For initial seeds, we report the best results by applying a range of thresholds to separate the foreground and background in the map $\mathcal{A}$, as following SEAM [56]. Our initial seeds are 6.8% better than the original CAMs [64], which provide a baseline, and this also outperforms the other methods. Note that Chang *et al.* [6] and SEAM [56] use Wide ResNet-38 [59], which provides better representation than ResNet-50 [18]. SEAM [56] also uses an auxiliary self-attention module that performs pixel-level refinement of the initial CAM by considering the relationship between pixels. We apply CRF, a widely used post-processing method, to the initial seeds of Chang *et al.* [6], SEAM [56], IRN [1], and our method. With the exception of SEAM, CRF improves the seed by more than 5% on average, but it improves the seed of SEAM only by 1.4%. We believe this is because the seed of SEAM is already refined by the self-attention module. Our seed after applying CRF is 5.3% better than that of SEAM.

We also compared pseudo ground truth masks, extracted after seed refinement, with existing methods. Most methods refine their initial seeds with PSA [2] or IRN [1]. For a fair comparison, we produced pseudo ground truth masks using both these seed refinement techniques. Table 1 shows that our method outperforms the others by a large margin, whichever seed refinement technique is used.

Weakly Supervised Semantic Segmentation: Table 2 compares our method with other recently introduced weakly supervised semantic segmentation methods with various levels of supervision: fully supervised pixel-level masks ($\mathcal{P}$), bounding boxes ($\mathcal{B}$) or image class labels ($\mathcal{I}$), with and without salient object masks ($\mathcal{S}$). All the results in Table 2

Table 1: mIoU (%) of the initial seed (Seed), the seed with CRF (+CRF), and the pseudo ground truth mask (Mask) on PASCAL VOC 2012 *train* images.

Method	Seed	+ CRF	Mask
Seed Refine with PSA [2]:			
PSA $_{\text{CVPR '18}}$ [2]	48.0	-	61.0
Mixup-CAM $_{\text{BMVC '20}}$ [5]	50.1	-	61.9
Chang *et al.* $_{\text{CVPR '20}}$ [6]	50.9	55.3	63.4
SEAM $_{\text{CVPR '20}}$ [56]	55.4	56.8	63.6
AdvCAM (Ours)	**55.6**	**62.1**	**68.0**
Seed Refine with IRN [1]:			
IRN $_{\text{CVPR '19}}$ [1]	48.8	54.3	66.3
MBMNet $_{\text{ACMMM '20}}$ [37]	50.2	-	66.8
CONTA $_{\text{NeurIPS '20}}$ [61]	48.8	-	67.9
AdvCAM (Ours)	**55.6**	**62.1**	**69.9**

were obtained using a ResNet-based backbone [18]. With image-level annotation alone, our method achieves mIoU values of 68.1 and 68.0 for the PASCAL VOC 2012 validation and test images respectively. This is significantly better than the other methods under the same level of supervision. In particular, the mIoU value for validation images is 4.6% higher than that for IRN [1], which is our baseline. CONTA [61], the best-performing method among our competitors, achieves an mIoU value of 66.1; but their method depends upon SEAM [56], which is known to outperform IRN [1]. If CONTA is implemented with IRN, the resulting mIoU value is 65.3, which is 2.8% worse than our method. Figure 5 presents examples of semantic masks produced by FickleNet [29], IRN [1], and our method.

Our method also outperforms other methods using auxiliary salient object mask supervision [35, 36] that provides exact boundary information of salient objects in an image, or extra web images or videos [30, 53]. The performance of our method is also comparable with that of methods [26, 31, 51] that use bounding box supervision.

Semi-Supervised Semantic Segmentation: Table 3 compares the mIoU scores of our method on the PASCAL VOC validation and test images with those of other recent semi-supervised segmentation methods, which use 1.5K images with fully supervised masks and 9.1K images with weak annotations. All the methods in Table 3 were implemented on the ResNet-based backbone [18], except that daggered (†) methods which used the VGG-based backbone [49]. We achieve mIoU values of 77.8 and 76.9 for the PASCAL VOC 2012 validation and test images respectively, which is better than the other methods under the same level of supervision. Specifically, the performance of our method on the validation images was 4.6% better than that of CCT [42], which is our baseline. Our method even outperforms Song *et al.* [51] which uses bounding box labels for 9.1K images, instead of class labels. Figure 5 presents examples of semantic masks produced by CCT [42] and our method.

Table 2: Weakly supervised semantic segmentation performance on PASCAL VOC 2012 *val* and *test* images.

Method	Sup.	*val*	*test*
Supervision: Stronger than image labels			
DeepLab $_{\text{TPAMI '17}}$ [8]	$\mathcal{P}$	76.8	76.2
SDI $_{\text{CVPR '17}}$ [26]	$\mathcal{B}$	69.4	-
Song *et al.* $_{\text{CVPR '19}}$ [51]	$\mathcal{B}$	70.2	-
BBAM $_{\text{CVPR '21}}$ [31]	$\mathcal{B}$	73.7	73.7
Supervision: Image-level tags			
Li *et al.* $_{\text{ICCV '19}}$ [34]	$\mathcal{I}, \mathcal{S}$	62.1	63.0
FickleNet $_{\text{CVPR '19}}$ [29]	$\mathcal{I}, \mathcal{S}$	64.9	65.3
Lee *et al.* $_{\text{ICCV '19}}$ [30]	$\mathcal{I}, \mathcal{S}, \mathcal{W}$	66.5	67.4
CIAN $_{\text{AAAI '20}}$ [13]	$\mathcal{I}, \mathcal{S}$	64.3	65.3
Zhang *et al.* $_{\text{ECCV '20}}$ [62]	$\mathcal{I}, \mathcal{S}$	66.6	66.7
Sun *et al.* $_{\text{ECCV '20}}$ [53]	$\mathcal{I}, \mathcal{S}$	66.2	66.9
Fan *et al.* $_{\text{ECCV '20}}$ [14]	$\mathcal{I}, \mathcal{S}$	67.2	66.7
Sun *et al.* $_{\text{ECCV '20}}$ [53]	$\mathcal{I}, \mathcal{S}, \mathcal{W}$	67.7	67.5
IRN $_{\text{CVPR '19}}$ [1]	$\mathcal{I}$	63.5	64.8
SSDD $_{\text{ICCV '19}}$ [48]	$\mathcal{I}$	64.9	65.5
SEAM $_{\text{CVPR '20}}$ [56]	$\mathcal{I}$	64.5	65.7
Chen *et al.* $_{\text{ECCV '20}}$ [7]	$\mathcal{I}$	65.7	66.6
Chang *et al.* $_{\text{CVPR '20}}$ [6]	$\mathcal{I}$	66.1	65.9
CONTA $_{\text{NeurIPS '20}}$ [61]	$\mathcal{I}$	66.1	66.7
AdvCAM (Ours)	$\mathcal{I}$	**68.1**	**68.0**

$\mathcal{P}$—pixel-level mask, $\mathcal{I}$—image class, $\mathcal{B}$—box, $\mathcal{S}$—saliency, $\mathcal{W}$—web

Table 3: Comparison of semi-supervised semantic segmentation methods on the PASCAL VOC 2012 *val* and *test* images.

Method	Training set	*val*	*test*
WSSL† [43]	1.5K $\mathcal{P}$ + 9.1K $\mathcal{I}$	64.6	66.2
MDC† [58]	1.5K $\mathcal{P}$ + 9.1K $\mathcal{I}$	65.7	67.6
Souly *et al.*† [52]	1.5K $\mathcal{P}$ + 9.1K $\mathcal{I}$	65.8	-
FickleNet† [29]	1.5K $\mathcal{P}$ + 9.1K $\mathcal{I}$	65.8	-
Song *et al.* [51]	1.5K $\mathcal{P}$ + 9.1K $\mathcal{B}$	71.6	-
Luo *et al.* [38]	1.5K $\mathcal{P}$ + 9.1K $\mathcal{I}$	76.6	-
CCT [42] (baseline)	1.5K $\mathcal{P}$ + 9.1K $\mathcal{I}$	73.2	-
AdvCAM (Ours)	1.5K $\mathcal{P}$ + 9.1K $\mathcal{I}$	**77.8**	**76.9**

$\mathcal{P}$—pixel-level mask, $\mathcal{I}$—image class label, $\mathcal{B}$—box, †— VGG backbone

5. Discussion

5.1. Iterative Adversarial Climbing

We analyzed the effectiveness of the iterative adversarial climbing and regularization technique introduced in Section 3.3 by evaluating the initial seed in terms of mIoU. Figure 6(a) shows the mIoU of the initial seed for each adversarial iteration. Initially, the mIoU rises steeply, with or without regularization; but without regularization the curves peaks around iteration 8.

To analyze this, we evaluate the truthfulness of the newly localized region at each adversarial climbing iteration in terms of the proportion of noise, which we define to be the proportion of pixels that are classified as foreground but are

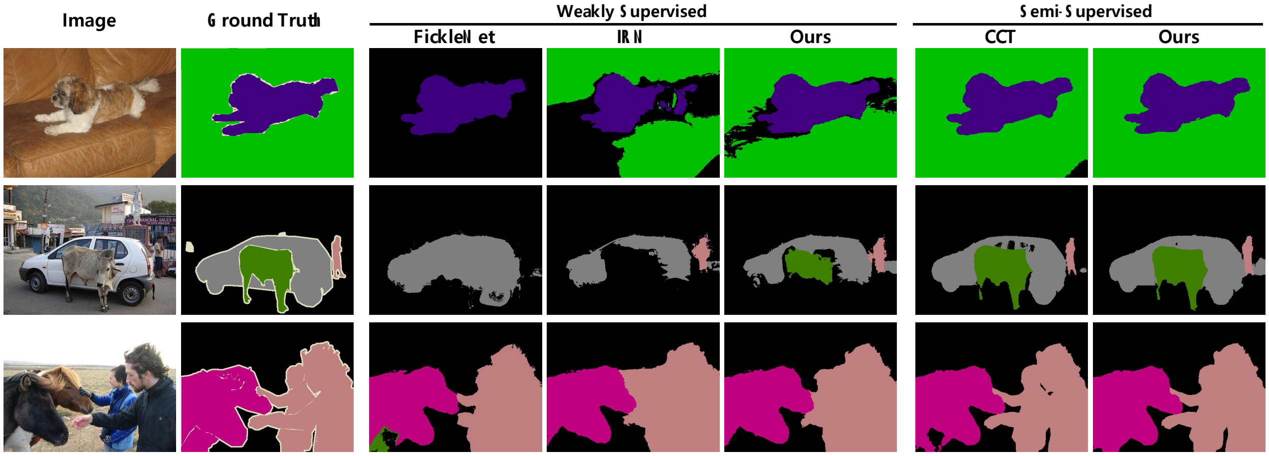

Figure 5: Examples of predicted semantic masks for PASCAL VOC *val* images in weakly and semi-supervised manner.

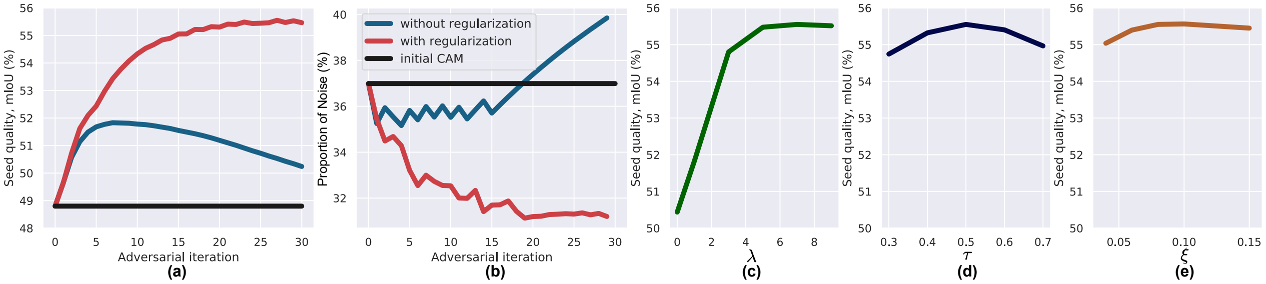

Figure 6: Effect of adversarial climbing and regularization on (a) the seed quality and (b) the proportion of noise. (c) Effect of the regularization coefficient λ. (d) Effect of the masking threshold τ. (d) Effect of the step size ξ.

Table 4: Effects of AdvCAM on different methods of generating the initial seed: mIoU of the initial seed (Seed) and of the pseudo ground truth mask (Mask), for the PASCAL VOC 2012 training images.

Method	Seed	Mask
Chang *et al.* [6]	50.9	63.4
+ AdvCAM	53.7 +2.8	67.5 +4.1
SEAM [56]	55.4	63.6
+ AdvCAM	58.6 +3.2	67.2 +3.6
IRN [2]	48.8	66.3
+ AdvCAM	55.6 +6.8	69.9 +3.6

actually background. Without regularization, the proportion of noise rises steeply after some iterations as shown in Figure 6(b), which means that new regions tend to be in the regions of background. Regularization allows new regions of the target object to be found in as many as 30 adversarial steps, keeping the proportion of noise much lower than that of initial CAM. Figure 7 shows examples of attribution maps at each adversarial iteration with and without regularization.

5.2. Hyper-Parameter Analysis

In the previous section, we looked at the effect of the number of adversarial iterations (Figures 6(a) and (b)). We also analyzed the sensitivity of the mIoU of the initial seed to the other three hyper-parameters used by AdvCAM.

Regularization Coefficient $\boldsymbol{\lambda}$: It controls the influence of the masking technique that limits the attribution scores of the regions that already have high scores during adversarial climbing, in Eq. 7. Figure 6(c) shows the mIoU of the initial seed for different values of λ. When $\lambda = 0$, there is no regularization. Masking technique improves performance by more than 5% (50.43 for $\lambda = 0$ *vs.* 55.55 for $\lambda = 7$). The flattening of the curve after $\lambda = 5$ suggests that it is not difficult to select a good value of λ.

Masking Threshold $\boldsymbol{\tau}$: It controls the size of the restricting mask $\mathcal{M}$ in Eq. 5, determining how many pixels' attribution values will remain similar to that of the original CAM during adversarial climbing. Figure 6(d) shows the mIoU of the initial seed for different values of τ. This parameter is even less sensitive than λ: varying τ between 0.3 and 0.7 produces less than 1% change in mIoU.

Step Size $\boldsymbol{\xi}$: It determines the extent of the manipulation to the image in Eq. 6. Figure 6(e) shows the mIoU of the initial seed for different values of ξ. In our system, changes in step size ξ are not particularly significant.

5.3. Generality of Our Method

In addition to IRN [1], we experimented with two state-of-the-art methods of generating an initial seed for weakly supervised semantic segmentation, namely Chang *et al.* [6]

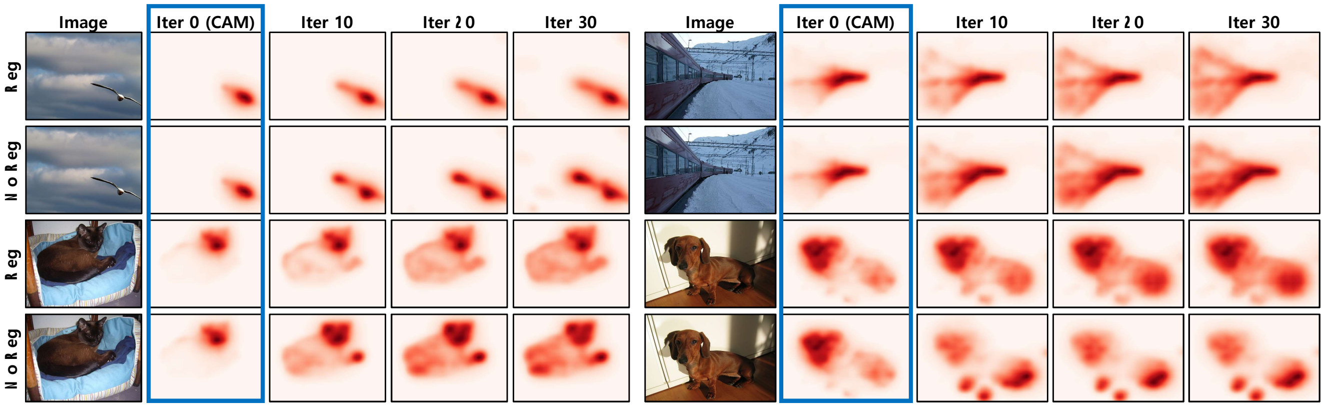

Figure 7: Examples of initial CAMs (the blue boxes) and successive localization maps obtained from images manipulated by iterative adversarial climbing, with the regularization procedure (*top*) and without (*bottom*).

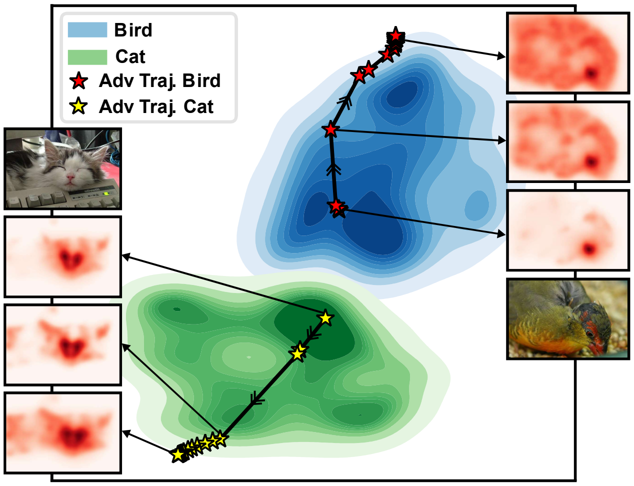

Figure 8: Feature manifold of images with “bird" (blue) and “cat" (green), and a trajectory of adversarial climbing for an image of each class. The dimensionality of the feature was reduced by t-SNE [39].

and SEAM [56]. We used the authors’ pre-trained classifier where possible, but we re-trained the classifier of IRN [1] since the authors do not provide pre-trained one. We also followed their experimental settings including the backbone networks and mask refinement methods, *i.e.,* we used PSA [2] to refine the initial seed from “Chang *et al.* + AdvCAM" or “SEAM + AdvCAM". Table 4 gives mIoU values for the initial seed and the pseudo ground truth mask obtained by combining each method with adversarial climbing. The use of AdvCAM improves the quality of the initial seed by an average of over 4%. Our approach does not require those initial seed generators to be modified or retrained.

5.4. Manifold Visualization

For visualizing a trajectory of adversarial climbing at a feature-level, we used t-SNE dimensional reduction [39]. We collect images that contain a single class of a cat or a bird and that are predicted by the classifier correctly. We then construct a set $\mathcal{F}$ containing the features of those images, before the final classification layer. We also choose a representative image of a cat, and another of a bird, and construct a set $\mathcal{F}'$ containing the features of those two images and their 20 manipulated images by adversarial climbing. Figure 8 presents t-SNE visualization of features in $\mathcal{F} \cup \mathcal{F}'$. We can see that adversarial climbing actually pushes the features away from the decision boundary boundary that separates the blue and green areas. In addition, despite 20 adversarial climbing steps, the manipulated features did not deviate significantly from the feature manifold of each class.

6. Conclusion

We have shown how adversarial manipulation can be used to expand the small discriminative regions of a target object, so as to obtain a better localization of that object. We manipulate images with a pixel-level perturbation, which is obtained from the gradient computed from the output of classifier with respect to the input image, which increase the classification score of the perturbed image. The attribution map of the manipulated image covers more of the target object. This is a post-hoc analysis of a trained classifier, and therefore no modification or re-training of the classifier is required. This allows AdvCAM to be readily integrated into existing methods. We have shown that AdvCAM can indeed be combined with recent weakly supervised semantic segmentation networks, and achieved new state-of-the-art performance on both weakly and semi-supervised semantic segmentation.

Acknowledgements: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) [2018R1A2B3001628], AIR Lab (AI Research Lab) in Hyundai & Kia Motor Company through HKMC-SNU AI Consortium Fund, and the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2021.

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *CVPR*, 2019.

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

[3] Anurag Arnab, Ondrej Miksik, and Philip HS Torr. On the robustness of semantic segmentation models to adversarial attacks. In *CVPR*, 2018.

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

[5] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Mixup-cam: Weakly-supervised semantic segmentation via uncertainty regularization. In *BMVC*, 2020.

[6] 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 *CVPR*, 2020.

[7] Liyi Chen, Weiwei Wu, Chenchen Fu, Xiao Han, and Yuntao Zhang. Weakly supervised semantic segmentation with boundary exploration. In *ECCV*, 2020.

[8] 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 TPAMI*, 2017.

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[11] Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In *NeurIPS*, 2019.

[12] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

[13] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Cian: Cross-image affinity net for weakly supervised semantic segmentation. *AAAI*, 2020.

[14] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Employing multi-estimations for weakly-supervised semantic segmentation. In *ECCV*, 2020.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

[16] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

[17] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *ICCV*, 2011.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[19] Juyeon Heo, Sunghwan Joo, and Taesup Moon. Fooling neural network interpretations via adversarial model manipulation. In *NeurIPS*, 2019.

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

[21] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In *NeurIPS*, 2018.

[22] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, 2019.

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

[24] Wei Chih Hung, Yi Hsuan Tsai, Yan Ting Liou, Yen Yu Lin, and Ming Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. In *BMVC*, 2018.

[25] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *NeurIPS*, 2019.

[26] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *CVPR*, 2017.

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

[28] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *ICLR*, 2017.

[29] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *CVPR*, 2019.

[30] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Frame-to-frame aggregation of active regions in web videos for weakly supervised semantic segmentation. In *ICCV*, 2019.

[31] Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon. Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. *arXiv preprint arXiv:2103.08907*, 2021.

[32] Sungmin Lee, Jangho Lee, Jungbeom Lee, Chul-Kee Park, and Sungroh Yoon. Robust tumor localization with pyramid grad-cam. *arXiv preprint arXiv:1805.11393*, 2018.

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

[34] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Attention bridging network for knowledge transfer. In *ICCV*, 2019.

[35] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *CVPR*, 2014.

[36] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *TPAMI*, 2010.

[37] Weide Liu, Chi Zhang, Guosheng Lin, Tzu-Yi HUNG, and Chunyan Miao. Weakly supervised segmentation with maximum bipartite graph matching. In *ACMMM*, 2020.

[38] Wenfeng Luo and Meng Yang. Semi-supervised semantic segmentation via strong-weak dual-branch network. In *ECCV*, 2020.

[39] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 2008.

[40] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*, 2016.

[41] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa. In *CVPR*, 2019.

[42] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *CVPR*, 2020.

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

[44] Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and Pushmeet Kohli. Adversarial robustness through local linearization. In *NeurIPS*, 2019.

[45] Shibani Santurkar, Andrew Ilyas, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Image synthesis with a single (robust) classifier. In *NeurIPS*, 2019.

[46] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

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

[48] Wataru Shimoda and Keiji Yanai. Self-supervised difference detection for weakly-supervised semantic segmentation. In *ICCV*, 2019.

[49] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[50] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *ICCV*, 2017.

[51] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In *CVPR*, 2019.

[52] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *ICCV*, 2017.

[53] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In *ECCV*, 2020.

[54] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised cnn segmentation. In *CVPR*, 2018.

[55] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *ICLR*, 2019.

[56] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *CVPR*, 2020.

[57] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *CVPR*, 2017.

[58] 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 *CVPR*, 2018.

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

[60] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *ICCV*, 2017.

[61] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xiansheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. In *NeurIPS*, 2020.

[62] Tianyi Zhang, Guosheng Lin, Weide Liu, Jianfei Cai, and Alex Kot. Splitting vs. merging: Mining object regions with discrepancy and intersection loss for weakly supervised semantic segmentation. In *ECCV*, 2020.

[63] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *CVPR*, 2018.

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