Adversarial Erasing Framework via Triplet with Gated Pyramid Pooling Layer for Weakly Supervised Semantic Segmentation

Sung-Hoon Yoon*, Hyeokjun Kweon*, Jegyeong Cho, Shinjeong Kim, and Kuk-Jin Yoon

Korea Advanced Institute of Science and Technology
{yoon307,0327june,j2k0618,aakseen,kjyoon}@kaist.ac.kr

Abstract. Weakly supervised semantic segmentation (WSSS) has employed Class Activation Maps (CAMs) to localize the objects. However, the CAMs typically do not fit along the object boundaries and highlight only the most-discriminative regions. To resolve the problems, we propose a Gated Pyramid Pooling (GPP) layer which is a substitute for a Global Average Pooling (GAP) layer, and an Adversarial Erasing Framework via Triplet (AEFT). In the GPP layer, a feature pyramid is obtained by pooling the CAMs at multiple spatial resolutions, and then be aggregated into an attention for class prediction by gated convolution. With the process, CAMs are trained not only to capture the global context but also to preserve fine-details from the image. Meanwhile, the AEFT targets an over-expansion, a chronic problem of Adversarial Erasing (AE). Although AE methods expand CAMs by erasing the discriminative regions, they usually suffer from the over-expansion due to an absence of guidelines on when to stop erasing. We experimentally verify that the over-expansion is due to rigid classification, and metric learning can be a flexible remedy for it. AEFT is devised to learn the concept of erasing with the triplet loss between the input image, erased image, and negatively sampled image. With the GPP and AEFT, we achieve new state-of-the-art both on the PASCAL VOC 2012 *val/test* and MS-COCO 2014 *val* set by 70.9%/71.7% and 44.8% in mIoU, respectively.

Keywords: Weakly supervised semantic segmentation

1 Introduction

Recently, semantic segmentation based on Deep Learning (DL) has been widely used in various applications such as autonomous driving and medical imaging. However, since the semantic segmentation model requires pixel-level labels, a considerable amount of cost and time is consumed to generate labels. To reduce this burden and make DL-based semantic segmentation more practically applicable in general tasks, many Weakly Supervised Semantic Segmentation (WSSS) studies that utilize only weak supervision such as image-level

* Equal contribution.

labels [2,50,56,15,1,7,30,27,58,35], scribble [36,48], bounding boxes [13,24,40,31], and points [4] have been proposed. In this work, we focus on WSSS with image-level labels, an especially challenging task among weakly-supervised ones.

To learn semantic segmentation with image-level labels only, most existing WSSS approaches follow the steps: (1) localize the objects through Class Activation Maps(CAMs) [61], (2) refine the CAMs and generate pseudo-labels in a pixel-level, and (3) train the semantic segmentation model with the pseudo-labels. Although the CAMs can localize the objects to some extent, they are not precise at an object boundary and only highlight the most discriminative pattern.

As far as we know, most CNN-based classifiers in WSSS employ a Global Average Pooling (GAP) layer to aggregate the feature map and predict the existing classes in the image. However, since the GAP layer *averages* all the features, even including ones from object-irrelevant regions, CAMs usually ignore small segments and do not fit with the object boundary (*i.e.* impreciseness). To overcome this innate limitation of the GAP layer, BES [50] and Araslanov *et al.* [3] utilize softmaxed-CAMs as a pooling weight while making the class prediction. Instead, in this paper, we propose a Gated Pyramid Pooling (GPP) layer that not only captures the global context but also localizes fine-details. In the proposed GPP layer, CAMs are average-pooled with various bin sizes (*e.g.* 8×8 or 16×16) and form a spatial pyramid. Then, this pyramid of the pooled features is aggregated sequentially through a gating mechanism [47] in a coarse-to-fine manner for better localization. The final output of the aggregation is used as a pixel-level weight that decides to either encourage or discourage the contribution of CAMs for predicting the image-level class. Here, by building the pyramid features with different spatial resolutions and using them as weights to generate the class prediction, CAMs are trained to capture not only global context (from low-scale bins) but also localize fine-details (from high-scale bins).

In addition to the GPP layer that effectively resolves the impreciseness problem in CAMs, we utilize the concept of Adversarial Erasing (AE) method to further guide CAMs to be activated even on the less-discriminative regions. AE methods [51,59,19,33,27], one of the most actively studied strategy in WSSS, extend CAMs to whole object regions by erasing the most-discriminative regions of an image or intermediate feature. However, since there is no explicit guidance regarding when to stop the erasing, CAMs generated from the AE approach usually suffer from an over-expansion [27]. To benefit from AE methods while preventing the over-expansion problem, we propose an Adversarial Erasing Framework via Triplet (AEFT) that reformulates the AE methods as triplet learning with the GPP feature. Here, we experimentally verify that imposing relatively rigid supervision (*e.g.* classification loss) on the AE framework leads the resulting CAMs to suffer from the over-expansion problem. Since we allow the framework to adjust its features according to the distance between them, this approach can be regarded as a more softened version compared to prior studies using rigid supervision [51,59,33,27]. For triplet learning, as shown in Fig 1, we define the original image as an *anchor* image. After masking the high-confidence

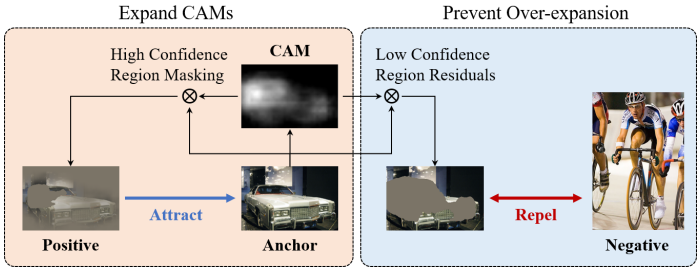


Fig. 1. Brief illustration of the proposed Adversarial Erasing Framework via Triplet (AEFT). An anchor image and a positive image denote the original image and the masked image according to CAMs, respectively. A negative image is sampled to do not have overlapping classes with the anchor image. In the feature space, we train the model to locate the positive image close to the anchor image while increasing the distance between the anchor image and the negative image.

regions of the CAMs from the anchor image, the remained image is regarded as a *positive* image. Finally, the other image, which has no overlapping classes with the anchor image, is used as a *negative* image. In AEFT, we minimize the distance between the anchor and positive in the GPP feature space (*i.e. Attract*) while maximizing the distance between the anchor and negative feature (*i.e. Repel*). While *Attract* guides the CAMs to explore less-discriminative region, *Repel* prevents the over-expansion problem. Since the distance between the anchor and negative is often already far enough, we intentionally exclude the high-confidence region from the anchor to impose a harder but helpful constraint for AEFT.

In summary, we propose (1) the Gated Pyramid Pooling (GPP) layer to resolve the architectural limitation of classifier (or GAP) and (2) the Adversarial Erasing Framework via Triplet (AEFT) to effectively prevent the over-expansion via triplet, while preserving the benefits of AE (expanding the CAMs to less-discriminative regions). With the proposed GPP and AEFT, we achieve state-of-the-art WSSS performance by a large gap on both the PASCAL VOC 2012 and MS-COCO 2014 sets, using the image-level labels only.

2 Related Works

Earlier works in WSSS WSSS with image-level labels generally utilizes CAMs to localize target objects on the images. However, as CAMs tend to only focus on the most discriminative parts and do not fit along the object boundary, subsequent works in WSSS have tried to generate high-quality pseudo-labels from the CAMs for training semantic segmentation. Many studies proposed to refine the CAMs using pixel-level affinity [2,16,43] or region growing [20,25]. Much research targeted to enhance the quality of localization of CAMs by using stochastic feature selection [29], attention map accumulation [22], and scale-invariance [50]. Also, lots of methods employ additional constrains such as sub-categorical classification [7], co-attention constraints [45,34], and complementary

patch loss [58]. Several studies [45,15,34,55,54,32] employ saliency map to indicate dominant foreground objects distinguished from the background. Despite the efficacy, neither saliency module nor an external dataset was adopted in the proposed method, in line with the objective of WSSS learning from only image-level labels. Like the proposed method, BES [10] and Araslanov *et al.* [3] utilize CAMs as weights for pooling when making class predictions. However, while both methods used softmaxed-CAMs as weights for pooling, the improvement of BES is marginal (within 1%), and Araslanov *et al.* [3] requires to define background constant. Unlike the previous methods, the proposed method not only preserves global context but also captures fine-details through sign-preserving gated convolution and pyramid pooling.

Adversarial erasing Adversarial Erasing (AE) [51,59,19,33,27] is widely used strategy in WSSS. By erasing the most discriminative region from the image, the AE method promotes the network to expand its CAMs to the less discriminative object region. The first AE method is proposed by Wei *et al.* [51], which is a recursive find-and-erase scheme. Zhang *et al.* [59] proposes an end-to-end feature-level erasing framework with complementary branches. However, if the initial classifier succeeded in completely erasing the object, the complementary classifier would suffer from the over-erasing. SeeNet [19] suggested using the ternary thresholding method for mask generation process to relieve the over-erasing, but it requires a pre-trained saliency detection module. In recent, GAIN [33] and OC-CSE [27] proposed soft erasing approaches that generate learnable masks. In these methods, the CAMs generation network is jointly trained by a classification loss and auxiliary loss regarding the adversarial erasing process. GAIN propose an attention mining loss to assure that the erased image does not contain any objects. However, this self-guidance makes the framework difficult to self-correct the over-expansion. In OC-CSE, the CAM of only one class is selected for erasing, and then the guidance from the pre-trained ordinary classifier is used to prevent the erasing network from erasing the objects of the not selected classes. But the usage of the pre-trained classifier limits the performance of this method. All of the aforementioned AE methods are based on imposing a classification loss on the erased image. In our view, forcing the network to make the prediction from the erased image according to the binary classification label (exist or not) is the main reason for over-expansion. Instead of this "rigid" constrain, we aim to let the network understand the concept of *erasing* in the form of triplet learning. This is a more softened approach compared to prior AE-based studies while not harming the benefit of AE methods.

Deep Metric Learning Deep metric learning has widely been used for resolving various computer vision tasks [38,11,18,12,39,21]. Generally, it aims to learn a metric that measures the semantic distance between instances. As a metric function, the embedding function is trained to map an instance to be close to the similar inputs than the dissimilar inputs. Contrastive loss [17] directly optimizes this goal by decreasing the distance between semantically close instances while increasing the distance between dissimilar instances. On the other hand, triplet loss [42] takes three inputs at once: anchor instance, positive instance,

and negative instance. Then, the loss minimizes the distance between the anchor instance and the positive object while it maximizes the distance between the anchor instance and the negative instance. For semantic segmentation, deep metric learning is used to improve performance in supervised learning [49] or to overcome the lack of data in challenging cases such as weak supervision [23] and open-world scenario [6]. Though we borrow the concept of the triplet loss, as far as we know, our method is the first AE-based method that incorporates metric learning in WSSS.

3 Proposed method

3.1 Overview

In this paper, we propose a Gated Pyramid Pooling (GPP) layer, which is a simple but effective replacement of the Global Average Pooling (GAP) layer widely used in WSSS. To fully utilize the outperforming localization ability of the GPP layer, we also devise a novel Adversarial Erasing Framework via Triplet (AEFT). The proposed framework mainly focuses on training the network to find less-discriminative regions while relieving the over-expansion problem. Note that our method only utilizes image-level labels.

3.2 CAMs Generation

Before discussing our main approaches, we briefly introduce the general process of generating CAMs. Let $f \in \mathbb{R}^{K \times h \times w}$ denote a feature map of the last convolution layer of the classifier, where K is the number of classes and h, w represent the spatial dimensions of the feature map, respectively. Then, an image-level class prediction p can be acquired by applying Global Average Pooling (GAP) on the feature map as $p = \sigma\left(\frac{1}{hw}\sum_{i,j} f(i,j)\right)$, where $f(i,j)$ denotes the feature vector at a location (i,j) and σ is a sigmoid function. By taking the Rectified Linear Unit (ReLU) to the feature map and normalizing it between 0 and 1 for each class, an activation map of k^{th} class (A^k) is generated as follows:

$$A^k \leftarrow \frac{ReLU(f^k)}{max(ReLU(f^k))}. \tag{1}$$

Here, we also apply bilinear upsampling on the CAMs to fit the spatial dimension of them with the input image.

Considering the formulation for the class prediction p, the GAP makes the feature contribute *equally* irrelevant to their location. As claimed by several works [3,10], the GAP increases a dependency on the context and misleads the classifier to learn erroneous correlations between image pixels and image-level class labels. Therefore, the resulting CAMs tend to be activated on highly-correlated background regions (*e.g.* railroad of *train* class, water of *boat* class) while ignoring the small objects. Since generating fine pixel-level pseudo-labels is crucial for WSSS, this is a critical disadvantage.

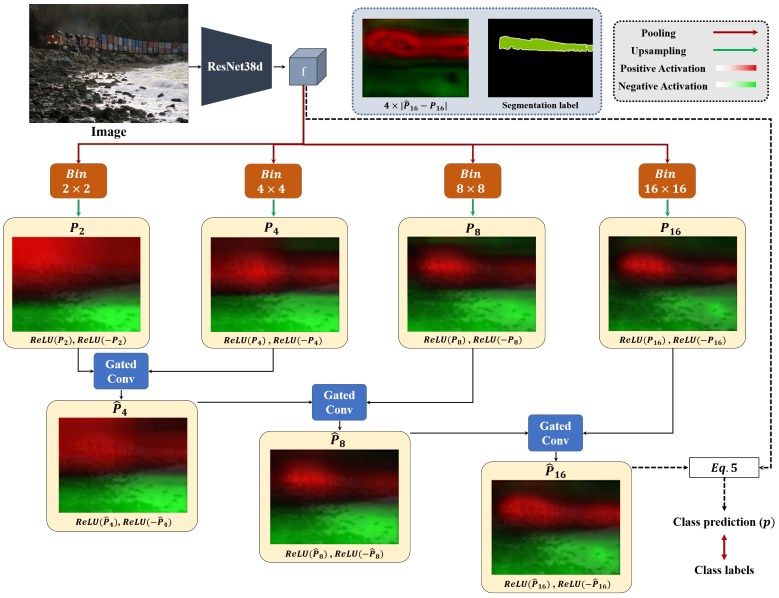


Fig. 2. Overview of the proposed Gated Pyramid Pooling (GPP). By applying pooling with different bin sizes to feature map f, we generate feature pyramid (P_2, P_4, P_8, P_{16}). Each component of the feature pyramid is interpolated to feature map size to apply gated convolution. Along with the feature pyramid, Gated features $(\hat{P}_4, \hat{P}_8, \hat{P}_{16})$ are visualized. *Red* and *green* color represent positive and negative activation, respectively. For simplicity only *train* class is visualized.

3.3 Gated Pyramid Pooling (GPP) Layer

To dispel the aforementioned problems of the GAP, we propose a Gated Pyramid Pooling (GPP) layer, which is a spatial-aware pooling method specialized for generating precise CAMs. Inspired by BES [10] and Araslanov *et al.* [3], we set a different pooling weight for each feature. Our main idea is that the weighting factor should be acquired multi-scale, which is crucial to handle the impreciseness problem of CAMs. Unlike the prior studies applying simple single-scale operations like softmax [3,10] on the CAMs, we pool the CAMs at multi-scale and build a feature pyramid. Then, as shown in Fig. 2, we employ a gating mechanism to aggregate the feature pyramid into a single multi-scale-aware prediction. From the low to the high scale, we sequentially refine the pooled feature map with multiple gated convolutional layers while preserving its sign, inspired by Takikawa *et al.* [47]. We experimentally verify that the proposed gated coarse-to-fine strategy outperforms naive averaging or scale-agnostic fusion.

In addition, we define sign-preserving attention operation $\mathcal{G}$ to deal with the nature of multi-label classification. Compared to the positive prediction which means "the existence of the class", the negative prediction is equally important for a model to decide the "non-existence of the class". Therefore, we devise

GPP to amplify the feature in both positive and negative directions. It can be formulated by taking ReLU and concatenating two different features with 3×3 convolution layer $Conv_{3\times3}$. The process can be defined as follows:

$$\mathcal{G}(x, y) = \sigma(Conv_{3\times3}(ReLU(x)||ReLU(y))). \tag{2}$$

Let P_γ denote the averaged result of feature map f with $\gamma \times \gamma$ pooling. By applying pooling with different sizes ($\gamma \in \{2, 4, 8, 16\}$) to f, we generate a feature pyramid (P_2, P_4, P_8, P_{16}) as in PSPNet [60]. Each component of the feature pyramid is upsampled to feature map f resolution. Then we obtain attention for positive (α) and negative (β) (where $\alpha, \beta \in \mathbb{R}^{2\times h\times w}$) maps as follows:

$$\alpha_n = \mathcal{G}_\alpha(\hat{P}_{2^n}, P_{2^{n+1}}), \quad \beta_n = \mathcal{G}_\beta(-\hat{P}_{2^n}, -P_{2^{n+1}}), \tag{3}$$

where σ is a sigmoid function and $n \in \{1, 2, 3\}$. And the gated feature $\hat{P}_{2^{n+1}}$ can be obtained as follows:

$$\begin{aligned}\hat{P}_{2^{n+1}} = &\left(ReLU(\hat{P}_{2^n}) \odot \alpha_{n,1} + ReLU(P_{2^{n+1}}) \odot \alpha_{n,2}\right)/2 \\ &- \left(ReLU(-\hat{P}_{2^n}) \odot \beta_{n,1} + ReLU(-P_{2^{n+1}}) \odot \beta_{n,2}\right)/2,\end{aligned} \tag{4}$$

where $\odot$ indicates element-wise product between tensors. Here, $\alpha_{n,1}$ and $\alpha_{n,2}$ are first and second channel of α_n, respectively. And $\hat{P}_{2^n}$ equals P_{2^n} only when $n = 1$. The final output of the Gated Pyramid Pooling (GPP) is $\hat{P}_{16} \in \mathbb{R}^{K\times h\times w}$. Here, K denotes the total number of classes.

$$\begin{aligned}p = \sigma(\frac{1}{hw}\sum\{&ReLU(\hat{P}_{16}) \odot ReLU(f)\} \\ &-\{ReLU(-\hat{P}_{16}) \odot ReLU(-f)\}),\end{aligned} \tag{5}$$

By decoupling the feature f with the feature pyramid and aggregating it with the proposed Gated Pyramid Pooling (GPP), regions that are not related to objects are penalized while regions that are highly related are encouraged. Since GPP aggregates features from coarse level (*i.e.* small bin size pooling) to fine level (*i.e.* large bin size pooling) thoroughly, the generated CAMs not only fit along the object boundary but also localize the whole object region. A more detailed ablation study regarding GPP will be discussed in Section 4.

3.4 Adversarial Erasing Framework via Triplet (AEFT) for WSSS

The proposed GPP layer enables the model to generate CAMs with a higher localization quality than the GAP layer. However, this architectural improvement is still insufficient to acquire dense pseudo-labels for semantic segmentation. Additional guidance is required to make the CAMs cover the less-discriminative regions, which are difficult to be activated by a mere classification task.

In the field of WSSS, an Adversarial Erasing (AE) is one of the most widely used approaches to mitigate this problem. For AE, the most-discriminative regions of the CAMs are intentionally erased from the image. Then, the model

is trained again to classify the erased image according to the original image-level classification labels. Continuously iterating this process make the model focus more on the less-discriminative regions, which were originally ignored, and thereby the resulting CAMs are also expanded. Because of its clear and intuitive strategy, plenty of WSSS studies [59,19,51,33,27] has been conducted based on the AE method. However, due to the lack of supervision for when to stop expanding, CAMs generated from the AE approach usually suffer from an over-expansion problem. To relieve such an over-expansion problem, a method using guidance from a pre-trained model [27] has been proposed recently. However, updating the guidance makes the training process unstable, and therefore the method has a limited performance due to the fixed classifier. As aforesaid, there are two main obstacles for the AE method to gently expand its CAMs while rejecting the undesired derailment. First, though the image-level classification labels are valuable supervision in WSSS, such supervision is often *too rigid* to follow and usually leads to the over-expansion problem when classifying the masked image/feature. Second, direct guidance from the AE branch to the CAMs makes the model be unstable in terms of the quality of the generated CAMs.

In this paper, we aim to train the model to understand the concept of *erasing* in a more flexible manner. To achieve this goal, we propose a novel Adversarial Erasing framework via Triplet (AEFT), the modified AE framework using a triplet loss between the images. In our framework, we make the representation of the masked image I_P embedded by the model to be close to that of the original input image I_A. To prevent the over-expansion, we maximize the feature-level distance between the I_A and the negative image I_N, an image that does not share any class with I_A. In other words, for the original image I_A, the masked image I_P and the negative image I_N are regarded as a positive sample and a negative sample, respectively. To avoid direct guidance to the CAMs, we utilize the feature space of the GPP layer as embedding space. Compared to using the rigid classification based on the binary label for the masked image in conventional works, the proposed metric-based approach helps the model flexibly adjust the distance between the features and the decision boundary.

Acquiring Masked Image To erase the highly activated regions from input image I_A and obtain corresponding masked image I_P, we generate a foreground map A^{fg} from the CAMs of I_A as follows:

$$A^{fg}(i,j) = \max\{A^k(i,j) : k = 1, \ldots, K\}, \tag{6}$$

where A^k is an activation map of k^{th} class and (i,j) denotes the pixel position. Then, according to the foreground map, we acquire the masked image I_P as follows:

$$I_P(i,j) = \begin{cases} 0, & \text{if } A^{fg}(i,j) \geq t_H \\ A^{fg}(i,j) I_A(i,j), & \text{otherwise.} \end{cases} \tag{7}$$

Note that we combine the hard-masking [51,59,19] and soft-masking [33,27] according to the threshold (which is denoted as t_H in Eq. 7). Since the regions with already higher activation are not the main target of learning in the proposed framework, we empirically find that this strategy is valid.

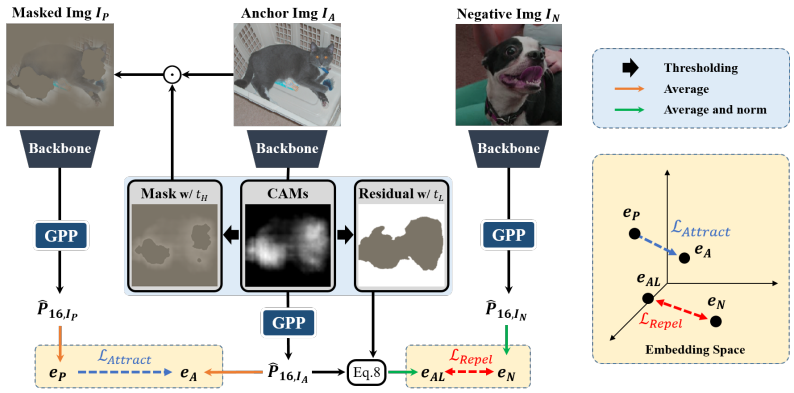


Fig. 3. Overview of the proposed Adversarial Erasing Framework via Triplet (AEFT). The weight of the networks is shared. In AEFT, we merged anchor image I_A, masked image I_P, and negative image I_N as a triplet set. We extract embedding from each member of the triplet set and impose loss relations based on their traits.

Adversarial Erasing via Metric Learning As depicted in Fig. 3, each image should be represented as a feature vector in the embedding space for metric learning. We map the anchor (input) image I_A and the negative image I_N to the anchor embedding $e_A = \frac{1}{hw}\sum \hat{P}_{16,I_A}$ and the negative embedding $e_N = \frac{1}{hw}\sum \hat{P}_{16,I_N}$, respectively. Likewise, embedding for the masked image e_P is obtained in the same manner. By minimizing the distance between the anchor embedding and positive embedding, the model focuses more on the less-discriminative regions. However, though we soften the loss as a form of metric learning, it is still true that minimizing the distance between two embeddings pushes the model to keep exploring the regions even after the complete erasing. Therefore, in AEFT, we devise another constraint to inhibit the over-expansion using negative image I_N. In specific, we intentionally aggregate the features on low confidence regions according to CAMs to acquire 'the embedding for low confidence region of the anchor image' (e_{AL}). Then, we maximize the distance between e_{AL} and negative embedding e_N. Once the CAMs are over-expanded, the embedding from the low confidence region would include less information regarding the objects in the image. Then it would be difficult for the networks to separate such less-information embedding with the negative embedding. Therefore, intuitively, the expansion of CAMs is suppressed while maximizing the distance between e_{AL} and the negative embedding. The embedding for the low confidence region of the anchor image (e_{AL}) can be acquired as follows:

$$e^k_{AL} = \frac{1}{N^k}\sum_{(i,j)} \mathbb{1}(A^k(i,j) < t_L) \cdot \hat{P}_{16,I_A}(i,j), \text{ where } N^k = \sum_{(i,j)} \mathbb{1}(A^k(i,j) < t_L). \tag{8}$$

Here, k and (i,j) denote the class order and the pixel index, respectively. $\mathbb{1}$ is an indicator function that returns 1 if the statement is true, otherwise 0.

The AEFT is composed of two metric losses: (1) $\mathcal{L}_{Attract}$ that minimizes the distance between anchor embedding and positive embedding and (2) $\mathcal{L}_{Repel}$ that maximizes the distance between the embedding for low confidence region of the anchor image and negative embedding. Each loss can be formulated as follows:

$$\mathcal{L}_{Attract} = ||e_A - e_P||^2, \tag{9}$$

$$\mathcal{L}_{Repel} = [-||e_{AL} - e_N||^2 + \epsilon]_+, \tag{10}$$

where $||\cdot||^2$ denotes mean squared error and ϵ denotes a fixed margin that constrains the maximum distance between embeddings. For the $\mathcal{L}_{Repel}$, we only consider positive distance ($[\cdot]_+$) to prevent unbounded embedding space. Each embedding is normalized before calculating the distance between them.

Our total loss function for training the AEFT is formulated as follows:

$$\mathcal{L}_{AEFT} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{Attract} + \lambda_2 \mathcal{L}_{Repel}, \tag{11}$$

where $\mathcal{L}_{cls}$ denotes the binary cross-entropy loss between the class prediction (p) and image-level labels.

In AEFT, we utilize features from the GPP layer to construct the embedding space for metric learning. The metric learning on the GPP feature can be interpreted as implicit learning of the pooling weight, which is an effective way to handle the CAMs while not interrupting the backbone features of the classifier themselves. We experimentally verify that using the GPP feature is beneficial than using the CAMs or the intermediate features of the classifier. Also, compared with methods directly using GT classification labels for training the AE branch, the proposed AEFT shows superior performance in both qualitative and quantitative manners. In specific, thanks to the metric-based approach of AEFT, learning a semantic distance between the images enables CAMs to explore the less-discriminative regions while preventing the over-expansion problem.

4 Experimental Results

4.1 Dataset and Evaluation Metric

Evaluation of the proposed method is conducted on the PASCAL VOC 2012 dataset [14] and MS-COCO 2014 dataset [37]. COCO dataset is is more challenging in WSSS since it contains more classes (81) with small objects than PASCAL VOC 2012 (21). For VOC dataset, the proposed framework is trained with the augmented *train* set (10,582), and evaluated using both *val* (1,449) and *test* sets (1,456). For COCO dataset, the proposed method is trained with *train* set (80k) and evaluated on *val* set (40k). We use the mean Intersection over Union (mIoU) for evaluating our methods as similar to many other WSSS studies. As pointed out in Lee *et al.* [32], we utilize GT segmentation labels from COCO-Stuff dataset [5] for evaluation since the ground truth segmentation labels of the MS-COCO 2014 dataset have some overlaps between objects.

Table 1. Ablation study of the Gated Pyramid Pooling, evaluated on the PASCAL VOC *train* set. Pooled feature maps of bin sizes $\{2 \times 2, 4 \times 4, 8 \times 8, 16 \times 16\}$ are listed and used as an weighting factor of Eq. 5. Aggregation methods are noted as: $\mathcal{A}$ (averaged), $\mathcal{G}$ (gated convolution, from coarse to fine), and $\mathcal{G}_I$ (gated convolution, from fine to coarse). The performance is evaluated with the PASCAL VOC 2012 *train* set.

2x2	4x4	8x8	16x16	Aggregation	mIoU (%)
✓				-	49.9
	✓			-	51.6
		✓		-	52.9
			✓	-	53.1
✓	✓	✓	✓	$\mathcal{A}$	53.3
✓	✓	✓	✓	$\mathcal{G}_I$	51.3
✓	✓	✓	✓	$\mathcal{G}$	**54.2**

4.2 Implementation Details

The proposed network is implemented with PyTorch. We employ ResNet38 [53] as a backbone and the network is initialized with ImageNet [41] parameters. The data is augmented using horizontal flipping, color jittering [26], and cropping. The model is trained on 4 RTX 3090 GPUs with batch size 32. We use a poly learning rate [9] with the initial learning rate of 0.01 and the power of 0.9. For the semantic segmentation network, we use Deeplab [8] with ResNet38 backbone as in [2,27,46,58,35] for fair comparison. Margin (ϵ) between anchor and negative feature is set to 0.5. Weights of loss terms, λ_1 and λ_2, are set to 0.15 and 0.15, respectively. Our code is available at https://github.com/KAIST-vilab/AEFT.

4.3 Ablation Studies

To evaluate the proposed GPP layer, we conduct experiments with several different pooling bin sizes and feature aggregation methods. We set our baseline as a mere classifier with the GAP layer (achieves 48.4% in mIoU). As shown in Table 1, the larger the bin size for pooling, the higher performance (mIoU) of CAMs can be achieved. Furthermore, when averaging $\mathcal{A}$ of different feature pyramid is used as a weighting factor of Eq. 5, the performance is higher than using one of pooled feature pyramid. The results also show that the direction of aggregation is important, since using gated convolution in fine-to-coarse direction ($\mathcal{G}_I$) shows lower performance than naive averaging ($\mathcal{A}$), while the proposed coarse-to-fine ($\mathcal{G}$) outperforms both. It supports our design intention: the global context and fine details are well preserved in both small and large bin sizes, and the coarse-to-fine aggregation can effectively exploit both information.

To clarify the source of improvements in AEFT, we conduct an ablation study as in Table 2. With the attraction loss ($\mathcal{L}_{Attract}$), the proposed AEFT achieves 55.0% in mIoU, while the repelling loss ($\mathcal{L}_{Repel}$) achieves 54.8%. Actually, we did not expect that the AEFT could increase the performance of the generated CAMs with the repelling loss only, which is designed to aid the attraction loss.

Table 2. Ablation study of the proposed AEFT. *Direct*: global average pooled result of CAMs is used as embedding for metric, *Indirect*: global average pooled result of GPP is used. The performance (mIoU,%) is evaluated with the VOC 2012 *train* set.

Distance	$\mathcal{L}_{Attract}$	$\mathcal{L}_{Repel}$	CAMs	CAMs w/ CRF
Indirect	✓		54.9	62.2
Indirect		✓	54.6	61.5
Direct	✓	✓	54.7	61.6
Indirect	✓	✓	**56.0**	**63.5**

Since the background regions around the foreground objects are sometimes activated by the CAMs, we interpret this result as the repelling loss successively penalizes such unwanted intrusion. It leads the framework to generate precise CAMs fit along the object boundary. When both loss functions are used, the proposed framework achieves 56.0% in mIoU, and these results represent that benefits from each loss function are synergistic. Furthermore, compare to using CAMs itself as a embedding space for triplet learning (*Direct*, 54.6%), using the GPP feature (*Indirect*, 56.0%) shows better performance. This result indirectly supports our hypothesis (the direct guidance from the AE branch to the CAMs makes the model be unstable). Moreover, if we maximize the distance between e_A and e_N instead of e_{AL} and e_N, the performance of AEFT decreases to 54.1%. This result implies that the distance between the two images without sharing class is already large enough, as we expected, and our strategy for using only the low-confidence region from the anchor image is effective in terms of the quality of the CAMs. Also, it is noteworthy that the gain of CRF is even larger in our method (↑7.5%) than vanilla CAMs (↑5.9%). It implies that the benefit of our method is not overlapped with that of CRF, which is advantageous in terms of generating high-quality pseudo-labels.

In addition, we experimentally verify that using rigid classification labels triggers an over-expansion problem in Adversarial Erasing (AE). To quantitatively compare the degree of over-expansion, we use *Precision* and *Recall* scores of the generated CAMs. Here, *Precision* means that the true activation over the whole activation and *Recall* is the true activation over the GT. Although *Precision* and *Recall* are not direct metrics for the over-expansion of CAMs, they could be reasonable measure for quantitative comparison. As shown in Table 3, we compare the *Precision*, *Recall*, and mIoU performances from the various settings in our proposed AEFT. Here, we quantitatively verify that forcing the network to make the prediction from the erased image I_P according to the binary classification label (exist or not) leads to over-expansion. When we use the classification label for guiding the classifier to explore the less-discriminative regions as conventional AE-based WSSS methods (*Attract(Rigid)* in Table 3), the performance becomes lower than using the GPP layer alone (54.2%). Though using the rigid classification labels increases *Recall* by 0.5%, it causes over-expansion and thereby harms *Precision* by -1.4%. Instead of using the rigid labels, by minimizing the distance between the anchor embedding e_A and positive embedding e_P in a soft

Table 3. Comparison of the precision, recall, and mIoU from the various settings in the AEFT. ***Attract (Rigid)***: uses rigid classification labels for the masked image I_P, ***Attract (Soft)***: minimizes the distance between the anchor e_A and positive e_P in a soft manner, ***Atrract (Soft)+Repel***: uses repelling loss $\mathcal{L}_{Repel}$ along with attraction loss $\mathcal{L}_{Attract}$ (our setting). The performance is evaluated on the VOC 2012 *train* set.

	Precision(%)	Recall(%)	mIoU(%)
GPP only	66.5	75.6	54.2
Attract (Rigid)	65.1 (-1.4)	76.1 (+0.5)	53.4 (-0.8)
Attract (Soft)	66.6 (+0.1)	77.2 (+1.6)	55.0 (+0.8)
Atrract (Soft)+Repel	68.4 (+1.9)	76.3 (+0.7)	56.0 (+1.8)

Table 4. Evaluation (mIoU,%) of the CAMs, the CAMs with CRF, and the CAMs with CRF and RW [2] on the PASCAL VOC 2012 *train* set.

Methods	*seed*	*w/ CRF*	*w/ CRF, RW*
CONTA[57] $_{NeurIPS20}$	56.2	65.4	66.1
EDAM[52] $_{CVPR21}$	52.8	58.2	68.1
AdvCAM[30] $_{CVPR21}$	55.6	62.1	68.0
ECS[46] $_{ICCV21}$	56.6	58.6	-
OC-CSE[27]$_{ICCV21}$	56.0	62.8	66.9
CDA[44] $_{ICCV21}$	58.4	-	66.4
PMM[35] $_{ICCV21}$	58.2	61.5	61.0
RIB[30] $_{NeurIPS}$	56.5	62.9	70.6
Ours	56.0	63.5	**71.0**

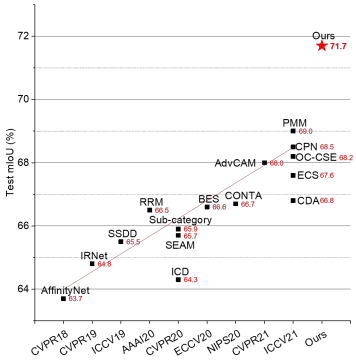
manner (*Attract(Soft)*), the proposed method increases *Recall* by 1.6% while not degrading the *Precision*. When we additionally employ repelling loss devised for preventing over-expansion (denoted as *Attract(Soft)+Repel*), both *Precision* and *Recall* are increased by 1.9% and 0.7%, respectively.

4.4 Comparison with State-of-the-arts

By applying a commonly used Random Walk (RW) approach [2] as in [50,7,30,58,46], we acquire further improved pixel-level pseudo labels for training the semantic segmentation model. As shown in Table 4, though the performance of CAMs of the proposed method is similar to the existing state-of-the-art, our method greatly benefits from CRF (about 7.5%). According to Kweon *et al.* [27], we can interpret this performance gain as a benefit from more precise CAMs that match along object boundaries while activating the whole object. The resulting performance of pseudo labels achieves 71.0% mIoU on PASCAL VOC 2012 *train* set. For a fair comparison with the current state-of-the-art, we train the Deeplab-LargeFOV [8] with the corresponding pseudo labels. The backbone of the segmentation model is ResNet38d. As shown in Table 5, the proposed AEFT achieves a state-of-the-art with 70.9% and 71.7% mIoU on PASCAL VOC 2012 *val* and *test* sets, respectively. Qualitative segmentation results of the proposed method can be found in the *Supplementary Material*, which depicts that the

Table 5. Performance (mIoU, %) comparison with other state-of-the-art WSSS methods on the PASCAL VOC 2012 and MS-MOCO 2014. Since we use neither saliency nor external dataset at all, we list the methods using image-level only in this table. **Bold** numbers represent the best results.

Methods	*Backbone*	VOC *val*	VOC *test*	COCO *val*
AffinityNet [2]$_{CVPR18}$	ResNet38	61.7	63.7	-
ICD [15]$_{CVPR20}$	ResNet101	64.1	64.3	-
IRNet [1]$_{CVPR19}$	ResNet50	63.5	64.8	32.6
SSDD [43]$_{ICCV19}$	ResNet38	64.9	65.5	-
SEAM [50]$_{CVPR20}$	ResNet38	64.5	65.7	31.9
Sub-category [7]$_{CVPR20}$	ResNet101	66.1	65.9	-
CONTA [57]$_{NIPS20}$	ResNet38	66.1	66.7	33.4
RRM [56]$_{AAAI20}$	ResNet101	66.3	66.5	-
BES [10]$_{ECCV20}$	ResNet101	65.7	66.6	-
CDA [44]$_{ICCV21}$	ResNet38	66.1	66.8	-
ECS [46]$_{ICCV21}$	ResNet38	66.6	67.6	-
AdvCAM [30]$_{CVPR21}$	ResNet101	68.1	68.0	-
OC-CSE[27]$_{ICCV21}$	ResNet38	68.4	68.2	36.4
CPN [58]$_{ICCV21}$	ResNet38	67.8	68.5	-
RIB [28]$_{NeurIPS21}$	ResNet101	68.3	68.6	43.8
PMM [35]$_{ICCV21}$	ResNet38	68.5	69.0	36.7
Ours	ResNet38	**70.9**	**71.7**	**44.8**



segmentation model can capture fine details as well, owing to the high quality pseudo labels used for training the model. We also evaluate our method in the MS-COCO 2014 dataset to show the superiority and versatility of the proposed framework. It achieves 44.8% on the MS-COCO *val* set, which is a new state-of-the-art, outperforming the other methods by a meaningful margin (1.0%).

5 Conclusions

To address the problems in weakly supervised semantic segmentation (WSSS), we propose a Gated Pyramid Pooling (GPP) layer that replaces the GAP layer by using a feature pyramid and a novel Adversarial Erasing framework via Triplet (AEFT) that incorporates metric learning for suppressing the over-expansion problem in AE. Extensive ablation studies support that the proposed GPP layer outperforms the conventional GAP layer while effectively resolving the impreciseness problem of CAMs with the help of the feature pyramid. In addition, the proposed AEFT succeeds in relieving the over-expansion problem of AE by exploiting the triplet loss as a softer criterion compared to classification loss conventionally used. With the proposed GPP and AEFT, we achieve the state-of-the-art performance both on the PASCAL VOC 2012 and MS-COCO 2014 *val* set with a great margin, only utilizing image-level supervision.

Acknowledgements This work was supported by Institute of Information and Communications Technology Planning & Evaluation(IITP) Grants 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, and No. 2014-3-00123, Development of High Performance Visual BigData Discovery Platform for Large-Scale Realtime Data Analysis.

References

1. Ahn, J., Cho, S., Kwak, S.: Weakly supervised learning of instance segmentation with inter-pixel relations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2209–2218 (2019)
2. Ahn, J., Kwak, S.: 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. pp. 4981–4990 (2018)
3. Araslanov, N., Roth, S.: Single-stage semantic segmentation from image labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4253–4262 (2020)
4. Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L.: What's the point: Semantic segmentation with point supervision. In: European conference on computer vision. pp. 549–565. Springer (2016)
5. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Computer vision and pattern recognition (CVPR), 2018 IEEE conference on. IEEE (2018)
6. Cen, J., Yun, P., Cai, J., Wang, M.Y., Liu, M.: Deep metric learning for open world semantic segmentation. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 15313–15322 (2021)
7. Chang, Y.T., Wang, Q., Hung, W.C., Piramuthu, R., Tsai, Y.H., Yang, M.H.: Weakly-supervised semantic segmentation via sub-category exploration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8991–9000 (2020)
8. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR (2015), `http://arxiv.org/abs/1412.7062`
9. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: 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)
10. Chen, L., Wu, W., Fu, C., Han, X., Zhang, Y.: Weakly supervised semantic segmentation with boundary exploration. In: European Conference on Computer Vision. pp. 347–362. Springer (2020)
11. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
12. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
13. Dai, J., He, K., Sun, J.: Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: Proceedings of the IEEE international conference on computer vision. pp. 1635–1643 (2015)
14. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
15. Fan, J., Zhang, Z., Song, C., Tan, T.: 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. pp. 4283–4292 (2020)

16. Fan, J., Zhang, Z., Tan, T., Song, C., Xiao, J.: Cian: Cross-image affinity net for weakly supervised semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 10762–10769 (2020)
17. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 2, pp. 1735–1742 (2006). https://doi.org/10.1109/CVPR.2006.100
18. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
19. Hou, Q., Jiang, P., Wei, Y., Cheng, M.M.: Self-erasing network for integral object attention. In: Advances in Neural Information Processing Systems. pp. 549–559 (2018)
20. Huang, Z., Wang, X., Wang, J., Liu, W., Wang, J.: Weakly-supervised semantic segmentation network with deep seeded region growing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7014–7023 (2018)
21. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning. pp. 4904–4916. PMLR (2021)
22. Jiang, P.T., Hou, Q., Cao, Y., Cheng, M.M., Wei, Y., Xiong, H.K.: Integral object mining via online attention accumulation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2070–2079 (2019)
23. Ke, T.W., Hwang, J.J., Yu, S.: Universal weakly supervised segmentation by pixel-to-segment contrastive learning. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=N33d7wjgzde
24. Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B.: Simple does it: Weakly supervised instance and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 876–885 (2017)
25. Kolesnikov, A., Lampert, C.H.: Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In: European conference on computer vision. pp. 695–711. Springer (2016)
26. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Communications of the ACM **60**(6), 84–90 (2017)
27. Kweon, H., Yoon, S.H., Kim, H., Park, D., Yoon, K.J.: Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6994–7003 (2021)
28. Lee, J., Choi, J., Mok, J., Yoon, S.: Reducing information bottleneck for weakly supervised semantic segmentation. Advances in Neural Information Processing Systems **34**, 27408–27421 (2021)
29. Lee, J., Kim, E., Lee, S., Lee, J., Yoon, S.: Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5267–5276 (2019)
30. Lee, J., Kim, E., Yoon, S.: Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4071–4080 (2021)
31. Lee, J., Yi, J., Shin, C., Yoon, S.: Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2643–2652 (2021)

32. Lee, S., Lee, M., Lee, J., Shim, H.: Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5495–5505 (2021)
33. Li, K., Wu, Z., Peng, K.C., Ernst, J., Fu, Y.: Tell me where to look: Guided attention inference network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9215–9223 (2018)
34. Li, X., Zhou, T., Li, J., Zhou, Y., Zhang, Z.: Group-wise semantic mining for weakly supervised semantic segmentation. arXiv preprint arXiv:2012.05007 (2020)
35. Li, Y., Kuang, Z., Liu, L., Chen, Y., Zhang, W.: Pseudo-mask matters in weakly-supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6964–6973 (2021)
36. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3159–3167 (2016)
37. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
38. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 212–220 (2017)
39. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. ArXiv **abs/1807.03748** (2018)
40. Papandreou, G., Chen, L.C., Murphy, K.P., Yuille, A.L.: Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: Proceedings of the IEEE international conference on computer vision. pp. 1742–1750 (2015)
41. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
42. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 815–823 (2015). https://doi.org/10.1109/CVPR.2015.7298682
43. Shimoda, W., Yanai, K.: Self-supervised difference detection for weakly-supervised semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5208–5217 (2019)
44. Su, Y., Sun, R., Lin, G., Wu, Q.: Context decoupling augmentation for weakly supervised semantic segmentation. arXiv preprint arXiv:2103.01795 (2021)
45. Sun, G., Wang, W., Dai, J., Van Gool, L.: Mining cross-image semantics for weakly supervised semantic segmentation. arXiv preprint arXiv:2007.01947 (2020)
46. Sun, K., Shi, H., Zhang, Z., Huang, Y.: Ecs-net: Improving weakly supervised semantic segmentation by using connections between class activation maps. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7283–7292 (2021)
47. Takikawa, T., Acuna, D., Jampani, V., Fidler, S.: Gated-scnn: Gated shape cnns for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5229–5238 (2019)
48. Vernaza, P., Chandraker, M.: Learning random-walk label propagation for weakly-supervised semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7158–7166 (2017)

49. Wang, W., Zhou, T., Yu, F., Dai, J., Konukoglu, E., Van Gool, L.: Exploring cross-image pixel contrast for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7303–7313 (2021)
50. Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X.: Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12275–12284 (2020)
51. Wei, Y., Feng, J., Liang, X., Cheng, M.M., Zhao, Y., Yan, S.: 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. pp. 1568–1576 (2017)
52. Wu, T., Huang, J., Gao, G., Wei, X., Wei, X., Luo, X., Liu, C.H.: Embedded discriminative attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16765–16774 (2021)
53. Wu, Z., Shen, C., Van Den Hengel, A.: Wider or deeper: Revisiting the resnet model for visual recognition. Pattern Recognition **90**, 119–133 (2019)
54. Xu, L., Ouyang, W., Bennamoun, M., Boussaid, F., Sohel, F., Xu, D.: Leveraging auxiliary tasks with affinity learning for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6984–6993 (2021)
55. Yao, Y., Chen, T., Xie, G.S., Zhang, C., Shen, F., Wu, Q., Tang, Z., Zhang, J.: Non-salient region object mining for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2623–2632 (2021)
56. Zhang, B., Xiao, J., Wei, Y., Sun, M., Huang, K.: Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12765–12772 (2020)
57. Zhang, D., Zhang, H., Tang, J., Hua, X., Sun, Q.: Causal intervention for weakly-supervised semantic segmentation. Advances in Neural Information Processing Systems (2020)
58. Zhang, F., Gu, C., Zhang, C., Dai, Y.: Complementary patch for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7242–7251 (2021)
59. Zhang, X., Wei, Y., Feng, J., Yang, Y., Huang, T.S.: Adversarial complementary learning for weakly supervised object localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1325–1334 (2018)
60. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)
61. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)