Self-supervised Equivariant Attention Mechanism for Weakly Supervised Semantic Segmentation

Yude Wang[1,2], Jie Zhang[1,2], Meina Kan[1,2], Shiguang Shan[1,2,3], Xilin Chen[1,2]
[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing, 100049, China
[3]CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai, 200031, China
yude.wang@vipl.ict.ac.cn, {zhangjie, kanmeina, sgshan, xlchen}@ict.ac.cn

Abstract

Image-level weakly supervised semantic segmentation is a challenging problem that has been deeply studied in recent years. Most of advanced solutions exploit class activation map (CAM). However, CAMs can hardly serve as the object mask due to the gap between full and weak supervisions. In this paper, we propose a self-supervised equivariant attention mechanism (SEAM) to discover additional supervision and narrow the gap. Our method is based on the observation that equivariance is an implicit constraint in fully supervised semantic segmentation, whose pixel-level labels take the same spatial transformation as the input images during data augmentation. However, this constraint is lost on the CAMs trained by image-level supervision. Therefore, we propose consistency regularization on predicted CAMs from various transformed images to provide self-supervision for network learning. Moreover, we propose a pixel correlation module (PCM), which exploits context appearance information and refines the prediction of current pixel by its similar neighbors, leading to further improvement on CAMs consistency. Extensive experiments on PASCAL VOC 2012 dataset demonstrate our method outperforms state-of-the-art methods using the same level of supervision. The code is released online[1].

1. Introduction

Semantic segmentation is a fundamental computer vision task, which aims to predict pixel-wise classification results on images. Thanks to the booming of deep learning researches in recent years, the performance of semantic segmentation model has achieved great progress [6, 23, 38], promoting many practical applications, *e.g.*, autopilot and medical image analysis. However, compared to other tasks such as classification and detection, semantic segmentation needs to collect pixel-level class labels which are time-consuming and expensive. Recently many efforts are devoted to weakly supervised semantic segmentation (WSSS) which utilizes weak supervisions, *e.g.*, image-level classification labels, scribbles, and bounding boxes, attempting to achieve equivalent segmentation performance of fully supervised approaches. This paper focuses on semantic segmentation by image-level classification labels.

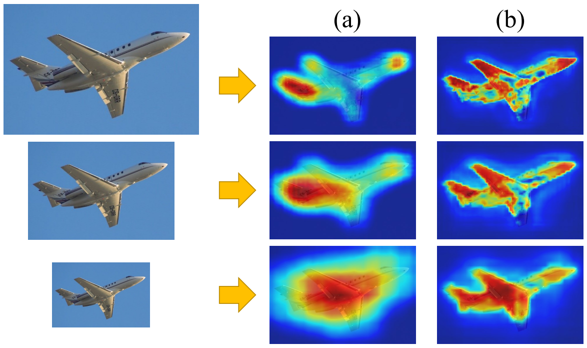

Figure 1. Comparisons of CAMs generated by input images with different scales. (a) Conventional CAMs. (b) CAMs predicted by our SEAM, which are more consistent over rescaling.

To the best of our knowledge, most of advanced WSSS methods are based on the class activation map (CAM) [39], which is an effective way to localize objects by image classification labels. However, the CAMs usually only cover the most discriminative part of the object and incorrectly activate in background regions, which can be summarized as under-activation and over-activation respectively. Moreover, the generated CAMs are not consistent when images are augmented by affine transformations. As shown in Fig. 1, applying different rescaling transformations on the same input images causes significant inconsistency on the

[1] https://github.com/YudeWang/SEAM

generated CAMs. The essential causes of these phenomena come from the supervision gap between fully and weakly supervised semantic segmentation.

In this paper, we propose a self-supervised equivariant attention mechanism (SEAM) to narrow the supervision gap mentioned above. The SEAM applies consistency regularization on CAMs from various transformed images to provide self-supervision for network learning. To further improve the network prediction consistency, SEAM introduces the pixel correlation module (PCM), which captures context appearance information for each pixel and revises original CAMs by learned affinity attention maps. The SEAM is implemented by a siamese network with equivariant cross regularization (ECR) loss, which regularizes the original CAMs and the revised CAMs on different branches. Fig. 1 shows that our CAMs are consistent over various transformed input images, with fewer over-activated and under-activated regions than baseline. Extensive experiments give both quantitative and qualitative results, demonstrating the superiority of our approach.

In summary, our main contributions:

- We propose a self-supervised equivariant attention mechanism (SEAM), incorporating equivariant regularization with pixel correlation module (PCM), to narrow the supervision gap between fully and weakly supervised semantic segmentation.
- The design of siamese network architecture with equivariant cross regularization (ECR) loss efficiently couples the PCM and self-supervision, producing CAMs with both fewer over-activated and under-activated regions.
- Experiments on PASCAL VOC 2012 illustrate that our algorithm achieves state-of-the-art performance with only image-level annotations.

2. Related Work

The development of deep learning has led to a series of breakthroughs on fully supervised semantic segmentation [6, 11, 23, 37, 38] in recent years. In this section, we introduce some works, including weakly supervised semantic segmentation and self-supervised learning.

2.1. Weakly Supervised Semantic Segmentation

Compared to fully supervised learning, WSSS uses weak labels to guide network training, *e.g.*, bounding boxes [7, 18], scribbles [22, 30] and image-level classification labels [19, 25, 27]. A group of advanced researches utilizes image-level classification labels to train models. Most of them refine the class activation map (CAM) [39] generated by the classification network to approximate the segmentation mask. SEC [19] proposes three principles, *i.e.*, seed, expand, and constrain, to refine CAMs, which are followed by many other works. Adversarial erasing [15, 32] is a popular CAM expansion method, which erases the most discriminative part of CAM, guides the network to learn classification features from other regions and expands activations. AffinityNet [2] trains another network to learn the similarity between pixels, which generates a transition matrix and multiplies with CAM several times to adjust its activation coverage. IRNet [1] generates a transition matrix from the boundary activation map and extends the method to weakly supervised instance segmentation. Here are also some researches endeavor to aggregate self-attention module [29, 31] in the WSSS framework, *e.g.*, CIAN [10] proposes cross-image attention module to learn activation maps from two different images containing the same class objects with the guidance of saliency maps.

2.2. Self-supervised Learning

Instead of using massive annotated labels to train network, self-supervised learning approaches aim at designing pretext tasks to generate labels without additional manual annotations. Here are many classical self-supervised pretext tasks, *e.g.*, relative position prediction [9], spatial transformation prediction [12], image inpainting [26], and image colorization [20]. To some extent, the generative adversarial network [13] can also be regarded as a self-supervised learning approach that the authenticity labels for discriminator do not need to be annotated manually. Labels generated by pretext tasks provide self-supervision for the network to learn a more robust representation. The feature learned by self-supervision can replace the feature pretrained by ImageNet [8] on some tasks, such as detection [9] and part segmentation [17].

Considering there is a large supervision gap between fully and weakly supervised semantic segmentation, it is an intuition that we should seek additional supervision to narrow the gap. Since image-level classification labels are too weak for network to learn segmentation masks which should well fit object boundary, we design pretext task using the equivariance of ideal segmentation function to provide additional self-supervision for network learning with only image-level annotations.

3. Approach

This section details our SEAM method. Firstly, we illustrate the motivation of our work. Then we introduce the implementation of equivariant regularization by a shared-weight siamese network. The proposed pixel correlation module (PCM) is integrated into the network to further improve the consistency of prediction. Finally, the loss design of SEAM is discussed. Fig. 2 shows our SEAM network structure.

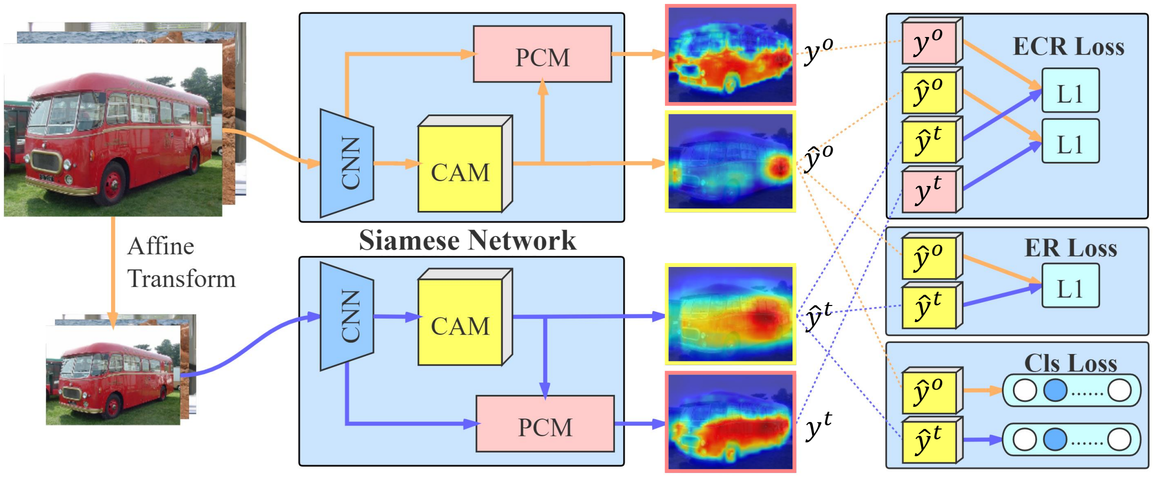

Figure 2. The siamese network architecture of our proposed SEAM method. The SEAM is the integration of equivariant regularization (ER) (Section. 3.2) and pixel correlation module (PCM) (Section. 3.3). With specially designed losses (Section 3.4), the revised CAMs not only keep consistent over affine transformation but also well fit the object contour.

3.1. Motivation

We denote ideal pixel-level semantic segmentation function as $F_{w_s}(\cdot)$ with parameters w_s. For each image sample I, the segmentation process can be formulated as $F_{w_s}(I) = s$, where s denotes pixel-level segmentation mask. The formulation is also consistent in classification task. With additional image-level label l and pooling function $\text{Pool}(\cdot)$, classification task can be represented as $\text{Pool}(F_{w_c}(I)) = l$ with parameters w_c. Most WSSS approaches are based on the hypothesis that the optimal parameters for classification and segmentation satisfy $w_c = w_s$. Therefore, these methods train a classification network firstly and remove pooling function to tackle segmentation task.

However, it is easy to find the properties of classification and segmentation function are different. Suppose there is an affine transformation $A(\cdot)$ for each sample, the segmentation function is more inclined to be equivariant, *i.e.*, $F_{w_s}(A(I)) = A(F_{w_s}(I))$. While the classification task focuses more on invariance, *i.e.*, $\text{Pool}(F_{w_c}(A(I))) = l$. Although the invariance of classification function is mainly caused by pooling operation, there is no equivariant constraint for $F_{w_c}(\cdot)$, which makes it nearly impossible to achieve the same objective of segmentation function during network learning. Additional regularizers should be integrated to narrow the supervision gap between fully and weakly supervised learning.

Self-attention is a widely accepted mechanism that can significantly improve the network approximation ability. It revises feature maps by capturing context feature dependency, which also meets the ideas of most WSSS methods using the similarity of pixels to refine the original activation map. Following the denotation of [31], the general self-attention mechanism can be defined as:

$$\mathrm{y}_i = \frac{1}{\mathcal{C}(\mathrm{x}_i)} \sum_{\forall j} f(\mathrm{x}_i, \mathrm{x}_j) g(\mathrm{x}_j) + \mathrm{x}_i, \tag{1}$$

$$f(\mathrm{x}_i, \mathrm{x}_j) = e^{\theta(\mathrm{x}_i)^{\mathrm{T}} \phi(\mathrm{x}_j)}. \tag{2}$$

Here x and y denote input and output feature, with spatial position index i and j. The output signal is normalized by $\mathcal{C}(\mathrm{x}_i) = \sum_{\forall j} f(\mathrm{x}_i, \mathrm{x}_j)$. Function $g(\mathrm{x}_j)$ gives a representation of input signal x_j at each position and all of them are aggregated into position i with the similarity weights given by $f(\mathrm{x}_i, \mathrm{x}_j)$, which calculates the dot-product pixel affinity in an embedding space. To improve the network ability for consistent prediction, we propose SEAM by incorporating self-attention with equivariant regularization.

3.2. Equivariant Regularization

During the data augmentation period of fully supervised semantic segmentation, the pixel-level labels should be applied with the same affine transformation as input images. It introduces an implicit equivariant constraint for the network. However, considering that the WSSS can only access image-level classification labels, the implicit constraint is missing here. Therefore, we propose equivariant regularization as follows:

$$\mathcal{R}_{ER} = ||F(A(I)) - A(F(I))||_1. \tag{3}$$

Here $F(\cdot)$ denotes the network, and $A(\cdot)$ denotes any spatial affine transformation, *e.g.*, rescaling, rotation, flip. To integrate regularization on the original network, we expand the network into a shared-weight siamese structure. One branch applies the transformation on the network output, the other branch warps the images by the same transformation before the feedforward of the network. The output activation maps from two branches are regularized to guarantee the consistency of CAMs.

3.3. Pixel Correlation Module

Although equivariant regularization provides additional supervision for network learning, it is hard to achieve ideal equivariance with only classical convolution layers. Self-attention is an efficient module to capture context information and refine pixel-wise prediction results. To integrate the classical self-attention module given by Eq. (1) and Eq. (2) for CAM refinement, the formulation can be written as:

$$\mathrm{y}_i = \frac{1}{\mathcal{C}(\mathrm{x}_i)} \sum_{\forall j} e^{\theta(\mathrm{x}_i)^{\mathrm{T}}\phi(\mathrm{x}_j)} g(\hat{\mathrm{y}}_j) + \hat{\mathrm{y}}_i, \tag{4}$$

where $\hat{\mathrm{y}}$ denotes the original CAM and y denotes the revised CAM. In this structure, the original CAM is embedded into residual space by function g. Each pixel aggregates with others with similarity given by Eq. (2). Three embedding functions θ, ϕ, g can be implemented by individual 1×1 convolution layers.

To further refine original CAMs by context information, we propose a pixel correlation module (PCM) at the end of the network to integrate the low-level feature of each pixel. The structure of PCM refers to the core part of the self-attention mechanism with some modifications and trained by the supervision from equivariant regularization. We use cosine distance to evaluate inter-pixel feature similarity:

$$f(\mathrm{x}_i, \mathrm{x}_j) = \frac{\theta(\mathrm{x}_i)^{\mathrm{T}}\theta(\mathrm{x}_j)}{||\theta(\mathrm{x}_i)|| \cdot ||\theta(\mathrm{x}_j)||}. \tag{5}$$

Here we take the inner-product in normalized feature space to calculate the affinity between current pixel i and others. The f can be integrated into Eq. (1) with some modifications as:

$$\mathrm{y}_i = \frac{1}{\mathcal{C}(\mathrm{x}_i)} \sum_{\forall j} \mathrm{ReLU}(\frac{\theta(\mathrm{x}_i)^{\mathrm{T}}\theta(\mathrm{x}_j)}{||\theta(\mathrm{x}_i)|| \cdot ||\theta(\mathrm{x}_j)||})\hat{\mathrm{y}}_j. \tag{6}$$

The similarities are activated by ReLU to suppress negative values. The final CAM is the weighted sum of the original CAM with normalized similarities. Fig. 3 gives an illustration of the PCM structure.

Compared to classical self-attention, PCM removes the residual connection to keep the same activation intensity of the original CAM. Moreover, since the other network branch provides pixel-level supervision for PCM, which is not as accurate as ground truth, we reduce parameters by removing embedding function ϕ and g to avoid overfitting on inaccurate supervision. We use ReLU activation function with L1 normalization to mask out irrelevant pixels and generate an affinity attention map which is smoother in relevant regions.

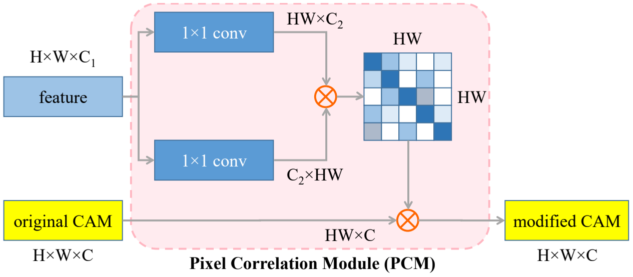

Figure 3. The structure of PCM, where $H, W, C/C_1/C_2$ denote height, width and channel numbers of feature maps respectively.

3.4. Loss Design of SEAM

Image-level classification label l is the only human-annotated supervision that can be used here. We employ the global average pooling layer at the end of the network to achieve prediction vector z for image classification and adopt multi-label soft margin loss for network training. The classification loss is defined for $C-1$ foreground object category as:

$$\begin{aligned} \ell_{cls}(\mathrm{z}, l) = -\frac{1}{C-1}\sum_{c=1}^{C-1}[l_c \log(\frac{1}{1+e^{-\mathrm{z}_c}}) \\ + (1-l_c)\log(\frac{e^{-\mathrm{z}_c}}{1+e^{-\mathrm{z}_c}})]. \end{aligned} \tag{7}$$

Formally we denote the original CAMs of siamese network as $\hat{\mathrm{y}}^o$ and $\hat{\mathrm{y}}^t$, where $\hat{\mathrm{y}}^o$ comes from the branch with original image input and $\hat{\mathrm{y}}^t$ stems from the transformed images. The global average pooling layer aggregates them into prediction vector z^o and z^t respectively. The classification loss is calculated on two branches as:

$$\mathcal{L}_{cls} = \frac{1}{2}(\ell_{cls}(\mathrm{z}^o, l) + \ell_{cls}(\mathrm{z}^t, l)). \tag{8}$$

The classification loss provides learning supervision for object localization. And it is necessary to aggregate equivariant regularization on original CAM to preserve the consistency of output. The equivariant regularization (ER) loss on original CAM can be easily defined as:

$$\mathcal{L}_{ER} = ||A(\hat{\mathrm{y}}^o) - \hat{\mathrm{y}}^t||_1. \tag{9}$$

Here $A(\cdot)$ is an affine transformation which has already been applied to the input image in the transformation branch of the siamese network. Moreover, to further improve the ability of network for equivariance learning, the original CAMs and features from the shallow layers are fed into PCM for refinement. The intuitive idea is introducing equivariant regularization between revised CAMs y^o and y^t. However, in our early experiments, the output maps of PCM fall into the local minimum quickly that all pixels in the image are predicted the same class. Therefore, we propose an equivariant cross regularization (ECR) loss as:

$$\mathcal{L}_{ECR} = ||A(\mathrm{y}^o) - \hat{\mathrm{y}}^t||_1 + ||A(\hat{\mathrm{y}}^o) - \mathrm{y}^t||_1. \quad (10)$$

The PCM outputs are regularized by the original CAMs on the other branch of the siamese network. This strategy can avoid CAM degeneration during PCM refinement.

Although the CAMs are learned by foreground object classification loss, there are many background pixels, which should not be ignored during PCM processing. The original foreground CAMs have zero vectors on these background positions, which cannot produce gradients to push feature representations closer between those background pixels. Therefore, we define the background score as:

$$\hat{\mathrm{y}}_{i,bkg} = 1 - \max_{1 \le c \le C-1} \hat{\mathrm{y}}_{i,c}, \quad (11)$$

where $\hat{\mathrm{y}}_{i,c}$ is the activation score of original CAM for category c at position i. We normalize the activation vectors of each pixel by suppressing foreground non-maximum activations to zeros and concatenate with additional background score. During inference, we only keep the foreground activation results and set the background score as $\hat{\mathrm{y}}_{i,bkg} = \alpha$, where α is the hard threshold parameter.

In summary, the final loss of SEAM is defined as:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{ER} + \mathcal{L}_{ECR}. \quad (12)$$

The classification loss is used to roughly localize objects and the ER loss is used to narrow the gaps between pixel- and image-level supervisions. The ECR loss is used to integrate PCM with the trunk of the network, in order to make consistent predictions over various affine transformations. The network architecture is illustrated in Fig. 2. We give the details of network training settings and carefully investigate the effectiveness of each module in the experiments section.

4. Experiments

4.1. Implementation Details

We evaluate our approach on PASCAL VOC 2012 dataset with 21 class annotations, *i.e.*, 20 foreground objects and the background. The official dataset separation has 1464 images for training, 1449 for validation and 1456 for testing. Following the common experimental protocol for semantic segmentation, we take additional annotations from SBD [14] to build an augmented training set with 10582 images. Noting that only image-level classification labels are available during network training. Mean intersection over union (mIoU) is used as a metric to evaluate segmentation results.

In our experiments, ResNet38 [35] is adopted as backbone network with $output\_stride = 8$. We extract the feature maps from stage 3 and stage 4, reduce their channel numbers into 64 and 128 respectively by individual 1×1 convolution layers. In PCM, these features are concatenated with images and fed into function θ in Eq. (5), which is implemented by another 1×1 convolution layer. The images are randomly rescaled in the range of [448, 768] by the longest edge and then cropped by 448×448 as network inputs. The model is trained on 4 TITAN-Xp GPUs with batch size 8 for 8 epochs. The initial learning rate is set as 0.01, following the poly policy $lr_{itr} = lr_{init}(1 - \frac{itr}{max\_itr})^\gamma$ with $\gamma = 0.9$ for decay. Online hard example mining (OHEM) is employed on the ECR loss remaining the largest 20% pixel losses.

During network training, we cut off gradients back-propagation at the intersection point between PCM stream and the trunk of the network to avoid the mutual interference. This setting simplifies the PCM into a pure context refinement module which still can be trained with the backbone of the network at the same time. And the learning of original CAMs will not be affected by PCM refinement process. During inference, since our SEAM is a shared-weight siamese network, only one branch needs to be restored. We adopt multi-scale and flip test during inference to generate pseudo segmentation labels.

4.2. Ablation Studies

To verify the effectiveness of our SEAM, we generate pixel-level pseudo labels from revised CAMs on PASCAL VOC 2012 *train* set. In our experiments, we traverse all background threshold options and give the best mIoU of pseudo labels, instead of comparing with the same background threshold. Because the highest pseudo label accuracy represents the best matching results between CAMs and ground truth segmentation masks. Specifically, the foreground activation coverage will expand with the increase of average activation intensity, while its matching degree with ground truth is not changed. And the highest pseudo label accuracy will not be improved when CAMs only increase average activation intensity rather than becoming more matchable with ground truth.

Comparison with Baseline: Tab. 1 gives an ablation study of each module in our approach. It shows that using the siamese network with equivariant regularization has a 2.47% improvement compared to baseline. Our PCM achieves significant performance elevation by 5.18%. After applying OHEM on equivariant cross regularization loss, the generated pseudo labels further achieve 55.41% mIoU on PASCAL VOC train set. We also test the baseline CAM with dense CRF to refine predictions. The results show that dense CRF improves the mIoU to 52.40%, which is lower than the SEAM result 55.41%. And our SEAM can further improve the performance up to 56.83% after aggregating dense CRF as post process. Fig. 4 shows that the CAMs

baseline	ER	PCM	OHEM	CRF	mIoU
√					47.43%
√				√	52.40%
√	√				49.90%
√	√	√			55.08%
√	√	√	√		55.41%
√	√	√	√	√	56.83%

Table 1. The ablation study for each part of SEAM. **ER**: equivariant regularization. **PCM**: pixel correlation module. **OHEM**: online hard example mining. **CRF**: conditional random field.

model	mIoU
CAM	47.43%
GradCAM	46.53%
GradCAM++	47.37%
CAM + SEAM	55.41%

Table 2. Evaluation of various weakly supervised localization methods with semantic segmentation metric (mIoU).

generated by SEAM have fewer over-activations and more complete activation coverage, whose shape is closer to the ground truth segmentation masks than baseline. To further verify the effectiveness of our proposed SEAM, we visualize the affinity attention maps generated by PCM. As shown in Fig. 5, the selected foreground and background pixels are very close in spatial, while their affinity attention maps are greatly different. It proves that the PCM can learn boundary sensitive features from self-supervision.

Improved Localization Mechanism: It is an intuition that improved weakly supervised localization mechanism will elevate mIoU of pseudo segmentation labels. To verify the idea, we simply evaluate GradCAM [28] and GradCAM++[3] before aggregating our proposed SEAM. However, the evaluation results given by Tab. 2 illustrates that both GradCAM and GradCAM++ cannot narrow the supervision gap between fully and weakly supervised semantic segmentation tasks, since the best mIoU results do not have improvement. We believe the improved localization mechanisms are only designed to represent object correlated parts without any constraints by low-level information, which is not suitable for the segmentation task. The CAMs generated by these improved localization methods are not becoming more matchable with ground truth masks. The following experiments further illustrate that our proposed SEAM can substantially improve the quality of CAM to fit the shape of object masks.

Affine Transformation: Ideally, the $A(\cdot)$ in Eq. (3) can be any affine transformation. Several transformations are conducted in the siamese network to evaluate the effect of them on equivariant regularization. As shown in Tab. 3, there are four candidate affine transformations: rescaling

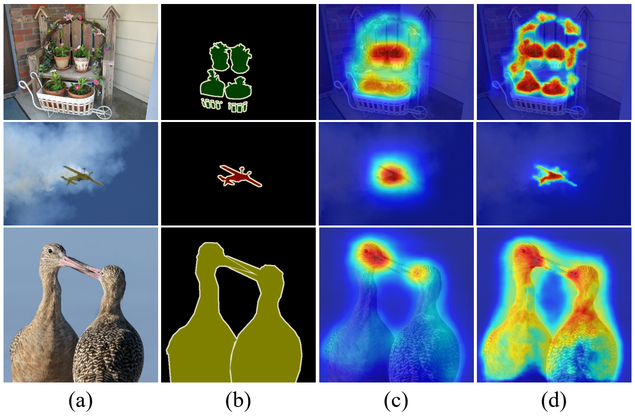

Figure 4. The visualization of CAMs. (a) Original images. (b) Ground truth segmentations. (c) Baseline CAMs. (d) CAMs produced by SEAM. The SEAM not only suppresses over-activation but also expands CAMs into complete object activation coverage.

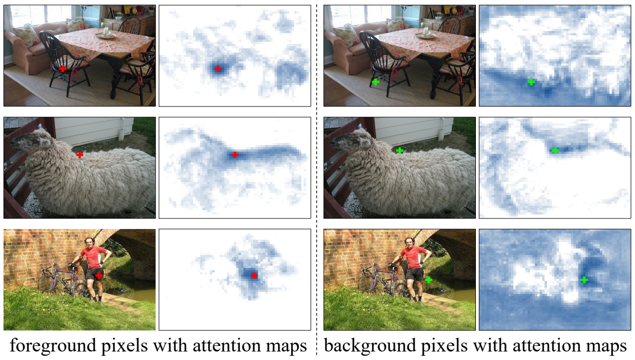
foreground pixels with attention maps
background pixels with attention maps

Figure 5. The visualization of affinity attention map on foreground and background. The red and green crosses denote the selected pixels, with similar feature representation in blue color.

with 0.3 down-sampling rate, random rotation in [-20, 20] degrees, translation by 15 pixels and horizontal flip. Firstly, our proposed SEAM simply adopts rescaling during network training. Tab. 3 shows that the mIoU of pseudo labels has significant improvement from 47.43% to 55.41%. Tab. 3 also shows that simply incorporating different transformations is not much effective. When rescaling transformation integrates with flip, rotation, and translation respectively, only flip makes tiny improvement. In our view, it is because the activation maps between flip, rotation, and translation are too similar to produce sufficient supervision. Without additional instructions, we only preserve rescaling as the key transformation with 0.3 down-sampling rate in our other experiments.

Augmentation and Inference: Compared to the original one-branch network, the siamese structure expands the augmentation range of image size in practice. To investigate whether the improvement stems from the rescaling range,

rescale	flip	rotation	translation	mIoU
				47.43%
√				55.41%
√	√			55.50%
√		√		53.13%
√			√	55.23%

Table 3. Experiments of various transformations on equivariant regularization. Simply aggregating different affine transformations cannot bring significant improvement.

model	random rescale	mIoU
baseline	[448, 768]	47.43%
baseline	[224, 768]	46.72%
SEAM	[448, 768]	53.47%

Table 4. Experiments of augmentation rescaling range. Here the rescale rate of SEAM is set to 0.5.

test scale	baseline (mIoU)	ours (mIoU)
[0.5]	40.17%	49.35%
[1.0]	46.10%	51.57%
[1.5]	47.51%	52.25%
[2.0]	46.12%	49.79%
[0.5, 1.0, 1.5, 2.0]	47.43%	55.41%

Table 5. Experiments with various single- and multi-scale test.

we evaluate the baseline model with a larger scale range and Tab. 4 gives the experiment results. It shows that simply increasing the rescaling range cannot improve the accuracy of generated pseudo labels, which proves that the performance improvement comes from the combination of PCM and equivariant regularization instead of data augmentation.

During inference, it is a common practice to employ multi-scale test by aggregating the prediction results from images with different scales to boost the final performance. It can also be regarded as a method to improve the equivariance of predictions. To verify the effectiveness of our propose SEAM, we evaluate the CAMs generated by both single-scale and multi-scale test. Tab. 5 illustrates that our proposed model outperforms baseline with higher peak performance in both single- and multi-scale test.

Source of Improvement: The improvement of CAM quality mainly stems from more complete activation coverage or fewer over-activated regions. To further analyze the improvement source of our SEAM, we define two metrics to represent the degree of under-activation and over-activation:

$$m_{FN} = \frac{1}{C-1}\sum_{c=1}^{C-1}\frac{FN_c}{TP_c}, \quad (13)$$

$$m_{FP} = \frac{1}{C-1}\sum_{c=1}^{C-1}\frac{FP_c}{TP_c}. \quad (14)$$

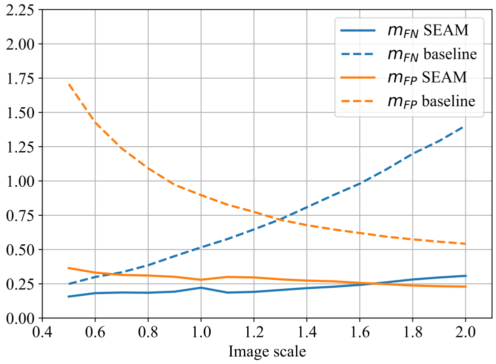

Figure 6. The curves of over-activation and under-activation. Lower m_{FN} curve represents fewer under-activation regions, and lower m_{FP} represents fewer over-activated regions.

Here TP_c denotes the pixel number of true positive prediction of class c, FP_c and FN_c denote false positive and false negative respectively. These two metrics exclude the background category since the prediction of background is inverse to the foreground. Specifically, if there are more false negative regions when CAMs do not have complete activation coverage, m_{FN} will have a larger value. Relatively, larger m_{FP} means there are more false positive regions, meaning that CAMs are over-activated.

Based on these two metrics, we collect the evaluation results from both baseline and our SEAM, then plot the curves in Fig. 6 which illustrates a large gap between baseline and our method. The SEAM achieves lower m_{FN} and m_{FP}, meaning that the CAMs generated by our approach have more complete activation coverage and fewer over-activated pixels. Therefore, the prediction maps of SEAM better fit the shape of ground truth segmentation. Moreover, the curves of SEAM are more consistent than baseline model over different image scales, which proves that the equivariance regularization works during network learning and contributes to the improvement of CAM.

4.3. Comparison with State-of-the-arts

To further elevate the accuracy of pseudo pixel-level annotations, we follow the work of [2] to train an AffinityNet based on our revised CAM. The final synthesized pseudo labels achieve 63.61% mIoU on PASCAL VOC 2012 train set. Then we train the classical segmentation model DeepLab [5] with ResNet38 backbone on these pseudo labels in full supervision to achieve final segmentation results. Tab. 6 shows the mIoU of each class on *val* set and Tab. 7 gives more experiment results of previous approaches. Compared to the baseline method, our SEAM significantly improves the performance on both *val* and *test* set with the same training setting. Moreover, our method presents the state-of-the-art performance using only image-level labels on PASCAL VOC 2012 *test* set. Noting that

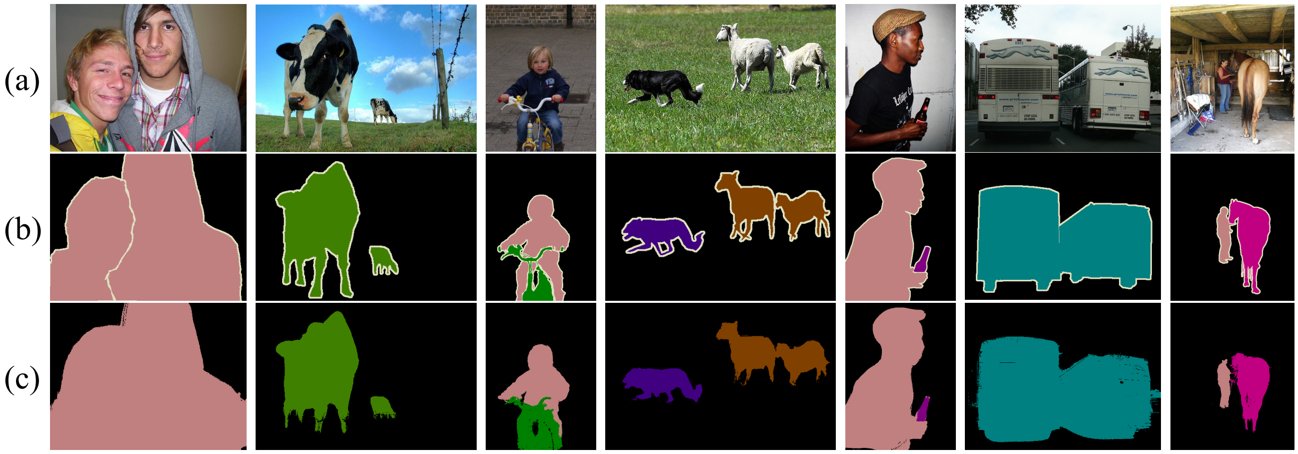

Figure 7. Qualitative segmentation results on PASCAL VOC 2012 *val* set. (a) Original images. (b) Ground truth. (c) Segmentation results predicted by DeepLab model retrained on our pseudo labels.

model	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbk	person	plant	sheep	sofa	train	tv	mIoU
CCNN [25]	68.5	25.5	18.0	25.4	20.2	36.3	46.8	47.1	48.0	15.8	37.9	21.0	44.5	34.5	46.2	40.7	30.4	36.3	22.2	38.8	36.9	35.3
MIL+seg [27]	79.6	50.2	21.6	40.9	34.9	40.5	45.9	51.5	60.6	12.6	51.2	11.6	56.8	52.9	44.8	42.7	31.2	55.4	21.5	38.8	36.9	42.0
SEC [19]	82.4	62.9	26.4	61.6	27.6	38.1	66.6	62.7	75.2	22.1	53.5	28.3	65.8	57.8	62.3	52.5	32.5	62.6	32.1	45.4	45.3	50.7
AdvErasing [32]	83.4	71.1	30.5	72.9	41.6	55.9	63.1	60.2	74.0	18.0	66.5	32.4	71.7	56.3	64.8	52.4	37.4	69.1	31.4	58.9	43.9	55.0
AffinityNet [2]	88.2	68.2	30.6	81.1	**49.6**	61.0	77.8	66.1	75.1	29.0	66.0	40.2	**80.4**	62.0	70.4	73.7	42.5	70.7	**42.6**	**68.1**	51.6	61.7
Our SEAM	**88.8**	**68.5**	**33.3**	**85.7**	40.4	**67.3**	**78.9**	**76.3**	**81.9**	**29.1**	**75.5**	**48.1**	79.9	**73.8**	**71.4**	**75.2**	**48.9**	**79.8**	40.9	58.2	**53.0**	**64.5**

Table 6. Category performance comparisons on PASCAL VOC 2012 *val* set with only image-level supervision.

Methods	Backbone	Saliency	*val*	*test*
CCNN [25]	VGG16		35.3	35.6
EM-Adapt [24]	VGG16		38.2	39.6
MIL+seg [27]	OverFeat		42.0	43.2
SEC [19]	VGG16		50.7	51.1
STC [33]	VGG16	√	49.8	51.2
AdvErasing [32]	VGG16	√	55.0	55.7
MDC [34]	VGG16	√	60.4	60.8
MCOF [36]	ResNet101	√	60.3	61.2
DCSP [4]	ResNet101	√	60.8	61.9
SeeNet [15]	ResNet101	√	63.1	62.8
DSRG [16]	ResNet101	√	61.4	63.2
AffinityNet [2]	ResNet38		61.7	63.7
CIAN [10]	ResNet101	√	64.1	64.7
IRNet [1]	ResNet50		63.5	64.8
FickleNet [21]	ResNet101	√	64.9	65.3
Our baseline	ResNet38		59.7	61.9
Our SEAM	ResNet38		64.5	65.7

Table 7. Performance comparisons of our method with other state-of-the-art WSSS methods on PASCAL VOC 2012 dataset.

our performance elevation stems from neither the larger network structure nor the improved saliency detector. The performance improvement mainly comes from the cooperation of additional self-supervision and PCM, which produces better CAMs for the segmentation task. Fig. 7 shows some qualitative results, which verify that our method works well on both large and small objects.

5. Conclusion

In this paper, we propose a self-supervised equivariant attention mechanism (SEAM) to narrow the supervision gap between fully and weakly supervised semantic segmentation by introducing additional self-supervision. The SEAM embeds self-supervision into weakly supervised learning framework by exploiting equivariant regularization, which forces CAMs predicted from various transformed images to be consistent. To further improve the ability of network for generating consistent CAMs, a pixel correlation module (PCM) is designed, which refines original CAMs by learning inter-pixel similarity. Our SEAM is implemented by a siamese network structure with efficient regularization losses. The generated CAMs not only keep consistent over different transformed inputs but also better fit the shape of ground truth masks. The segmentation network retrained by our synthesized pixel-level pseudo labels achieves state-of-the-art performance on PASCAL VOC 2012 dataset, which proves the effectiveness of our SEAM.

Acknowledgement: This work was partially supported by National Key R&D Program of China (No. 2017YFA0700800), CAS Frontier Science Key Research Project (No. QYZDJ-SSWJSC009) and Natural Science Foundation of China (Nos. 61806188, 61772496).

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[2] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[3] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. 2018.

[4] Arslan Chaudhry, Puneet K Dokania, and Philip HS Torr. Discovering class-specific pixels for weakly-supervised semantic segmentation. In *Proc. British Machine Vision Conference (BMVC)*, 2017.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *Proc. International Conference on Learning Representations (ICLR)*, 2015.

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

[7] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2015.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[9] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2015.

[10] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Cian: Cross-image affinity net for weakly supervised semantic segmentation. *arXiv preprint arXiv:1811.10842*, 2018.

[11] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[12] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. Neural Information Processing Systems (NIPS)*, 2014.

[14] Bharath Hariharan, Pablo Arbelaez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2011.

[15] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In *Proc. Neural Information Processing Systems (NIPS)*, 2018.

[16] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[17] Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, and Jan Kautz. Scops: Self-supervised co-part segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[18] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[19] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *Proc. European Conference on Computer Vision (ECCV)*, 2016.

[20] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *Proc. European Conference on Computer Vision (ECCV)*, 2016.

[21] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[22] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[23] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[24] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2015.

[25] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2015.

[26] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[27] Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[28] 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 *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. Neural Information Processing Systems (NIPS)*, 2017.

[30] Paul Vernaza and Manmohan Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[31] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[32] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[33] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE Transactionson Pattern Analysis and Machine Intelligence (TPAMI)*, 39(11):2314–2320, 2017.

[34] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

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

[36] Wang Xiang, You Shaodi, Li Xi, and Ma Huimin. Weakly-supervised semantic segmentation by iteratively mining common object features. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[37] Yuhui Yuan and Jingdong Wang. Ocnet: Object context network for scene parsing. *arXiv preprint arXiv:1809.00916*, 2018.

[38] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[39] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.