Self-supervised Scale Equivariant Network for Weakly Supervised Semantic Segmentation

Yude Wang[1,2], **Jie Zhang**[1], **Meina Kan**[1], **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, jie.zhang}@vipl.ict.ac.cn, {kanmeina, sgshan, xlchen}@ict.ac.cn

Abstract

Weakly supervised semantic segmentation has attracted much research interest in recent years considering its advantage of low labeling cost. Most of the advanced algorithms follow the design principle that expands and constrains the seed regions from class activation maps (CAM). As well-known, conventional CAM tends to be incomplete or over-activated due to weak supervision. Fortunately, we find that semantic segmentation has a characteristic of spatial transformation equivariance, which can form a few self-supervisions to help weakly supervised learning. This work mainly explores the advantages of scale equivariant constrains for CAM generation, formulated as a self-supervised scale equivariant network (SSENet). Specifically, a novel scale equivariant regularization is elaborately designed to ensure consistency of CAMs from the same input image with different resolutions. This novel scale equivariant regularization can guide the whole network to learn more accurate class activation. This regularized CAM can be embedded in most recent advanced weakly supervised semantic segmentation framework. Extensive experiments on PASCAL VOC 2012 datasets demonstrate that our method achieves the state-of-the-art performance both quantitatively and qualitatively for weakly supervised semantic segmentation. Code has been made available[1].

Introduction

Deep convolutional neural networks have achieved remarkable successes in recent years with the support of massive labeled data. While the research moves forward slowly burdened with expensive data annotation processes, especially for the semantic segmentation, whose annotation requirement is much more complex and expensive than the classification and detection tasks. Therefore, some weakly supervised semantic segmentation works focus on training network with lower-level supervision, such as bounding boxes (Dai, He, and Sun 2015; Khoreva et al. 2017), scribbles (Lin et al. 2016; Vernaza and Chandraker 2017) and points (Bearman et al. 2016), which are much cheaper to be labeled than pixel-level segmentation mask. Image-level category label is the most common used supervision since there are already many large-scale classification datasets, such as ImageNet (Deng et al. 2009). To the best of our knowledge, almost all the research of image-level weakly supervised semantic segmentation methods are based on Class Activation Maps (CAM) (Zhou et al. 2016), which generate a roughly activated feature map to locate objects spatial positions. The original version of CAM has relatively complete coverage on small objects, while with the increasing of object size, the activated regions shrink into the most discriminative part, e.g. the head of a dog and the wheel of a car. The incomplete activation maps as pseudo segmentation labels heavily damage the training of segmentation network, leading to severe performance degradation.

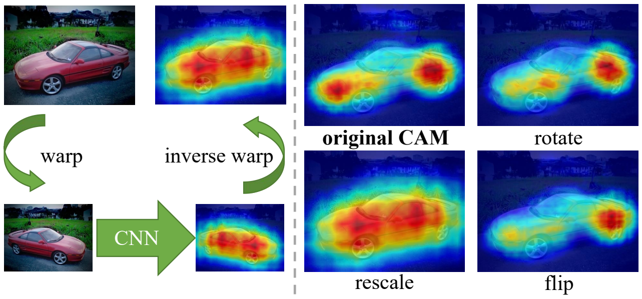

Figure 1: Comparing the inconsistency between original CAM with other CAMs warped by various spatial transformations, following the pipeline on the left.

The ideal segmentation network should be an affine transformation equivariant mapping function. As shown in Fig. 1, the input image is warped by some affine transformations and then fed into CNN to achieve CAMs. The generated CAMs are then inverse warped to meet with the CAM generated from the original image. The wrapped CAMs should keep the same with the original CAM. However, Fig. 1 illustrates that there is an inconsistency between the wrapped CAMs and the origin one, especially by the rescaling operation. The equivariant constraint has been implicitly used in fully supervised segmentation that the pixel-level labels always keep consistent with corresponding pixels during spatial transformation augmentation. However, the class activation maps learned by image-level supervision are not equiv-

[1] https://github.com/YudeWang/SSENet-pytorch

ariant in most cases, which may hurt the generalization performance of CAM based weakly supervised semantic segmentation methods.

Considering that the conventional CAM has severe inconsistency by rescaling transformation that it covers more background regions on small object and covers fewer foreground regions on large objects, we turn to regularize the class activation map of different scales to refine each other. In this paper, we propose a novel two branch self-supervised scale equivariant network (SSENet) to overcome the drawback of CAM mentioned above by a self-supervision framework. The network constrains the activated feature maps to be scale equivariant, i.e. the activation of images keep consistent on various scales. With the scale equivariant regularization (SER), the CAM consistency is significantly improved as shown in Fig. 2. The regularization is effective and easy to be employed on any CAM-based algorithm for weakly supervised semantic segmentation. Benefited from the improved CAM, the performance of weakly supervised semantic segmentation will be further improved.

The main contributions can be summarized as follows:

- We propose a novel scale equivariant regularization (SER) to narrow the consistency gap between the CAMs generated from various scale images, leading to significant improvement on CAMs.
- We propose a novel self-supervised scale equivariant network (SSENet) architecture, which is the first try to utilize self-supervised learning for image-level weakly supervised semantic segmentation.
- Experiments on PASCAL VOC 2012 dataset (Everingham et al. 2015) demonstrate the outstanding performance of our SSENet comparing with state-of-the-arts.

Related Work

Weakly Supervised Semantic Segmentation

Although fully supervised semantic segmentation algorithms (Chen et al. 2015; Long, Shelhamer, and Darrell 2015) have achieved great successes in recent years, the pixel-level annotations are expensive to collect, resulting in that more and more weakly supervised approaches are proposed and studied to alleviate this practical problem.

Image-level supervision means only object category labels are available during training time. The most fundamental work CAM (Zhou et al. 2016), trains an image classification network and multiples the weight of fully connect layer on the feature map during inference for roughly object localization. Early approaches adopt various strategies, such as EM algorithm (Papandreou et al. 2015) and multiple instance learning (Pinheiro and Collobert 2015) to achieve pseudo segmentation labels. SEC (Kolesnikov and Lampert 2016) proposes three principles for the task, which selects confident initial seeds from CAM, expands activate regions by global weighted rank pooling and constrains segmentation boundary considering color information. The CAM generation network always activates on the discriminative parts of the objects, remaining a challenge to predict segmentation mask covering the entire foreground object. The adversarial erasing strategy is widely employed in (Wei et al. 2017a; Hou et al. 2018) to solve this case by erasing discriminative regions and mining others. FickleNet (Lee et al. 2019) randomly dropout the weight of convolution at each position to discover more class activated regions. Another expanding method from reliable seed regions is random walk with transition matrix (Ahn and Kwak 2018; Ahn, Cho, and Kwak 2019). The matrix can be derived from AffinityNet supervised by classified pixel pair (Ahn and Kwak 2018) or learned from the class boundary map (Ahn, Cho, and Kwak 2019).

Self-supervised learning

Comparing to fully supervised network training with massive annotated data, self-supervised learning is a candidate solution to learn more robust visual features without any additional annotation cost. Most of the self-supervised learning algorithms propose pretext tasks, which are predefined to generate controlled supervision signals from the input domain as optimization directions for deep neural networks. The task-related feature representations will be learned through this process (Jing and Tian 2019). To some extent, These self-supervised pre-trained features bring comparable performance improvement with ImageNet pre-trained model (Doersch, Gupta, and Efros 2015).

Here are some self-supervised research works based on various pretext tasks. The most famous ones are generative models, e.g. generative adversarial networks (Goodfellow et al. 2014). These models learn the feature distribution from a large unlabeled dataset by adversarial training. There are also some discriminative models learning with pretext tasks on static images, e.g. relative position prediction (Doersch, Gupta, and Efros 2015) and spatial transformation prediction (Gidaris, Singh, and Komodakis 2018).

Scale Equivariance

As for the research of scale invariance and scale equivariance, most of the works focus on designing special network architecture to preserve the scale invariance or scale equivariance (Kanazawa, Sharma, and Jacobs 2014; Worrall and Welling 2019). Since the sizes of convolution kernels are discrete, it is hard to perfectly achieve scale equivariance by refining network architecture. It still leaves a long way to go. In this paper, our proposed SSENet resorts to constraining the activation map consistency on various scales during network training instead of designing the equivariant module. The proposed framework effectively preserves the scale equivariance of CAM, which significantly improves the generated pseudo labels for weakly supervised semantic segmentation problem.

Approach

In this section, we will carefully present the idea of the proposed two-branch self-supervised scale equivariant network (SSENet). Firstly we will illustrate the scale inconsistency problem of class activation map in weakly supervised semantic segmentation and analyze the essential cause

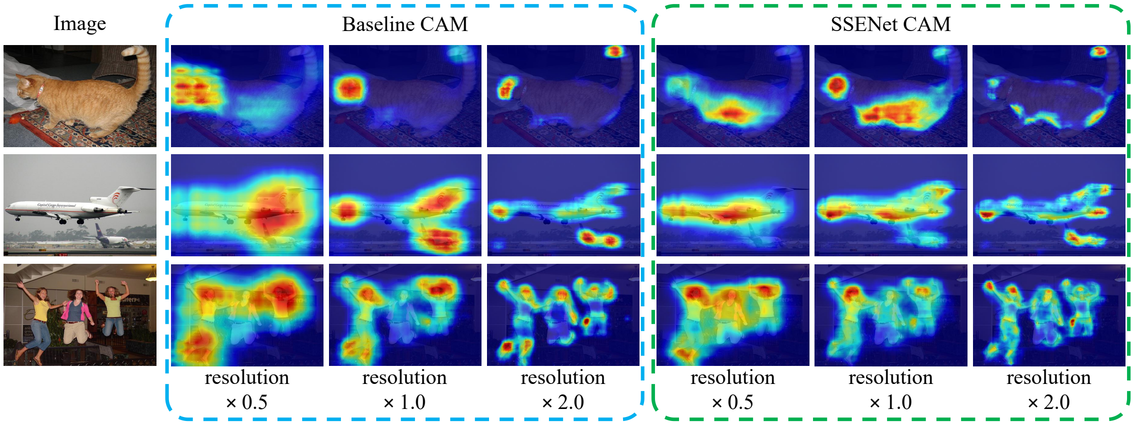

Figure 2: Visualization of CAMs on various scales. The left column shows the original images. The middle part shows the conventional CAMs and the right part is the CAMs generated by our SSENet. As shown in the figure, the CAMs generated by our SSENet have better consistency on various scales.

of it. Secondly, we will describe the details of the self-supervised regularization which improves the scale consistency for class activation map. Finally, we will show how to integrate scale equivariant regularization into a two-branch network for weakly supervised semantic segmentation.

Observation

To figure out the semantic regions in a static image, most of the approaches follow the work (Zhou et al. 2016) by training a CNN with the image-level label to get class activation map (CAM). It is well known to all that CAMs always located in the most discriminative part, especially for the large objects. It is hard to cover the entire object regions. Meanwhile, as shown in Fig. 2, CAM highlights a relatively expanded activation map for small object covering too many background regions. The inconsistency extremely damages the network performance for solving weakly supervised semantic segmentation problem. Although researches are focusing on finding more discriminative regions, e.g. adversarial erasing strategies (Wei et al. 2017a) remove the most highlighted parts to activate more class-related feature regions. While it remains a challenge to reduce those over-activated background regions around small objects. Comparing with fully supervised semantic segmentation, image-level supervision is too weak to determine the object boundaries, additional supervisions or regularizations should be employed on the network to elevate performance.

For the reason that the training process of CAM always excludes the background category, it becomes a tricky problem to generate pseudo segmentation mask which contains the background. One of the simplest methods is using a hard threshold to separate the foreground and the background. However, as shown in Fig. 2 that the best threshold of CAM is different on various scales. The threshold parameter should be large enough to exclude those over-activated regions on small objects, while it should be relatively small to include more activation parts on incomplete CAM for large objects. Although there are some post-processing approaches such as dense CRF (Chen et al. 2018) fine-tuning the contour of activation regions to some extent by color constraints, it can not solve the problem fundamentally. It is hoped that the contour should be well determined by a scale equivariant activation map, making it more general for objects of various scales. Therefore, it is reasonable to employ self-supervised scale equivariance constrain as auxiliary supervision to help image-level weakly supervised network training.

Self-supervised Regularization

The fully supervised semantic segmentation task can be generally formulated as $f_{\theta_s}(x) = y$, where x denotes the input image with the corresponding segmentation ground truth mask y. $f_\theta(\cdot)$ denotes the CNN based nonlinear mapping function. While for the image-level weakly supervised semantic segmentation, the network is attached with additional global pooling function $P(\cdot)$ to solve classification task as $P(f_{\theta_c}(x)) = l$, where l is image category label of x. Ideally, the network parameters θ_s and θ_c should keep the same if the fully supervision task and weakly supervision task have the same optimization objective.

Random scale input images is a common practice of data augmentation for network training, which can always improve the inference performance. Suppose there is an affine transformation matrix A, the fully supervised segmentation task requires the equivariance of affine transformation as $f_{\theta_s}(Ax) = Ay$, while the weakly supervision task focusing on the mapping invariance as $P(f_{\theta_c}(Ax)) = l$. The invariance mainly caused by pooling function $P(\cdot)$, but there is no explicit equivariance constrain for $f_{\theta_c}(\cdot)$. The different optimization objectives make it hard to guarantee the networks with different supervisions have the same convergence point, and it is nearly impossible to achieve θ_s with only image-level label l. Image-level weakly supervised semantic segmentation needs additional supervisions to guide the optimization direction.

Under the scenario of weakly supervised learning, only

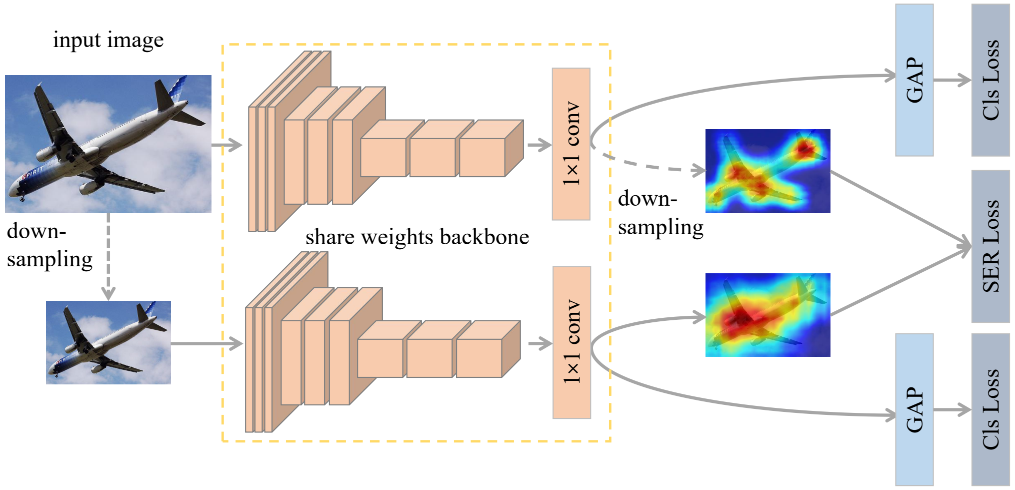

Figure 3: The two-branch architecture of SSENet. The SSENet takes the original and the downsampled version of images as input, endeavoring to preserve the consistency of CAMs from each branch by scale equivariant regularization loss (**SER Loss**). **GAP** denotes the global average pooling layer. **Cls Loss** denotes the multi-label classification loss.

self-supervised labels and handcrafted constraints are available to narrow the gap of optimal solution between classification and semantic segmentation tasks. We resort to the self-supervised learning, proposing scale equivariant regularization (SER) to improve the CAM quality.

$$\min_{\theta} \sum_{i} \frac{||P(f_\theta(x_i)) - l_i|| + ||P(f_\theta(Ax_i)) - l_i||}{2} + \eta R_i, \tag{1}$$

$$R_i = ||f_\theta(Ax_i) - Af_\theta(x_i)||. \tag{2}$$

Our proposed cost function with scale equivariant regularization is given as Eq. 1. i denotes the sample index and η is the parameter to control the influence of regularization which is set to 1 in the following experiments without careful tuning. The formulation of the scale equivariant regularization is defined in Eq. 2, where A is the matrix form of bilinear interpolation. The elements in matrix A is predefined, e.g. $1/2$ downsampling operation, and fixed in the training process. Since the input images are randomly rescaled during data augmentation preprocessing, SER endeavors to preserve the scale equivariance on the overall scale range.

Moreover, the extended version of scale equivariant regularization is that A can be any spatial linear transformation, e.g. flip and rotation. In order to verify the effectiveness of the proposed equivariant regularization, our work only focuses on scaling inconsistency. As shown in Fig. 2, training with scale equivariant regularization obviously increase the similarity of class activation map on various scales. And the trend of CAM contour turns to meet the object boundary closely. At the same time, the regularized activation maps have more complete coverage of the entire object comparing to the unconstrained ones, which demonstrates that the scale equivariance regularization has the same effect as the methods of object region mining.

Self-supervised Scale Equivariant Network

We propose a novel weight-shared two-branch network to employ the scale equivariant regularization (SER) for network training. As shown in Fig. 3, the randomly selected RGB image, and its downsampled copy are fed into the two branch network respectively. At the end of the backbone networks, the $C - 1$ channels feature maps are achieved, where $C - 1$ is the number of categories excluding the background. And the feature maps are also known as class activation maps. Global average pooling layer is attached to the feature map with multi-label classification loss as supervision. Considering that the scales of CAM from two branch are not the same, we downsample the CAM output from the large branch by bilinear interpolation to keep scale consistent with the one from the other branch. L_2 distance is used to measure the gap between these two branch CAMs, working as the scale equivariant regularization to constrain the activation consistency during the training process. Finally, the loss for network training is the weighted sum of classification losses mean and SER item with $\eta = 1$ in Eq. 1.

During the test phase, we preserve one branch as the inference network to obtain the final class activation maps. Noting that the two branches of the network have shared weights, it is equivalent to preserve the parameters from the large or small branch. Moreover, the additional background score maps will be concatenated with $C - 1$ channel feature maps manually to form preliminary segmentation results. The background score maps are defined as follows.

$$M_c = \begin{cases} \alpha & c \text{ is background} \\ \frac{ReLU(\hat{y}_c - \epsilon)}{\max ReLU(\hat{y}_c) + \epsilon} & \text{others} \end{cases} \tag{3}$$

Where $\hat{y}_c$ denotes the predicted activation map of class c and ϵ is set as 10^{-5} to avoid dividing zero. M_c is the normalized activation map of class c. α denotes the parameter used to control the background confidence score which is set to 0.2 in our experiments. We choose ResNet-38 as the backbone network in all experiments, following the setting of (Ahn and Kwak 2018). Dense CRF (Chen et al. 2018) is attached as a post-processing step to refine the contour of CAM to be more close with the object boundary.

To further improve the performance of weakly supervised semantic segmentation, we follow the pipeline of (Ahn and Kwak 2018). In shortly, several pixel pairs are sampled based on the improved CAM $(h \times w)$ to train an AffinityNet, and the pixel affinity matrix $(hw \times hw)$ is calculated by the feature map of AffinityNet. Then the improved CAM is resized into $1 \times hw$, multiplied with pixel affinity matrix several times, which is named as random walk step, and CAM vector is resized back to $h \times w$ eventually as the pseudo label. Finally, a classical semantic segmentation model DeepLab is trained by these pseudo labels. We will carefully investigate the performance improvement of each step in next section to show the benefits of our SSENet to weakly supervised semantic segmentation.

Experiments

In order to illustrate the superiority of our proposed SSENet, we conduct some weakly supervised semantic segmentation experiments from several aspects. SSENet is also embedded into advanced weakly supervised semantic segmentation framework without using any additional data, achieving outstanding performance with other state-of-the-art methods.

Implementation Details

To evaluate the effectiveness of our SSENet, we adopt PASCAL VOC 2012 benchmark (Everingham et al. 2015) which contains 20 object categories and another background category. With the additional annotation of SBD (Hariharan et al. 2011), the common setting of fully supervised semantic segmentation task takes 10,582 images as the augmented training set, 1,449 for validation and 1,456 for testing. Our experiments follow the same dataset partition while only image-level classification labels are provided during network training, and take the mean intersection of union (mIoU) as evaluation metric as well as other previous works.

As for training settings, the backbone network used in our experiments is the modified version of ResNet38[2] pretrained on ImageNet, which takes two 3×3 convolution layers in each residual block instead of the bottleneck structure, removing original global average pooling and fully connected layers. As well known that ResNet groups several residual blocks into one level, using stride 2 convolution layer at the beginning. We replace the stride convolution of the last two levels by dilation convolution with rate 2 and 4 in last two levels respectively to keep the same network receptive field. Additional 1×1 convolution is attached to the end of the network as the pixel-wise classifier, followed by a global average pooling layer which pools the feature map into feature vector for the classification task. The network is trained on 4 Titan-xp GPUs with batch size 8 for 15 epochs. The initial learning rate is 0.01 and the training schedule follows the poly policy that $lr = lr_{init} * (1 - \frac{itr}{max\_itr})^{\gamma}$, where $\gamma = 0.9$ in our experiments. The input images are randomly rescaled into [448, 768] on the longest edge, then randomly cropped by 448×448 and fed into the large branch of the network. The regularization weight in Eq. 1 is set as $\eta = 1$ in all our experiments. During inference, only one branch network remains since these two branches have shared weights. Flip and multi-scale test are adopted to improve performance.

Model	CAM (mIoU)	CAM+rw (mIoU)
Baseline	47.3%	58.8%
SSENet (0.6)	48.5%	61.5%
SSENet (0.5)	48.9%	61.7%
SSENet (0.4)	49.4%	61.8%
SSENet (0.3)	**49.8%**	**62.1%**
SSENet (0.2)	49.4%	61.7%

Table 1: Comparison between baseline model and SSENet with various branch downsampling rates (given in parentheses). We evaluate the generated pseudo labels from CAM generation step (**CAM**) and the following random walk step (**CAM+rw**) on PASCAL VOC 2012 train set.

Model	scaling range	CAM (mIoU)
Baseline	[448, 768]	47.3%
Baseline	[224, 768]	46.6%
SSENet (0.5)	[448, 768]	48.9%

Table 2: Pseudo label comparison between baseline model and SSENet (0.5 downsampling rate) with different scale augmentation ranges on PASCAL VOC 2012 train set.

Ablation Study

Branch Downsampling Rate Branch downsampling rate of the SSENet is a significant super parameter to control the effectiveness of scale equivariance regularization. We make some experiments on various branch downsampling rate which are evaluated by multi-scale and flip test. Tab. 1 shows networks trained with SER work better on the performance of CAM than baseline models over all branch downsampling rates. And the mIoU improvements are further boosted by employing random walk step. When selecting branch downsampling rate as 0.3, the mIoU of CAM produced by SSENet achieves 2.5% improvement and 3.3% after random walk step. Besides, they are further utilized as the pseudo labels to train advanced fully-supervised semantic segmentation network, leading to the performance improvement of the whole solution framework.

Scale Augmentation Range Rescaling is a basic data augmentation method for network training, which elevates the robustness of the network to different image scales. In our SSENet, the downsampling branch resizes the input images into a smaller scale, enlarging the scale augmentation range to some extent. To verify whether the performance improvement comes from the larger scaling range, we train the baseline model and SSENet with different scale augmentation range, and the Tab. 2 summarizes the experiment results. It shows that SSENet with 0.5 downsampling rate works better

[2] Model A1 version in (Wu, Shen, and Van Den Hengel 2019)

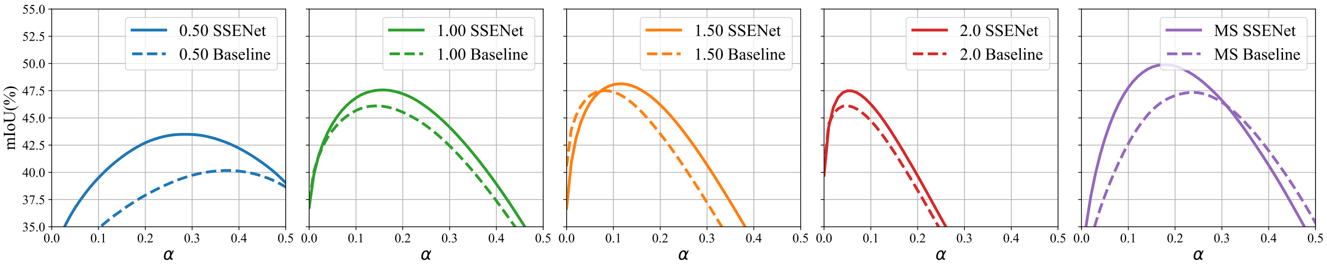

Figure 4: The evaluation of generated pseudo labels on PASCAL VOC 2012 train set by baseline and SSENet. The decimals in the legends are the scaling rate of single-scale test, **MS** denotes multi-scale test, and α is the background confidence threshold.

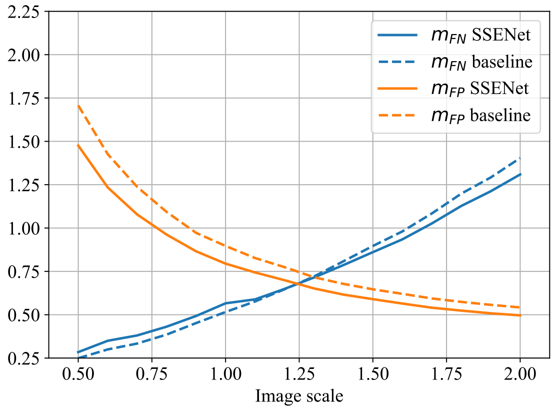

Figure 5: The model with scale equivariant regularization activates fewer background pixels (lower m_{FP}) and covers more object parts (lower m_{FN} at right side) than traditional CAM method. The figure is best viewed on screen.

than baseline model when scaling range is [448, 768], noting that the scale range of SSENet downsampling branch is [224, 384] at the same time. Besides, we train a baseline model with [224, 768] scale augmentation, which expands the range to include the image scales from SSENet downsampling branch. As seen in the Tab. 2, the baseline model degenerates by 0.7% mIoU comparing to that trained with [448, 768] scale augmentation, illustrating that the performance contribution of SSENet mainly comes from scale equivariant regularization rather than larger scaling range.

Multi-scale Test In many previous works, it is a common practice to compute multiple CAMs from multiple rescaled images and aggregate them to produce more accurate activation maps during test. In this paragraph, we evaluate SSENet and baseline model with both various single-scale test and multi-scale test. As shown in Fig. 4, the CAMs of SSENet achieve higher mIoU than the baseline model with all kinds of single-scale test setting. Moreover, with multi-scale aggregation, our SSENet achieves a further improvement which also significantly beats the baseline. It demonstrates that our SSENet is effective for various scales and can be further improved with multi-scale aggregation.

Source of Performance Improvement Besides the visualization results shown in Fig. 2, we prefer quantitative evaluations to investigate the source of performance improvement. Considering the evaluation metric is that

$$mIoU = \frac{1}{C}\sum_{c=0}^{C}\frac{TP_c}{TP_c + FN_c + FP_c}, \tag{4}$$

where C is category number and TP_c, FN_c, FP_c denotes the true positive, false negative, false positive predicted regions of each class respectively. The phenomenon in Fig. 2 shows that for small input images, the activation maps overcover the object regions, leading to a lower proportion of FN/FP. When the input images are resized into large scale, the activation regions shrink into the most discriminative parts, causing a higher proportion of FN/FP. To further analyze the contributions of these two parts, we define another two metrics

$$m_{FN} = \frac{1}{C-1}\sum_{c=1}^{C-1}\frac{FN_c}{TP_c}. \tag{5}$$

$$m_{FP} = \frac{1}{C-1}\sum_{c=1}^{C-1}\frac{FP_c}{TP_c}. \tag{6}$$

Note that the background category is excluded since the background activation region is reverse to the foreground categories. To simplify the analysis, we remove the background in these evaluation metrics. The Fig. 5 shows the curve of m_{FN} and m_{FP}, based on the CAMs generated by the baseline model and SSENet in terms of various single-test scales. The tendency of these curves meets the claim that CNN generates rough CAM over-covering object when the input image is small, i.e., the prediction contains more false positive regions out of object regions. While the activation zone shrinks into the discriminative part with a larger input image and most parts of the object body are not activated, i.e., the prediction contains more false negative regions. Comparing the baseline and SSENet, the m_{FP} curve demonstrates that the CAM generated by SSENet is more compact with fewer over-cover regions. Besides SSENet accurately activates more foreground regions only on large-size images since its m_{FN} curve is higher than baseline at the right side of the axis while keeping lower at the left side. In shortly, the source of performance improvement mainly

Methods	Supervision	Saliency	mIoU(val)	mIoU(test)
EM-Adapt (Papandreou et al. 2015)	Image-level	-	38.2%	39.6%
MIL (Pinheiro and Collobert 2015)	Image-level	-	42.0%	40.6%
SEC (Kolesnikov and Lampert 2016)	Image-level	-	50.7%	51.7%
AffinityNet (Ahn and Kwak 2018)	Image-level	-	**61.7%**	**63.7%**
STC (Wei et al. 2017b)	Image-level	√	49.8%	51.2%
AdvErasing (Wei et al. 2017a)	Image-level	√	55.0%	55.7%
SeeNet (Hou et al. 2018)	Image-level	√	63.1%	62.8%
DSRG (Huang et al. 2018)	Image-level	√	61.4%	63.2%
FickleNet (Lee et al. 2019)	Image-level	√	64.9%	65.3%
What's Point (Bearman et al. 2016)	Point	-	46.0%	43.6%
RAWK (Vernaza and Chandraker 2017)	Scribble	-	61.4%	-
ScribbleSup (Lin et al. 2016)	Scribble	-	63.1%	-
BoxSub (Dai, He, and Sun 2015)	Bbox	-	62.0%	64.6%
SDI (Khoreva et al. 2017)	Bbox	-	65.7%	67.5%
SSENet	Image-level	-	**63.3%**	**64.9%**

Table 3: Comparison with state-of-the-art weakly supervised approaches on both PASCAL VOC validation and test set. Our method is learned by image-level labels without extra supervisions.

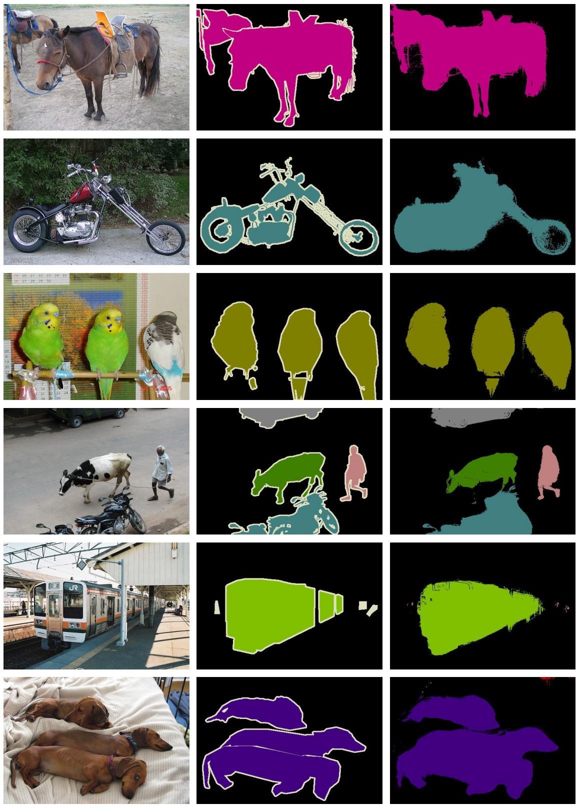

Figure 6: More segmentation results based on our approach. For each tuple, the left one is the original image, the middle is ground truth label and the right is the prediction of the final segmentation model based on our method. Our proposed weakly supervised method not only has complete segmentation coverage of large objects but also meets object boundary details.

comes from less over-activated regions by scale equivariant regularization.

Comparisons with State-of-the-arts

To further elevate the weakly supervised semantic segmentation network performance, we follow the work of Affinity (Ahn and Kwak 2018) to expand and constrain the class activation maps achieved by our proposed SSENet. Moreover, we reimplement DeepLab with ResNet38 as the backbone, using modified activation maps as pseudo labels for semantic segmentation training. The Tab. 3 illustrates that the final result of our SSENet has significant improvement than AffinityNet baseline. The performance elevation mainly stems from the improved CAMs and the more accurate pseudo segmentation labels. Besides, the method based on SSENet even beats some advanced approaches which are embedded with additional off-the-shelf saliency methods. Moreover, our method also achieves comparable performance with the state-of-the-art weakly supervised semantic segmentation approaches based on stronger supervisions like point, scribble and bounding box.

Conclusion

In this paper, we resort to the self-supervision regularization for weakly supervised semantic segmentation. We propose scale equivariant regularization (SER) to deal with the inconsistency of network activation map on various image sizes. With the SER, the class activation maps from scale augmented images keep the same after rescaled into the same size. Based on this regularization, we design a two-branch self-supervised scale equivariant network (SSENet) for class activation map learning with only image-level supervision. The network learns more discriminative regions on large objects and overcomes the phenomenon of over-activated on small objects. We evaluate the proposed method on PASCAL VOC 2012 dataset and the results demonstrate that our approach has outstanding performance than other state-of-the-art weakly supervised methods.

References

[Ahn and Kwak 2018] Ahn, J., and Kwak, S. 2018. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *CVPR*, 4981–4990.

[Ahn, Cho, and Kwak 2019] Ahn, J.; Cho, S.; and Kwak, S. 2019. Weakly supervised learning of instance segmentation with inter-pixel relations. In *CVPR*.

[Bearman et al. 2016] Bearman, A.; Russakovsky, O.; Ferrari, V.; and Fei-Fei, L. 2016. Whats the point: Semantic segmentation with point supervision. In *ECCV*, 549–565. Springer.

[Chen et al. 2015] Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2015. Semantic image segmentation with deep convolutional nets and fully connected crfs. *ICLR*.

[Chen et al. 2018] Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2018. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI* 40(4):834–848.

[Dai, He, and Sun 2015] Dai, J.; He, K.; and Sun, J. 2015. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, 1635–1643.

[Deng et al. 2009] Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255. Ieee.

[Doersch, Gupta, and Efros 2015] Doersch, C.; Gupta, A.; and Efros, A. A. 2015. Unsupervised visual representation learning by context prediction. In *ICCV*, 1422–1430.

[Everingham et al. 2015] Everingham, M.; Eslami, S. A.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2015. The pascal visual object classes challenge: A retrospective. *IJCV* 111(1):98–136.

[Gidaris, Singh, and Komodakis 2018] Gidaris, S.; Singh, P.; and Komodakis, N. 2018. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.

[Goodfellow et al. 2014] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NIPS*, 2672–2680.

[Hariharan et al. 2011] Hariharan, B.; Arbelaez, P.; Bourdev, L.; Maji, S.; and Malik, J. 2011. Semantic contours from inverse detectors. In *ICCV*.

[Hou et al. 2018] Hou, Q.; Jiang, P.; Wei, Y.; and Cheng, M.-M. 2018. Self-erasing network for integral object attention. In *NIPS*, 549–559.

[Huang et al. 2018] Huang, Z.; Wang, X.; Wang, J.; Liu, W.; and Wang, J. 2018. Weakly-supervised semantic segmentation network with deep seeded region growing. In *CVPR*, 7014–7023.

[Jing and Tian 2019] Jing, L., and Tian, Y. 2019. Self-supervised visual feature learning with deep neural networks: A survey. *arXiv preprint arXiv:1902.06162*.

[Kanazawa, Sharma, and Jacobs 2014] Kanazawa, A.; Sharma, A.; and Jacobs, D. 2014. Locally scale-invariant convolutional neural networks. In *NIPS*.

[Khoreva et al. 2017] Khoreva, A.; Benenson, R.; Hosang, J.; Hein, M.; and Schiele, B. 2017. Simple does it: Weakly supervised instance and semantic segmentation. In *CVPR*, 876–885.

[Kolesnikov and Lampert 2016] Kolesnikov, A., and Lampert, C. H. 2016. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*, 695–711. Springer.

[Lee et al. 2019] Lee, J.; Kim, E.; Lee, S.; Lee, J.; and Yoon, S. 2019. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *arXiv preprint arXiv:1902.10421*.

[Lin et al. 2016] Lin, D.; Dai, J.; Jia, J.; He, K.; and Sun, J. 2016. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *CVPR*, 3159–3167.

[Long, Shelhamer, and Darrell 2015] Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*.

[Papandreou et al. 2015] Papandreou, G.; Chen, L.-C.; Murphy, K. P.; and Yuille, A. L. 2015. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *ICCV*, 1742–1750.

[Pinheiro and Collobert 2015] Pinheiro, P. O., and Collobert, R. 2015. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, 1713–1721.

[Vernaza and Chandraker 2017] Vernaza, P., and Chandraker, M. 2017. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *CVPR*, 7158–7166.

[Wei et al. 2017a] Wei, Y.; Feng, J.; Liang, X.; Cheng, M.-M.; Zhao, Y.; and Yan, S. 2017a. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *CVPR*, 1568–1576.

[Wei et al. 2017b] Wei, Y.; Liang, X.; Chen, Y.; Shen, X.; Cheng, M.-M.; Feng, J.; Zhao, Y.; and Yan, S. 2017b. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *TPAMI* 39(11):2314–2320.

[Worrall and Welling 2019] Worrall, D. E., and Welling, M. 2019. Deep scale-spaces: Equivariance over scale. *arXiv preprint arXiv:1905.11697*.

[Wu, Shen, and Van Den Hengel 2019] Wu, Z.; Shen, C.; and Van Den Hengel, A. 2019. Wider or deeper: Revisiting the resnet model for visual recognition. *PR* 90:119–133.

[Zhou et al. 2016] Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *CVPR*, 2921–2929.