Deform-CAM: Self-attention Based on Deformable Convolution for Weakly Supervised Semantic Segmentation

Feihong Huang[1,2], Da-Han Wang[1,2](✉), Hai-Li Ye[1,2], and Shunzhi Zhu[1,2]

[1] School of Computer and Information Engineering, Xiamen University of Technology, Xiamen, China
hfh@stu.xmut.edu.cn, {wangdh,szzhu}@xmut.edu.cn, yehl@mail.sustech.edu.cn

[2] Fujian Key Laboratory of Pattern Recognition and Image Understanding, Xiamen, China

Abstract. Weakly-supervised semantic segmentation (WSSS) receives increasing attentions from the community in recent years as it leverages the weakly annotated data to solve the problem of lacking of fully annotated data. Among them, the WSSS method based on image-level annotation is the most direct and effective while the image-level annotation is easy to obtain. Most advanced methods use class activation maps (CAM) as initial pseudo-labels, however, they only identify local regions of the target, while ignoring the context information among local regions. To solve this problem, this paper proposes a deformable convolution based self-attention module (DSAM), which introduces a pixel relationship matrix, to learn the contextual information of the image. A regularization loss is introduced to narrow the distance between the DSAM and the CAM. Compared to the base CAM method, our method can identify more target features and robustly improve the performance of WSSS without training the classifier multiple times. Our proposed method achieves the mIoU of 65.5% and 66.8% on the Pascal VOC 2012 val and test sets, respectively, demonstrating the feasibility of the method.

Keywords: Deformable convolution · Self-attention · Convolutional neural network · Weakly-supervised semantic segmentation

1 Introduction

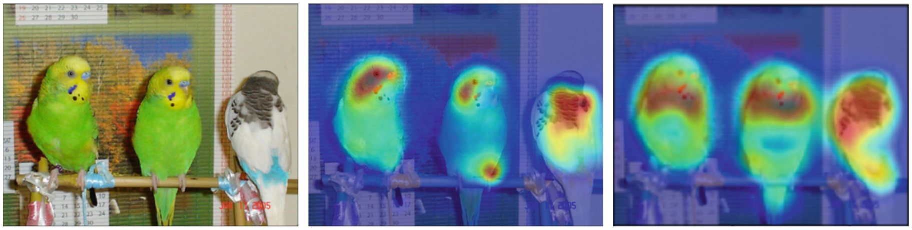
Recently, the semantic segmentation model [1–3] based on deep learning has achieved significant progress due to the power of feature learning. However, fully supervised learning [4, 5] has the major limitation of relying on pixel-level annotations, which is especially expensive for annotating and organizing pixel-based semantic segmentation. Hence, current research attempt to use some of the more accessible annotations rather than pixel-level annotations, such as bounding-box [6], graffiti [7], dot [8], image-level label [9], etc. These different types of weak labels are used for semantic segmentation. Among them, image-level tags require the least amount of annotation work and have been

Y. Wang et al. (Eds.): IGTA 2022, CCIS 1611, pp. 145–155, 2022.
https://doi.org/10.1007/978-981-19-5096-4_11

popularly used. This paper mainly studies weakly-supervised semantic segmentation based on image-level labels.

Weakly-supervised semantic segmentation (WSSS) methods with image-level supervised labels mainly learn visual features to generate pseudo labels of pixels, such as Class Activation Maps (CAM) [10], which adds a global average pooling (GAP) on top of a fully convolutional network to obtain the class localization map. However, this network structure only recognizes the most discriminative object regions and tends to obtain incorrect pixel labels for boundary pixels of objects or different regions. To solve this problem, Alexander Kolesnikov *et al.* [11] improve the CAM through three principles of "seed", "expand" and "constrain". Yunchao Wei *et al.* [12] proposed an adversarial erasing (AE) method, which completes pseudo-pixel-level labels by stitching erased images. Jiwoon Ahn *et al.* [13] proposed the AffinityNet network structure to effectively exploit the semantic similarity between adjacent coordinate pixel pairs in an image. These methods can improve the quality of CAM effectively, but they also mark the background area.

The latest research trend is to add auxiliary tasks such as consistency regularization, sub-category classification, and cross-image semantic mining, and jointly train with the classification network to make the network focus on more pixels [14–16]. Yude Wanget *et al.* [14] adopted the idea of sharing weights in Siamese networks, and proposed a SEAM network framework. Yu-Ting Chang *et al.* [15] clustered image features, generated pseudo-subclass labels for each parent class label. Guolei Sun *et al.* [16] took the cross-image as the starting point, and proposed a co-attention module and an adversarial attention module. Tong Wu *et al.* [17] proposed EDAM, which learns collaborative features for the same set of input images. However, these methods involve a complex training phase or require the introduction of additional information, such as saliency maps.



（a）　（b）　（c）

Fig. 1. Visualizations of CAMs. (a) input image. (b) conventional CAMs. (c) the Deform-CAM

To address the above problems, this paper proposes a novel framework called Deform-CAM that introduces a deformable convolution based self-attention module (DSAM) to CAM to learn the contextual information of the image and hence generate more robust pixel classification results, as shown in Fig. 1. The main characteristics of DSAM is to generate an image pixel relationship matrix based on the learned pixel context features using the self-attention mechanism. By minimizing the distance between the pixel relationship matrix and the CAM, the background noise is reduced and the target boundary is refined, leading to higher performance of semantic segmentation. Experiments on public datasets demonstrate the effectiveness of our method.

Our main contributions are as follows:

1. We propose a deformable convolution self-attention module DSAM to explore the context information of image pixels with the self-attention mechanism, which effectively reduces the background noise and refines the target boundary.
2. We propose a novel WSSS framework called Deform-CAM that combines the DSAM and CAM. The proposed Deform-CAM effectively improve the quality of CAM without complex training and the introduction of additional information.

2 Methodology

This section details the proposed Deform-CAM method. Figure 2 shows the network structure of Deform-CAM. Besides the backbone network, our network structure contains two branches: one branch is traditional CAM, and the other branch introduces DSAM to learn the correlation between image pixels. The feature maps of the stage3 and stage4 output by the backbone network, and the original image are concatenated to form the input of DSAM, which ensures that the features are more abundant. Then, DSAM uses deformable convolution to add offsets to reduce the influence of background noise at the target boundary, and applies a self-attention mechanism to explore the synergistic information between feature maps at different stages and the original image, which we call the pixel relationship matrix. Finally, the gap between the pixel relation matrix and the CAM is reduced via minimizing a contrastive loss. Compared to the based CAM, the proposed Deform-CAM method covers more target area and reduces the boundary noise.

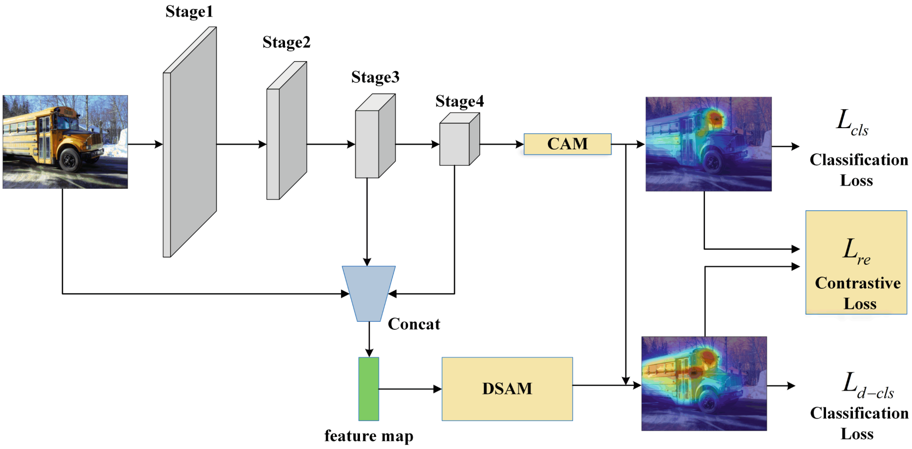


Fig. 2. The network architecture of the proposed Deform-CAM method. *Stage1-Stage4* represent the four stages of the backbone network respectively.

2.1 Class Activation Maps

First, we introduce the traditional methods of generating attention maps, CAMs. For the input image $I \in R^{3\times H\times W}$, the image I is passed to the multi-label classification network. Under the action of the feature extractor, the extracted image feature $F(I) \in R^{C\times H\times W}$ is obtained after passing through the final classifier, a set of CAMs of class activation maps can be obtained. The formula for CAMs is as follows.

$$A_c = W_c * F(I), \tag{1}$$

where $A_c \in R^{N\times H\times W}$ are the resulting CAMs. W_c is the weight of the last fully connected layer in class c.

2.2 Deformable Convolution Based Self-attention Module

In Fig. 3, we introduce the structure of the Deformable Convolution based Self-Attention Module (DSAM). DSAM consists of three parts, deformable convolution, pixel relationship matrix and channel attention module. The pixel relationship branch and the channel relationship branch are two parallel branches to capture the context information of pixels and channels, respectively.

Since the self-attention mechanism can well capture the contextual information of pixels, this paper performs self-attention processing on the underlying features of the backbone network. In addition to building the affinity matrix between pixels, this module can also extract high-level features of the image. The self-attention module formula is as follows:

$$y_i = \frac{1}{C(x_i)} \sum_{\forall j} e^{\theta(x_i)^{\mathrm{T}}\phi(x_j)} \cdot g\left(\hat{x}_j\right) + x_i, \tag{2}$$

where i, j represent the position index, x is the input feature,y represents the obtained pixel relationship matrix, and $g\left(\hat{x}_j\right)$ gives the representation of the input feature x_j of each location, all the signal are all aggregated to position j, and the three embedding functions θ, ϕ, g can be implemented by a 1×1 convolutional layer. The response is normalized by a factor $C(x_i)$.

To obtain a richer pixel relationship map, we concatenate the original image $I \in R^{3\times H\times W}$ to the input feature map to integrate the underlying features of the image. Meanwhile, in order to reduce the influence of background noise, we use another branch of CAM as pixel-level supervision to perform training modification on the pixel relation matrix.

Although the affinity between pixels is more obvious, the traditional convolution kernel has limitations that it cannot accurately locate the target during the convolution process and it also makes the target boundary challenging to distinguish. On this basis, we add the deformable convolution. By adding a learnable offset, it is not limited to the regular grid points of traditional convolution so that it can focus on the image texture boundary. The deformable convolution formula is as follows:

$$y(p) = \sum_{k=1}^{K} w_k \cdot x(p + p_k + \Delta p_k) \cdot \Delta m_k, \tag{3}$$

where $x(p)$ and $y(p)$ represent the feature at position p in the input feature x and the output feature y, respectively. K is the convolution kernel of K sampling locations, and w_k and p_k represent the weight and pre-specified offset of the k-th location. Δp_k and Δm_k are the learnable offset and modulation scalar for the k-th position. The range of Δm_k is [0, 1]. When computing $p + p_k + \Delta p_k$, we used bilinear interpolation.

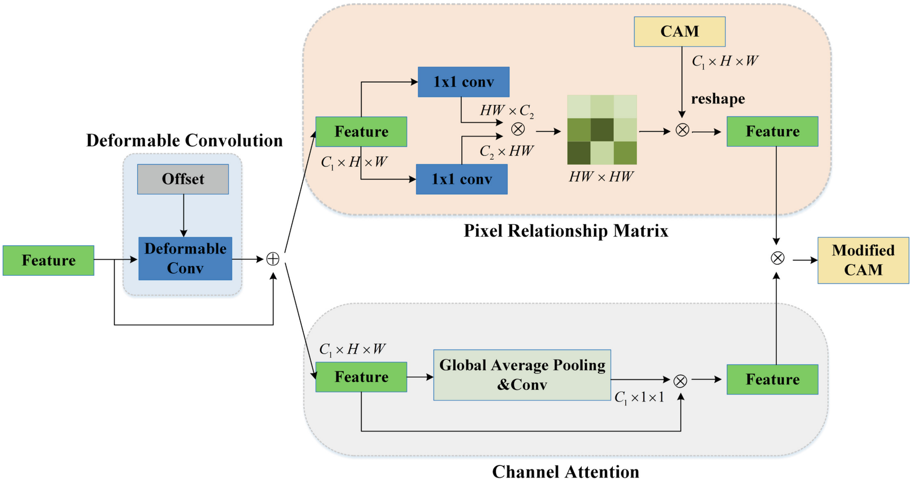


Fig. 3. The structure of DSAM

Since the channel map of each high-level feature can be viewed as the response of a specific class, we exploit the interdependence between channels to emphasize the irrelevant feature maps, thereby reducing the noise effect. Therefore, we introduce the channel attention module, and we directly perform a global average pooling operation on the feature map $I \in R^{C \times H \times W}$ obtained by deformable convolution. By the above operations, we can get the attention vector $I_2 \in R^{C_1 \times 1 \times 1}$ that contains the semantic dependencies between channels. Each vector in I_2 aggregates the contextual information of the image. The channel attention formula is:

$$Y_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_c(i, j), \tag{4}$$

where i, j represent the position index, H, W represent the size of the feature map, X_c represents the input feature map X of the c-th channel, Y_c represents the channel attention vector of the c-th channel.

Compared with the traditional self-attention module, the DSAM adds deformable convolution and applies the residual structure, so that the edge texture information of different targets can be adaptively learned. In the pixel relationship matrix, we use the ReLU activation function with L1 normalization to mask irrelevant pixels, generating an affinity attention map that is smoother in relevant regions. A channel attention mechanism is also introduced to further subdivide the pixel relationship through the inter-channel interdependence to reduce the interference of background noise.

2.3 Loss Design of Deform-CAM

In this paper, only image-level labels are used in our experiments, as well as in the loss design. We perform the GAP processing at the end of the network, using a multi-label classification loss for classification. The classification loss can be expressed as:

$$L_{cls} = -log(\sigma(G(A_c))), \tag{5}$$

$$L_{deform-cls} = -log\big(\sigma\big(G\big(A_{deform-c}\big)\big)\big), \tag{6}$$

where A is the class activation map, $G(\cdot)$ represents the global average pooling. $\sigma(\cdot)$ represents activation function. These two classification losses can improve the performance of object localization. In order to maintain the consistency of the output, the relationship between pixels needs to be aggregated on the original CAM to minimize the effect of background noise. We added reconstruction regularization to make it correspond the original CAM and the modified CAM. The loss can be easily defined as.

$$L_{re} = \left|A_{cls} - A_{deform-cls}\right|. \tag{7}$$

In short, the final loss can be expressed as:

$$L_{all} = L_{cls} + L_{deform-cls} + \alpha \cdot L_{re}, \tag{8}$$

where α is the balance of weights for different losses. Coarse localization of target is performed using classification losses L_{cls} and $L_{deform-cam}$. The reconstruction loss L_{re} is used to bridge the gap between pixel-level and image-level supervised processes and integrate DSAM with the network. We give details of the network training setup and study the effectiveness of each modules in the experimental section.

3 Experiments

3.1 Implementation Details

In this section, we present the implementation details of our method. In the official *PASCAL VOC 2012* dataset, there are 1464 images for training, 1449 for validation, and 1456 for testing. We set up one background class and 20 foreground classes to evaluate our method. Following the commonly used experimental protocol for semantic segmentation, we extract additional annotations from SBD [18] to construct an augmented training set containing 10582 images. However, during network training we only use image-level labels.

In the experiments, we use ResNet38 with output stride = 8 as the backbone network. During training, we crop all images to 448 × 448 as network input. The model is trained on Tesla V100-PCIE-32 GB. *batch_size* is set to 8, *epoch* is 15, the *learning* rate is 0.01, and the learning rate policy uses $lr_{itr} = lr_{init}(1 - itr\big/(\max - itr))^{\gamma}$, where $\gamma = 0.9$.

3.2 Ablation Studies

We conducted ablation experiments on DSAM, the main module of Deform-CAM. Here, we still used the mIoU as the evaluation index. As shown in Table 1, the CAM accuracy of the baseline is 48.1%. After the adjustment of the DSAM module, we improved the accuracy to 50.5%. Based on the baseline, we can see that adding the pixel relationship matrix (PRM) can enrich the semantic information between pixels, and the accuracy is improved by 1.2%. By applying the deformable convolution (DC) again to refine the image boundaries and reduce the background noise at the boundaries, the accuracy is further improved by 0.5%. Finally, by adding the channel attention (CA) branch to enhance intra-class features between channels, the generated pseudo-labels achieve 50.5% accuracy on the *PASCAL VOC* validation set.

Table 1. The ablation study for each part of DSAM. **CAM:** Class Activation Maps. **PRM:** pixel relationship matrix. **DC:** deformable convolution. **CA:** channel attention.

CAM	PRM	DC	CA	mIoU
√				48.1%
√	√			49.3%
√	√	√		49.8%
√	√	√	√	50.5%

Table 2 shows the ablation results of the network loss. Baseline accuracy is 48.1%. When applying the classification loss only to the output of the DSAM module, the accuracy instead drops to 47.3%, this is because the DSAM module can acquire more target areas, but it introduces some background noise for some classes, which will affect the quality of the CAM. By reconstructing the regularization loss L_{re}, the network expands the correct local features, increasing the accuracy by 0.5%. When we introduce classification loss and reconstruction loss together, the accuracy rises to 50.5%, not only the noise information is reduced, but the features on the boundary are also more precise.

Table 2. The ablation study of the network loss.

L_{cls}	L_{deform}	L_{re}	mIoU
√			48.1%
√		√	48.6%
√	√		47.3%
√	√	√	50.5%

Table 3. Performance comparisons of our method with other methods on *PASCAL VOC 2012* dataset. *: the segmentation results with post-processing. Indicate: *I*-image-level, *S*-external saliency maps.

Method	Backbone	Sup	val	test
AffinityNet [13]	ResNet-38	*I*	61.7	63.7
SEAM [14]	ResNet-38	*I*	64.5	65.7
RRM [19]	ResNet-38	*I*	62.6	62.9
BES* [20]	ResNet-101	*I*	65.7	66.6
OOA [21]	ResNet-101	*I*	65.2	66.4
IRNet [22]	ResNet-50	*I*	63.5	64.8
CIAN [23]	ResNet-101	*S*	64.1	64.7
CDA [24]	ResNet-38	*I*	64.2	65.8
H-DSRG [25]	ResNet-38	*I*	64.6	65.2
Ours	ResNet-38	*I*	**65.5**	**66.8**

3.3 Comparison with Existing State-of-the-Art Methods

To further improve the accuracy of pseudo-pixel-level annotations, we follow the work of IRNet [22] and add a boundary branch to the modified CAM. According to our generated CAM, the boundary branch is trained, and the semantic segmentation task is completed by generating random seeds and performing a random walk strategy. The final generated pseudo-labels achieve 66.5% accuracy on the *val* set of *PASCAL VOC 2012*.

In Table 3, the *mIoU* comparison between our method and previous methods is shown. We can find that on the validation set, the accuracy of our method is almost the same as that of BES [20] using denseCRF post-processing. And on the test set, we are even higher than that. Compared with other baseline methods, Deform-CAM achieves significant performance improvements on both *val* and *test* sets under the same training settings. Notably, our accuracy gains do not come from large network results, but through an efficient combination of variable convolution, pixel relations, and channel attention. Figure 4 shows qualitative results on the *val* set, we can find that compared to IRNet, our proposed method can identify more accurate regions in columns 1 and 7. In columns 2 and 3, we can better identify the boundary details. And from columns 4 to 6, we can see that our network did not segment the background region. In conclusion, our results are closer to the GT than IRNet, illustrating that the proposed method achieves good results on both large and small objects.

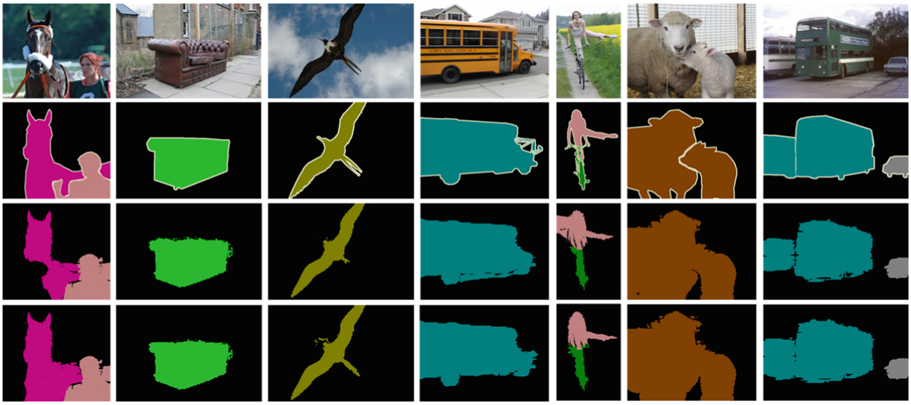


Fig. 4. Qualitative segmentation results on *PASCAL VOC* 2012 *val* set. *From top to bottom*: input images, ground-truths, IRNet [22] and our segmentation results.

4 Conclusion

This paper designs a Deform-CAM network structure to leverage image-level labels to close the supervision gap between FSSS and WSSS. DSAM expands the correct local feature range through deformable convolution and an efficient combination of pixel-to-pixel relationships. Our Deform-CAM is implemented with an efficient reconstruction loss network structure, and the generated CAM not only has less background noise, but also better approximates the shape of GT. According to the generated pixel-level pseudo-labels, combined with the random walk strategy, a good improvement is achieved on the *PASCAL VOC 2012* dataset, proving the effectiveness of Deform-CAM.

Acknowledgments. This work was supported by the Natural Science Foundation of China (No. 61773325 and 61806173), the Industry–University Cooperation Project of Fujian Science and Technology Department (No. 2021H6035), the Natural Science Foundation of Fujian Province (No. 2021J011191), the Joint Funds of the 5th Round of Health and Education Research Program of Fujian Province (No. 2019-WJ-41), the Science and Technology Planning Project of Fujian Province (No. 2020H0023), and the Young Teacher Education Research Project of Fujian Province (No. JT180435).

References

1. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
2. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
3. Chen, L.C., Papandreou, G., Kokkinos, I., et al.: DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2018)

4. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 833–851. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_49
5. Fu, J., et al.: Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
6. Papandreou, G., et al.: Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: Proceedings of the IEEE International Conference on Computer Vision (2015)
7. Lin, D., et al.: Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
8. Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L.: What's the point: semantic segmentation with point supervision. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 549–565. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_34
9. Pathak, D., et al.: Fully convolutional multi-class multiple instance learning. arXiv preprint arXiv:1412.7144 (2014)
10. Zhou, B., et al.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
11. Kolesnikov, A., Lampert, C.H.: Seed, expand and constrain: three principles for weakly-supervised image segmentation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 695–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_42
12. Wei, Y., et al.: 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 (2017)
13. 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 (2018)
14. Wang, Y., et al.: Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
15. Chang, Y.-T., et al.: Weakly-supervised semantic segmentation via sub-category exploration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
16. Sun, G., Wang, W., Dai, J., Van Gool, L.: Mining cross-image semantics for weakly supervised semantic segmentation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12347, pp. 347–365. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58536-5_21
17. Wu, T., et al.: Embedded discriminative attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
18. Hariharan, B., et al.: Semantic contours from inverse detectors. In: 2011 International Conference on Computer Vision. IEEE (2011)
19. Zhang, B., et al.: Reliability does matter: an end-to-end weakly supervised semantic segmentation approach. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, issue number 07 (2020)
20. Chen, L., Wu, W., Fu, C., Han, X., Zhang, Y.: Weakly supervised semantic segmentation with boundary exploration. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020.

LNCS, vol. 12371, pp. 347–362. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58574-7_21

21. Jiang, P.-T., et al.: Integral object mining via online attention accumulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2019)
22. Ahn, J., Cho, S., Kwak, S.: Weakly supervised learning of instance segmentation with inter-pixel relations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
23. Fan, J., et al.: CIAN: cross-image affinity net for weakly supervised semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, issue number 07 (2020)
24. Su, Y., et al.: Context decoupling augmentation for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
25. Chong, Y., et al.: Erase then grow: generating correct class activation maps for weakly-supervised semantic segmentation. Neurocomputing **453**, 97–108 (2021)