AN IMPROVED APPROACH TO WEAKLY SUPERVISED SEMANTIC SEGMENTATION

Lian Xu[1⋆], *Mohammed Bennamoun*[1⋆], *Farid Boussaid*[2⋆], *Senjian An*[3†], *Ferdous Sohel*[4‡]

[1]School of Computer Science and Software Engineering
[2]School of Electrical, Electronics and Computer Engineering
[3]School of Electrical Engineering, Computing and Mathematical Sciences
[4]School of Engineering and Information Technology
[⋆]The University of Western Australia, [†]Curtin University, [‡] Murdoch University

ABSTRACT

Weakly supervised semantic segmentation with image-level labels is of great significance since it alleviates the dependency on dense annotations. However, it is a challenging task as it aims to achieve a mapping from high-level semantics to low-level features. In this work, we propose a three-step method to bridge this gap. First, we rely on the interpretable ability of deep neural networks to generate attention maps with class localization information by back-propagating gradients. Secondly, we employ an off-the-shelf object saliency detector with an iterative erasing strategy to obtain saliency maps with spatial extent information of objects. Finally, we combine these two complementary maps to generate pseudo ground-truth images for the training of the segmentation network. With the help of the pre-trained model on the MS-COCO dataset and a multi-scale fusion method, we obtained mIoU of 62.1% and 63.3% on PASCAL VOC 2012 *val* and *test* sets, respectively, achieving new state-of-the-art results for the weakly supervised semantic segmentation task.

Index Terms— Semantic segmentation, attention maps, object saliency, weakly supervised learning, deep convolutional neural network

1. INTRODUCTION

Human cognition is formed gradually through the process of continuous exploration of our surroundings. This process involves significant supervisory information provided by external feedback. A large body of research has focused on developing machines with the learning ability of humans. For semantic segmentation, the prediction accuracy relies heavily on a large number of pixel-level labels, which comes at a prohibitively high human annotation cost. In contrast, humans perform semantic segmentation without the need of such fine pixel-level supervision and information. They instead rely only on weak supervision. Inspired by this observation, a number of weakly supervised semantic segmentation approaches have been proposed, with the following weak supervision categories: bounding boxes [1], scribbles [2], points [3] and image-level labels [4, 5, 6]. Among them, image-level labels are the most popular and economical setting, as they are simple and easy to collect.

Weakly supervised semantic segmentation is commonly achieved in a two-step pipeline. Pseudo or approximate ground-truth training images are first generated and then used to train a segmentation model. For the first step, most approaches rely on a technique called Class Activation Map (CAM) [7] to produce object localization cues. CAM modifies image classification convolutional neural network (CNN) architectures by removing the fully-connected (FC) layers. It instead adds a Global Average Pooling (GAP) layer in the penultimate layer before the last prediction layer. Therefore, the modified networks are capable of localizing class-discriminative regions in the image. Similar techniques based on Global Max Pooling (GMP) [8] and log-sum-exp pooling [9] have also been investigated. However, such approaches sacrifice model complexity and performance for an improved transparency into the working of the model. Ramprasaath [10] recently proposed a more generalized class-discriminative object localization technique "Grad-CAM". This approach can highlight in the image the regions which are important for prediction by using the gradients of the target on the final convolutional layer. Without a change in architecture, Grad-CAM can be applied to a wide variety of CNN models.

Once object localization cues are obtained, prior works expand the sparse cues to achieve better-quality pseudo ground-truth mainly in three ways: **i**) Discovering the non-discriminative regions, based on common object features. In [11], a region classification network trained on superpixels, labeled with the initial object localization cues, is proposed. This network can be used to predict classes of unlabeled regions. In [12], based on the traditional seeded region growing method, the regions are grown from the initial object local-

This research is partially supported by China Scholarship Council funds (CSC, 201607565016) and Australian Research Council Grants (DP150104251 and DE120102960)

ization cues depending on their proposed similarity criterion, in which output probabilities from the segmentation network are used as pixel features. **ii**) Modifying image classification CNN architecture for improving performance of CAM and Grad-CAM. Wei *et al.* [6] take advantage of "dilated convolution" with enlarged receptive field to incorporate context and achieve dense object localization maps, by adding multiple dilated convolutional blocks of different dilated rates to the image classification network. Li *et al.* [13] designed guided attention inference networks with two identical branches. One branch takes as input the image, of which the most discriminative regions located by the other branch are erased. As this process is constrained by an attention mining loss, it guides the other branch to discover the whole object of interest. **iii**) Using objectness priors. Off-the-shelf saliency detectors are commonly used to provide information about the extent of objects, which is complementary to object localization cues. In [14], a hierarchical saliency detection method is proposed based on [15], to generate better object saliency maps, thus achieving improved-quality pseudo ground-truth when combined with object localization cues.

After obtaining pseudo segmentation ground-truth, general semantic segmentation models can be trained in a fully supervised manner. Specially, in order to ensure that segmentation prediction is consistent with the image-level labels, Kolesnikov *et al.* [4] proposed a global weighted rank-pooling to aggregate segmentation scores into classification scores and apply a standard classification loss. Similarly, in [5, 6], segmentation predictions are aggregated via GAP to produce classification scores, which are further used as weights for the segmentation score maps. Both classification scores and weighted segmentation maps are used auxilliarily to optimize the segmentation network.

In this work, we propose an improved approach based on Grad-CAM and class-agnostic saliency detection to generate pseudo segmentation ground-truth with image-level labels. Our work achieves a mIoU of 62.1% and 63.3% on *val* and *test* of the PASCAL VOC 2012 benchmarking, respectively, achieving new state-of-the-art results.

2. APPROACH

In this section, we describe the pseudo segmentation ground-truth generation, and the multi-scale image segmentation for weakly supervised semantic segmentation with image-level labels.

2.1. Pseudo segmentation ground-truth generation

To generate pseudo segmentation ground-truth, we rely on attention maps generated from an image classification network and class-agnostic saliency maps obtained from an object saliency detector. The attention maps provide class information of sparse object regions in the image, while the saliency maps inform the spatial extent of objects. Given that

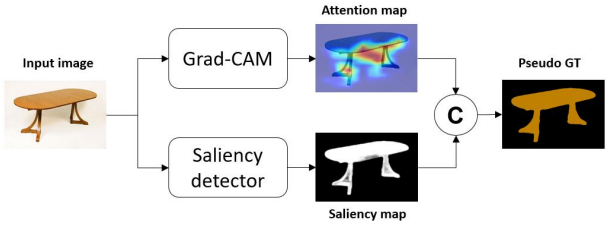

Fig. 1. Pseudo segmentation ground-truth generation.

these two sources of information are complementary, they can thus be merged to produce pixel-wise annotation maps, which can be used as pseudo segmentation ground-truth images. This procedure is illustrated in Figure 1.

2.1.1. Attention maps generation

Among existing class-discriminative localization techniques, Grad-CAM was chosen because it provides superior interpretability and faithfulness to the original model. Therefore, we employed Grad-CAM to generate our attention maps.

In order to obtain the attention map of a target class c for a given image I, we need to first compute the gradients of the target class score y^c (before the softmax layer) to the k-th activation map A^k from a convolutional layer of an image classification network. Then the importance of this activation map to the target class c can be computed by globally averaging the gradients:

$$\alpha_k^c = \frac{1}{N} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{1}$$

where N denotes the number of pixels in the activation map.

Finally, the Grad-CAM attention map is computed by applying a ReLU to the weighted linear combination of activation maps:

$$A^c = ReLU(\sum_k \alpha_k^c A^k) \tag{2}$$

Note that the ReLU operation ensures that only features with a positive influence on the target class are taken into account.

2.1.2. Saliency maps generation

In this work, we use an off-the-shelf saliency object detector, DHSNet [15] which was trained on images with corresponding ground-truth saliency masks, to generate saliency maps of training images. Considering that common saliency object detectors often fail to detect occurrences of multiple object instances in an image, we follow the iterative erasing strategy that was used in [5, 14] to discover more regions of salient objects. Specifically, we first obtain the initial saliency confidence map by feeding the input image through the DHSNet. Then, we define highly salient regions by setting a threshold to the saliency confidence map and erase the corresponding regions in the input image. Next, we repeat the aforementioned steps but replace the input image with the erased image

Algorithm 1 Pseudo segmentation ground-truth generation

Input: C: Image labels; A: Attention maps; S: Saliency maps; γ_1, γ_2: Thresholds.

Output: G: Pseudo segmentation ground-truth.

for each c in C and each location (i, j) **do**
  $M(c, i, j) = (A^c(i, j) + S(i, j))/2$
end for
for each location (i, j) **do**
  if $M(c, i, j) < \gamma_1$ **or** $S(i, j) < \gamma_2$ **then**
    $G(i, j) = 0$
  else
    $G(i, j)$ = argmax $(M(c, i, j))$
  end if
end for

from the last iteration, thus producing multiple saliency object maps. For instance, in the following experimental section 3, we applied the erasing step twice and obtained three saliency maps $S_0(i, j), S_1(i, j)$ and $S_2(i, j)$ from the original image, the image after the first erasing, and the image after the second erasing, respectively. The final saliency object map was generated by computing the pixel-wise maximum value of these three saliency maps:

$$S(i, j) = Max(S_1(i, j), S_2(i, j), S_3(i, j)) \quad (3)$$

Finally, we compute the pixel-wise average of the normalized attention maps and saliency maps, followed by a hard thresholding operation to obtain the final pseudo segmentation ground-truth. This procedure is summarized in Algorithm 1.

2.2. Multi-scale image segmentation

For image segmentation, we use the latest most popular system "DeepLab". It re-purposes image classification networks to the task of semantic segmentation by applying the "atrous convolution" with enlarged *field-of-views* of filters for dense feature extraction.

To improve the network's ability to handle objects of varying sizes, multi-scale processing is often effective. DeepLab v2 [16] uses two methods to deal with scale variability in semantic segmentation. In the **first** method, multiple (three, i.e., 0.5, 0.75 and 1, in the experiments of section 3) re-scaled versions of the original images are fed into parallel DCNNs with shared weights to produce multiple score maps. The final results are obtained by fusing multiple score maps, which are linearly interpolated to the same resolution, by taking the maximum response at each location. In the **second** approach, "Atrous Spatial Pyramid Pooling (ASPP) " is used and implemented through multiple parallel atrous convolutional layers with different sampling rates. For an input image, multiple feature maps extracted from separate branches are further fused to form the final results.

In this work, we simultaneously use these two multi-scale strategies. For training, the loss function to be optimized is the sum of four losses: three losses associated to three multi-scale outputs and the loss corresponding to the maximum output. In the testing phase, we take the maximum response as the final predicted output.

3. EXPERIMENTAL RESULTS

Datasets and evaluation metrics. We evaluated our approach on the PASCAL VOC 2012 segmentation benchmark [17]. The original dataset has 20 categories and one background category, and it includes training, validation and testing splits with 1,464, 1,449 and 1456 images, respectively. Following common practice, we use the augmented dataset of pixel-wise annotated 10,582 images provided by [18] for training. The mean Intersection-over-Union (mIoU) of all 21 categories between outputs from the segmentation network and pixel-level ground-truth is used to evaluate the performance. The results on the *test* set are obtained from the official PASCAL VOC online evaluation server.

Training/Testing Settings. We use PyTorch [19] to implement our approach. To generate our attention maps using Grad-CAM, we adopt VGG-16 pre-trained on ImageNet [20] except for the last classification layer that we changed to 20 nodes. We optimize the image classification network by minimizing a multi-label soft margin loss. For data augmentation, input images are randomly cropped to 224 $\times$ 224, and then randomly horizontally flipped. We train the network for 15 epochs with a batch size of 30. The initial learning rate is set to 0.001 (0.01 for the last layer), which is decreased by a factor of 10 every 6 epochs. In order to mine more salient objects in the process of generating saliency maps, we perform the erasing step twice with thresholds of 0.7 and 0.8, successively. For each time, the image pixels whose saliency scores are greater than the threshold are erased and replaced with the average pixel value. To obtain pseudo segmentation ground-truth, we set γ_1 and γ_2 in Algorithm 1 as 0.2 and 0.06, respectively, to select background pixels. For the segmentation framework, we adopted DeepLab-ASPP model with four branches and large atrous rates (r = 6, 12, 18, 24) built on ResNet-101, which is initialized with weights from a model pre-trained on the MS-COCO dataset. We randomly crop the input images to 321 $\times$ 321. For single-scale inputs, we train the network for 8000 iterations with a mini-batch of 10 images. The initial learning rate is set to 0.001 and is decreased by a factor of 10 every 2000 iterations. For multi-scale inputs, we set the batch size to 1 and train the network for 20K iterations. The initial learning rate is set to 0.00025 and a "poly" learning rate policy is used as suggested in [16]. At test time, Conditional Random Fields (CRF) with default parameters provided in [21] are used to refine the predicted masks from the segmentation network.

Table 1 reports the effects of the following factors on the segmentation performance on the PASCAL VOC 2012 *val* set: **i**) transferring weights from the model pre-trained on

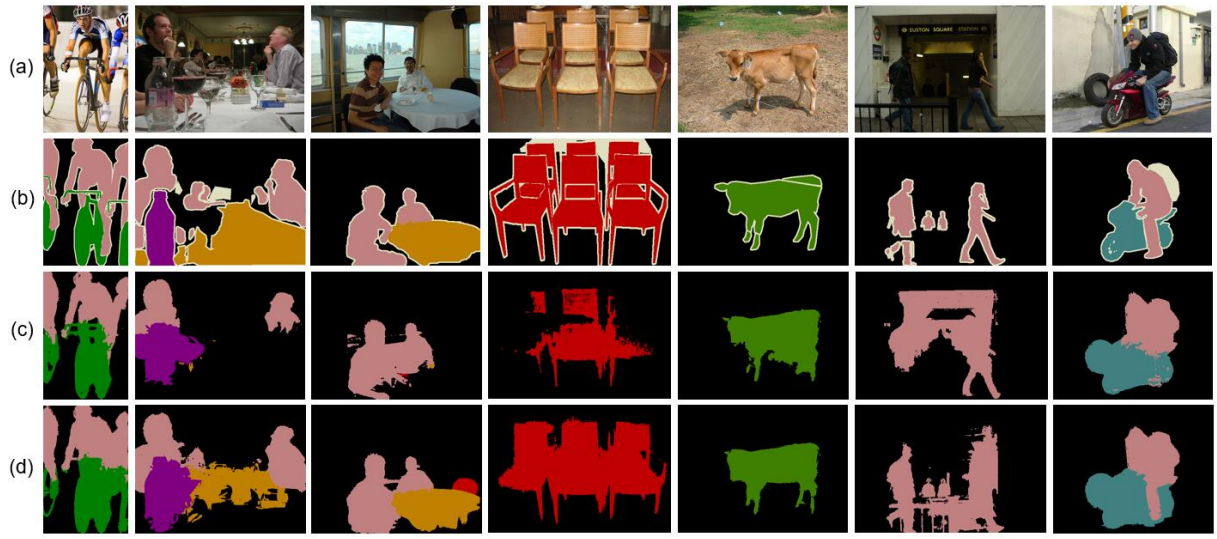

Fig. 2. Qualitative results on the PASCAL VOC 2012 *val* set. (a) Input images. (b) Ground-truth segmentation. (c) Results obtained with single-scale inputs. (d) Results obtained with multi-scale fusion. Multi-scale fusion achieves much better results than single scale when segmenting multiple object instances. It produces accurate boundaries for objects of various sizes.

Table 1. Employing DeepLab-ASPP built on ResNet-101 on PASCAL VOC 2012 *val* set. COCO: transferring weights from the model pre-trained on the MS-COCO dataset; MSC: training on multi-scale inputs with max fusion.

ASPP	COCO	MSC	CRF	mIoU(%)
✓				56.6
✓			✓	58.7
✓	✓			58.5
✓	✓		✓	60.7
✓	✓	✓		60.0
✓	✓	✓	✓	62.1

Table 2. Comparison of image-level weakly supervised semantic segmentation methods on PASCAL VOC 2012 *segmentation val* and *test* sets. All with CRF. * Use VGG-16 [22] in segmentation network and others use ResNet-101.

Methods	mIoU(%) (*val*)	mIoU(%) (*test*)
[6]*	60.4	60.8
GAIN [13]*	55.3	56.8
MCOF [11]	60.3	61.2
DSRG [12]	61.4	63.2
DCSP [14]	60.8	61.9
ours	**62.1**	**63.3**

the MS-COCO dataset, **ii**) multi-scale inputs and **iii**) using CRF for post-processing. The results in Table 1 demonstrate that transferring knowledge from the model pre-trained on the MS-COCO improves performance from 56.6% to 58.5%, an improvement of 1.5% is achieved when using multi-scale inputs, and with CRF post-processing, the performance gain is 2.1%. Figure 1 shows qualitative visual comparison of the model's results using single-scale inputs and multi-scale inputs for training. We can observe that multi-scale fusion yields better results in segmenting multiple object instances such as "bike" and "chair" in the first and 4-th columns and producing more accurate boundaries of the "cow" and "person" as shown in the rightmost three columns. Moreover, multi-scale image representations can discriminate some hard classes, such as "table" which is prone to be misclassified into "background" using single-scale representation as shown in the second and third columns. In particular, for the third image, the multi-scale approach can even correctly segments a small region of "chair" that is not marked in the ground-truth.

Table 2 compares the proposed method to state-of-the-art methods on mIoU on the PASCAL VOC 2012 *val* and *test* sets. Our proposed approach achieves the best results. This demonstrates the importance of transfer learning and of a multi-scale strategy for training the weakly supervised segmentation model.

4. CONCLUSION

In this work, we propose an improved method to generate pseudo ground-truth for training a segmentation network with image-level labels. We have shown the effectiveness of transfer learning from other datasets and training with multi-scale inputs on obtaining a well-performing segmentation model. The proposed method achieves state-of-the-art results for the weakly supervised semantic segmentation task on PASCAL VOC 2012 segmentation benchmark.

5. ACKNOWLEDGEMENTS

The authors acknowledge NVIDIA for providing a Titan X GPU for the experiments involved in this research.

6. REFERENCES

[1] Anna Khoreva, Rodrigo Benenson, Jan Hendrik Hosang, Matthias Hein, and Bernt Schiele, "Simple does it: Weakly supervised instance and semantic segmentation.," in *IEEE CVPR*, 2017, vol. 1, p. 3.

[2] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *IEEE CVPR*, 2016, pp. 3159–3167.

[3] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei, "Whats the point: Semantic segmentation with point supervision," in *ECCV*. Springer, 2016, pp. 549–565.

[4] Alexander Kolesnikov and Christoph H Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *ECCV*. Springer, 2016, pp. 695–711.

[5] 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 *IEEE CVPR*, 2017, vol. 1, p. 3.

[6] 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 *IEEE CVPR*, 2018, pp. 7268–7277.

[7] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *IEEE CVPR*, 2016, pp. 2921–2929.

[8] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic, "Is object localization for free?-weakly-supervised learning with convolutional neural networks," in *IEEE CVPR*, 2015, pp. 685–694.

[9] Pedro O Pinheiro and Ronan Collobert, "From image-level to pixel-level labeling with convolutional networks," in *IEEE CVPR*, 2015, pp. 1713–1721.

[10] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, et al., "Grad-cam: Visual explanations from deep networks via gradient-based localization.," in *ICCV*, 2017, pp. 618–626.

[11] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma, "Weakly-supervised semantic segmentation by iteratively mining common object features," in *IEEE CVPR*, 2018, pp. 1354–1362.

[12] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang, "Weakly-supervised semantic segmentation network with deep seeded region growing," in *IEEE CVPR*, 2018, pp. 7014–7023.

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

[14] Arslan Chaudhry, Puneet K Dokania, and Philip HS Torr, "Discovering class-specific pixels for weakly-supervised semantic segmentation," *BMVC*, 2017.

[15] Nian Liu and Junwei Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *IEEE CVPR*, 2016, pp. 678–686.

[16] 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 transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[17] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[18] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik, "Semantic contours from inverse detectors," 2011.

[19] Adam Paszke, Sam Gross, Soumith Chintala, and Gregory Chanan, "Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration," 2017.

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE CVPR*. Ieee, 2009, pp. 248–255.

[21] Philipp Krähenbühl and Vladlen Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in neural information processing systems*, 2011, pp. 109–117.

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