Weakly Supervised Semantic Segmentation by Multiple Group Cosegmentation

1st Kunming Luo
School of information and Communication Engineering
University of Electronic Science and Technology of China
Chengdu, China
kimluo1993@gmail.com

2nd Fanman Meng
School of information and Communication Engineering
University of Electronic Science and Technology of China
Chengdu, China
fmmeng@uestc.edu.cn

3rd Qingbo Wu
School of information and Communication Engineering
University of Electronic Science and Technology of China
Chengdu, China
qbwu@uestc.edu.cn

4th Hongliang Li
School of information and Communication Engineering
University of Electronic Science and Technology of China
Chengdu, China
hlli@uestc.edu.cn

***Abstract*—Weakly supervised semantic segmentation aims at segmenting images by image-level labels. The existing methods try to train an end-to-end CNN network, which needs to handle multiple classes that is difficult. In addition, the existing methods are sensitive to the image-level cues such as discriminative regions and the pseudo-annotations. To avoid these drawbacks, this paper proposes a new strategy, which first obtains the foregrounds of each class by multiple group cosegmentation, and then combines the results to form the semantic segmentation. In our method, three new aspects are considered. 1) we solve semantic segmentation by each class that is easy to handle. 2) we extract discriminative regions more globally by context analysis. 3) we learn local-to-global segmentation network to segment the object from local discriminative priors. A new CNN network for multiple group cosegmentation is proposed. Two subnetworks such as global context based discriminative region extraction network and local-to-global segmentation network are designed. A simple combination method based on the discriminative map is proposed to finally obtain the semantic segmentation results. We verify the proposed method on Pascal VOC dataset. The experimental results show that the proposed method can obtain mIOU value 0.563 and 0.603 (without CRF post-processing) on the validation and test dataset that outperforms many existing weakly supervised semantic segmentation methods.**

***Index Terms*—Weakly Supervised Semantic Segmentation, Image Segmentation, Discriminative Region Extraction**

I. Introduction

Weakly supervised semantic segmentation uses image-level labels to segment semantic regions, which is a challenging task. Several deep learning based weakly supervised semantic segmentation methods have been proposed. These methods can be classified into two classes: one-stage based and two-stage based methods.

One-stage-based method [1]–[4] solves semantic segmentation by constructing end-to-end CNN network. The discriminative regions are used as the object priors, which is the essential step for the successful segmentation. Researchers paid much attention on designing efficient losses that lead to better discriminative priors. Several useful losses by the image-level labels are proposed. The results show that the discriminative regions can be captured well by CNN network. Two-stage-based method solves weakly supervised semantic segmentation by two steps. The first step automatically generates the pseudo-annotations from images with image-level labels. The second step then learns the semantic segmentation model by the pseudo-annotations. Better results can be obtained by the two-stage-based method since more annotations are used. Meanwhile, generating sufficient and accurate pseudo-annotations is a challenging task. Several types of pseudo-annotation generation strategy have been proposed [5], [6], such as the generation by simple images, user interactions, and web videos. It is verified that the semantic regions can be successfully segmented once the pseudo-annotations are well enough.

The existing weakly supervised semantic segmentation methods can be considered as the modification of fully supervised semantic segmentation by replacing the pixel-level annotation with the image-level annotation. Due to the rough priors provided by the image-level labels, the existing methods have two challenging problems. Firstly, the discriminative priors captured by the CNN networks are usually the local parts rather than the entire objects, which leads to the under segmentation of the objects. Secondly, the existing methods are sensitive to the quality of the pseudo-annotations, while the generation of good pseudo annotations by weak image-level label is still hard to be guaranteed. It is useful to generate discriminative priors with more global cues, and is also useful to avoid the process of pseudo-annotation generation.

This paper proposes a new and simple strategy to accomplish the weakly supervised semantic segmentation by cosegmentation, which first divides the images into multiple classes by the image-level labels, and then generates the common foregrounds of each class by multiple group cosegmentation. Finally, the cosegmentation results are combined according to the image labels of each image to form the final results. To

978-1-5386-4458-4/18/$31.00 ©2018 IEEE

fulfill the semantic segmentation, a new multiple group cosegmentation is proposed. Two cues such as the discriminative cue inter group (to describe both the intra-class common priors and inter-class discriminative priors), and the local-to-global segmentation cue (to segment object from local priors) are considered. A corresponding CNN network for multiple group cosegmentation is constructed, where two subnetworks such as the discriminative region generation network by considering global context, and local-to-global segmentation network by segmenting global region from local prior are proposed. A simple fusion method based on the discriminative map is proposed to finally generate the semantic segmentation results. We verify the proposed method on Pascal Voc 2012 Dataset. The experimental results show that the proposed method can obtain 0.563 and 0.603 mIOU value on the validation and test dataset without CRF post-processing, which is superior to many existing methods.

II. The Proposed Method

A. Overview of The Proposed Method

The flowchart of the proposed method is shown in Fig. 1, where the proposed method consists of two steps: multiple group cosegmentation step to generate initial foregrounds for each class, and the fusion step to generate the semantic segmentation from the initial foregrounds.

The essential step is the cosegmentation step. Here, we employ multiple group cosegmentation, because it can obtain better segmentation results by capturing various cues from single image, multiple images, and multiple groups. However, the existing multiple group cosegmentation methods are based on single class, while multiple classes are considered here. Therefore, we extend the multiple group cosegmentation to multiple classes, and propose a new multiple image group cosegmentation to fulfill our semantic segmentation task.

B. Multiple Group Cosegmentation Step

The multiple group cosegmentation is built by three considerations. 1) The discriminative regions that carry the segmentation cues inter and intra image groups are considered. 2) The method of capturing more global discriminative information is considered. 3) The segmentation from local part prior to global object is also considered to help the segmentation of global objects.

Based on the three considerations, the proposed multiple group cosegmentation network is shown in Fig. 1, which consists of two steps. The first step generates discriminative probability map from image classification network. The second step then sends the discriminative map and the original image into the local-to-global segmentation network, and obtains the segmentation results. The segmentation results are then used as new discriminative map to update the segmentation results further. The two processes are iteratively implemented until the convergency of the segmentation.

1) The Discriminative Probability Map: The discriminative probability map contains the probability value of each pixel belonging to the discriminative region. It is used to describe the cues inter and intra image classes, where the cue inter class is the regions that distinguish different classes, and the cue intra class is the common regions within a class. The two cues are the critical information to model cosegmentation. Here, we generate the discriminative probability map by the classification strategy, i.e., using the rough image-level labels to first learn the CNN-based classification model, and then extracting the discriminative probability map from the CNN network by the back propagation. Here, a new generation method for discriminative probability map is proposed. Our idea is to use the context information to generate more global discriminative regions, which is verified to has the capability of improving the final semantic segmentation results by our experiments.

Our purpose is to extract the discriminative region more global. The traditional methods consider one class cue for back propagation. However, one class cue is easily to result in a local solution that is enough to distinguish the objects. Here, we consider the discriminative cues of all foreground classes and the background to generate the discriminative region of one class, i.e., the context information around the object is also considered. In our discriminative region extraction network, a 1×1 Conv layer is added after the convolution process to transform the feature map into the first discriminative map. Two branches are then connected. One is the GAP layer that transforms the discriminative map to the classification scores. It is used to force the extraction of discriminative regions. The other is comprised of three sub-branches: two residual branches, and one BN and Relu branch. We use a 1×1 Conv layer and a 3×3 Conv layer each with a Batch Normalization layer to form the two residual branches, with the purpose of integrating different scales. The three branches are then added to generate the second discriminative map. We also use a GAP layer to convert the second discriminative map into classification scores. After training the network by classification task, the second discriminative map is then used as the discriminative probability map directly.

2) The local-to-global segmentation network: Local-to-global segmentation network aims at segmenting object from the priors that may be a part of the object. We design the local-to-global segmentation network based on FCN semantic segmentation model, but some essential improvements are performed. First, we change the input as (I_i, M') that is a combination of the image I_i and its discriminative map M'. The inputs are then sent into the convolutional neural network for the feature extraction. VGG-16 network is used as the backbone network. Finally, we change the output to be a binary segmentation layer.

Since we lack pixel-level annotations for test images, and we hope to generate a general foreground segmentation network that can be used in widespread applications, we use the existing weak-related pixel-level annotations to train the local-to-global network. Here, the local-to-global segmentation

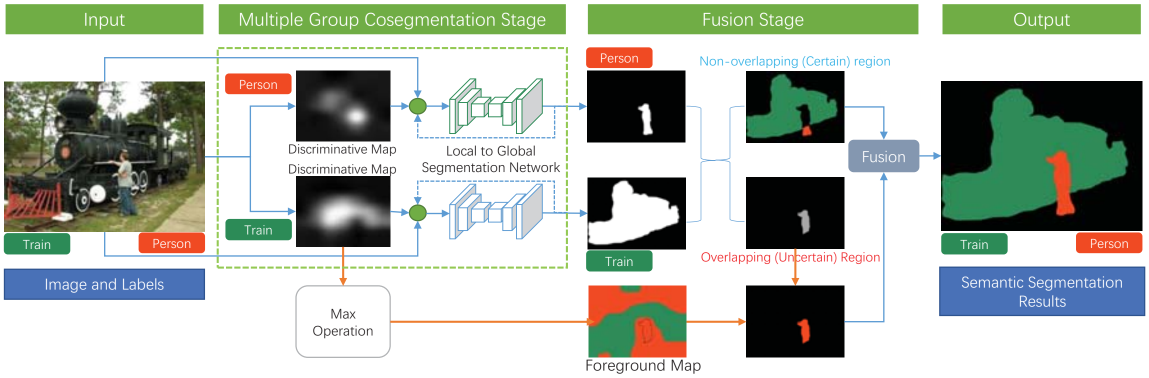
Input
Multiple Group Cosegmentation Stage
Fusion Stage
Output
Person
Discriminative Map
Discriminative Map
Local to Global Segmentation Network
Train
Train
Person
Image and Labels
Max Operation
Person
Non-overlapping (Certain) region
Train
Overlapping (Uncertain) Region
Foreground Map
Fusion
Train
Person
Semantic Segmentation Results

Fig. 1. The flowchart of the proposed method. The proposed method consists of two steps: the multiple group cosegmentation step and the fusion step. The first step is comprised by two subnetworks, i.e., discriminative region generation network, and the local-to-global segmentation network.

network is trained by considering all the classes in the COCO dataset except the 20 classes in the Pascal Voc dataset. It is proofed by the experimental results that such local-to-global segmentation network has a good generalization to the new classes.

3) The iteration of the local-to-global segmentation: Based on the initial segmentation results, we use the initial results as the new discriminative map, and re-implement the local-to-global segmentation to obtain better segmentation results. Similar to EM algorithm, such two processes are iteratively implemented until the convergency of the segmentation results.

C. The Fusion Step

After obtaining the foregrounds by the multiple group cosegmentation, we next fuse these foregrounds to generate the final semantic segmentation. The details can be found in Fig. 1. Specifically, given an image and its labels such as “Person” and “Train”, the initial discriminative probability maps M_{person} and M_{train} are first obtained in the cosegmentation step. Then, the foregrounds F_{person} and F_{train} for “Person” and “Train” obtained by local-to-global segmentation are then classified into non-overlapping region R_c and overlapping region R_u. The pixels in non-overlapping region have unique labels that can be used directly as the labels L_c of semantic segmentation result. The overlapping regions have multiple labels, and need to determine further. Here, we simply compare the probability values of the overlapping pixels in M_{person} and M_{train}, and select the labels of larger value as the semantic segmentation results. In Fig. 1, such determination is implemented by the max operation on the M_{person} and M_{train}. By denoting the semantic segmentation result of overlapping regions as L_u, the semantic segmentation results are finally obtained by combining L_c and L_u simply.

III. Experimental Results

A. Dataset

We verify the proposed method on Pascal Voc 2012 dataset. All the images in the validation and test dataset are used. The training dataset is also used to generate the discriminative map of the validation and test dataset. The ground-truths of the validation dataset and the evaluation website of Pascal Voc 2012 are used to calculate the objective results.

B. Experimental Setup

We train our discriminative region extraction network by considering all image classes, i.e., 20 classes in Pascal Voc Dataset. For the images of validation and test dataset, the images in train dataset are also used to train the model and to generate the discriminative probability map.

The local-to-global segmentation network is trained from the training images in COCO dataset. The images of the rest 60 classes apart from the 20 classes in Pascal Voc dataset are used for the training, i.e., the training data does not contain any images belonging to the 20 class in VOC dataset. The training dataset is constructed by the original images, the foreground probability maps generated from groundtruth and gaussian filtering, and the groundtruth. We set the iteration number of our local-to-global segmentation as two empirically.

C. Subjective Results

Some semantic segmentation results are displayed in Fig. 2. The original images with one class tag, the discriminative maps by the proposed discriminative region extraction network, our semantic segmentation results and the ground-truths are displayed. Some images contain multiple object instances such as the second and the fifth images. It is seen that our method successfully segments the semantic regions from these images.

More semantic segmentation results are displayed in Fig. 3, where the images with two class tags are shown. The original images, the two discriminative maps by the proposed discriminative region extraction network, the corresponding segmentation results by the proposed local-to-global segmentation, our semantic segmentation results and the ground-truths are displayed. It is seen that our method performs well on these images that contain more than one class.

D. Objective Results

We next display the objects results, where the intersection-over-union (IOU) value that is usually used to measure the

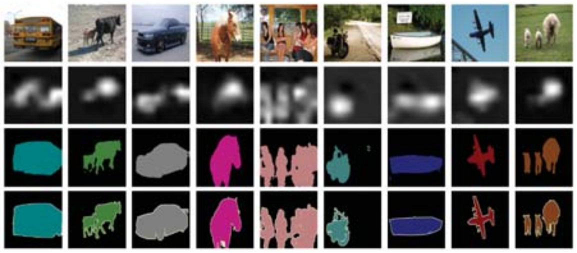

Fig. 2. The original images with one class tag, the discriminative probability map by the proposed discriminative region extraction network, our semantic segmentation results, and the groundtruth are displayed.

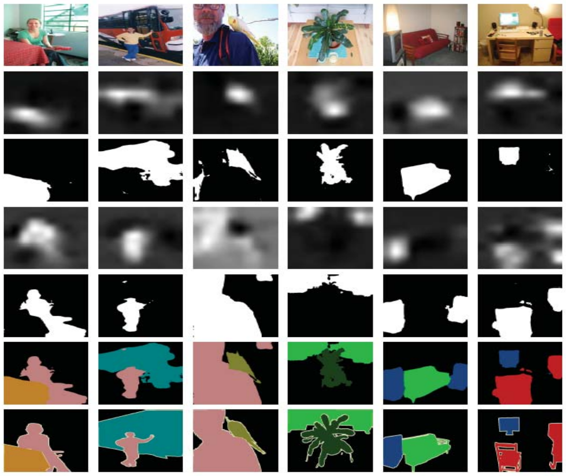

Fig. 3. Our semantic segmentation results on images with two class tags.

semantic segmentation method is used for the verification. The IOU values by our method on both the validation and test dataset are displayed in Table I. The mean IOU values (mIOU) over 20 classes are shown. The mIOU values by several recent weakly supervised semantic segmentation results are also displayed in Table I for comparison. It is seen that our method obtains 56.3 and 60.3 for the validation set and test set respectively, which is better than the existing methods. Note that CRF post-processing is not used in our method. This further demonstrates the effectiveness of the proposed method.

TABLE I
THE mIOU VALUES ON PASCAL VOC 2012 DATASET

Method	CRF	Validation	Test
SEC [7]	Yes	50.7	51.7
TransferNet [4]	Yes	52.1	51.2
C-BT-S [1]	Yes	52.8	53.7
Built-in [8]	Yes	44.8	45.8
Two-Phase [9]	Yes	53.1	53.8
Multi-Evidence [10]	Yes	–	55.6
Adversarial Erasing [11]	Yes	55.0	55.7
TMWL [12]	Yes	55.3	56.8
Ours	No	56.3	60.3

IV. Conclusion

This paper proposes a new weakly supervised semantic segmentation strategy, which first obtains foregrounds of each class by multiple group cosegmentation, and then combines the cosegmentation results to obtain the semantic segmentation. A new multiple group cosegmentation based CNN network is constructed. Two new subnetworks such as the discriminative region extraction network and the local-to-global segmentation network are proposed. A simple combination method based on the discriminative map is finally proposed to generate the semantic segmentation. The proposed method is verified on the Pascal VOC dataset. The experimental results demonstrate the effectiveness of the proposed method.

Acknowledgement

This work was supported in part by the National Nature Science Foundation of China under Grant 61502084, 61525102, 61871087, and 61831005.

References

[1] A. Roy and S. Todorovic, “Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 00, July 2017, pp. 7282–7291.

[2] T. Durand, T. Mordan, N. Thome, and M. Cord, “Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 5957–5966.

[3] S. Kwak, S. Hong, B. Han *et al.*, “Weakly supervised semantic segmentation using superpixel pooling network.” in *AAAI*, 2017, pp. 4111–4117.

[4] S. Hong, J. Oh, H. Lee, and B. Han, “Learning transferrable knowledge for semantic segmentation with deep convolutional neural network,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 00, June 2016, pp. 3204–3212.

[5] P. Hu and D. Ramanan, “Bottom-up and top-down reasoning with hierarchical rectified gaussians,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 00, June 2016, pp. 5600–5609.

[6] S. Hong, D. Yeo, S. Kwak, H. Lee, and B. Han, “Weakly supervised semantic segmentation using web-crawled videos,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 00, July 2017, pp. 2224–2232.

[7] A. Kolesnikov and C. H. Lampert, “Seed, expand and constrain: Three principles for weakly-supervised image segmentation,” in *European Conference on Computer Vision (ECCV)*. Cham: Springer International Publishing, 2016, pp. 695–711.

[8] F. Saleh, M. S. Aliakbarian, M. Salzmann, L. Petersson, S. Gould, and J. M. Alvarez, “Built-in foreground/background prior for weakly-supervised semantic segmentation,” in *European Conference on Computer Vision (ECCV)*. Cham: Springer International Publishing, 2016, pp. 413–432.

[9] D. Kim, D. Cho, and D. Yoo, “Two-phase learning for weakly supervised object localization,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, vol. 00, Oct. 2018, pp. 3554–3563.

[10] W. Ge, S. Yang, and Y. Yu, “Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning,” *CoRR*, vol. abs/1802.09129, 2018. [Online]. Available: http://arxiv.org/abs/1802.09129

[11] Y. Wei, J. Feng, X. Liang, M. Cheng, Y. Zhao, and S. Yan, “Object region mining with adversarial erasing: A simple classification to semantic segmentation approach,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 00, July 2017, pp. 6488–6496.

[12] K. Li, Z. Wu, K. Peng, J. Ernst, and Y. Fu, “Tell me where to look: Guided attention inference network,” *CoRR*, vol. abs/1802.10171, 2018. [Online]. Available: http://arxiv.org/abs/1802.10171