COSEGMENTATION LOSS: ENHANCING SEGMENTATION WITH A FEW TRAINING SAMPLES BY TRANSFERRING REGION KNOWLEDGE TO UNLABELED IMAGES

Wataru Shimoda & Keiji Yanai
Department of Informatics,
The University of Electro-Communications, Tokyo
1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585 Japan
shimoda-k@mm.inf.uec.ac.jp

ABSTRACT

We treat semi-supervised semantic segmentation where a few pixel-wise labeled samples and a large number of unlabeled samples are available. For this situation we propose cosegmentation loss which enables us to transfer the knowledge of a few pixel-wise labeled samples to a large number of unlabeled images. In the experiments, for the task of human-part segmentation with a few pixel-wise labeled images and 1715 unlabeled images, and proved that the proposed co-segmentation loss helped make effective use of unlabeled images.

1 INTRODUCTION

Recently, weakly-supervised semantic segmentation has drawn a large attention to reduce costs to prepare pixel-wise annotated training images which are time-consuming and costly to obtain. In weakly-supervised segmentation, we use only image-level labels for training CNNs. However, most of the existing methods on weakly-supervised segmentation have fatal drawbacks that they cannot train a segmentation network for the targets having strong co-occurrency to each other such as “train” with “railroad”, “chair” with “table”, and “boat” with “sea”. Strong co-occurrence objects will be recognized as the same class objects, since weakly-supervised approach relies on only image-labels. If all the image in a dataset contains both a car and road, it is impossible to discriminate between car regions and road regions. In fact, “desk” and “chair” are included in the Pascal VOC dataset which is widely used as a benchmark dataset in the weakly-supervised semantic segmentation, and their segmentation accuracy tends to be relatively lower.

As other datasets, there are more strong co-occurrence between targets in the images of road scene segmentation which is known as one of the practical application of semantic segmentation. Since automotive images always include “road” and “sky”, it is extremely difficult to extract road and sky regions using only image-level labels. Similarly, human-parts segmentation with weakly-supervision is extremely difficult. Instead, semi-supervision is promising where a small part of training samples are fully annotated (pixel-wise label in case of segmentation) and the rest has no labels.

In this work, we treat a semi-supervised approach to reduce pixel-wise annotation cost of semantic segmentation for the situation where strong co-occurrence between targets exists. As recent semi-supervised works for semantic segmentation, Hong et al. (2015) and Papandreou et al. (2015) exists both of which assumed the setting of weakly-supervised segmentation where image-level class labels were available for all the training samples. On the other hand, in our method, we assume that the training samples excepts a few fully-annotated samples having pixel-level labels have no class labels. Our setting is more realistic and practical than the setting of Hong et al. (2015) and Papandreou et al. (2015) where even the samples having no pixel-wise labels have image-label class labels.

In our method, we follow co-segmentation approach which estimates similar regions between two training images. We detect similar regions between fully-labeled samples and unlabeled samples, and use similar regions in the unlabeled samples to the regions in fully annotated samples for training of unlabeled samples. By using region masks in the fully-annotated samples as seeds, we search unlabeled samples for similar regions to the seed regions.

In this work, we use the PASCAL VOC human part dataset (Chen et al. (2014)) as a dataset which have co-occurrency between target classes. Especially, we use four class in the dataset, “head”, “body”, “hand”, and “leg parts”, and made segmentation experiments on these four classes. For simplicity, we use bounding box annotations of human regions for all the training images, which can be substituted in the practical situations by the results of state-of-the-art human detectors such as Faster RCNN.

Our contributions in this paper are as follows:

- We propose a semi-supervised method for semantic segmentation which requires no image-level class labels for unlabeled samples.
- We propose a novel co-segmentation loss for region knowledge transfer which takes account of region similarities between pairwise images.
- The proposed method achieved comparable or better results to the existing semi-supervised methods for semantic segmentation without using image-level class label annotation of the unlabeled samples.

2 Method

We denote an input image as x, human region mask as y, and human part region mask as z. Let $X_f = \{x_i^f, y_i^f, z_i^f\}_{i=0}^{n_f}$ be a fully annotated set of n_f samples and let $X_u = \{x_i^u, y_i^u\}_{i=0}^{n_u}$ be a unlabeled set of n_u samples. We assume that both set X_f and X_u have human region mask y in this setting.

We define the target image x_i^t which is selected from X_u as the most similar image to the x_i^f. To calculate image similarity, we use the output of a segmentation network, $h^f(\theta)$ and $h^u(\theta)$, where θ is a parameter set in a segmentation network to be optimized. We adapt masking layer (Dai et al. (2015)) to the hidden representation in order to focus on parts segmentation. The masked feature is given by element-wise product $m^f(\theta) = h^f(\theta) \cdot \hat{y}^f$, $m^u(\theta) = h^u(\theta) \cdot \hat{y}^u$, where $\hat{y}^f$ and $\hat{y}^u$ are the same size tensor to the hidden feature which is obtained by resizing and stacking y^f and y^u.

To extract feature representation of an image, we convert the heat map $m^f(\theta), m^u(\theta)$ to vectors by Global Average Pooling (GAP). Let $G : \mathbb{R}^{C \times W \times H} \to \mathbb{R}^C$ be a GAP function. We calculate the similarity for an image pair from converted feature $g^f(\theta) = G(m^f), g^u(\theta) = G(m^u(\theta))$. Target image x_i^t is obtained by $x_i^t = \arg\max_{\forall j \in n_u} D(g_i^f(\theta), g_j^u(\theta))$, where $D(.,.)$ is Euclidean distance function. We select the x_i^t for each of x_i^f during training time.

Co-segmentation is an algorithm to segment corresponding regions between pairs of images simultaneously. The insight of cosegmentation loss approach is to transfer fully annotated knowledge x_i^f to unlabeled x_i^t by mining similar regions. The network architecture used in our cosegmentation approach is shown in Fig.1 in Appendix. First, we define distinct representation for each of the human-part classes. Let $p_c^f(\theta)$ and $p_c^t(\theta)$ be the softmax output of segmentation network for class $c \in \mathcal{C}$. The maximum point for p_c can be regarded as important location for each class. $(a_c, b_c) = \arg\max_{\forall k \in z_c} p_c^f(a_k, b_k; \theta)$. We provide the location (a_c, b_c) to the masked feature $m^f(\theta)$. We define the representation of x_i^f for class c as $r_{c,i}^t(\theta) = m_i^f(a_c, b_c; \theta)$. We explore similar regions to the r_c^f from $m_c^t(\theta)$. Vector distance $d_c(a, b)$ can be computed by $d_c(a, b; \theta) = D(r_c^f, m_c^u(a, b; \theta))$. Let $\hat{d}$ be sorted result for d. We define the space $s \in \mathcal{S}$ which satisfies following condition: $\hat{d}_c(a_s, b_s; \theta) > \hat{d}_c(a_{s+1}, b_{s+1}; \theta)$. Cosegmentation loss is represented by:

$$\mathcal{L}_{coseg} = -\frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{1}{n_c^f/2} \sum_{s=0; s \in \mathcal{S}}^{n_c^f/2} \log p^t(\theta) \tag{1}$$

where n_c^f is the number of pixels z. It is better than fix value or random value, since the image similarity is related to the region size. Cosegmentation loss relies on $r_c^f(\theta)$, so that we need to combine this loss: $\mathcal{L}_{humanparts} = -\frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{1}{|z_c^f|} \sum_{k \in z_c^f} \log p_c^f(a_k, b_k; \theta)$。 For additional loss, we use $\mathcal{L}_{human}$ loss, since masking layer is not enough for eliminating background factor. $\mathcal{L}_{human} = -\frac{1}{|y_1^t|} \sum_{k \in y_1^t} \log(1 - p_0^t(a_k, b_k; \theta))$。 Our final loss is as follows: $\mathcal{L}_{all} = \mathcal{L}_{humanparts} + \alpha \mathcal{L}_{human} + \beta \mathcal{L}_{coseg}$. We set the parameter α, β with 0.001 for generating better representation $r^f(\theta)$ before decreasing the loss $\mathcal{L}_{coseg}$.

3 Experimental Results

We performed experiments on human-part segmentation with the Pascal VOC human-part dataset (Chen et al. (2014)). The dataset contains 1715 training images and 1818 validation images with human-part annotation. We randomly selected a few (10 to 100) images as a few fully-annotated samples from 1715 training images. Fig. in Appendix shows examples of 10 selected fully-annotated samples. We selected four human parts, “head”, “body”, “hand” and “leg”, as the target part classes by merging some classes provided by the authors. In the experiments, we trained the segmentation network for each of the human-part models separately. That is, the segmentation network outputs segmentation masks of three classes, human part, other region and backgrounds. The evaluation protocol is based on a simple mean intersection over union (IOU). In evaluation, we do not take care of the background class. We used the DeepLab large fov model (Chen et al. (2015)) as a segmentation network, and initialized parameters provided by the authors which was trained with ImageNet. We set the batch size to 24 which was the maximum size for our GPU environment, and in each mini-batch we assigned the size of the full supervised data to 1 and the size of the unlabeled data size to 23 for a large search space.

We show the experimental results in Table 1. As baselines we prepared two simple full supervised model, (a) and (c). The difference in the models is only the number of images. We trained (a) with 10 selected images and (c) with all 1715 images. On the other hand, for our proposed method (b), we used 10 pixel-wise labeled images and 1715 unlabeled images. Our method improved the results from (a) consistently for each of the target parts. Particularly, our method enhanced the result on “hand” in which the category score is only 2.5% in the (a) method. Fig.4 in Appendix shows qualitative results. In many cases, our method improved the segmentation results compared to the result of (a), and some results became close to the result of (c). These results indicated that co-segmentation loss has the ability to transfer the region knowledge to unlabeled images. The reason why the scores on ”others” were lower than (a) is that the method of (a) had the strong tendency that segments are classified as “other regions” in general, which brought better results on the scores of “others”.

Table 2 in Appendix shows the result of simultaneous estimation of four-class human parts. The evaluation was performed in the same way as Table 1. The graph in Fig. in Appendix shows the Mean IoUs of the same experiments regarding the proposed method and three baselines including EM-adapt (Papandreou et al. (2015)), only using fully-supervised samples and the method based on Global Max Pooling. From these results, the proposed method is still effective for simultaneous estimation. This tables also shows the results when varying the number of pixel-wise labeled images from 10 to 100. In the most cases, the results by the proposed method using unlabeled images were comparable or superior to the baseline results.

4 Conclusions

We have proposed cosegmentation loss which is a novel region knowledge transferring method. We showed that the approach was effective for enhancing a few full-supervised data with a large amount of unlabeled data. For the dataset where strong co-occurrency between the target classes exists, we achieved comparable or better results to the existing semi-supervised methods for semantic segmentation without using image-level class label annotation of the unlabeled samples.

We expect that the recent method on weakly supervised segmentation which is combination of generating region seeds and re-training segmentation results using the seeds is effective even for semi-supervised semantic segmentation. For future work, we plan to introduce this method to semi-supervised segmentation with the same setting as this paper.

Table 1: Mean IoU score (%) for the each human part segmentation results.

Method	n^f	n^u	head		body		hand		leg	
			head	others	body	others	hand	others	leg	others
(a) Fully supervised only	10	0	43.7	67.2	13.8	50.2	2.5	60.4	15.0	62.7
(b) Cosegmentation loss	10	1715	50.0	68.4	20.6	44.8	11.7	51.7	29.9	60.4
(c) Fully supervised only	1715	0	56.1	71.3	36.4	54.2	23.0	63.1	39.2	67.8

References

L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and Yuille A. L. Semantic image segmentation with deep convolutional nets and fully connected CRFs. In *Proc. of International Conference on Learning Representations*, 2015.

X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2014.

J. Dai, K. He, and J. Sun. Convolutional feature masking for joint object and stuff segmentation. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2015.

S. Hong, H. Noh, and B. Han. Decoupled deep neural network for semi-supervised semantic segmentation. In *Advances in Neural Information Processing Systems*, 2015.

G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille. Weakly-and semi-supervised learning of a dcnn for semantic image segmentation. In *Proc. of IEEE International Conference on Computer Vision*, 2015.

D. Pathak, E. Shelhamer, J. Long, and T. Darrell. Fully convolutional multi-class multiple instance learning. In *Proc. of International Conference on Learning Representations*, 2015.

Appendix

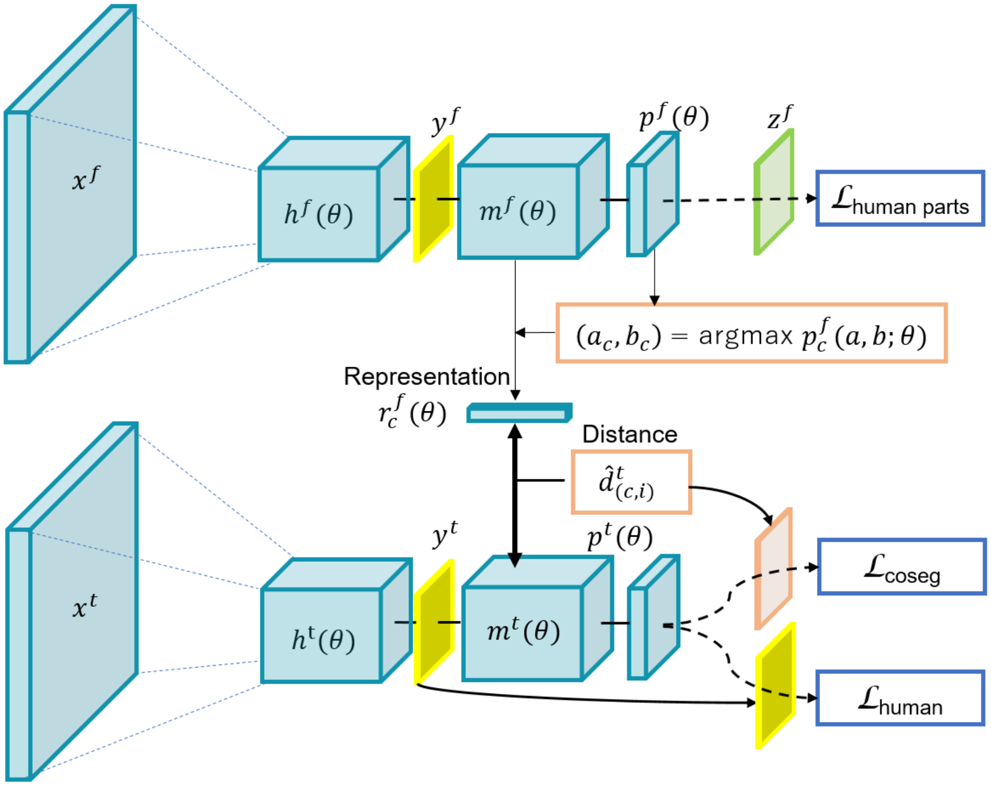

Figure 1: The proposed network architecture. We mined the point (a_c, b_c) from the softmax output $p_c^f(\theta)$. $r_c^f(\theta)$ is the representation for the input x^f. We defined the $r_c^f(\theta)$ as the point of masked feature $m^f(\theta)$. Cosegmentation loss used similarity between $r_c^f(\theta)$ and $m^t(\theta)$.

Figure 2: 10 selected pixel-wise labeled images as a few fully annotated samples.

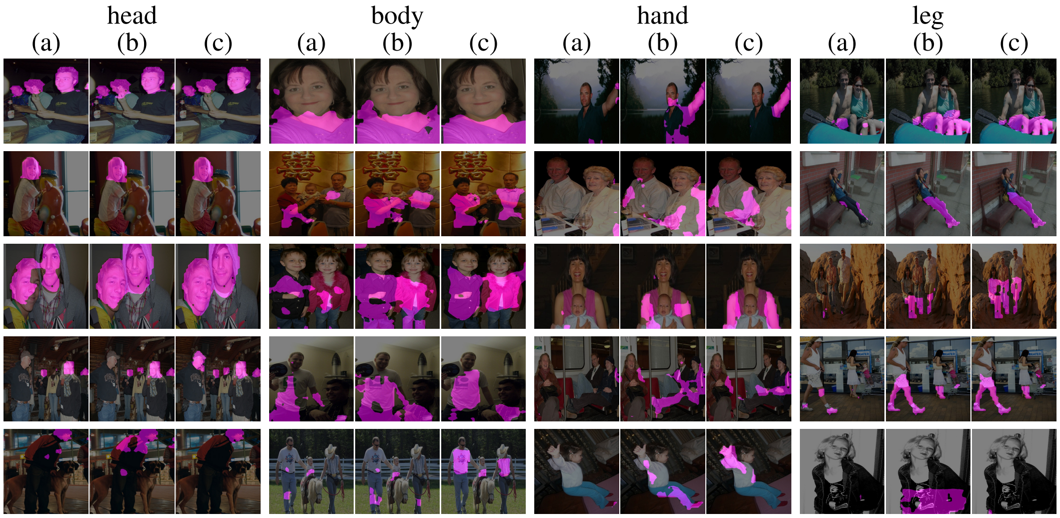

Figure 3: Qualitative results for the proposed method (b) and the baselines (a),(c). In the top four rows, successful results are shown, and in the bottom row, failure results are shown.

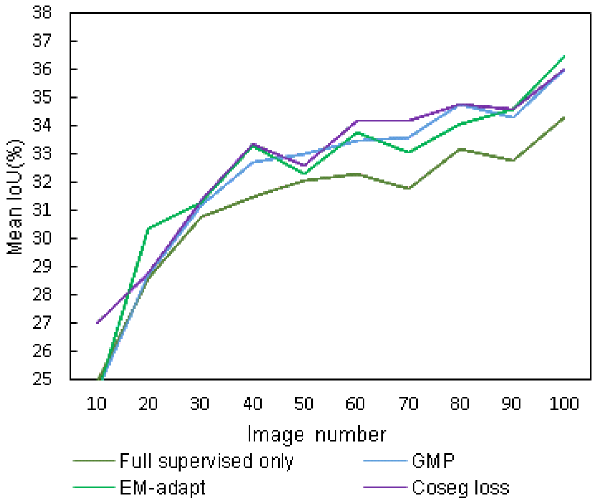

Figure 4: The average Mean IoU for each of the number of fully-annotated samples for the proposed method and the baselines.

Table 2: Mean IoU score (%) for the each human part segmentation results.

Method	n^f	n^u	head	body	hand	leg	mean
Fully supervised only	10	0	42.2	26.4	7.6	23.6	24.9
Pathak et al. (2015) (with image-label)	10	1715	35.8	21.9	**14.8**	**26.0**	24.6
Papandreou et al. (2015) (with image-label)	10	1715	42.5	**30.3**	6.7	18.2	24.4
Cosegmentation loss	10	1715	**44.1**	25.5	12.8	25.6	**27.0**
Fully supervised only	20	0	44.8	22.7	14.3	**32.7**	28.6
Global Max Pooling (with image-label)	20	1715	42.0	24.0	**18.2**	30.6	28.7
Em-adapt (with image-label)	20	1715	**46.6**	**27.6**	16.1	31.3	**30.4**
Cosegmentation loss	20	1715	45.0	23.0	15.3	31.7	28.8
Fully supervised only	30	0	46.5	27.3	16.8	**32.7**	30.8
Global Max Pooling (with image-label)	30	1715	46.3	29.2	18.6	30.6	31.2
Em-adapt (with image-label)	30	1715	**48.4**	**29.5**	18.4	28.9	31.3
Cosegmentation loss	30	1715	47.1	28.4	**18.8**	31.3	**31.4**
Fully supervised only	40	0	48.1	29.5	18.5	30.0	31.5
Global Max Pooling (with image-label)	40	1715	47.5	28.8	**21.6**	32.7	32.7
Em-adapt (with image-label)	40	1715	**49.7**	**31.8**	20.8	30.7	33.3
Cosegmentation loss	40	1715	49.3	31.3	19.7	**33.1**	**33.4**
Fully supervised only	50	0	47.5	27.9	20.3	32.8	32.1
Global Max Pooling (with image-label)	50	1715	**48.8**	**30.4**	**22.6**	30.2	**33.0**
Em-adapt (with image-label)	50	1715	47.7	30.3	21.5	29.5	32.3
Cosegmentation loss	50	1715	48.2	29.0	22.2	**30.9**	32.6
Fully supervised only	60	0	50.1	29.3	21.1	28.5	32.3
Global Max Pooling (with image-label)	60	1715	50.8	31.4	21.3	**30.7**	33.5
Em-adapt (with image-label)	60	1715	**52.2**	**32.7**	21.9	28.4	33.8
Cosegmentation loss	60	1715	**52.2**	31.7	**22.5**	30.4	**34.2**
Fully supervised only	70	0	50.2	31.1	18.4	27.6	31.8
Global Max Pooling (with image-label)	70	1715	**51.2**	31.0	**22.0**	30.1	33.6
Em-adapt (with image-label)	70	1715	50.0	**32.6**	19.7	30.2	33.1
Cosegmentation loss	70	1715	49.9	31.7	21.7	**31.3**	**33.7**
Fully supervised only	80	0	49.7	28.9	20.5	33.6	33.2
Global Max Pooling (with image-label)	80	1715	**51.9**	30.9	**22.3**	33.5	**34.8**
Em-adapt (with image-label)	80	1715	50.7	31.1	21.1	33.4	34.1
Cosegmentation loss	80	1715	51.6	**31.2**	21.0	**35.4**	**34.8**
Fully supervised only	90	0	50.0	30.6	21.6	29.1	32.8
Global Max Pooling (with image-label)	90	1715	51.5	31.5	23.1	31.1	34.3
Em-adapt (with image-label)	90	1715	52.1	**32.4**	22.7	31.3	**34.6**
Cosegmentation loss	90	1715	**52.6**	30.7	**23.3**	**31.8**	**34.6**
Fully supervised only	100	0	50.5	31.0	22.8	33.0	34.3
Global Max Pooling (with image-label)	100	1715	52.8	31.9	**25.3**	33.9	36.0
Em-adapt (with image-label)	100	1715	**53.0**	**33.9**	24.7	34.3	**36.5**
Cosegmentation loss	100	1715	51.7	32.5	24.2	**35.6**	36.0
Fully supervised only	1715	0	55.8	41.6	25.6	39.5	40.6