

Predicting Segmentation “Easiness” from the Consistency for Weakly-Supervised Segmentation

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, yanai}@mm.inf.uec.ac.jp

Abstract

Weakly-supervised segmentation has come to draw a lot of attention, since it costs very high to create pixel-wise annotated image datasets for fully-supervised segmentation. Recently, it has achieved great progress by the method of (re-)training of a fully-supervised segmentation model with roughly estimated initial masks which is proposed by Wei et al. [25]. However, the initial estimated masks tend to include some noise, which sometimes causes erroneous results. Therefore in this paper we focus on improving of quality of initial estimated masks for (re-)training of a fully-supervised segmentation model. We propose a novel algorithm to retrieve “good seeds” by predicting segmentation “Easiness” of images based on consistency among the outputs with different conditions. We show that there is a trade-off between training data quality and the number of selected images, and our proposed method can improved the trained model using data augmentation. We have achieved state-of-the-art in a weakly-supervised segmentation setting on Pascal VOC 2012 segmentation benchmark dataset.

1. Introduction

Due to recent progress of convolutional neural network (CNN), the accuracy of semantic segmentation has been much improved, especially in fully-supervised semantic segmentation which requires pixel-wise annotation as training data. However, pixel-wise annotation is very costly to obtain in general. On the other hand, collecting images with image-level annotation is much easier than those with pixel-level annotation, since many images attached with tags are available on hand-crafted open image datasets such as ImageNet as well as on the Web. Thus, weakly-supervised semantic segmentation which requires not pixel-wise annotation nor bounding box annotation but only image-level annotation has been explored actively. Before the CNN era, the performance of weakly-supervised semantic segmentation was too low to regard it as being practical approaches. CNN changed such situation greatly. Some re-

cent methods [25, 10] achieved more than 50% as IoU accuracy for PASCAL 2012 dataset, which has outperformed the performance of fully-supervised semantic segmentation methods [2] proposed in 2014 which used training data with pixel-wise annotation. As CNN-based methods on weakly supervised semantic segmentation which employs the localization ability of CNN for target objects, there exist a feed-forward based method and a backward based method. As a different approach from them, in 2016, Wei et al. [25] proposed iterative training of CNN-based fully-supervised segmentation models for weakly supervised segmentation. In the first step, they estimated object regions for easier training images using a method on unsupervised saliency map estimation. Then, they trained state-of-the-art fully-supervised model, DeepLab [4], with training images and object masks estimated by saliency maps. After the second step, they re-estimated object masks with the supervised segmentation model trained in the previous step for more diverse training images including more complex ones, and (re)-trained DeepLab again with the estimated masks. They showed that repeating estimation and training like EM algorithm helped create more robust segmentation CNN model which has tolerance to noise in pseudo-pixel-wise training data consisting of image-level-annotated images and artificially generated segmentation masks. Their method proved that state-of-the-art fully-supervised methods with pseudo-pixel-wise training samples worked as capable weakly-supervised segmentation methods which sometimes outperformed the existing state-of-the-art methods. This is a kind of a break-through in weakly-supervised semantic segmentation.

At present, the difference on segmentation performance between state-of-the-art fully-supervised methods and weakly-supervised methods is still about 20% regarding the PASCAL 2012 dataset. This comes from erroneous segmentation masks estimated by saliency maps or trained CNNs. Especially, noise in the initial seeds affects quality of final results greatly. From this observation, we think it is the most important to obtain better initial seeds for better final results. Therefore, in this paper, we focus on improving

a method to produce initial seeds.

In this paper, we propose a novel algorithm to estimate the accuracy of weakly supervised segmentation results with unsupervised approach and retrieved “good seeds” by the evaluation. For the accuracy estimation of segmentation, we consider “consistency” among the results with different conditions. To do that, we use two kinds of weakly-supervised segmentation method, a back-propagation-based region mask estimation method proposed by Simonyan et al [22] and an improved method proposed by Shimoda et al [21]. By evaluating the consistency between the results by the two methods, we estimate segmentation “Easiness” of each of training image, select the easier ones as “seed images” and regards their estimated masks as “seed masks”. In addition, it is also effective to introduce data augmentation to make seed images more various after selecting seed images.

To summarize our contributions in this paper, they are as follows:

- We propose a novel algorithm to estimate “Easiness” by consistency among results of different conditions.
- We show that the retrieved images by our proposed method is effective for data augmentation.
- We achieved state-of-the-art on the Pascal VOC 2012 benchmark dataset in the weakly supervised settings without additional supervision nor additional training samples.

2. Related Work

As CNN-based weakly-supervised semantic segmentation methods, three approaches exists.

1. Feed-forward based method. Feature map activations are used for estimating target object regions.
2. Backward based method. Back-propagation to the input image can estimate rough object regions, which was originally proposed as a visualization method of CNNs by Simonyan et al [22].
3. EM-like method employing a fully-supervised method. Alternate iteration of segmentation mask estimation and (re-)training of a fully-supervised segmentation method enables a weakly-supervised segmentation with a fully-supervised method.

2.1. Feed-Forward Based Methods

Sermanet et al [20] proposed fully convolutional networks (FCN) which accept input images of arbitrary sizes by replacing fully-connected layers with convolutional layers. A FCN generates object heatmaps which indicate rough object regions. Oquab et al.[14] proposed to use Global

Max Pooling (GMP) in the last layer to allow input images of arbitrary sizes to be used even in training time, which improved accuracy of weakly-supervised object localization. After that, some derived methods employing GMP were proposed by Pinheiro et al. [18] and Zhou et al.[26].

2.2. Backward-Based Methods

Simonyan et al. [22] showed that object segmentation without pixel-wise training data can be done by using back-propagation processing. Springenberg et al. [23] also proposed a method for object localization by back-propagating the derivatives of a maximum loss value of the object detected in the feed-forward computation. Jianming et al. [9] showed that their method outperformed feed-forward-based methods employing Global Average/Max Pooling, while Shimoda et al. [21] achieved the state-of-the-art results in weakly supervised semantic segmentation for the PASCAL2012 dataset except for iterative methods by using estimated class maps as prior of dense CRF.

2.3. EM-Like Methods Employing Fully-Supervised Methods

Pathak et al. [17, 16] and Papandreou et al. [15] proposed weakly-supervised semantic segmentation by adapting CNN models for fully-supervised segmentation to weakly-supervised segmentation. Both CCNN and EM-adopt generated pseudo-pixel-level labels from image-level labels using constraints and EM algorithms to train FCN and DeepLab which were originally proposed for fully supervised segmentation, respectively. Both showed Dense CRF [11] were helpful to boost segmentation performance even in the weakly supervised setting.

Recently Wei et al.[25] proposed iterative process of region mask estimation and (re-)training CNN by changing training samples from easy images to complex images. They showed that their methods achieved around 50% mean IoU at the PASCAL VOC 2012 dataset for the first time, which was almost equivalent to the performance of fully supervised methods proposed several years ago, although this method used 41,625 extra training images in addition to the commonly-used augmented Pascal VOC training dataset [6]. After that, Kolesnikov et al. [10] achieved robust seed generation by using Global Average Pooling [26].

Though most of weakly supervised segmentation methods use correlation of the different task “Classification” and “Segmentation”, some recent weakly-supervised approaches achieved boosting accuracy by additional supervision signals. Chen et al.[15] proposed to use bounding box annotation for weakly-supervised semantic segmentation, which can be regarded as being less costly than pixel-wise annotation but still costly than only image-level annotation. As cheaper additional annotation, point annotation [1] and checking generated initial masks by crowd-

sourcing [19] were proposed, which utilized minimal additional supervision by human. Tokmakov et al. [24] proposed to use motion segmentation of videos as additional training information for weakly supervised segmentation. On the other hand, SEC [10] showed a high performance with only unsupervised techniques of loss function for neural network. The research for the unsupervised approach will be worth for the further weakly supervised segmentation research progress. We consider that human sometimes makes decision from consistency of the observations for the ambiguous problem. Actually, the decision by majority is widely used in real world. In this paper, we explored a weakly-supervised way without additional supervision nor additional training samples and estimated priority of weakly supervised segmentation result from consistency and retrieved “good seeds”.

3. Method

In this paper, we basically adopt iterative approach of mask estimation and training of fully-supervised semantic segmentation model in the similar way to [25, 10] for the weakly-supervised semantic segmentation tasks. To estimate initial masks which need training of a fully-supervised model in the weakly-supervised task, we use Distinct Class-specific Saliency Maps (DCSM) proposed by Shimoda et al. [21] which achieved the state-of-the-art in the PASCAL VOC 2012 dataset except for iterative methods. In this paper, in order to obtain better initial masks, we pay attention on “consistency” among the results of different processing for selecting “good seeds”.

3.1. Estimation of “Easiness” of Training Images

We estimate “Easiness” of training images at first. We paid attention to the following two points:

1. Correlation on “Easiness” between classification and segmentation.
2. Coherence on the segmentation results between one obtained by a sophisticated method and one obtained by a simpler method.

From the two assumptions, we select easier images from the training dataset, and give priority to them in the initial training phase.

3.1.1 Difference between DCSM and DCSM without subtraction

It is easy to imagine that the images to be classified is hard to be segmented. However, the easy-classified images are not always easy for segmentation. Therefore it is difficult to estimate “Easiness” on segmentation directly from classification results. Thus, in this paper, we utilize the BP-based object-specific saliency map estimation, DCSM [21].

In the method proposed by Simonyan et al. [22], their class saliency maps are relatively vague and not distinct. In addition, when different kinds of target objects are included in the image, the maps tend to respond to all the object regions. To resolve the weaknesses of their method, Shimoda et al. [21] proposed a method, DCSM, to generate more distinct class saliency maps which discriminate the regions of a target class from the regions of the other classes by subtracting saliency maps of the other classes from saliency maps of the target class to differentiate target objects from other objects.

Here, we think about the difference between the original DCSM and the DCSM without subtraction. If no difference appears in both results, the input images can be regarded as being simple images containing single kinds of objects. On the other hand, if both results are largely different, the input images can be regarded as being complex images containing multiple kinds of images.

For image x , let $V_o(x)$ be segmentation result without subtraction, $V_w(x)$ be segmentation result with subtraction. “Easiness” for subtraction $R_{sub}(x)$ is calculated as:

$$R_{sub}(x) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} IoU(V_o^c(x), V_w^c(x)) \quad (1)$$

where $IoU(\cdot, \cdot)$ is a function which returns the Intersection over Union (IoU) for two regions, \mathcal{C} is the set for the difference input image sizes.

3.1.2 Coherence on size change of input images

As additional measurements on “Easiness”, we also think coherence of the estimated DCSM maps when varying the size of input images. If the coherence is kept widely in terms of size change, the given images can be regarded as being simpler than the images the results of which are changed for size change. In this paper, we use this as the second evaluation measure of “Easiness” of training images.

In the experiments, we used the three sizes, $s_n = 320, 416, 512 (n = 0, 1, 2)$. We represent DCSM maps before adapting CRF as $M^{s_n}(x)$. We obtain aggregated maps, $M^b(x)$ with

$$M^b(x) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} M^{s_c}(x)$$

$V^b(x)$ represents the CRF result of $M^b(x)$ after applying the dense CRF-based refinement. Then, we compute the coherence on size change, $R_{size}(x)$, by the following equation:

$$R_{size}(x) = \frac{1}{2|\mathcal{C}|} \sum_{c \in \mathcal{C}} IoU(V_o^b(x), V_o^c(x)) + IoU(V_w^b(x), V_w^c(x)) \quad (2)$$

Finally we combine two kinds of the reliable scores in the following equation:

$$score = \lambda_1 \cdot R_{size} + \lambda_2 \cdot R_{sub} \quad (3)$$

where λ_1 and λ_2 are pre-defined constant values. In this paper we simply set λ_1 and λ_2 to 0.5. Fig 1 shows the example of the results of the estimated mask in different conditions.








































	input size	320	416	512	320	416	512
	visualization						
	CRF result						
		w/o sub			w/ sub		
	input size	320	416	512	320	416	512
	visualization						
	CRF result						
		w/o sub			w/ sub		
	input size	320	416	512	320	416	512
	visualization						
	CRF result						
		w/o sub			w/ sub		

Figure 1. Examples of visualization on the different conditions. In the top case, we define this sample is “good seeds” so that almost predicted regions have consistency. In the middle case, the visualization results have large difference in the simple visualization and visualization with subtraction, hence we estimate the result of this sample is bad for the (re-)training. In the bottom case, visualization results have corresponding region, but some results have inconsistency in the results of varying input size, thus our proposed method generate a not good score for this sample.

3.2. Generating of segmentation mask

We generate segmentation masks of the training images with only image-level annotation by using DCSM [21]. The final results are obtained after applying dense CRF. On the contrary to the original DCSM [21], we used single-class classifier CNNs as well and we generate the final mask by integrating single-class classifier results with the multi-class classifier results. In case of PASCAL dataset, we train each single-class CNNs with softmax cross entropy loss. Fig 2 shows the examples of the generated mask. Note that we used only corresponding regions for results of different conditions as the training data such like localization cue used in [10].



Figure 2. (1)input image, (2)estimated mask by single-class model, (3)estimated mask by multi-class model, (4)integrated mask

4. Experiment

4.1. Dataset

We evaluated the proposed method on PASCAL VOC 2012 segmentation benchmark[5]. We followed the common practice to augment the training data provided by [6]. There are 10,582 training images, 1,449 validation images and 1,456 test images. In this benchmark, although the PASCAL VOC dataset contains 20 classes, we need to classify 21 classes including the background class.

4.2. Experimental setup

For the setup of both of the multi-class and single-class classification model, we followed [21]. To train fully-supervised segmentation model with the estimated masks we used DeepLab-CRF model of [3]. To optimize the model we used SGD for 10000 iteration, the batch size is 16, momentum parameter is 0.9 and a weight decay is 0.0005. We set the learning rate to 0.001 except for the last layer which learning rate is 0.01. We decrease the learning rate by 0.1 for 2000 iteration. Each model is trained with 7-8 hours by a NVIDIA GeForce Titan GPU with 12GB memory. All the experiments are conducted using DeepLab code [3], which is implemented based on the publicly available Caffe framework [8].

4.3. Evaluation for the estimation of “Easiness”

Fig 3 shows top5 retrieval results of each class obtained by our proposed algorithm “Easiness”. Although our proposed model is based on unsupervised approach, in most cases “good seeds” are retrieved. For example, in case of aeroplane, car, cow and dog, retrieved seeds are close to the ground truth. On the other hand, the results of sofa, chair and table include some noise. In general segmentation results of these class show low performance, hence the retrieved results are affected by the low quality of prediction directly.

Data augmentation potential is well known as the way for avoiding overfitting and improving accuracy on test data. However, it is expected that augmented data including noise will be not effective for improving accuracy. Therefore, we augmented training data for only “good seeds” retrieved by our proposed algorithm. As the data augmentation method, we referred approach of Liu et al [12], which is mentioned on their poster in the conference. In fully supervised detection they changed training data greatly by data augmentation, then we followed their approach [12] and augmented data by dynamic random cropping and random padding. For cropped images, we recognized the images by the multi-class classification model and used only the images which results corresponding to the class label. For each training data and each augmentation process we augmented 10 images. Table 1 shows result of combination of “Easiness” with data augmentation, where the image number is defined by the threshold of equation 2. Base image N represents the image number which is used as the training data without data augmentation, while aug image N indicates the number of image used for data augmentation. As the results, setting (c) achieved the best accuracy 51.3 pt, this setting limits both base training image number and augmented image number. Our proposed method improved the simple approach certainly, considering result of setting (f) score is 48.8 pt which was trained with all training images and augmented all images, this is the lowest score in all settings. The table also show that training data selection for base images is effective constantly. In setting (a) and (b), in order to collect the training data which has further quality, we limited augmented number of image to 780, but we got the worse results. In training deep CNN, the number of image and training data quality is trade off, however, these results indicate that the accuracy can be boost by data selection and even though the training data size is small, quality of data is important and can be used for the data augmentation effectively.

setting	Base image N	Aug image N	mIoU
(a)	8760 (th ≥ 0.3)	730 (th ≥ 0.8)	50.1
(b)	10582 (all)	730 (th ≥ 0.8)	48.9
(c)	8760 (th ≥ 0.3)	2105 (th ≥ 0.7)	51.3
(d)	10582 (all)	2105 (th ≥ 0.7)	49.9
(e)	8760 (th ≥ 0.3)	8760 (th ≥ 0.3)	49.7
(f)	10582 (all)	10582 (all)	48.8

Table 2, Table 3 show comparison with the other weakly supervised segmentation methods. Our method achieved state-of-the-art on the same condition using only image-level-label as training data. Especially our proposed method achieved better result than F/B prior[19], STC [25], SEC [10] score which methods employ (re)-trained DeepLab again with the estimated masks. Our approach also outperformed SDS [7] which is based on fully supervised method

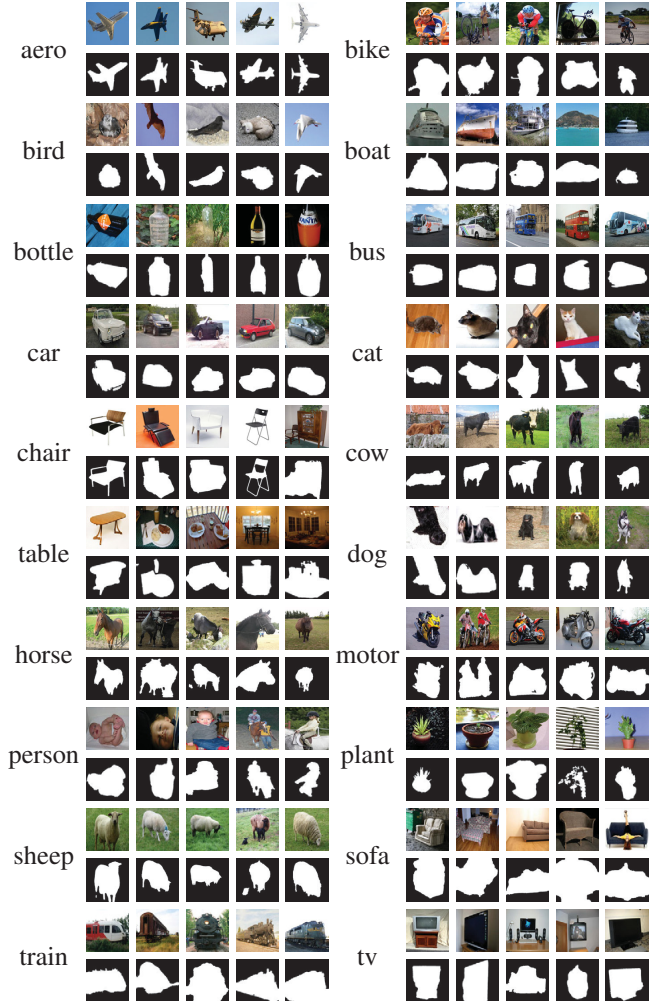


Figure 3. Top5 retrieval results obtained by our proposed “Easiness” score on Pascal VOC 2012 train_aug dataset.

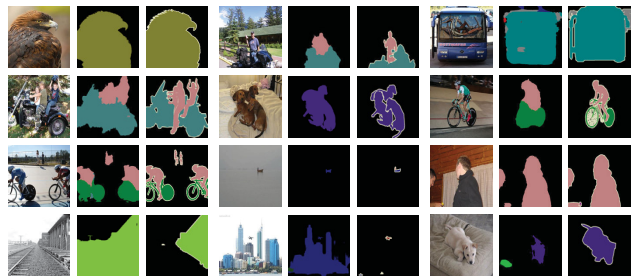


Figure 4. From left to right, input image, prediction and ground truth

proposed in 2014. We expect that our proposed method will be well combine with some other weakly supervised techniques, for example, self-paced learning [25] or seed expanding and constrain approach [10], so that our proposed approach is different from those in terms of adapting data selection. Fig 4 shows examples of our proposed method.

Table 2. Results on PASCAL VOC 2012 *val set*.

Methods	additional images																					mIoU	
		bg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train		tv
MIL-FCN [17]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	25.7
EM-Adapt [15]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	38.2
CCNN [16]	-	65.9	23.8	17.6	22.8	19.4	36.2	47.3	46.9	47.0	16.3	36.1	22.2	43.2	33.7	44.9	39.8	29.9	33.4	22.2	38.8	36.3	34.5
MIL-sppxl [18]	✓	77.2	37.3	18.4	25.4	28.2	31.9	41.6	48.1	50.7	12.7	45.7	14.6	50.9	44.1	39.2	37.9	28.3	44.0	19.6	37.6	35.0	36.6
MIL-bb [18]	✓	78.6	46.9	18.6	27.9	30.7	38.4	44.0	49.6	49.8	11.6	44.7	14.6	50.4	44.7	40.8	38.5	26.0	45.0	20.5	36.9	34.8	37.8
MIL-seg [18]	✓	79.6	50.2	21.6	40.6	34.9	40.5	45.9	51.5	60.6	12.6	51.2	11.6	56.8	52.9	44.8	42.7	31.2	55.4	21.5	38.8	36.9	42.0
DCCSM w/ CRF [21]	-	76.7	45.1	24.6	40.8	23.0	34.8	61.0	51.9	52.4	15.5	45.9	32.7	54.9	48.6	57.4	51.8	38.2	55.4	32.2	42.6	39.6	44.1
F/B prior [19]	-	79.2	60.1	20.4	50.7	41.2	46.3	62.6	49.2	62.3	13.3	49.7	38.1	58.4	49.0	57.0	48.2	27.8	55.1	29.6	54.6	26.6	46.6
STC [25]	✓	84.5	68.0	19.5	60.5	42.5	44.8	68.4	64.0	64.8	14.5	52.0	22.8	58.0	55.3	57.8	60.5	40.6	56.7	23.0	57.1	31.2	49.8
SEC [10]	-	82.4	62.9	26.4	61.6	27.6	38.1	66.6	62.7	75.2	22.1	53.5	28.3	65.8	57.8	62.3	52.5	32.5	62.6	32.1	45.4	45.3	50.7
Ours	-	81.6	64.9	25.8	71.4	29.2	57.8	75.2	68.0	72.7	15.2	46.6	33.8	56.7	57.1	60.9	60.7	24.1	65.4	31.5	43.9	35.3	51.3

Table 3. Results on PASCAL VOC 2012 *test set*.

Methods																					mIoU	
	bg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train		tv
Fully Supervised:																						
O2P [2]	85.4	69.7	22.3	45.2	44.4	49.6	66.7	57.8	56.2	13.5	46.1	32.3	41.2	59.1	55.3	51.0	36.2	50.4	27.8	46.9	44.6	47.6
SDS [7]	86.3	63.3	25.7	63.0	39.8	59.2	70.9	61.4	54.9	16.8	45.0	48.2	50.5	51.0	57.7	63.3	31.8	58.7	31.2	55.7	48.5	51.6
FCN-8s [13]	-	76.8	34.2	68.9	49.4	60.3	75.3	74.7	77.6	21.4	62.5	46.8	71.8	63.9	76.5	73.9	45.2	72.4	37.4	70.9	55.1	62.2
Deeplab Large FOV [3]	92.6	83.5	36.6	82.5	62.3	66.5	85.4	78.5	83.7	30.4	72.9	60.4	78.5	75.5	82.1	79.7	58.2	82.0	48.8	73.7	63.3	70.3
Using Additional Supervision:																						
CCNN w/size [16]	-	42.3	24.5	56.0	30.6	39.0	58.8	52.7	54.8	14.6	48.4	34.2	52.7	46.9	61.1	44.8	37.4	48.8	30.6	47.7	41.7	45.1
One point[1]	80.6	50.2	23.9	38.4	33.1	38.5	52.0	50.9	55.4	18.3	38.2	37.7	51.0	46.1	54.7	43.2	35.4	45.1	33.0	49.6	40.0	43.6
F/B prior + CheckMask [19]	87.4	65.7	26.0	64.2	43.7	53.2	72.6	63.6	59.5	17.1	48.0	43.7	61.2	52.0	69.3	54.8	43.0	50.3	34.6	59.2	42.0	52.9
Weakly Supervised:																						
MIL-FCN [17]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	24.9
EM-Adapt [15]	76.3	37.1	21.9	41.6	26.1	38.5	50.8	44.9	48.9	16.7	40.8	29.4	47.1	45.8	54.8	28.2	30.0	44.0	29.2	34.3	46.0	39.6
CCNN [16]	-	21.3	17.7	22.8	17.9	38.3	51.3	43.9	51.4	15.6	38.4	17.4	46.5	38.6	53.3	40.6	34.3	36.8	20.1	32.9	38.0	35.5
MIL-ILP-seg [18]	78.7	48.0	21.2	31.1	28.4	35.1	51.4	55.5	52.8	7.8	56.2	19.9	53.8	50.3	40.0	38.6	27.8	51.8	24.7	33.3	46.3	40.6
DCCSM w/ CRF [21]	78.1	43.8	26.3	49.8	19.5	40.3	61.6	53.9	52.7	13.7	47.3	34.8	50.3	48.9	69.0	49.7	38.4	57.1	34.0	38.0	40.0	45.1
F/B prior [19]	80.3	57.5	24.1	66.9	31.7	43.0	67.5	48.6	56.7	12.6	50.9	42.6	59.4	52.9	65.0	44.8	41.3	51.1	33.7	44.4	33.2	48.0
STC [25]	85.2	62.7	21.1	58.0	31.4	55.0	68.8	63.9	63.7	14.2	57.6	28.3	63.0	59.8	67.6	61.7	42.9	61.0	23.2	52.4	33.1	51.2
SEC [10]	83.5	56.4	28.5	64.1	23.6	46.5	70.6	58.5	71.3	23.2	54.0	28.0	68.1	62.1	70.0	55.0	38.4	58.0	39.9	38.4	48.3	51.7
Weakly Supervised:																						
Ours	83.0	67.5	29.7	69.7	28.8	59.7	71.2	66.4	69.8	18.6	49.8	44.7	49.4	60.5	73.5	61.8	32.7	62.7	39.0	34.3	36.5	52.8

5. Conclusion

In this work, we estimate “Easiness” of prediction for segmentation from visualization results and we trained fully supervised segmentation model with retrieved prediction results by “Easiness”. In training deep CNN, the image number and training data quality is trade off, however we showed that we can boost the segmentation accuracy by combining data selection with data augmentation. In addition, we achieved state-of-the-art in weakly-supervised segmentation setting on the Pascal VOC 2012 bench mark.

References

- A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei. What’s the point: Semantic segmentation with point supervision. In *ECCV*, 2016. 2, 6
- J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*, 2012. 1, 6
- L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and Y. A. L. Semantic image segmentation with deep convolutional nets and fully connected CRFs. In *ICLR*, 2015. 4, 6
- L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015. 1
- M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. 4
- B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and M. J. Semantic contours from inverse detectors. In *ICCV*, 2011. 2, 4
- B. Hariharan, P. Arbelaez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *ECCV*, 2014. 5, 6
- Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding. <http://caffe.berkeleyvision.org/>, 2013. 4
- Z. Jianming, L. Zhe, B. Jonathan, S. Xiaohui, and S. Sclaroff. Top-down neural attention by excitation backprop. In *ECCV*, 2016. 2
- A. Kolesnikov and C. H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*, 2016. 1, 2, 3, 4, 5, 6
- P. Krahenbuhl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in Neural Information Processing Systems*, 2011. 2
- W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 5
- J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 6
- M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014. 2
- G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille. Weakly-and semi-supervised learning of a dcnn for semantic image segmentation. In *ICCV*, 2015. 2, 6
- D. Pathak, P. Krahenbuhl, and T. Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*, 2015. 2, 6
- D. Pathak, E. Shelhamer, J. Long, and T. Darrell. Fully convolutional multi-class multiple instance learning. In *ICLR*, 2015. 2, 6
- P. Pedro and C. Ronan. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, 2015. 2, 6
- F. Saleh, M. Akbarian, M. Salzmann, L. Petersson, S. Gould, and J. M. Alvarez. Built-in foreground/background prior for weakly-supervised semantic segmentation. In *ECCV*, 2016. 3, 5, 6
- P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *Proc. of International Conference on Learning Representations*, 2014. 2
- W. Shimoda and K. Yanai. Distinct class saliency maps for weakly supervised semantic segmentation. In *ECCV*, 2016. 2, 3, 4, 6
- K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR WS*, 2014. 2, 3
- J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR*, 2015. 2
- P. Tokmakov, K. Alahari, and C. Schmid. Weakly-supervised semantic segmentation using motion cues. In *ECCV*, 2016. 3
- Y. Wei, X. Liang, Y. Chen, X. Shen, M. Cheng, Y. Zhao, and S. Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. In *ECCV*, 2016. 1, 2, 3, 5, 6
- B. Zhou, A. Khosla, A. Lapedriza, and A. Oliva, A. Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 2