Supplementary Material: STC: A Simple to Complex Framework for Weakly-supervised Semantic Segmentation

Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, *Senior Member, IEEE* and Shuicheng Yan *Senior Member, IEEE*

1 Justifications of Different Saliency Detection Methods

TABLE 1
Comparison of I-DCNN models by using different saliency maps for training on VOC 2012 *val* set.

Saliency method	bkg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mIoU
HS	81.1	48.1	20.8	53.3	30.9	38.5	45.0	45.1	51.6	14.0	57.6	14.9	50.1	54.5	37.6	51.8	29.1	54.9	33.6	49.1	31.6	42.5
DRFI	81.3	62.2	24.8	49.2	36.4	45.0	61.3	54.3	50.8	13.2	50.2	17.1	48.9	52.2	54.6	56.5	24.1	49.9	26.3	53.6	31.5	44.9

2 Justifications of the Number of Training Images

TABLE 2
Comparison of I-DCNN models trained on different numbers of training images on the VOC 2012 *val* set.

Training set	bkg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mIoU
1/16	80.5	58.2	18.2	56.5	25.9	43.5	51.7	50.7	43.7	4.1	41.6	10.4	38.8	40.5	40.1	40.9	31.9	46.7	21.6	49.7	39.6	39.8
1/8	81.4	64.3	12.8	62.0	34.2	45.3	55.5	51.7	44.9	6.8	44.2	12.5	45.3	47.3	48.3	47.4	27.3	45.2	18.7	51.5	37.0	42.1
1/4	82.7	65.0	22.3	60.7	39.8	47.3	59.7	54.6	53.3	12.8	47.5	12.3	45.7	50.9	51.2	52.9	28.5	49.9	26.8	55.5	40.4	45.7
1/2	82.0	64.4	24.6	54.6	37.9	45.8	60.8	54.9	52.6	13.4	48.0	17.1	49.6	52.4	54.2	56.1	25.6	46.9	27.4	55.4	33.4	45.6
All	81.3	62.2	24.8	49.2	36.4	45.0	61.3	54.3	50.8	13.2	50.2	17.1	48.9	52.2	54.6	56.5	24.1	49.9	26.3	53.6	31.5	44.9

3 Justifications of the Necessity of Simple Images

To validate the necessity of using simple images, we conduct experiments using complex images from Pascal VOC to learn the I-DCNN. We first produce those saliency maps for Pascal training images using DRFI. Then, we follow the same operation detailed in Section 3.1 to produce foreground/background maps for the single label images. For those multi-label images (it is unknown which category their foreground maps belong to), we adopt the following strategy to produce multiple foreground maps for the annotated classes. In particular, we use v to denote the pixel value of the foreground map (generated by DRFI) of a training image and n_c denotes the number of annotated labels of the training image. The value of this foreground pixel is assigned to each class with confidence value of v/n_c. By producing foreground/background maps on Pascal training images in this way for supervision, we train the I-DCNN and the E-DCNN for semantic segmentation. As shown in Table 3, the performance of the I-DCNN is very bad, owing to the low quality of the produced saliency maps for complex images. In addition, the improvement brought by the E-DCNN is limited using those bad segmentation masks produced by the I-DCNN for training. Therefore, it is significantly important for STC to use simple images to obtain an initial segmentation DCNN capable of producing sufficiently good segmentation masks.

TABLE 3
Semantic segmentation results of I-DCNN and E-DCNN on the VOC 2012 *val* set using complex images from VOC 2012 *train_aug* for training.

Training set	bkg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mIoU
I-DCNN	75.9	38.3	0.5	37.4	19.8	23.0	10.3	15.0	7.7	2.4	22.4	0.3	13.3	11.7	10.8	9.3	20.3	19.6	1.0	12.5	18.8	17.6
E-DCNN	75.9	42.3	0.4	37.4	18.7	21.2	10.8	15.3	9.5	2.1	27.5	0.1	15.8	9.8	8.2	8.8	16.4	22.6	0.2	10.2	18.5	17.7

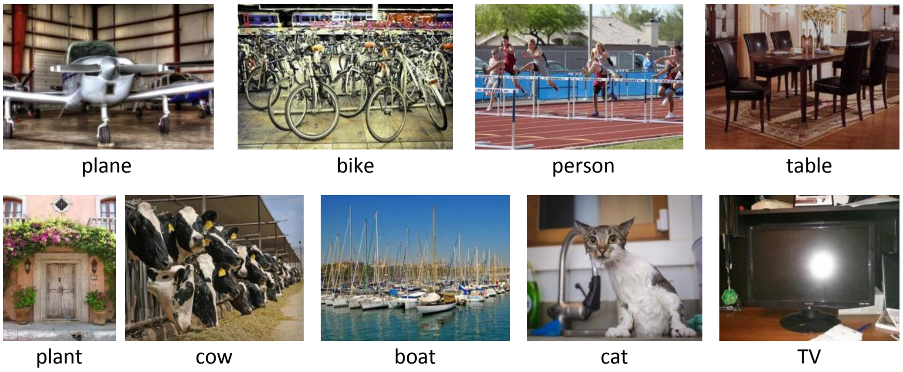
plane bike person table
plant cow boat cat TV

Fig. 1. Examples of noisy images from Flickr-Clean.

4 Semi-supervised Semantic Segmentation based on STC

To investigate the performance of the simple to complex framework under the semi-supervised setting, we train another I-DCNN model using 1,464 images with full-supervision information by adopting the Large-FOV network architecture in [1]. Then, the I-DCNN is employed to predict segmentation masks of images from Flickr-Clean based on image-level labels, and we train the E-DCNN based on 1,464 fully-annotated images and 40K predicted segmentation masks. Finally, we train a P-DCNN for semantic segmentation by incorporating complex samples (*i.e.,* 9K VOC images) whose segmentation masks are predicted by E-DCNN.

TABLE 4
Comparison of semi-supervised semantic segmentation results on the VOC 2012 *val* set.

Training strategy	bkg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mIoU
I-DCNN	91.2	79.8	27.5	74.3	58.5	64.0	76.7	76.0	73.9	19.0	57.4	48.5	65.5	59.1	68.8	77.1	34.9	65.5	39.7	68.7	58.3	61.2
E-DCNN (+Flickr-Clean)	90.1	80.0	27.0	81.2	62.2	58.6	81.6	73.7	75.2	22.8	75.6	47.6	66.5	73.6	65.9	72.3	40.0	74.8	44.7	74.8	54.5	63.9
P-DCNN (Semi- w/ STC)	90.5	80.2	27.0	78.5	64.8	59.9	81.8	73.5	74.6	32.2	73.4	50.2	67.4	75.0	66.6	74.7	43.4	75.7	45.0	75.1	56.9	65.1
Semi- w/o STC	90.0	79.5	29.2	77.9	63.9	58.3	82.8	73.5	74.8	26.0	74.2	43.5	64.5	74.8	67.5	70.3	42.6	75.3	43.7	73.4	57.0	63.9

Table 4 shows the experimental results of this semi-supervised setting. It can be observed that the mIoU of the I-DCNN achieves 61.2% on the PASCAL VOC 2012 *val* set, and can be further boosted to 65.1% (around 4% improvement) by the proposed STC framework. To see whether the STC strategy is beneficial for semi-supervised semantic segmentation, we additionally train a segmentation DCNN based on 1,464 strong supervised masks and (40K+9K) masks predicted by the I-DCNN in one step without the progressive simple to complex scheme. Without the STC framework, the segmentation performance will drop by more than 1% as shown in Table 4. As reported by [1], the performance of using all strong supervised images (10,582) from VOC 2012 for training is 67.6%, which is only 2.5% better than our semi-supervised scheme.

In addition, semi-supervised STC also gives better result than that reported in [1] (*i.e.,* 65.1% *vs.* 64.6%). This demonstrates the effectiveness of introducing a large number of weakly-supervised images. However, one may observe that the improvement is kind of marginal. One reason is that those simple images from Flickr-Clean are crawled and filtered automatically, thus some images present multiple objects and cluttered background. Those images will harm the training of segmentation DCNNs and finally decrease the mIoU scores. We visualize some noisy samples from Flick-Clean in Figure 1. One may assume that the performance of our STC-based semi-supervised approach can be further improved with cleaner simple images,

In summary, the proposed STC framework can also boost the semantic segmentation performance when only a small number of strong supervised images is available.

References

[1] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille. Weakly-and semi-supervised learning of a dcnn for semantic image segmentation. *arXiv preprint arXiv:1502.02734*, 2015.