

## Human-in-the-loop image segmentation and annotation

Xiaoya ZHANG<sup>1</sup>, Lianjie WANG<sup>2</sup>, Jin XIE<sup>1\*</sup> & Pengfei ZHU<sup>2</sup>

<sup>1</sup>*Department of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China;*  
<sup>2</sup>*College of Intelligence and Computing, Tianjin University, Tianjin 300072, China*

Received 31 August 2019/Revised 5 November 2019/Accepted 23 December 2019/Published online 24 February 2020

**Citation** Zhang X Y, Wang L J, Xie J, et al. Human-in-the-loop image segmentation and annotation. *Sci China Inf Sci*, 2020, 63(11): 219101, <https://doi.org/10.1007/s11432-019-2759-y>

Dear editor,

Semantic segmentation aims to assign the category information to all pixels of an image and plays a vital role in image understanding. In the past few years, deep convolutional neural networks have achieved great success in a large variety of computer vision tasks. Inspired by the advances of CNN in recognition, a fully convolutional network (FCN) is developed in an end-to-end, pixel-to-pixel training manner for semantic segmentation. Owing to the computational efficiency for dense prediction and end-to-end learning manner, numerous variants of FCN are then proposed to boost the performance of semantic segmentation.

The excellent performance of deep models, however, highly relies on expensive and laborious label annotations of massive images [1]. Actually, most existing deep learning models for semantic segmentation are firstly pre-trained on millions of images with sample-level annotations, e.g., ImageNet, and then fine-tuned with thousands of pixel-wise annotated images [1]. There remain three issues for semantic segmentation.

- The annotation for semantic segmentation has to be conducted pixel by pixel, which is labor intensive.

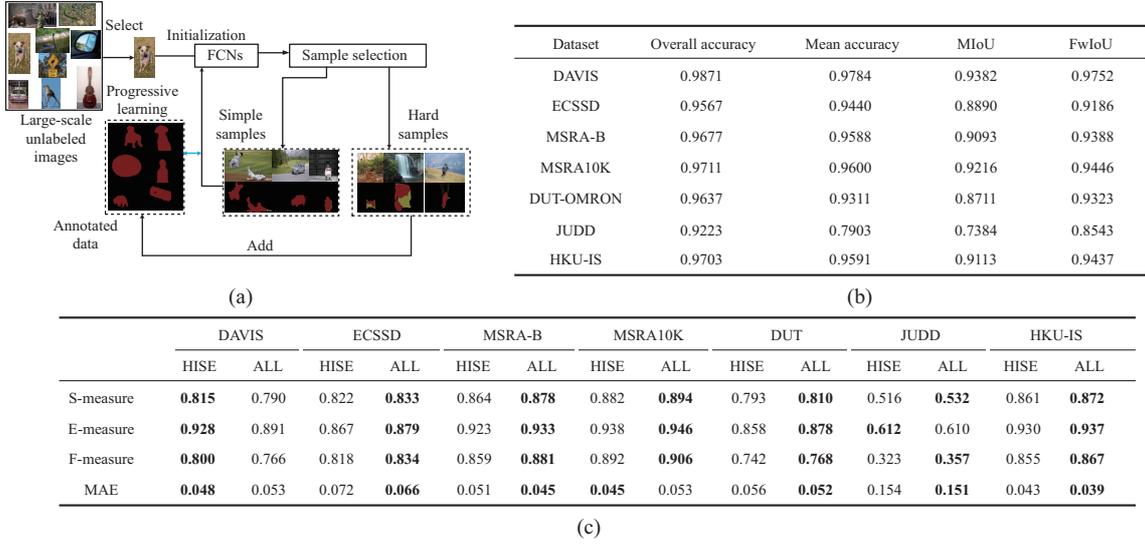
- There are inexhaustible unlabeled or partially labeled images in the wild. The recent advances in object detection [2] and image classification [3] show that large-scale unlabeled data can be made good use of to boost the model performance.

- Almost all existing benchmarks ignore the difference of images and only provide pixel-level annotations. There are a large number of images that contribute little to the learning of the segmentation models.

To reduce the labor cost for image annotation, several interactive segmentation models and tools have been developed by using weakly supervised information, e.g., clickpoints, lines, curves or bounding boxes [4]. Nevertheless, these studies are proposed for interactive annotation for a single image rather than annotating images in batch. To exploit the informative images in the wild, researchers have introduced active learning [5], semi-supervised learning [6], uncertainty learning [7], incremental learning [8], context learning [9] and self-supervised learning [1] for model enhancement. To sum up, we wonder whether we can annotate the unlabeled image with the least human labor and train a state-of-the-art segmentation model using the least data.

In this study, we propose a human-in-the-loop segmentation (HISE) framework, which is combined with a classic semantic segmentation model, i.e., FCN. We conduct experiments on seven benchmark datasets: DAVIS2016, MSRA-B, MSRA10K, ECSSD, DUT-OMRON, HKU-IS, and JUDD, to verify the effectiveness of the proposed HISE framework. Experimental results show that HISE can achieve comparable performance with much fewer human annotations and output a seg-

\* Corresponding author (email: csjxie@njust.edu.cn)



**Figure 1** (Color online) (a) Illustration of our proposed HISE framework. HISE can mine hard samples for human annotation by active learning. The reliable regions of images annotated by the machine and manually annotated images are fed to progressively finetune the FCNs. HISE can finally output both a deep model and a well annotated dataset. (b) The quality evaluation of machine annotations by semantic segmentation metrics. (c) The evaluation results by four salient object metrics. The better results are shown in bold.

mentation dataset with rich annotations.

*Framework overview.* The proposed HISE framework is shown in Figure 1(a). HISE firstly pre-trains a deep model (e.g., FCN) for segmentation on a small set of images with accurate pixel-wise labels. Then a large number of unlabeled images can be collected and initially annotated by FCN. An active learning strategy is used to judge whether the image should be automatically annotated. Images with low local confidence should be manually annotated. The images with high confidence are annotated by the machine. Because the edges in the foreground cannot be accurately segmented, we propose a tripartition pseudo labeling method for machine annotation. As the deep model is incrementally boosted with iterations, the parameters for active learning are progressively updated.

*Active learning.* In unsupervised learning, it is very challenging to evaluate the importance of an unlabeled sample. Active learning usually uses different kinds of evaluation metrics, e.g., uncertainty and diversity, to mine hard samples. Assume that there are  $c$  classes (the background is also included) in the image. FCN can predict the probability  $p(y_i = j|x_i; \mathbf{W})$  of a pixel  $x_i$  to the  $j$ th class, where  $\mathbf{W}$  is the network parameter. The confidence score for pixel  $x_i$  is defined as

$$\mu_i = \max_j p(y_i = j|x_i; \mathbf{W}). \quad (1)$$

$\mu_i$  is the largest probability of the pixel  $x_i$  to all classes. In unsupervised setting, the label of  $x_i$  is

unknown and therefore a larger  $\mu_i$  reflects higher confidence for the prediction.

In order to accurately mine hard samples, we simply define the selection criteria for hard samples in the segmentation task. Local consistency  $\tau_l$  is defined as the average confidence score of the pixels that are predicted as the object but the probability is less than a threshold  $\beta$ .

$$\tau_l = \frac{\sum_{i=1}^{n_f} \mu_i \theta(\mu_i)}{\sum_{i=1}^{n_f} \theta(\mu_i)}, \quad (2)$$

where  $n_f$  is the number of pixels predicted as the object. If  $\mu_i < \beta$ , then  $\theta(\mu_i)$  is 1. Otherwise,  $\theta(\mu_i)$  is zero. If the local consistency is low, the edge of the object is not reliable and therefore cannot be automatically annotated.

The active learning strategy is designed according to local consistency. All images are ranked in descent order according to  $\tau_l$  of each sample.

The pseudo-label is defined as follows:

$$j^* = \arg \max_j p(y_i = j|x_i; \mathbf{W}), \quad (3)$$

$$y_i = \begin{cases} j^*, & \mu_i > \delta, \\ \text{inf}, & \text{otherwise,} \end{cases} \quad (4)$$

where  $\delta$  is a threshold, and inf denotes infinite. If the confidence score  $\mu_i$  of the pixel  $x_i$  is less than  $\delta$ , the label assigned by the machine is unreliable and therefore pixel  $x_i$  will not be automatically annotated. In this way, only the pixels with high confidence scores are automatically annotated and

participate in the finetuning process of the training model.

*Progressive learning.* As the performance of the deep model is incrementally improved with iterations, we embed progressive learning into the proposed HISE framework. The loss function integrated with progressive learning is defined as follows:

$$\min_{\mathbf{W}} l(\mathbf{X}, \mathbf{W}, \beta, \gamma). \quad (5)$$

The parameters  $\beta$  and  $\gamma$  for active learning are updated during the training process.  $\beta$  is a threshold to ensure that the selected pixel is the object edge, and this edge is not trusted.  $\gamma$  represents the proportion of machine annotations in the iteration. The updating strategy is defined as

$$\beta = \begin{cases} \beta_0, & t = 0, \\ \beta_0 - \Delta\beta t, & t > 0, \end{cases} \quad (6)$$

$$\gamma = \begin{cases} \gamma_0, & t = 0, \\ \gamma_0 + \Delta\gamma t, & t > 0, \end{cases} \quad (7)$$

where  $\gamma_0$  and  $\beta_0$  are the initial thresholds.  $\Delta\gamma$  and  $\Delta\beta$  control threshold variation.  $t$  denotes the number of iterations.

*Experiments.* We evaluate our HISE framework on seven public datasets, i.e., DAVIS2016, MSRA-B, MSRA10K, ECSSD, DUT-OMRON, HUK-IS, and JUDD. We evaluate the quality of the machine annotations by four semantic segmentation metrics. As shown in Figure 1(b), the scores are generally higher than 90%. We adopt four evaluation metrics specially designed for salient object detection, i.e., S-measure, E-measure, F-measure, and mean absolute error (MAE). The results are summarized in Figure 1(c). For our proposed HISE, the results of the final iterations are shown and compared with FCN directly trained on all training images with ground truth. From the results, we can conclude that HISE can achieve comparable performance in terms of salient object metrics using much fewer human annotations.

*Conclusion and future work.* In this study, we proposed an HISE framework. HISE can output both annotated data and segmentation models at

the same time. To mine hard samples, an active learning strategy is proposed by defining local consistency. Progressive learning is introduced to incrementally boost the segmentation model. Experiments on seven public datasets show that HISE can achieve comparable performance using much fewer human annotations. Compared with the classic human annotation, HISE can also report the segmentation difficulty for each image, which can be used as prior for the development of advanced segmentation models. In the future work, we will consider more complex segmentation tasks with many classes.

**Acknowledgements** This work was supported by National Natural Science Foundation of China (Grant Nos. 61876084, 61876127, 61732011). The authors would like to greatly appreciate all the anonymous reviewers for their comments.

## References

- 1 Zhan X, Liu Z, Luo P, et al. Mix-and-match tuning for self-supervised semantic segmentation. 2017. ArXiv: 1712.00661
- 2 Wang K, Lin L, Yan X, et al. Cost-effective object detection: active sample mining with switchable selection criteria. IEEE Trans Neural Netw Learn Syst, 2019, 30: 834–850
- 3 Wang K, Zhang D, Li Y, et al. Cost-effective active learning for deep image classification. IEEE Trans Circ Syst Video Technol, 2017, 27: 2591–2600
- 4 Acuna D, Ling H, Kar A, et al. Efficient interactive annotation of segmentation datasets with Polygon-RNN++. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 859–868
- 5 Jain S D, Grauman K. Active image segmentation propagation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 2864–2873
- 6 Papandreou G, Chen L-C, Murphy K P, et al. Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, 2015. 1742–1750
- 7 Zhao C R, Chen K, Zang D, et al. Uncertainty-optimized deep learning model for small-scale person re-identification. Sci China Inf Sci, 2019, 62: 220102
- 8 Liu X, Kan M, Shan S, et al. Noisy face image sets refining collaborated with discriminant feature space learning. In: Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition, 2017. 544–550
- 9 Huang T T, Xu Y C, Bai S, et al. Feature context learning for human parsing. Sci China Inf Sci, 2019, 62: 220101