WEAKLY SUPERVISED INSTANCE SEGMENTATION USING HYBRID NETWORKS

Shisha Liao[⋆] *Yongqing Sun*[†] *Chenqiang Gao*[⋆]
Pranav Shenoy K P[‡] *Song Mu*[⋆] *Jun Shimamura*[†] *Atsushi Sagata*[†]

⋆ School of Communication and Information Engineering,
Chongqing University of Posts and Telecommunications, Chongqing, China
† NTT Media Intelligence Laboratories, Japan
‡ Georgia Institute of Technology, GA, USA

ABSTRACT

Weakly-supervised instance segmentation, which could greatly save labor and time cost of pixel mask annotation, has attracted increasing attention in recent years. The commonly used pipeline firstly utilizes conventional image segmentation methods to automatically generate initial masks and then use them to train an off-the-shelf segmentation network in an iterative way. However, the initial generated masks usually contains a notable proportion of invalid masks which are mainly caused by small object instances. Directly using these initial masks to train segmentation models is harmful for the performance. To address this problem, we propose a kind of hybrid networks in this paper. In our architecture, there is a principle segmentation network which is used to handle the normal samples with valid generated masks. In addition, a complementary branch is added to handle the small and dim objects without valid masks. Experimental results indicate that our method can achieve significantly performance improvement both on the small object instances and large ones, and outperforms all state-of-the-art methods.

Index Terms— Weakly-supervised, Instance Segmentation, FPN

1. INTRODUCTION

Instance segmentation, which serves as a fundamental task for a broad set of vision applications, such as remote sensing [1], medical imaging [2], and automatic drive [3], has attracted extensive attention and made large progress in the recent years. Most state-of-the-art methods [4, 5, 6] rely on large-scale dense annotations for training deep networks and show promising performances among the challenging benchmark datasets, including COCO [7], CityScapes [8] and PASCAL VOC [9]. However, annotating pixel-level labels for object instances is particularly expensive and time-consuming [10]. Comparing with complex and enormous pixel-level masks, some weakly annotations are much easier to obtain, e.g., points, scribbles, bounding boxes and image-level labels.

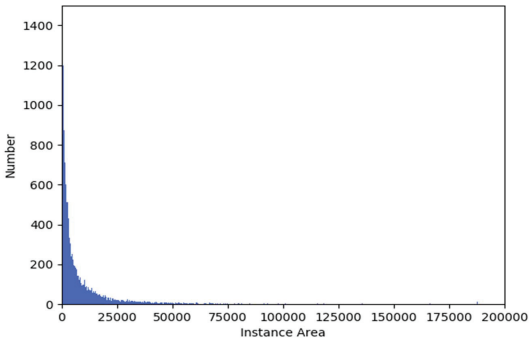

Fig. 1. The distribution of different sizes (areas) within initial invalid object instances from GrabCut [11]. The invalid masks concentrate on small object instances

Therefore, investigating the potentials of weakly supervised instance segmentation can effectively mitigate the labor cost, showing great practical significance.

In the weakly supervised instance segmentation realm, bounding box annotations are widely utilized due to two aspects. On one hand, bounding boxes provide precise position and category information. On the other hand, they can be used as a prior information for conventional methods, e.g., GrabCut [11] and MCG [12], to generate initial pixel-level mask labels, abbreviated as *mask* in the follows. BoxSup [13] generated initial masks using MCG based on bounding boxes and then proposed an iterative training procedure to obtained a good instance segmentation model. Khoreva et al. [14] employed GrabCut and MCG to generate delicate fake masks from the given box-level annotations and then adopted off-the-shelf segmentation network to implement weakly-supervised instance segmentation. In [15], the masks came from the intersection of the labels generated by GrabCut and MCG. Like BoxSup [13], the generated masks were refined in an iterative training fashion. Among aforementioned methods, the quality of initial segmentation masks from GrabCut, as well as MCG, relies on the sizes of bounding boxes, namely the sizes of object instances. The masks from large

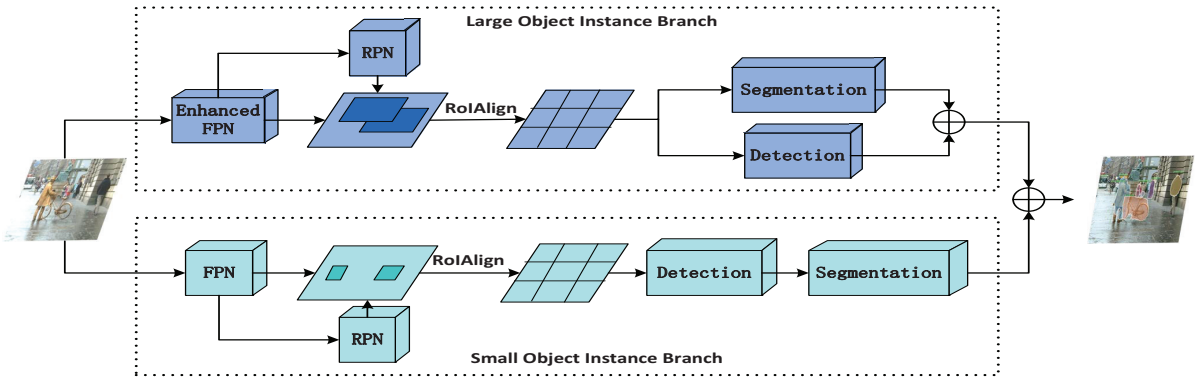

Fig. 2. Illustration of our framework.

object instances routinely are of good quality, while small object instances tend to be of poor quality. Because we have no ground-truth for the object instance mask, the quality is roughly estimated by the intersection-over-union (IoU) between the bounding box of the automatically generated object instance mask and the ground-truth bounding box. In this paper, one generated object instance mask is considered to be invalid if the IoU is under 0.5. Figure 1 shows the distribution of object instances in invalid masks obtained by GrabCut. It can be obviously observed that the scales of objects with invalid masks are of a wide range, but mainly concentrates on small object instances. Statistically, invalid small object instances whose size is less than 64×64 pixels, accounts for 55.54% in invalid object instances, but accounts for 76.48% in all small object instances with both invalid and valid ones. In contrast, invalid large object instances whose size is bigger than 64×64 pixels only accounts for 23.20% in all large instances with both invalid and valid ones. The total invalid object instances within the initial masks can reach 30% of all masks. This notable proportion of invalid masks is harmful for directly training models, even using the iterative training technique as done by some of above methods.

Based on above statistical analysis and observations, we propose a hybrid instance segmentation network. In this architecture, a principle segmentation network is trained using only the samples with valid masks. Noteworthy, according to our statistics, the majorities of valid samples are large object instances. Thus, in this network, the training samples are mostly unified and pure, which is beneficial to the overall training. In addition, an Enhanced-FPN architecture is added to this branch to reduce the transfer distance of low-level feature, providing more localization information. For the invalid object instance masks which have correct bounding boxes, we design a complementary branch to handle these hard samples which mainly consist of small and dim object instances as discussed before. The proposed architecture is evaluated on the validation set of PASCAL VOC 2012, and the experimental results reveal that our method can achieve significant improvement on the aforementioned difficult samples, showing the effectiveness of the complementary framework.

2. THE PROPOSED METHOD

As illustrated in Figure 2, the framework of our method contains two branches: a large object instance branch and a small object instance branch. The large object instance branch cooperates detection and segmentation simultaneously to handle the large object instance segmentation, while the small object instance branch sequentially conducts detection and segmentation to avoid the omission of small object instances.

In the training stage, we firstly use GrabCut [11] to automatically obtain initial object instance masks based on given bounding box annotations. Then, we divide the initial masks into two groups according to the IoU as described previously: valid masks and invalid masks. The former is used to train the large object instance branch, while the latter is used to train the small object instance branch. In the test stage, images are simultaneously fed into two branches and the segmentation results from both branches are fused to obtain the final results. Specifically, all the object instance segmentation results with size less than 64×64 pixels come from Non-maximum Suppression (NMS) results of both small and big object instance branches, and all the object instance segmentation results with size more than 64×64 pixels come from the big object instance branch.

2.1. Large object instance branch

The large object instance branch is built on Mask R-CNN and improved based on our weakly-supervised task. It is composed of four components, including raw feature extraction (Enhanced-FPN module), proposal generation (RPN module), bounding-box recognition (Detection module) and mask prediction (Segmentation module). First, We adopt ResNet-50 with Enhanced-FPN as the backbone. Specifically, the conventional Feature Pyramid Network (FPN) [16] architecture is replaced by our Enhanced-FPN to improve

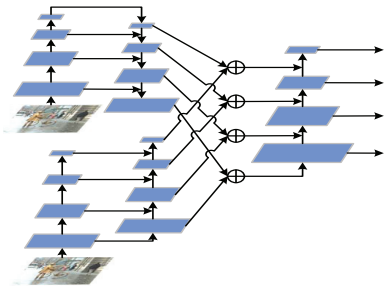

Fig. 3. Illustration of our Enhanced-FPN structure.

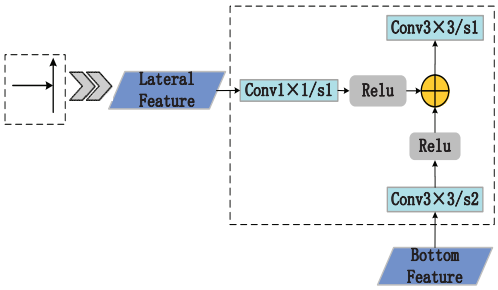

Fig. 4. Illustration the connection type of our Enhanced-FPN. (Note: conv $m \times m/sn$ means a convolution whose size is $m \times m$ and whose stride is n. Relu means a rectified linear unit)

the performance. The Enhanced-FPN and implementation details will be introduced in the section 2.3. Then, RPN [17] is utilized to generate proposals. After handled by the RoIAlign [5] operation, each proposal becomes a fixed-size feature map. Finally, bounding-box recognition and mask prediction are implemented simultaneously through this feature map. The detection branch conforms to the spirit of the fast R-CNN [18] to realize localization and classification, and the segmentation branch uses FCN [19] to achieve pixel-level prediction.

2.2. Small object instance branch

Small object instance branch served as an important complementary module of large objects branch, focuses on small object instances segmentation. Without the existence of noisy labels, this branch can get better detection performance. In addition, the morphological characteristic ensures better segmentation performance of small object instances. Under the supervision of the box-level annotations, this branch first does object detection by Faster R-CNN [17]. The detection results offer the bounding boxes information for GrabCut to obtain final segmentation. For these segmentation results with poor quality, we replace them with ellipses. In this process, these basic modules are similar to that of the large object branch. The main difference between these two branches is the execution fashion of detection module and segmentation modules. The large object instance branch conducts detection and segmentation simultaneously, whereas the small object instance branch conducts them sequentially.

2.3. Enhanced-FPN

Enhanced-FPN is improved based on FPN. There exists two problems in conventional FPN. On the one hand, FPN treats different feature maps unfairly. It is well-known that FPN adopts the top-down fusion pattern to increase localization accuracy. High-level semantic information is gradually transferred to the low-level feature map, so each low-level feature map includes the information of high-level feature map. However, the high-level feature is not enhanced adequately. On the other hand, there is a long path from low-level to topmost features and this introduces difficulty to access accurate localization information [6]. Therefore, we propose an Enhance-FPN to balance the enhancement function of every feature maps, and shorten the propagation distance of bottom feature maps. The framework of Enhanced-FPN is illustrated in Figure 3. There are multiple structures that the lateral feature map connects with bottom feature map. The specific connection type is shown in Figure 4.

2.4. Implementation Details

Most hyper-parameters in Mask R-CNN are applied to our first branch. Specifically, We train on 3 GPUs (so effective minibatch size is 6) for 30k iterations, with a learning rate of 0.005 which is decreased by 10 at the 20k and 26K iteration, respectively. We use a weight decay of 0.0002. In addition, we use images with shorter edge randomly sampled from 600, 800 for training and with shorter edge 600 for inference. The longer edge of the images is 1000 for both training and inference. For the second branch, it shares the identical hyper-parameters with the first branch except the learning rate. Its learning rate is 0.0075.

3. EXPERIMENTS

3.1. Dataset and metrics

The PASCAL VOC dataset involves 20 semantic categories of objects, which is extensively used in the field of weakly-supervised tasks. Following the previous work [14, 20], we utilize additional images from the SBD dataset [21] to obtain a training set of 10582 images, and report all of the experimental results on the validation set, including 1449 images. We adopt the widely used metrics in instance segmentation community, including $\mathrm{mAP}^{\mathrm{r}}_{0.5}$ and $\mathrm{mAP}^{\mathrm{r}}_{0.75}$. And the Average Best Overlap (ABO) [22] metric is also employed for evaluation to give a different perspective.

Table 1. Results of different methods on the PASCAL VOC 2012 val.

Supervision	Methods	$\mathrm{mAP}^{\mathrm{r}}_{0.5}$	$\mathrm{mAP}^{\mathrm{r}}_{0.75}$	ABO
box-level	DeepMask[14]	39.4	8.1	45.8
box-level	DeepLabBOX[14]	44.8	16.3	49.1
box-level	Ours	**51.3**	**22.4**	**51.9**
pixel-level	Ours	56.5	29.6	57.4

3.2. Comparison with state-of-the-art methods

Four state-of-the art methods are selected for comparison, including DeepMask [14] and DeepLabBOX [14], soft proposal networks (SPN) [23] and peak response maps (PRM) [20]. The former two methods are based on bounding-box level with the same configuration for weak supervised instance segmentation, while the latter two ones are based on the image-level label. The results of different methods are shown in Table 1. It can be observed that our method obviously outperforms all state-of-the-art methods in terms of all metrics. We can also see that the bounding-box-based methods are totally better than the image-label-based methods. This is because bounding boxes offer more precise information than image labels for instance segmentation. Among three bounding-box-based methods, the performance of our method is evident, especially in terms of metrics of $\mathrm{mAP}^{\mathrm{r}}_{0.5}$ and $\mathrm{mAP}^{\mathrm{r}}_{0.75}$. This is due to that our method not only improves the instance segmentation of large objects, but also improves the instance segmentation of small objects. From the late row of the Table 1 where "pixel-level" means supervised instance segmentation, we can see that the In addition, the performance of our weakly supervised method with only bounding box information is close to our supervised version with precise pixel mask information for training. This is further verify the effectiveness of our weakly supervised method.

3.3. Effect of two-branch structure

To understand the effect of two-branch structure, we conduct a series of experiments on weakly-supervised instance segmentation task. Firstly, Mask R-CNN serves as the baseline network architecture for this task. Secondly, our two-branch network is decomposed into two one-branch structures,including small-branch structure and big-branch structure, to conduct controlled experiments. Finally, the performance our two-branch structure is reported. To be fair, all of these networks take ResNet-50 as the backbone. And each network is evaluated on the PASCAL VOC 2012 val in terms of AP, AP_S and AP_L. Corresponding results are shown in Table 2. They clearly show that our two-branch structure achieve the best performance on all metrics. Among them, small-branch structure has performance advantage on small object instance, while big-branch structure has performance advantage on big object instance. So the small-branch structure and the big-branch structure present obvious complementary superiority, which demonstrated the reasonableness our two-branch structure.

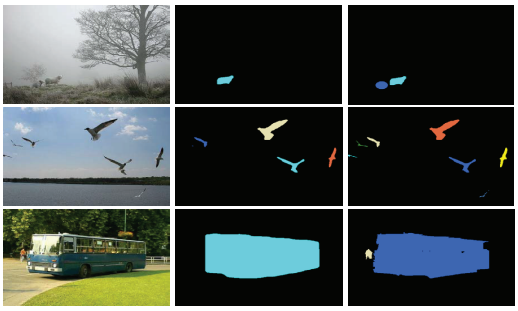

Fig. 5. Visual results on PASCAL VOC validation set. The first column is the original images, the second column are the results of the single branch method, and the last column is the results of our method.

Table 2. Effect of two-branch structure on the PASCAL VOC 2012 val in terms of AP, AP_S and AP_L.

Methods	AP	AP_S	AP_L
baseline	23.73	6.75	27.75
samll-branch	18.91	8.42	22.24
big-branch	25.14	7.11	29.45
ours	**25.20**	**8.52**	**29.48**

3.4. Quality analysis

Figure 5 shows some representative results of our methods. Note that the second column are results of our large object instance branch trained by all initial generated masks including both invalid and valid masks. This is similar to commonly used weakly supervised instance segmentation pipeline. From Figure 5, we can observe that our method achieves significant improvement on the small object instance segmentation through adding a complementary small object instance branch.

4. CONCLUSION

We propose a novel hybrid segmentation network to handle the invalid mask problem in initial generated masks in the weakly supervised instance segmentation task. The proposed hybrid network consists of two branches. One branch cooperates detection and segmentation simultaneously to handle the large object instance segmentation, while the other sequentially conducts detection and segmentation to avoid the omission of small object instances. Experimental results reveal that our method outperforms state-of-the-art methods, and has obvious advantage on the small object instance segmentation.

5. REFERENCES

[1] L. Mou and X.X. Zhu, "Vehicle instance segmentation from aerial image and video using a multi-task learning residual fully convolutional network," *arXiv preprint arXiv:1805.10485*, 2018.

[2] H. Chen, X. Qi, L. Yu, Q. Dou, J. Qin, and P. Heng, "Dcan: Deep contour-aware networks for object instance segmentation from histology images," *Medical image analysis*, vol. 36, pp. 135–146, 2017.

[3] Z. Zhang, S. Fidler, and R. Urtasun, "Instance-level segmentation for autonomous driving with deep densely connected mrfs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 669–677.

[4] L. Chen, A. Hermans, G. Papandreou, F. Schroff, P. Wang, and H. Adam, "Masklab: Instance segmentation by refining object detection with semantic and direction features," *arXiv preprint arXiv:1712.04837*, vol. 2, 2018.

[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2980–2988.

[6] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8759–8768.

[7] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C.L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[8] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M.s Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[9] M. Everingham, Van G.L., C.K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[11] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *ACM transactions on graphics (TOG)*. ACM, 2004, vol. 23, pp. 309–314.

[12] J. Pont-Tuset, P. Arbelaez, J. T Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping for image segmentation and object proposal generation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 1, pp. 128–140, 2017.

[13] J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1635–1643.

[14] A. Khoreva, R. Benenson, J.H. Hosang, M. Hein, and B. Schiele, "Simple does it: Weakly supervised instance and semantic segmentation.," in *CVPR*, 2017, vol. 1, p. 3.

[15] Q. Li, A. Arnab, and P.H. Torr, "Weakly-and semi-supervised panoptic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 102–118.

[16] T. Lin, P. Dollár, R.B. Girshick, K. He, B. Hariharan, and S.J. Belongie, "Feature pyramid networks for object detection.," in *CVPR*, 2017, vol. 1, p. 4.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, , no. 6, pp. 1137–1149, 2017.

[18] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[20] Y. Zhou, Y. Zhu, Q. Ye, Q. Qiu, and J. Jiao, "Weakly supervised instance segmentation using class peak response," *arXiv preprint arXiv:1804.00880*, 2018.

[21] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," 2011.

[22] J. Pont-Tuset and Van G.L., "Boosting object proposals: From pascal to coco," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1546–1554.

[23] Y. Zhu, Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao, "Soft proposal networks for weakly supervised object localization," in *Proc. IEEE Int. Conf. Comput. Vis.(ICCV)*, 2017, pp. 1841–1850.