Supplementary for BoxInst: High-Performance Instance Segmentation with Box Annotations

Zhi Tian, Chunhua Shen*, Xinlong Wang, Hao Chen
The University of Adelaide, Australia

1. Experiments on Cityscapes

We also conduct experiments on Cityscapes [1]. The Cityscapes benchmark includes 2975, 500 and 1525 images with fine instance segmentation annotations for training, validation and testing, respectively. It also has 20K images with coarse annotations, which we do not use. Except that we do not use the mask annotations, we use the same training settings for Cityscapes as in `Detectron2` [2]. The model is initialized with the BoxInst model pre-trained on COCO. During training, we only update the weights in FCOS classification branch to avoid overfitting. The other hyper-parameters for the loss terms are the same as in COCO. The results on the `val` split are shown in Table 1.

AP	AP_{50}	person	rider	car	truck	bus	train	mcycle	bicycle
24.9	51.4	24.4	10.3	37.0	25.8	50.1	28.9	12.8	10.2

Table 1: Box-supervised instance segmentation results on Cityscapes `val` split. ResNet-50 is used as the backbone.

2. Box-supervised Character Segmentation

In order to demonstrate the generality of BoxInst, we conduct experiments to obtain the character masks with character box annotations. Our experiments are conducted on the ICDAR 2019 ReCTS dataset [3], which contains 20K training images and 5K testing images. These images are annotated with text-line and character-level boxes. We train our model with the character boxes. All the training settings are the same as that of COCO. Since we do not have mask annotations for the testing set, it is impossible to report the mask AP. We instead show some qualitative results in Fig. 1, demonstrating that BoxInst can obtain high-quality character masks. The text masks might provide useful cues for detecting and recognising text of arbitrary shapes. We believe that the ability of BoxInst generating character masks automatically may inspire new applications on this task.

*Corresponding author.

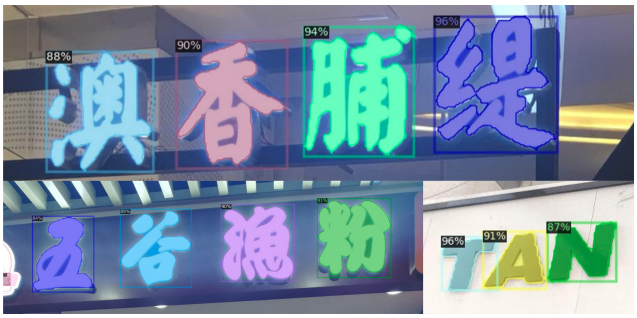

Figure 1: Character masks predicted by BoxInst. No mask annotations are used for training.

3. Video Demo

A video demo of BoxInst can be found at https://www.youtube.com/watch?v=NuF8NAYf5L8.

References

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[2] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 1

[3] Rui Zhang, Yongsheng Zhou, Qianyi Jiang, Qi Song, Nan Li, Kai Zhou, Lei Wang, Dong Wang, Minghui Liao, Mingkun Yang, et al. ICDAR 2019 robust reading challenge on reading chinese text on signboard. In *Proc. Int. Conf. Document Analysis Recogn.*, pages 1577–1581, 2019. 1