BlendMask: Top-Down Meets Bottom-Up for Instance Segmentation

Hao Chen[1]* Kunyang Sun[2,1]*, Zhi Tian[1], Chunhua Shen[1], Yongming Huang[2], Youliang Yan[3]
[1] The University of Adelaide, Australia [2] Southeast University, China [3] Huawei Noah's Ark Lab

Appendix A: Panoptic Segmentation

We use the semantic segmentation branch of Panoptic-FPN [1] to extend BlendMask to the panoptic segmentation task. We use annotations of COCO 2018 panoptic segmentaiton task. All models are trained on `train2017` subset and tested on `val2017`. We train our model with the default FCOS [2] 3× schedule with scale jitter (shorter image side in $[640, 800]$. To combine instance and semantic results, we use the same strategy as in Panoptic-FPN, with instance confidence threshhold 0.2 and overlap threshhold 0.4.

Results are reported in Table 1. Our model is consistently better than its Mask R-CNN counterpart, Panoptic-FPN. We assume there are three reasons. First, our instance segmentation is more accurate, this helps with both thing and stuff panoptic quality because instance masks are overlaid on top of semantic masks. Second, our pixel-level instance prediction is also generated from a global feature map, which has the same scale as the semantic prediction, thus the two results are more consistent. Last but not least, since the our bottom module shares structure with the semantic segmentation branch, it is easier for the network to share features during the closely related multi-task learning.

Appendix B: More Qualitative Results

We visualize qualitative results of Mask R-CNN and BlendMask on the validation set in Fig. 1. Four sets of images are listed in rows. Within each set, the top row is the Mask R-CNN results and the bottom is BlendMask. Both models are based on the newly released `Detectron2` with use R101-FPN backbone. Both are trained with the 3× schedule. The Mask R-CNN model achieves 38.6% AP and ours 39.5% AP.

Since this version of Mask R-CNN is a very strong baseline, and both models achieve very high accuracy, it is very difficult to tell the differences. To demonstrate our advantage, we select some samples where Mask R-CNN has trouble dealing with. Those cases include:

- Large objects with complex shapes (Horse ears, human poses). Mask R-CNN fails to provide sharp borders.
- Objects in separated parts (tennis players occluded by nets, trains divided by poles). Mask R-CNN tends to include occlusions as false positive or segment targets into separate objects.
- Overlapping objects (riders, crowds, drivers). Mask R-CNN gets uncertain on the borders and leaves larger false negative regions. Sometimes, it assigns parts to the wrong objects, such as the last example in the first row.

Our BlendMask performs better on these cases. 1) Generally, BlendMask utilizes features with higher resolution. Even for the large objects, we use stride-8 features. Thus details are better preserved. 2) As shown in previous illustrations, our bottom module acts as a class agnostic instance segmenter which is very sensitive to borders. 3) Sharing features with the bounding box regressor, our top module is very good at recognizing individual instances. It can generate attentions with flexible shapes to merge the fine-grained segments of bottom module outputs.

Method	Backbone	PQ	SQ	RQ	PQ^{Th}	PQ^{St}	mIoU	AP^{box}	AP^{Th}
Panoptic-FPN [1]	R-50	41.5	79.1	50.5	48.3	31.2	42.9	40.0	36.5
BlendMask		**42.5**	**80.1**	**51.6**	**49.5**	**32.0**	**43.5**	**41.8**	**37.2**
Panoptic-FPN [1]	R-101	43.0	80.0	52.1	49.7	32.9	44.5	42.4	38.5
BlendMask		**44.3**	**80.1**	**53.4**	**51.6**	**33.2**	**44.9**	**44.0**	**38.9**

Table 1: Panoptic results on COCO `val2017`. Panoptic-FPN results are from the official `Detectron2` implementation, which are improved upon the original published results in [1].

*equal contribution. This work was done when K. Sun was visiting The University of Adelaide. Correspondence should be addressed to C. Shen and Y. Huang.

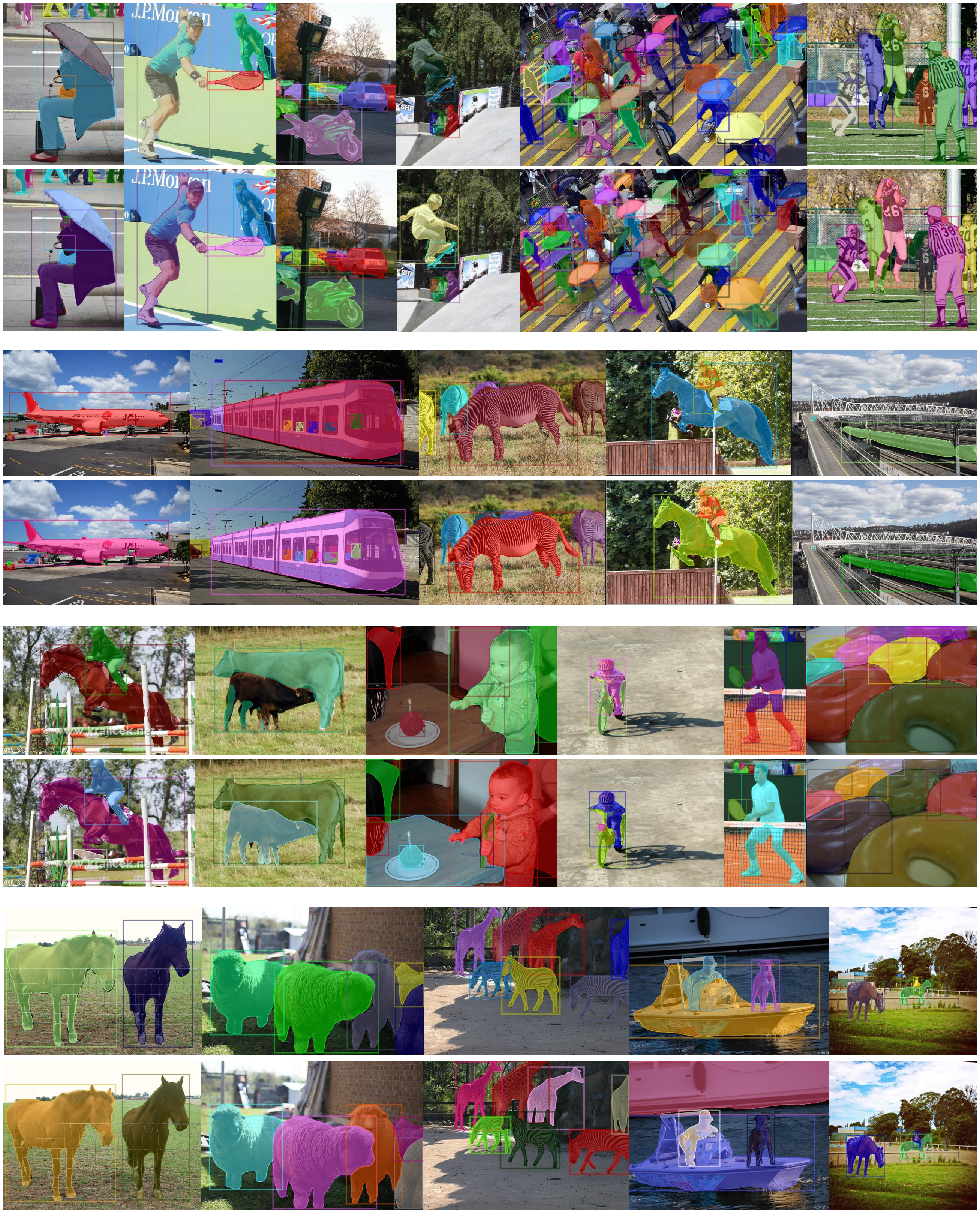

Figure 1: Selected results of Mask R-CNN (top) and BlendMask (bottom). Both models are based on `Detectron2`. The Mask R-CNN model is the official 3× R101 model with 38.6 AP. BlendMask model obtains 39.5 AP. Best viewed in digital format with zoom.

References

[1] Alexander Kirillov, Ross B. Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 6399–6408, 2019.

[2] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: fully convolutional one-stage object detection. *Proc. Int. Conf. Comp. Vis.*, abs/1904.01355, 2019.