A. CrowdHuman Testing Benchmark

The *CrowdHuman*[4] testing subset has $5,000$ images and the annotations of testing subset have not yet been released. To push the upper-bound of object detection research, the *Detection In the Wild Challenge* was held in CVPR 2019. The *CrowdHuman* testing subset is served as a benchmark in this challenge and this allows us to compare our approach to state-of-the-art methods on *CrowdHuman*.

To improve the performance of our approach, we replace the ResNet-50[1] with a larger model: SEResNeXt101[2, 5], the short edge of test images are resized to 1200 pixels and all other settings are the same as described in our paper. We then submit our result to the test server and find that our method outperforms all the results in this challenge. The results are shown in Table.1, and the full leaderboard is accessible on the official website of CrowdHuman Track Leaderboard [1].

B. Ablation on Number of Heads

For completeness, we further explore the only hyper-parameter of our method: K in this section. In our paper, we let $K = 2$ because we find it is satisfied for almost all the images and proposals in *CrowdHuman*. If we make the K larger, the network will be able to detect instances in more crowded scenes. To explore the performance under the different values of the K, an experiment is conducted on the *CrowdHuman* dataset and all the settings remain the same as described in our paper except the value of K is changed. We show the results of different K values in the Table.2.

C. More Results of Our Method

In this section we will show more results of our method on a video from YouTube and the *CrowdHuman* validation dataset. The visualization thresh is set to 0.7 to remove the redundant boxes in the results.

The video is in the attached file, and the results on the *CrowdHuman* validation dataset is shown in the Figure.1.

Rank	Team Name	Institution	mJI/%
1	zack0704	Tencent AI Lab	77.46
2	boke	Sun Yat-Sen University	75.25
3	ZNuanyang	Zhejiang University	74.46
Method		Backbone	mJI/%
Baseline		ResNet-50	72.20
Ours		ResNet-50	76.60
Ours		SEResNeXt101	**77.74**

Table 1. Part of the leaderboard and our results. The baseline model is our reimplemented FPN[3].

[1] https://www.objects365.org/crowd_human_track.html

	AP/%	MR^{-2}/%	JI/%
$K = 1$	85.8	42.9	79.8
$K = 2$	**90.7**	**41.4**	**82.3**
$K = 3$	**90.7**	41.6	82.1

Table 2. Ablation experiments evaluated on the *CrowdHuman* validation set. It worth noting that if K = 1, the architecture is the same as the *single-instance-prediction* baseline.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[2] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 1

[3] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 1

[4] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 1

[5] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 1

Figure 1. Visual comparison of the baseline, GossipNet, RelationNet, and our approach. The blue boxes are the detection results, the white boxes are the missed instances, and the orange boxes are redundant boxes. The green boxes in our method are multiple predictions form one proposal.