Rethinking Counting and Localization in Crowds: A Purely Point-Based Framework

Qingyu Song[1*] Changan Wang[1*] Zhengkai Jiang[1] Yabiao Wang[1]
Ying Tai[1] Chengjie Wang[1] Jilin Li[1] Feiyue Huang[1†] Yang Wu[2]
[1]Tencent Youtu Lab, [2]Applied Research Center (ARC), Tencent PCG
qingyusong@zju.edu.cn, {changanwang, zhengkjiang, caseywang}@tencent.com
{yingtai, jasoncjwang, jerolinli, garyhuang, dylanywu}@tencent.com

Supplementary

1. Counting Evaluation Metrics

Similar to previous works in crowding counting, we adopt Mean Absolute Error (MAE) and Mean squared error (MSE) as our evaluation metrics which are defined as:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i}^{N}\left|\hat{z}_i - z_i\right|, \tag{1}$$

$$\mathrm{MSE} = \sqrt{\frac{1}{N}\sum_{i}^{N}(\hat{z}_i - z_i)^2}, \tag{2}$$

where $\hat{z}_i$ and z_i represent estimated crowd number and ground-truth crowd number of the i-th image, respectively. N denotes the total number of test images.

2. Discussion on Spatial Scale Problem

Despite its superior performance, the proposed P2PNet did not explicitly deal with the scale variation problem. Actually, different from bounding boxes, the head points themselves are scale ignorant in nature. In other words, the one-one matching ensures that no matter which scale the head is, only one optimal predicted proposal will be chosen as its prediction. Thus, some implicit scale cues might be learned automatically during the training process. Besides, the proposed framework is orthogonal to some previous works dealing with scale variations, such as FPN [8], PGCNet [15], CSRNet [7], MCNN [16], *etc*.

3. Hyperparameters Analysis

We set the number of reference points (K) based on the nearest neighbour distance distribution of ground truth points. Specifically, based on the observation that nearly 95% (SHTech PartA) of the head points are within the nearest neighbour distance of 4 pixels, we set the number of the reference points K as 4 on the feature map with stride 8. We experimentally analyze the accuracy sensitivity of this parameter in Table 1. As shown from the results, the model with K=1 still achieves state-of-the-art accuracy, although it's reference points are too few to cover all the heads in congested areas. Setting K to a value greater than 4 leads to inferior accuracy, which might be caused by the increase of negative samples.

K	MAE	MSE	nAP$_\delta$
K=1	54.08	**84.37**	60.1
K=4	**52.74**	85.06	**64.4**
K=8	53.43	87.57	58.8
K=12	54.13	87.9	58.6
K=16	53.47	86.1	58.3

Table 1. The performance change w.r.t. the number K for reference points. For an overall comparison, we use $\delta = \{0.05 : 0.05 : 0.50\}$.

4. Localization Performance

Method	F1-Measure	Precision	Recall
FasterRCNN [11]	0.068	0.958	0.035
TinyFaces [3]	0.567	0.529	0.611
RAZ [9]	0.599	0.666	0.543
Crowd-SDNet [14]	0.637	0.651	0.624
PDRNet [6]	0.653	0.675	0.633
TopoCount [1]	0.692	0.683	**0.701**
D2CNet [2]	<u>0.700</u>	**0.741**	0.662
Ours	**0.712**	<u>0.729</u>	<u>0.695</u>

Table 2. Comparison for the localization performance on NWPU.

Thanks to the scarce yet valuable box annotations provided by the NWPU-Crowd dataset [13], we could com-

*Equal contribution. †Corresponding author.

pare the localization performance of our P2PNet with other competitors using their metrics. As shown in Table 2, our P2PNet achieves the best F1 score among the published methods with similar computation complexity.

Among a few existing localization-based methods, almost none of them have official codes or third-party re-implementations except for [12]. So for a fair comparison, we evaluate the $\text{nAP}_{0:05:0:05:0:50}$ of [12] on SHTech PartA, SHTech PartB and QNRF, which are 33.2%, 45.8% and 8.9% respectively. As shown from the results, our P2PNet achieves significantly higher localization performance in terms of nAP, especially on the challenging QNRF dataset.

5. Visual Results for Qualitative Evaluation

In Figure 1-13, we exhibit the results of several example images with different densities from sparse, medium to dense. As seen from these results, our P2PNet achieves impressive localization and counting accuracy under various crowd density conditions.

Additionally, from these qualitative results, we also find that P2PNet may fail on some extreme large heads and gray images (old photos). But similar failure cases could also be found in other top methods, such as ASNet (CVPR'20) [5], AMSNet (ECCV'20) [4], SDANet (AAAI'20) [10], *etc*. Fortunately, these might be alleviated to some extent by adding more relevant training data.

References

[1] Shahira Abousamra, Minh Hoai, Dimitris Samaras, and Chao Chen. Localization in the crowd with topological constraints. 2021. 1

[2] Jian Cheng, Haipeng Xiong, Zhiguo Cao, and Hao Lu. Decoupled two-stage crowd counting and beyond. *IEEE Transactions on Image Processing*, 30:2862–2875, 2021. 1

[3] Peiyun Hu and Deva Ramanan. Finding tiny faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 951–959, 2017. 1

[4] Yutao Hu, Xiaolong Jiang, Xuhui Liu, Baochang Zhang, Jungong Han, Xianbin Cao, and David Doermann. Nas-count: Counting-by-density with neural architecture search. In *European Conference on Computer Vision*, 2020. 2

[5] Xiaoheng Jiang, Li Zhang, Mingliang Xu, Tianzhu Zhang, Pei Lv, Bing Zhou, Xin Yang, and Yanwei Pang. Attention scaling for crowd counting. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[6] Chongyi Li, Chunle Guo, Jichang Guo, Ping Han, Huazhu Fu, and Runmin Cong. Pdr-net: Perception-inspired single image dehazing network with refinement. *IEEE Transactions on Multimedia*, 22(3):704–716, 2019. 1

[7] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1

[8] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1

[9] Chenchen Liu, Xinyu Weng, and Yadong Mu. Recurrent attentive zooming for joint crowd counting and precise localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1

[10] Yunqi Miao, Zijia Lin, Guiguang Ding, and Jungong Han. Shallow feature based dense attention network for crowd counting. In *Association for the Advancement of Artificial Intelligence*, 2020. 2

[11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 1

[12] Deepak Babu Sam, Skand Vishwanath Peri, Mukuntha Narayanan Sundararaman, Amogh Kamath, and Venkatesh Babu Radhakrishnan. Locate, size and count: Accurately resolving people in dense crowds via detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[13] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1

[14] Yi Wang, Junhui Hou, Xinyu Hou, and Lap-Pui Chau. A self-training approach for point-supervised object detection and counting in crowds. *IEEE Transactions on Image Processing*, 30:2876–2887, 2021. 1

[15] Zhaoyi Yan, Yuan Yuchen, Zuo Wangmeng, Tan Xiao, Wang Yezhen, Wen Shilei, and Ding Errui. Perspective-guided convolution networks for crowd counting. In *IEEE International Conference on Computer Vision*, 2019. 1

[16] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1

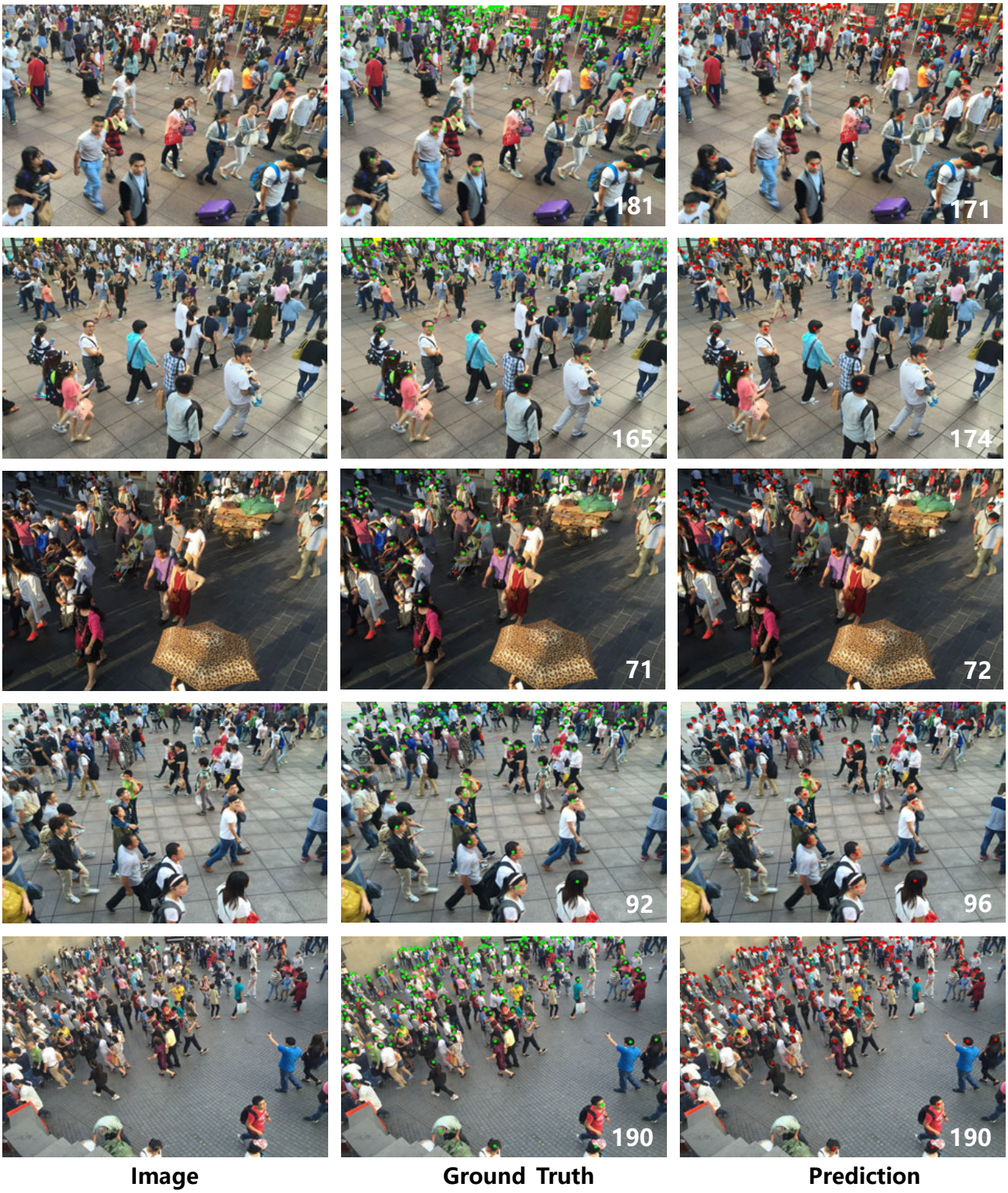

Figure 1. Visual results of sparse scenes (1).

Figure 2. Visual results of sparse scenes (2).

Figure 3. Visual results of sparse scenes (3).

Figure 4. Visual results of sparse scenes (4).

Figure 5. Visual results of moderately congested scenes (1).

Figure 6. Visual results of moderately congested scenes (2).

Figure 7. Visual results of moderately congested scenes (3).

Figure 8. Visual results of moderately congested scenes (4).

Figure 9. Visual results of moderately congested scenes (5).

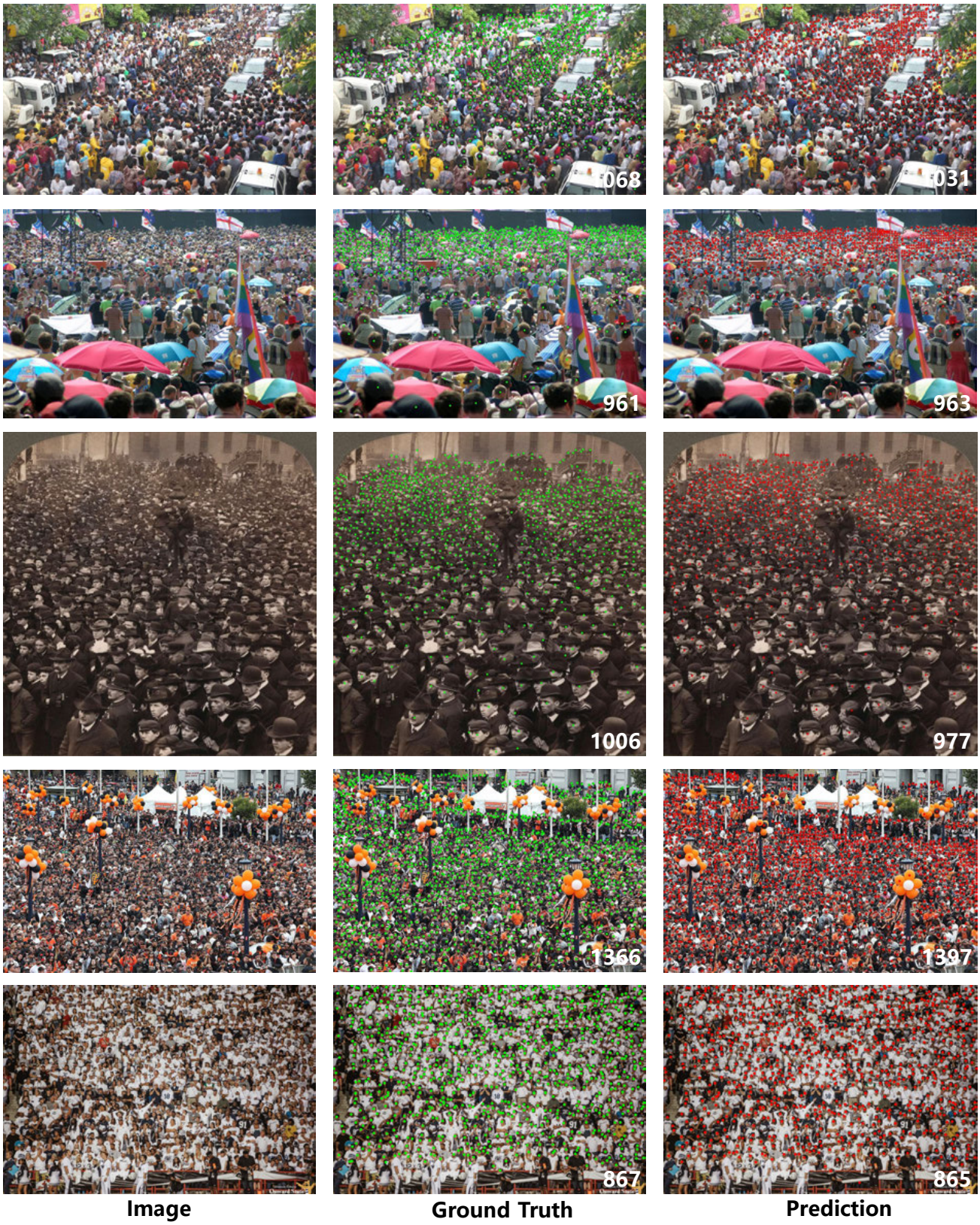

Figure 10. Visual results of congested scenes (1).

Figure 11. Visual results of congested scenes (2).

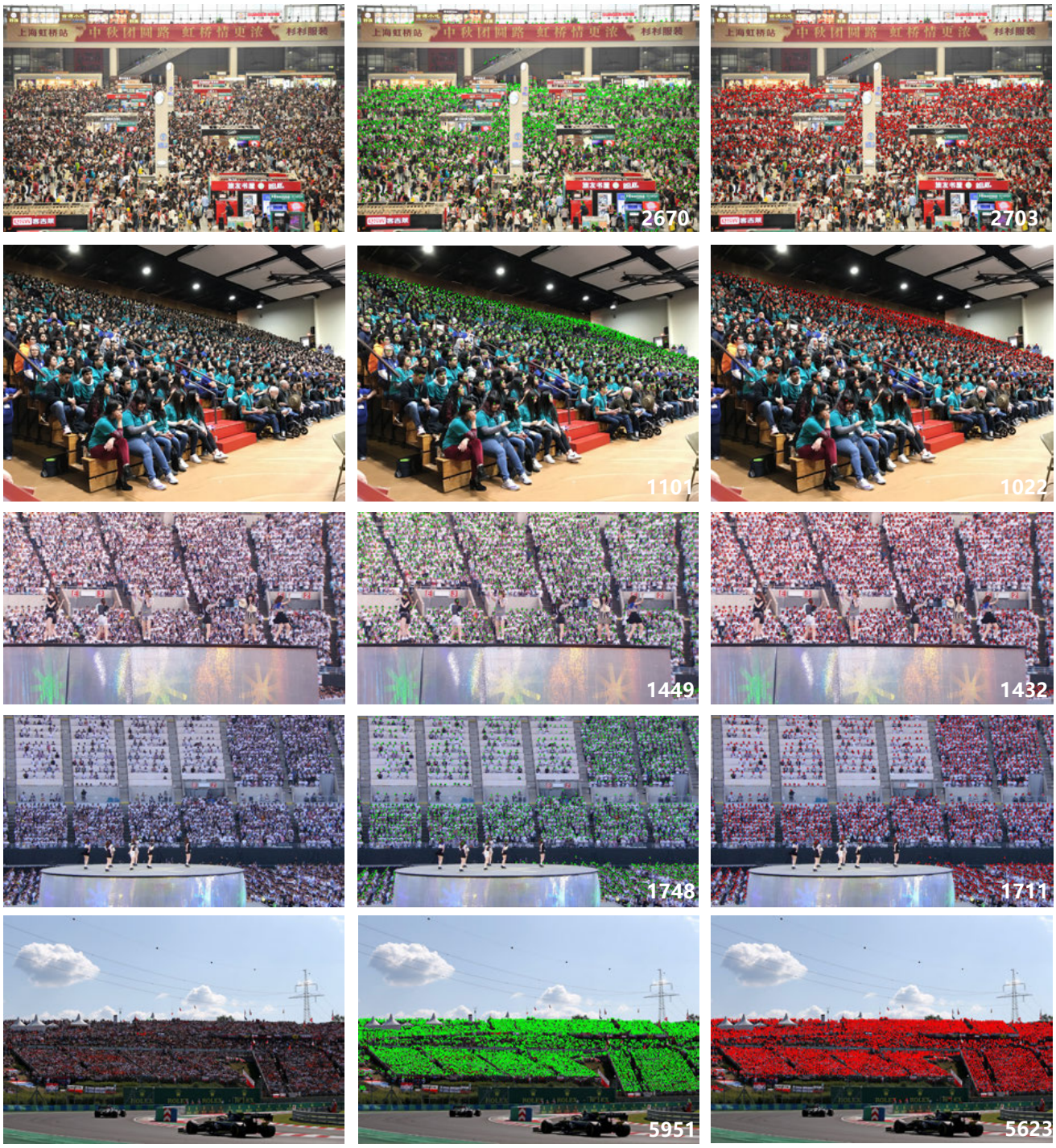

Figure 12. Visual results of congested scenes (3).

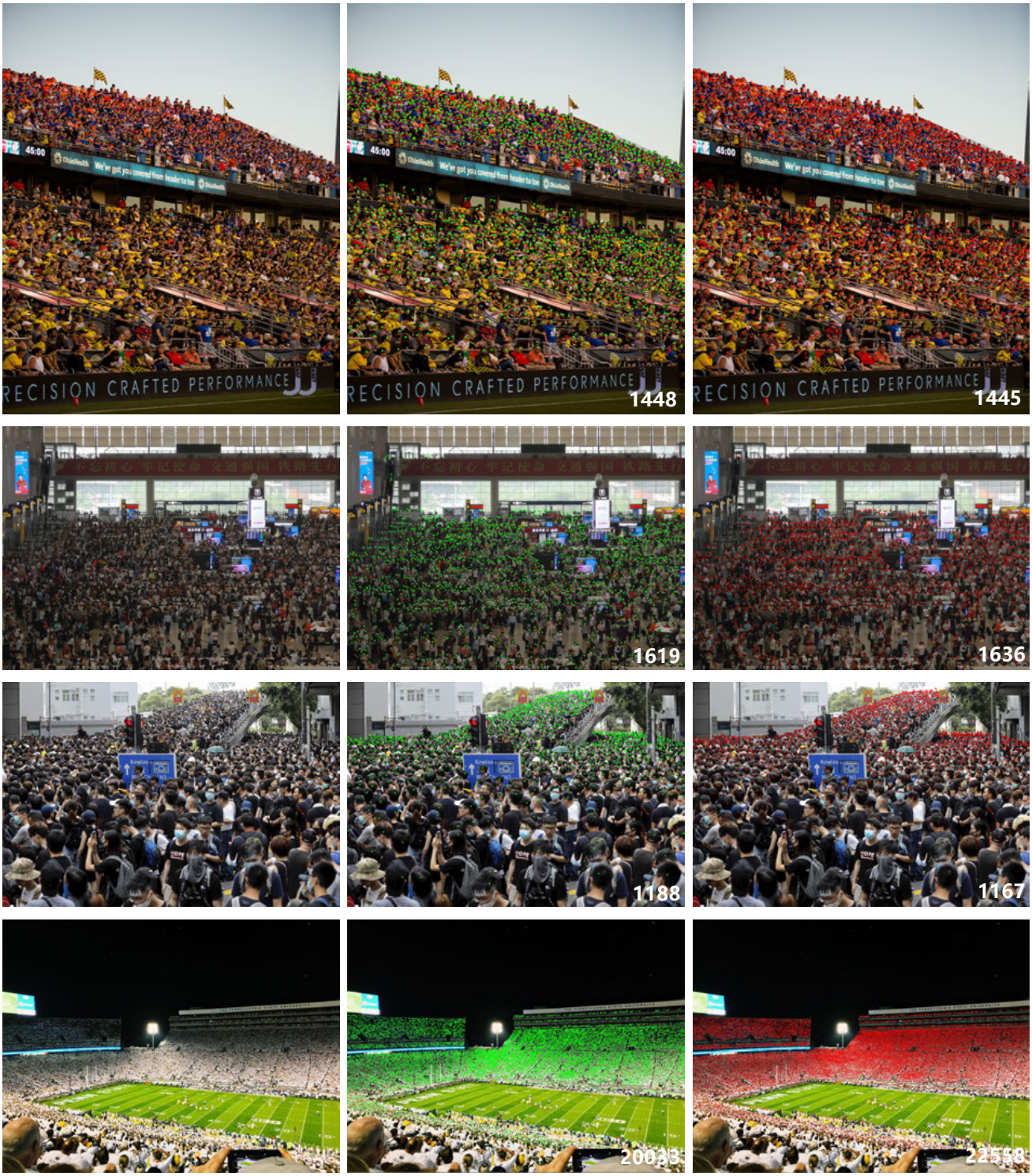

Figure 13. Visual results of congested scenes (4).