From Open Set to Closed Set: Counting Objects by Spatial Divide-and-Conquer Supplementary Materials

Haipeng Xiong†, Hao Lu‡, Chengxin Liu†, Liang Liu†, Zhiguo Cao†, Chunhua Shen‡
†Huazhong University of Science and Technology, China
‡The University of Adelaide, Australia
{hpxiong,zgcao}@hust.edu.cn, hao.lu@adelaide.edu.au

In this Supplement, we provide further clarifications and discussions on the motivation of S-DCNet, compare S-DCNet with other related ideas, and show qualitative results on evaluated datasets.

1. The Open-Set Problem of Density Maps

Density maps are actually in the open set as well. As shown in Fig. 1(b) (top), for a single point, different kernel sizes lead to different density values. When multiple objects exist and are close, density patterns are even much diverse as in Fig. 1(b) (bottom). Since observed samples are limited, density maps are certainly in an open set.

We add another baseline of CSRNet [3] to the toy experiment in Fig. 1(a). CSRNet also performs worse than S-DCNet in the open set (> 10), which implies the open-set problem also exists in density map based methods.

Furthermore, density map cannot be used in S-DCNet, because it is not spatially divisible. This is determined by its physical definition. However, local counts can. Thus we adopt local counts in S-DCNet rather than density maps.

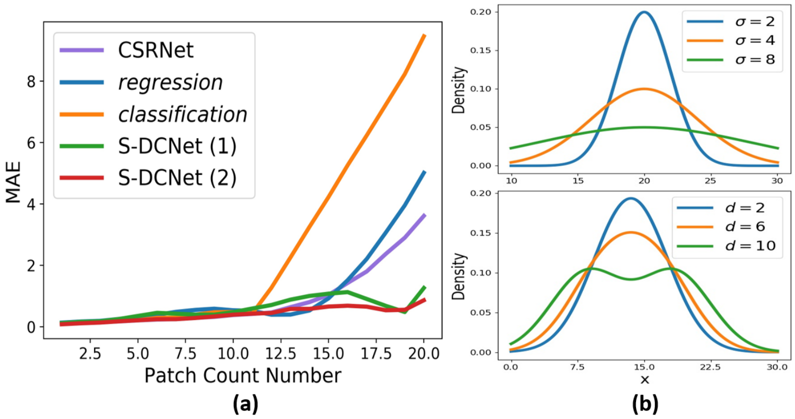

Figure 1. (a) The toy-level experiment with an extra "CSRNet" baseline. (b) Density values along one axis with various kernels (top), and with two kernels with different relative distances.

2. Relation to Other Methods

IG-CNN [1] IG-CNN drew inspirations from ensemble learning and trained a series of networks to tackle different scenes. While our S-DCNet focuses on inducing and utilizing physical laws, such as the "open set" problem in counting and the spatial divisibility of local counts. We propose to transform the open-set counting into a closed-set problem via spatial divide-and-conquer.

Attention Mechanisms Despite it is possible to provide explicit supervision to W_i, we find that S-DCNet already can produce reasonably good divisions with the implicit supervision provided by L_R^2. This has another benefit, the network can learn when to divide not just in counts larger than C_{max}. The visualizations of W_is in Fig. 2 further justify our point. To highlight the difference against attention, we remove the division decider and generate a three-channel output conditioned on F_2, then process it with softmax to obtain $W_0^{att}, W_1^{att}, W_2^{att}$. The final count is merged as $W_0^{att} * upsample(C_0) + W_1^{att} * upsample(C_1) + W_2^{att} * C_2$. In SHTech PartA, it has 64.1 MAE and 109.9 MSE (worse than S-DCNet). As per the visualization of W_i^{att} in Fig. 2, we find the attention only focuses on the highest resolution and no effect of division is observed. In addition, S-DCNet executes fusion progressively, while attention fuses the prediction in a single step.

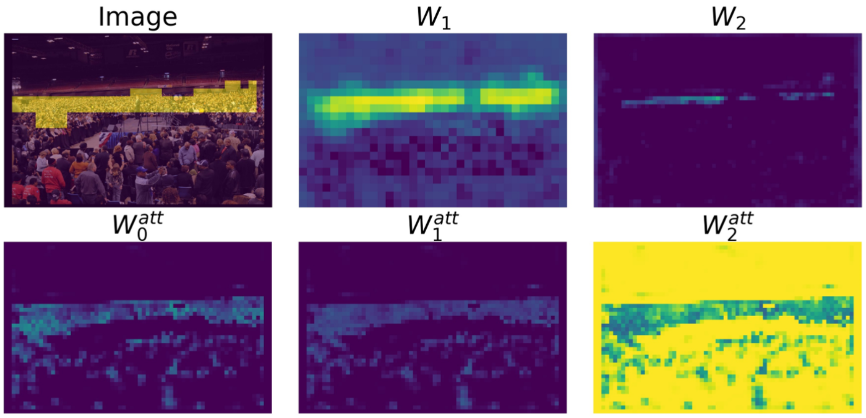

Figure 2. Visualization of W_i for S-DCNet (top) and the attention baseline (bottom). The lighter the image is, the greater the values are. In the input image, count values greater than C_{max} are indicated by yellow regions.

3. Further Discussions on S-DCNet

The necessity to distinguish counting task into open set and closed set scenarios One may raise the concern like: the relevance of distinguishing counting to an open set and closed set is unnecessary if each data point (head) is treated separately and the network learns to count each data point. If the network can count each head well, counting should already be addressed by detection networks. However, detection performs poorly when objects seriously overlap. This is why the notion of density map is introduced in [2], and density-based networks beat detection networks in counting. It is thus not suitable to treat each point separately, and distinguishing counting to an open set and closed set makes sense.

Generating ground-truth local counts Generating local counts directly from point annotations does not take partial objects cropped in patches into account. Density maps naturally tackle this situation. Thus we generate ground-truth counts of local patches by integrating over the density maps. This strategy is only utilized during training, while the point annotations are still used to calculate errors during validation.

If one position in W_1 is 0, which means the initial prediction should not be replaced. Is it possible that the same position in W_2 is 1? In theory, it is possible, because each division decision is independent. However, in practice, we do not observe such a behaviour of W (Fig. 2). Even this situation appears, we do not think it will be a problem. W_2 gives the second chance for division if the division decider makes a wrong decision in W_1.

Why C_2 is performing much worse than C_1 and C_0 in S-DCNet? C_0, C_1 and C_2 are trained jointly in S-DCNet and greatly influenced by the loss of L_R^2. As shown in Fig. 2, W_2 focus on local patches with high density, which means L_R^2 will push C_2 to predict well on these patches and ignore others. High density patches, however, only occupy a small fraction. C_2 thus tends to predict worse than C_0 and C_1. This may also explain why three-stage/four-stage S-DCNet performs worse than two-stage S-DCNet.

4. Qualitative Results of S-DCNet

We present some qualitative results of two-stage S-DCNet on five benchmarks (ShanghaiTech, UCF_CC_50, UCF-QNRF, TRANCOS and MTC) in Fig. 3 to 8. S-DCNet predicts the local count map conditioned on the input image, where each element denotes a count value of the corresponding 16×16 local area. Meanwhile, since the output stride of S-DCNet is 64, we pad the original image with zeros to ensure that the length and width are multiples of 64.

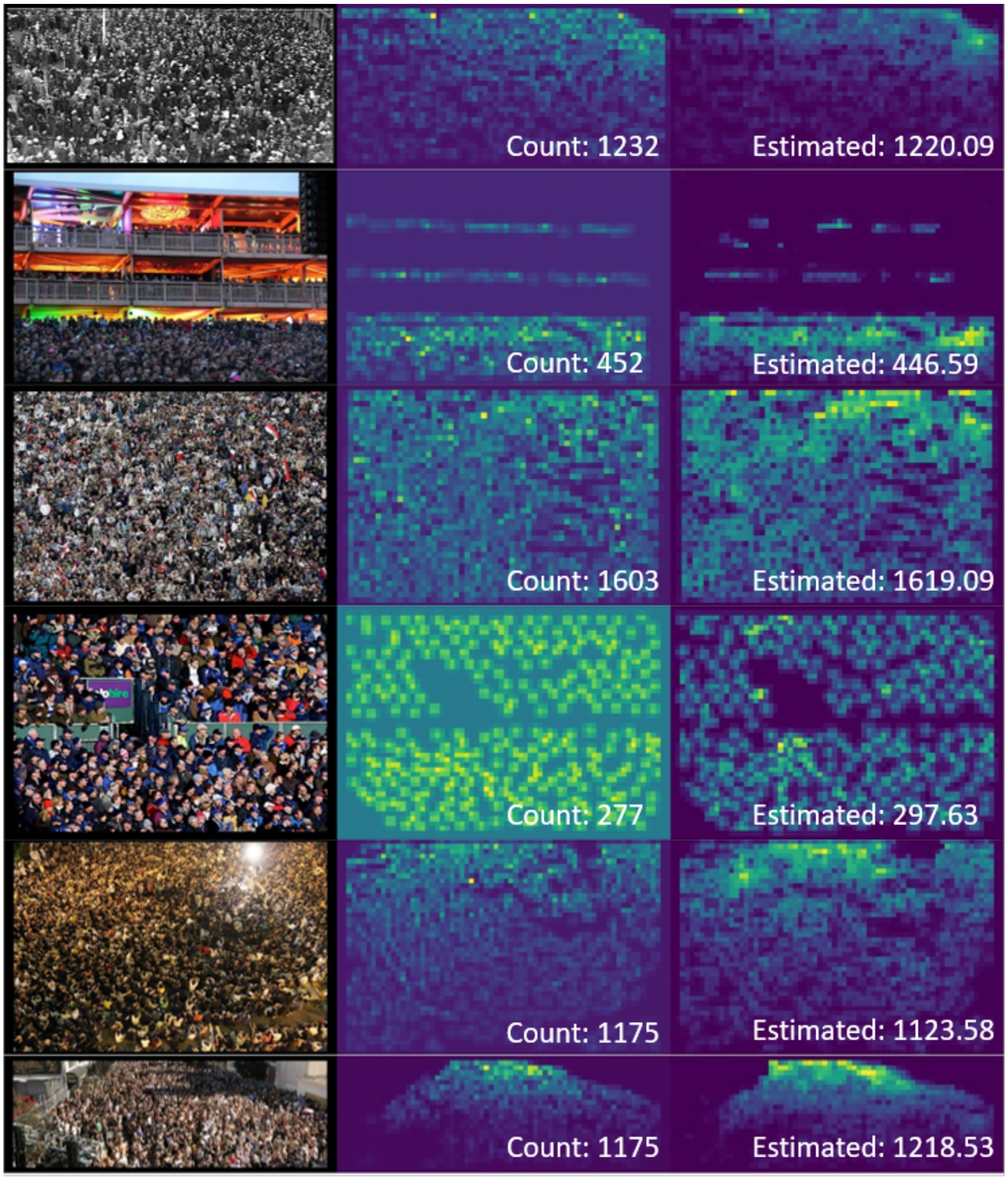

Figure 3. Some samples generated by S-DCNet from the test set of ShanghaiTech Part_A dataset. The left column shows the original images, while the middle and right columns display the ground truth and predicted count maps respectively.

References

[1] Deepak Babu Sam, Neeraj N. Sajjan, R. Venkatesh Babu, and Mukundhan Srinivasan. Divide and grow: Capturing huge diversity in crowd images with incrementally growing cnn. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3618–3626, 2018. 1

[2] Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1324–1332, 2010. 2

[3] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1091–1100, 2018. 1

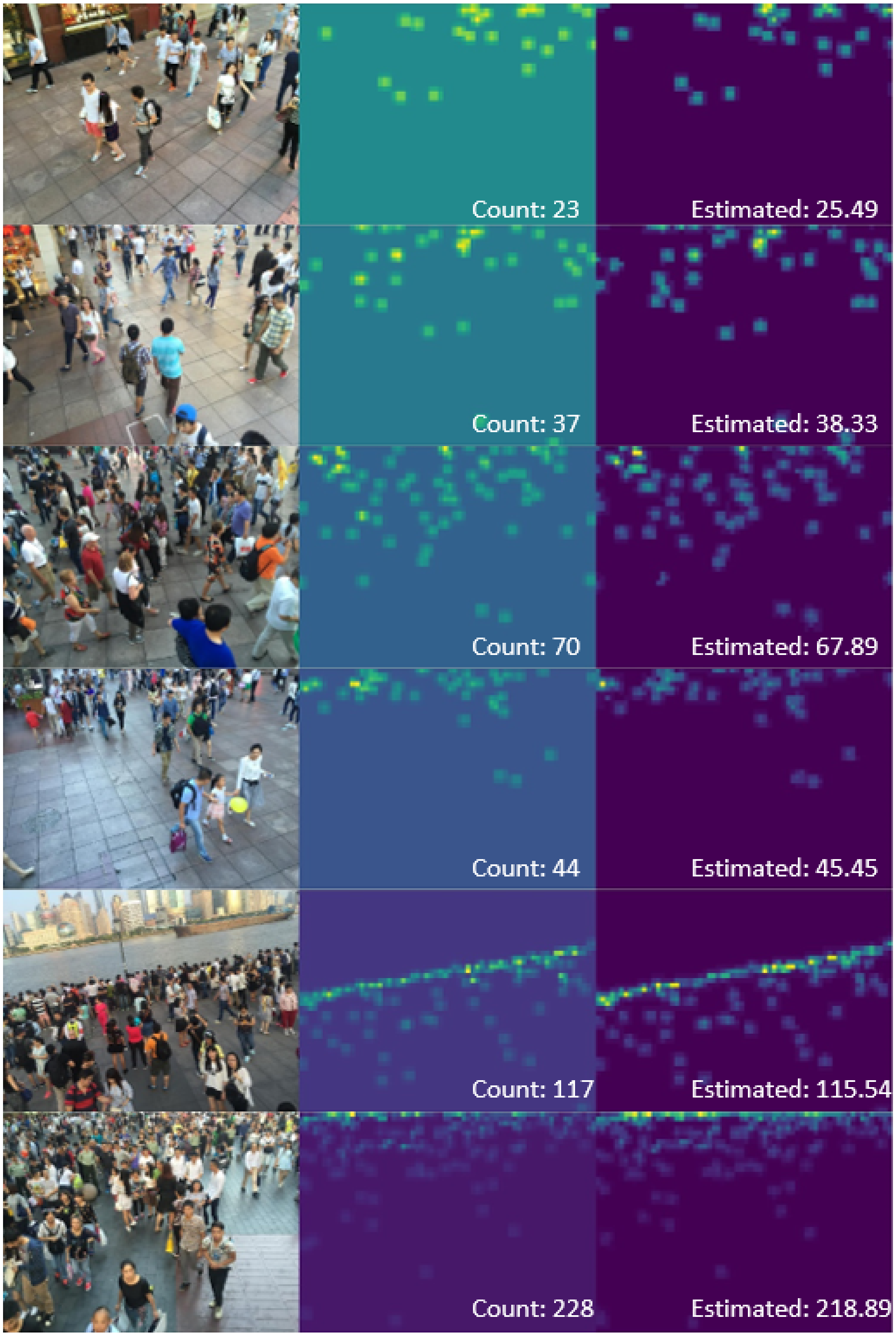

Figure 4. Some samples generated by S-DCNet from the test set of ShanghaiTech Part_B dataset. The left column shows the original images, while the middle and right columns display the ground truth and predicted count maps respectively.

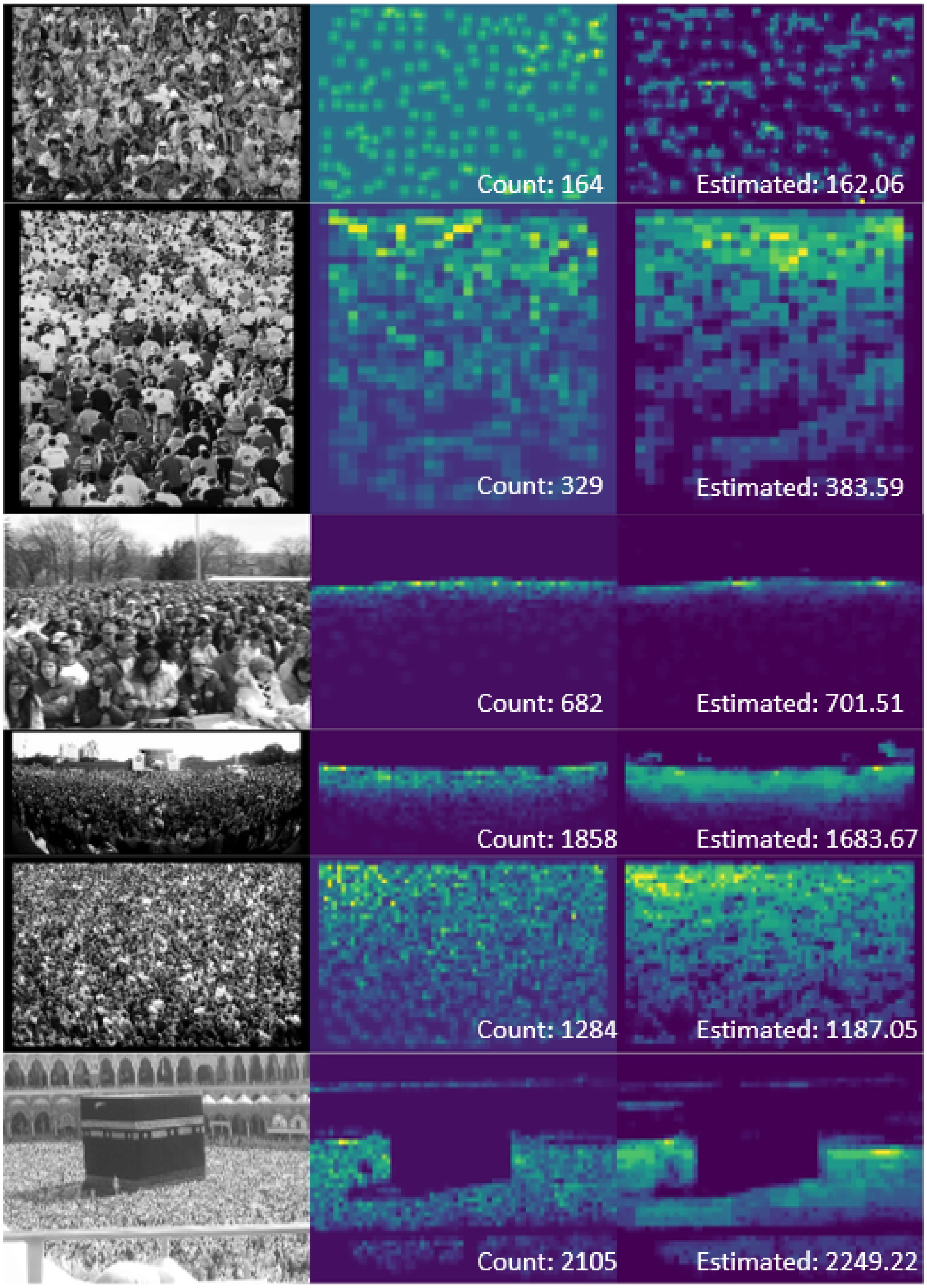

Figure 5. Some samples generated by S-DCNet from the test set of UCF_CC_50 dataset. The left column shows the original images, while the middle and right columns display the ground truth and predicted count maps respectively.

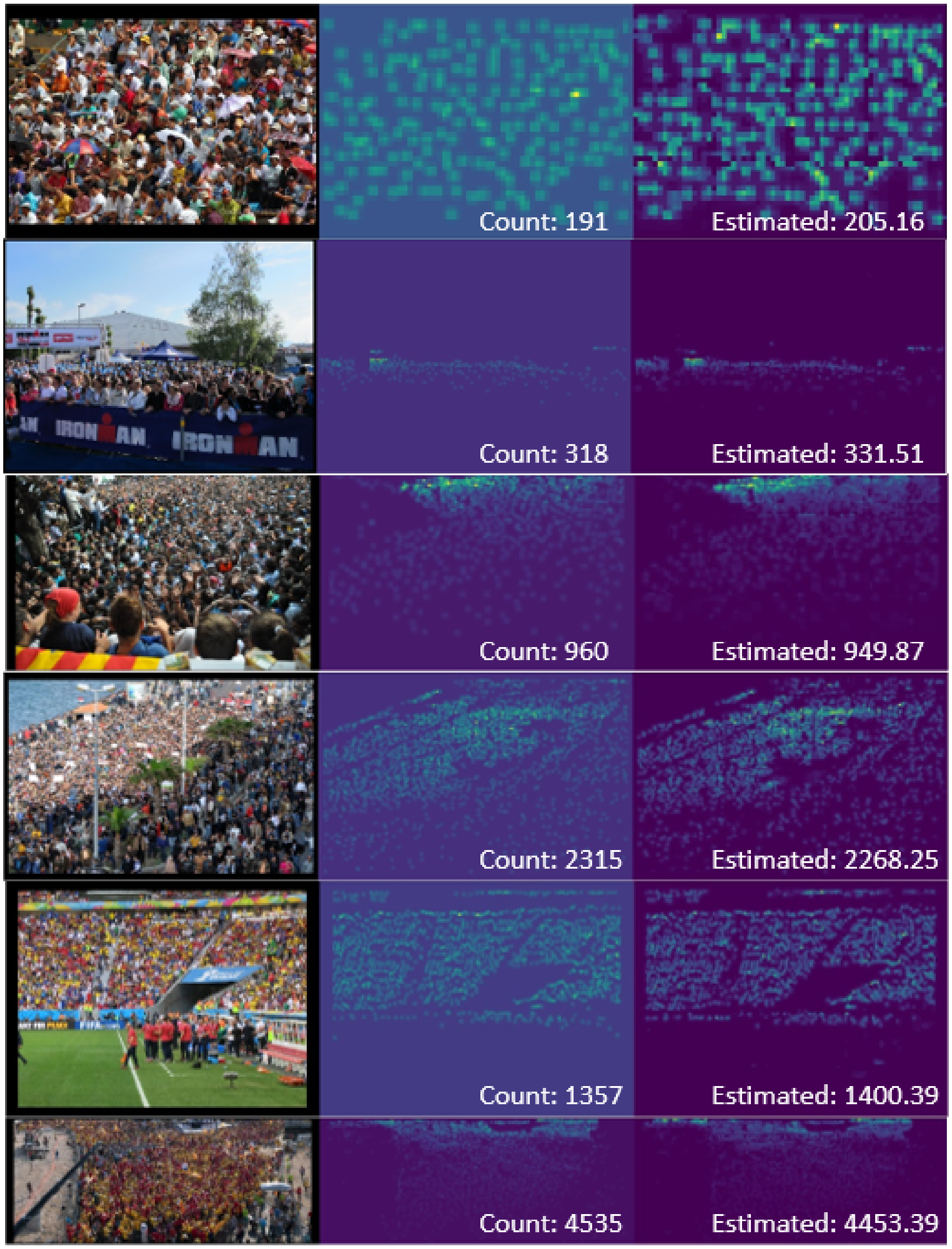

Figure 6. Some samples generated by S-DCNet from the test set of UCF-QNRF dataset. The left column shows the original images, while the middle and right columns display the ground truth and predicted count maps respectively.

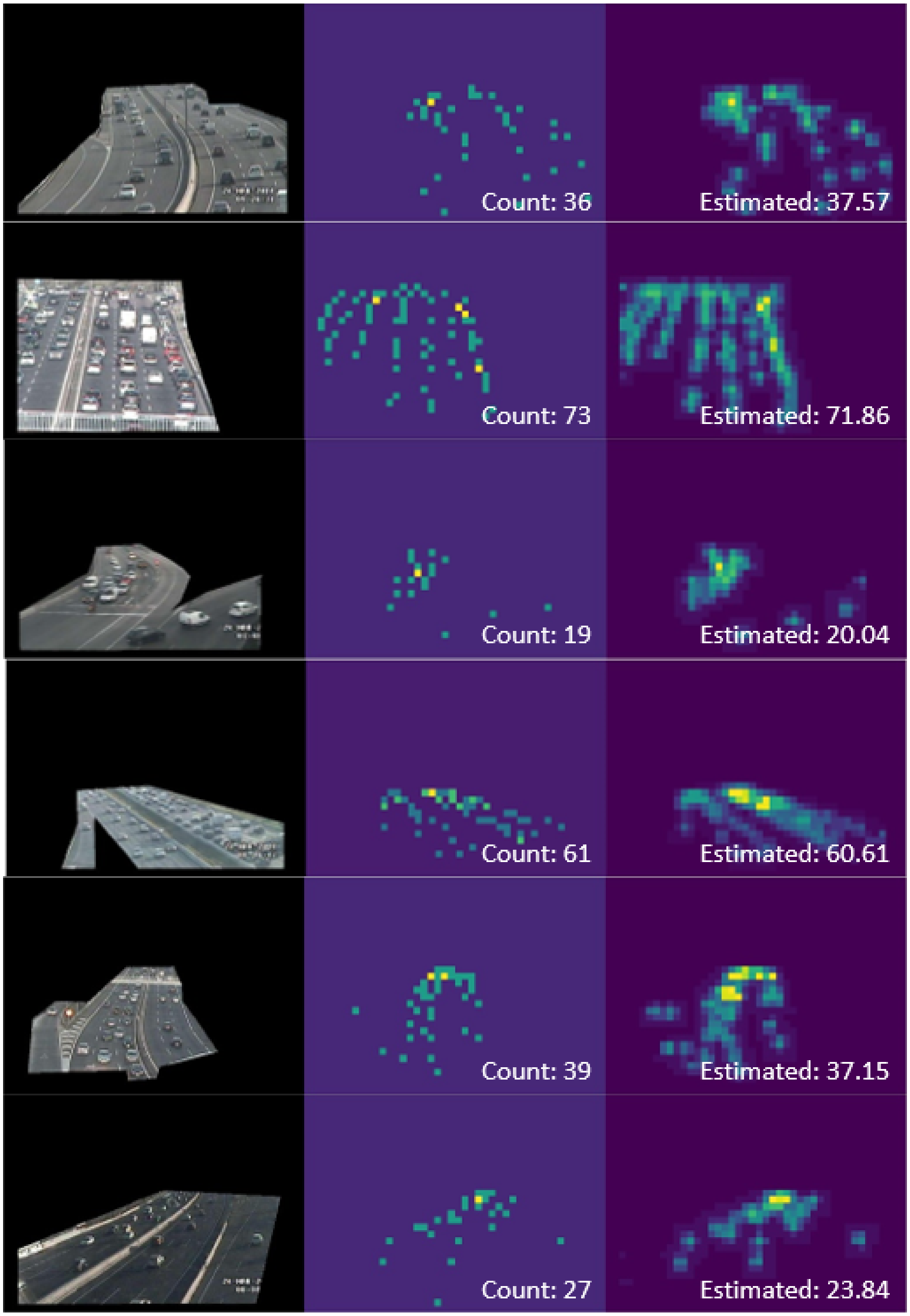

Figure 7. Some samples generated by S-DCNet from the test set of TRANCOS dataset. The left column shows the original images, while the middle and right columns display the ground truth and predicted count maps respectively.

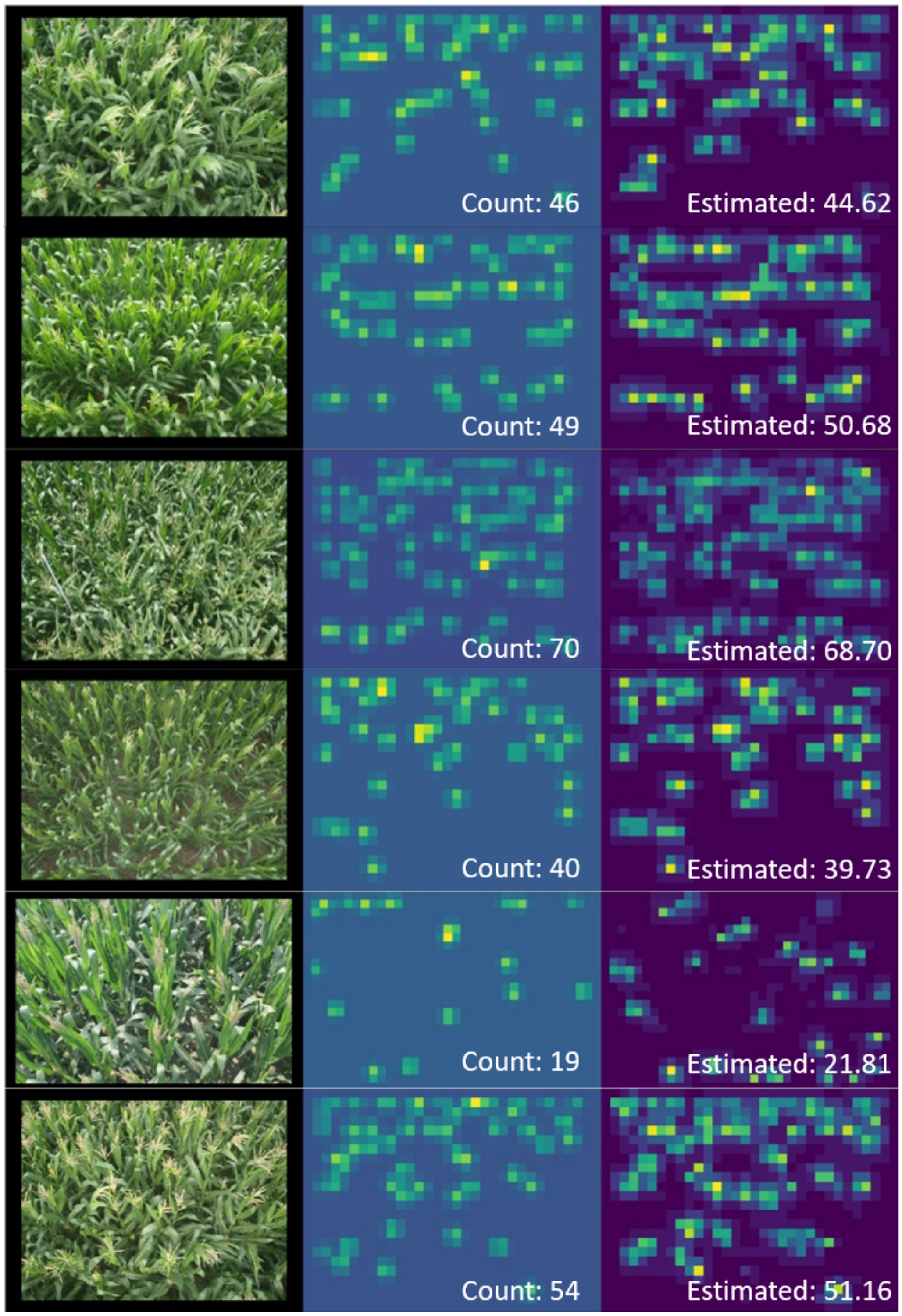

Figure 8. Some samples generated by S-DCNet from the test set of MTC dataset. The left column shows the original images, while the middle and right columns display the ground truth and predicted count maps respectively.