Counting with Focus for Free

Zenglin Shi, Pascal Mettes, and Cees G. M. Snoek
University of Amsterdam

Appendix A

In this section, we provide the architecture and ablation study of encoder-distiller-decoder network, the benefit of non-uniform kernel estimation across counting networks, and additional qualitative examples of (i) our encoder-distiller-decoder network, (ii) the effect of focus from segmentation, focus from global density and our combined focus, and (iii) success and failure cases for six benchmark datasets to better understand the benefits and limitations of the proposed method.

A.1. Encoder-Distiller-Decoder Network

The proposed encoder-distiller-decoder network (Section 3.4 in the main paper) is visualized in Fig. 1, and an ablation study on it is elaborated next.

We perform an ablation study on ShanghaiTech Part_A to analyze the encoder-distiller-decoder network configuration. We vary the architecture by including and excluding the distiller and decoder. When relying on the encoder and distiller only, the predicted density maps are upsampled to full resolution using bilinear interpolation. Results are in Table 1.

Encoder-Distiller. Adding a distiller module on top of the encoder reduces the MAE from 114.8 to 82.5. The distiller module fuses different features from multiple convolution layers with varying dilation rates, which is beneficial when counting multiple objects which appear in multiple scales in the image.

Encoder-Decoder. A traditional encoder-decoder network gives a better count than just encoder and an encoder-distiller network. An encoder-only network would compress the target objects to smaller size resulting in loss of fine details. Moreover, it produces density maps with a reduced resolution due to the downsample strides in the convolution operations. The distiller can compete with the decoder to some extent, but it cannot recover the spatial resolution and important details as well as the decoder.

Encoder-Distiller-Decoder. Incorporating the distiller in between an encoder and decoder into a single network gives the best counting results on all metrics due to the merits of both scale invariance and detail-preserving density maps. In Fig. 2 we show qualitatively that the network obtains a lower count error and generates higher quality density maps with less noise.

Table 1: **Ablation study of encoder-distiller-decoder network** on ShanghaiTech Part_A. Incorporating the proposed distiller module improves the performance of both an encoder-only network and an encoder-decoder network.

Encoder-distiller-decoder			**Metrics**	
Encoder	Distiller	Decoder	MAE	RMSE
✓			114.8	178.2
✓	✓		82.5	140.6
✓		✓	78.8	137.4
✓	✓	✓	**74.8**	**131.0**

A.2. Benefit of non-uniform kernel across counting networks

Next, we study the benefit of our non-uniform kernel estimation for existing counting methods. Apart from our own network, we also evaluate the benefit on two other counting networks, *i.e.* [2] and [1], for which code is available. Results in Table 2 demonstrate the proposed kernel has a better MAE and RMSE performance than the commonly used geometry-adaptive kernel [2] for all three networks. It demonstrates our non-uniform kernel is independent of the counting model.

A.3. Qualitative Results for Segment-, Density- & Combined-Focus

To illustrate the beneficial effect of the proposed focuses for reducing the counting error and suppressing background noise, we refer to Fig. 3. As shown in Fig. 3 (c) and Fig. 3 (d) compared to Fig. 3 (b), both segmentation focus and global-density focus show the ability to suppress noise and reduce the counting error. The combination of these two focuses leads to the lowest counting error and higher quality density maps with less noise as shown in Fig. 3 (e).

A.4. Success and Failure Cases

We have showed some success and failure results (Section 5.5 in the main paper). Finally we provide more quali-

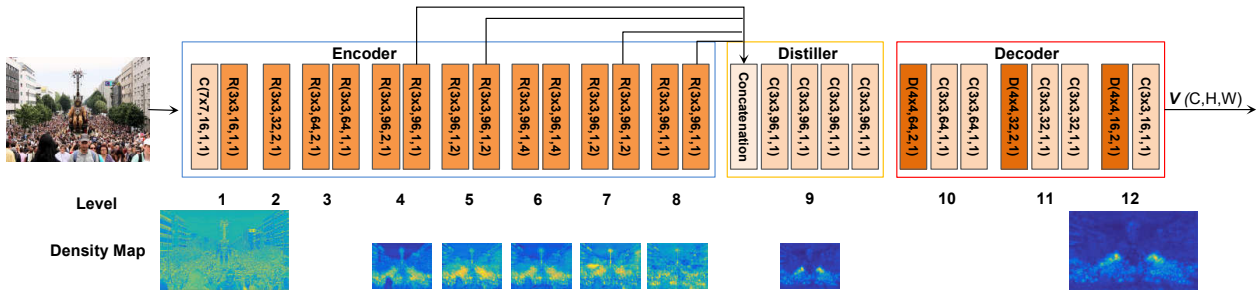

Figure 1: **Encoder-distiller-decoder network**. The network consists of convolution layers (C), residual blocks (R) and deconvolution layers (D) with parameters $(k \times k, c, s, d)$, where $k \times k$ is the kernel size, c is the number of channels, s is the stride size and d is the dilation size. Each convolution layer is followed by a ReLU activation layer and a batch normalization layer. The network is divided into several levels, such that all layers within a level have the same dilation and spatial resolution. The bottom row visualizes the mean feature map from different levels. The distiller module integrates the features from several encoder levels by attending to different parts of the image content for a better overall representation.

Table 2: **Benefit of non-uniform kernel estimation** on ShanghaiTech Part_A. Relying on a ground truth density map generated by the proposed kernel, rather than GAK [2], lowers the counting error for our method as well as alternatives.

	Zhang *et al.* [2]		Shi *et al.* [1]		*This paper*	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
GAK [2]	110.2	173.2	73.5	112.3	67.9	115.6
This paper	107.0	156.5	71.7	109.5	**65.2**	**109.4**

tative results on all six datasets. Even in challenging scenes our method is able to achieve an accurate count, as shown in the first two rows of Fig. 4, 5, 6, 7, 8 and 9. From the failure cases, as shown in the last two rows of Fig. 4, 5, 6, 7, 8 and 9, we can see that scenes with extremely dense small objects are still a big challenge, opening up opportunities for future work.

References

[1] Zenglin Shi, Le Zhang, Yun Liu, Xiaofeng Cao, Yangdong Ye, Ming-Ming Cheng, and Guoyan Zheng. Crowd counting with deep negative correlation learning. In *CVPR*, 2018. 1, 2

[2] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *CVPR*, 2016. 1, 2

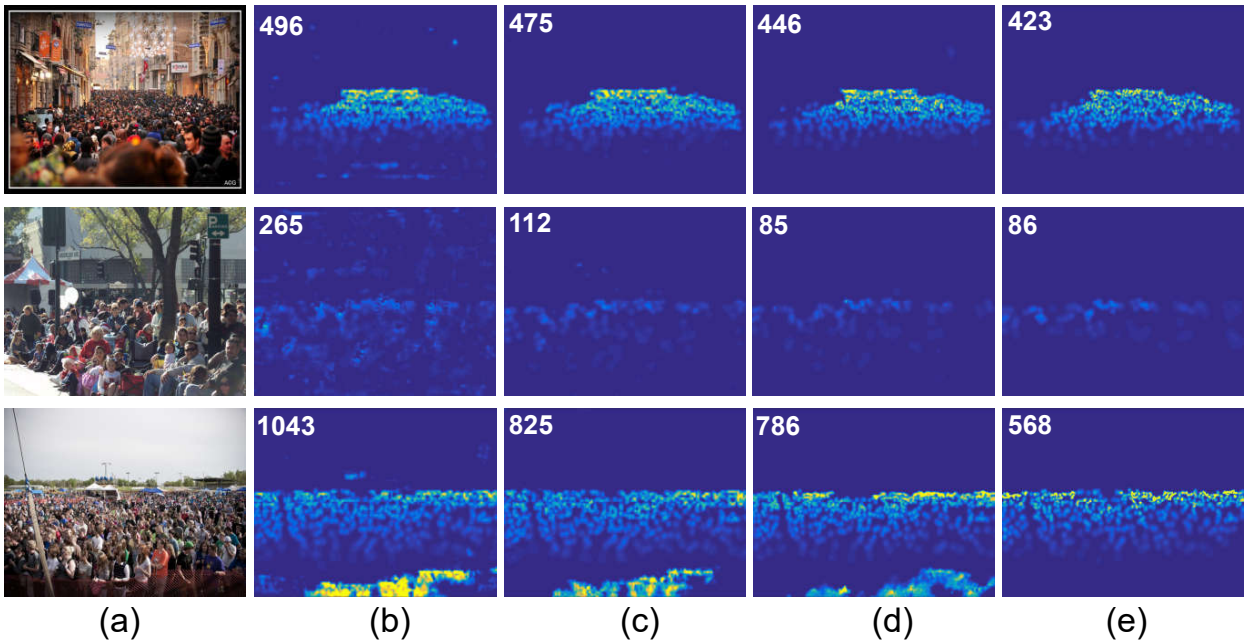

Figure 2: **Ablation study of encoder-distiller-decoder network.** (a) Sample images from ShanghaiTech Part_A and (b) predicted density map by encoder. Effect of (c) encoder-distiller and (d) encoder-distiller-decoder. For comparison, we show the ground truth for each sample in (e).

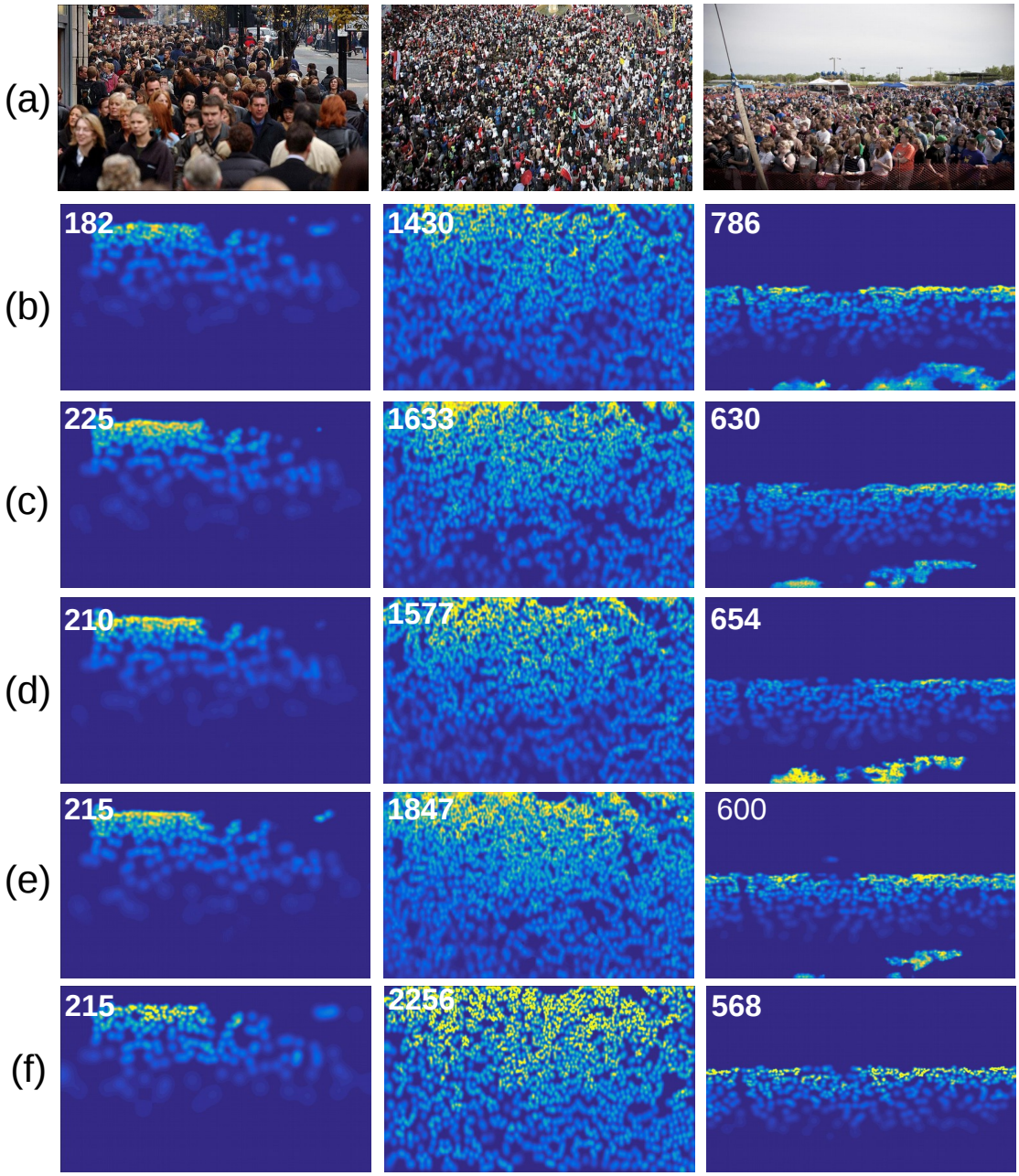

Figure 3: **Effect of segment-, density- & combined-focus** (a) Sample images from ShanghaiTech Part_A and (b) predicted density map without focus. Effect of (c) focus from segmentation, (d) focus from global density, and (e) our combined focus. For comparison, we show the ground truth for each sample in (f).

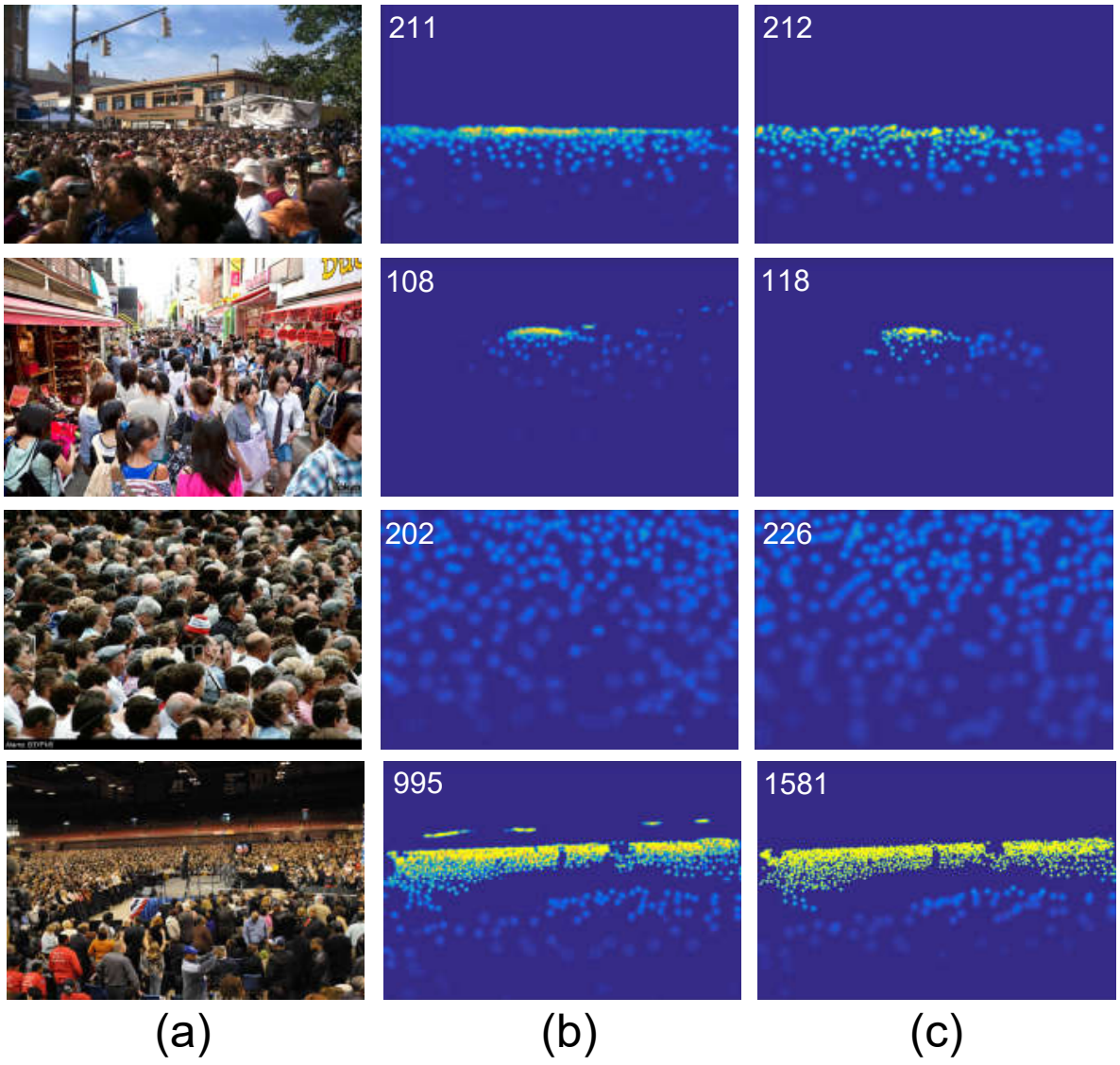

Figure 4: **Qualitative results for ShanghaiTech PartA.** (a) Sample images, (b) predicted density map, and (c) the ground truth.

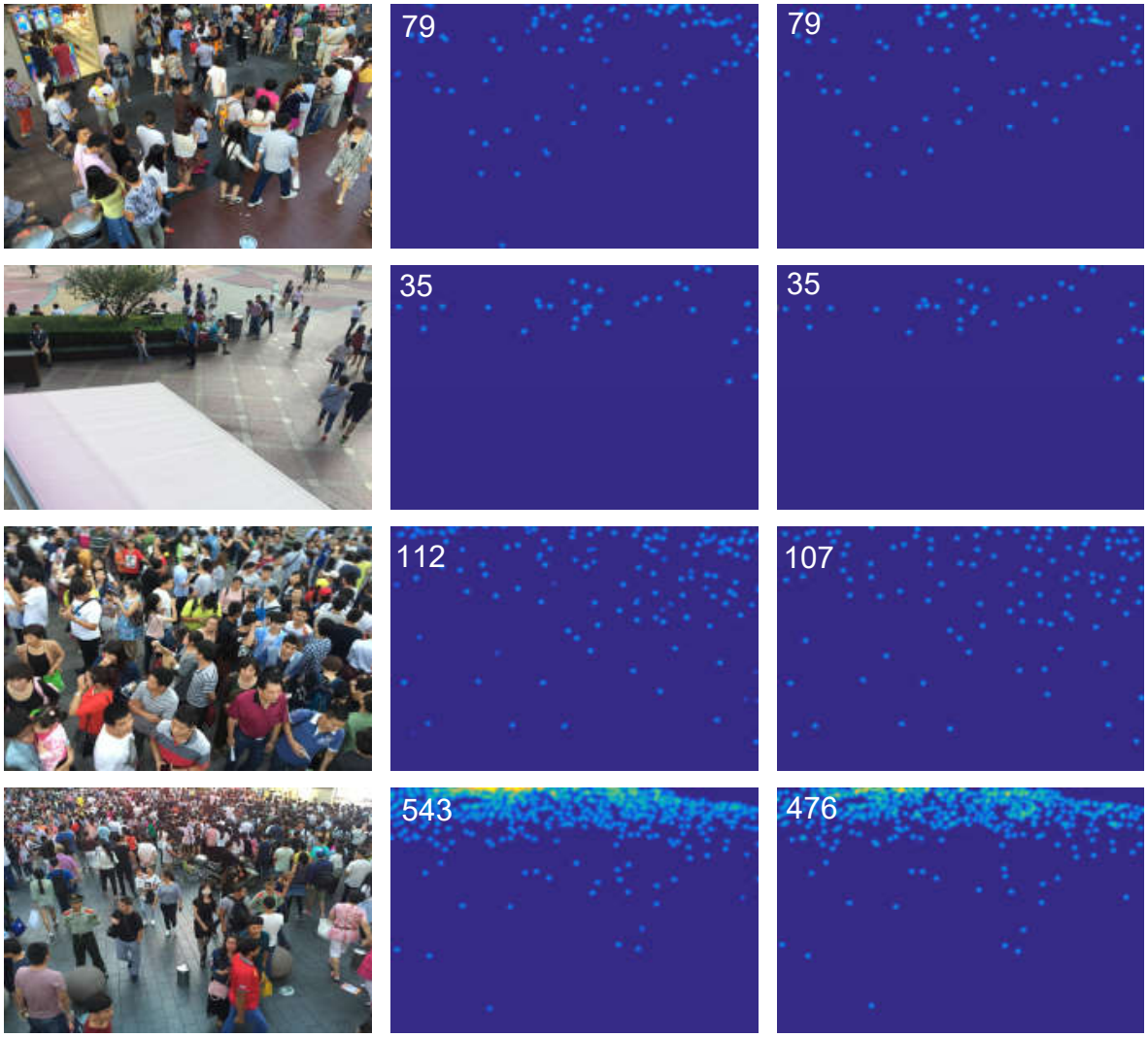

Figure 5: **Qualitative results for ShanghaiTech PartB.** (a) Sample images, (b) predicted density map, and (c) the ground truth.

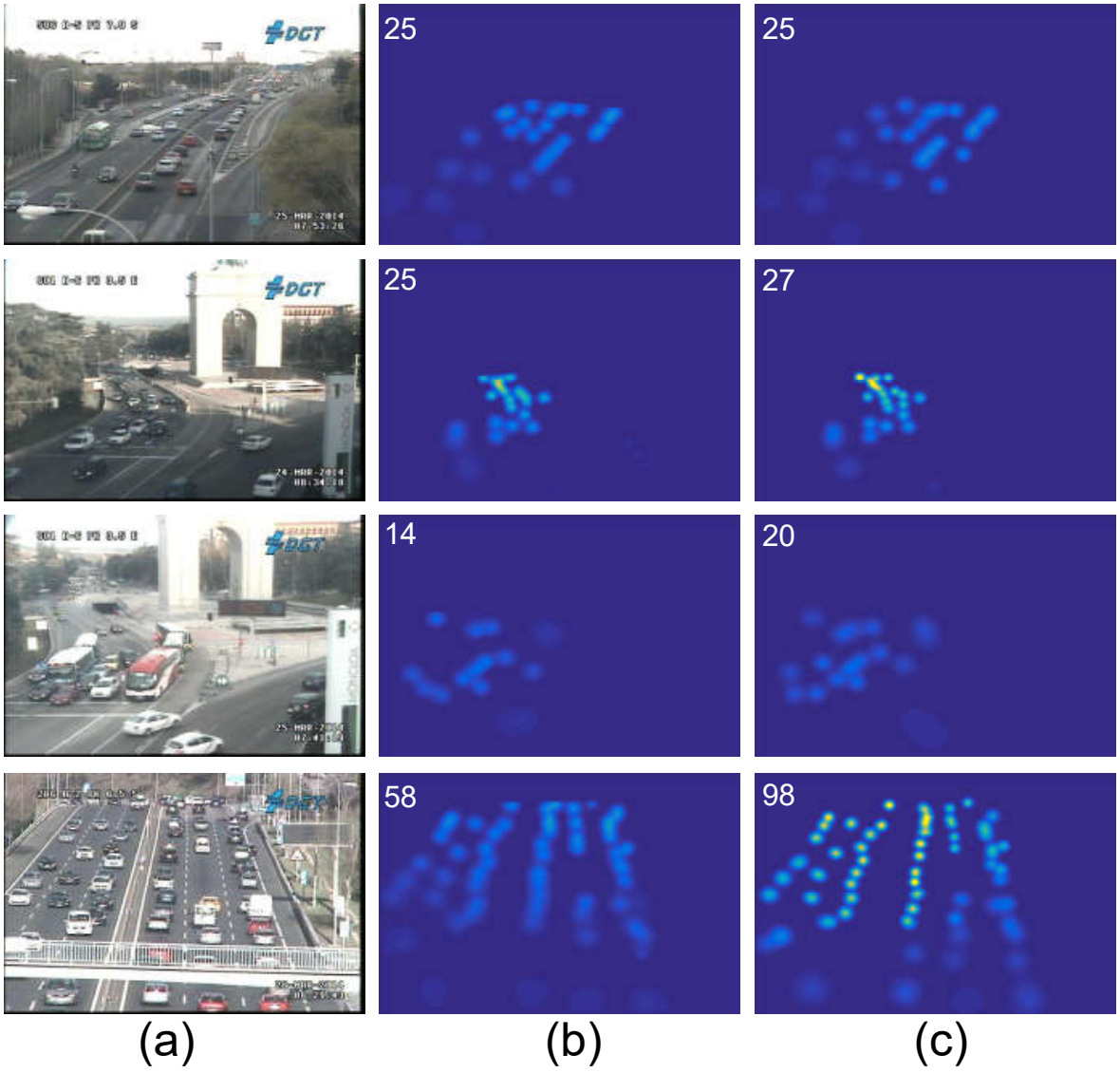

Figure 6: **Qualitative results for TRANCOS.** (a) Sample images, (b) predicted density map, and (c) the ground truth.

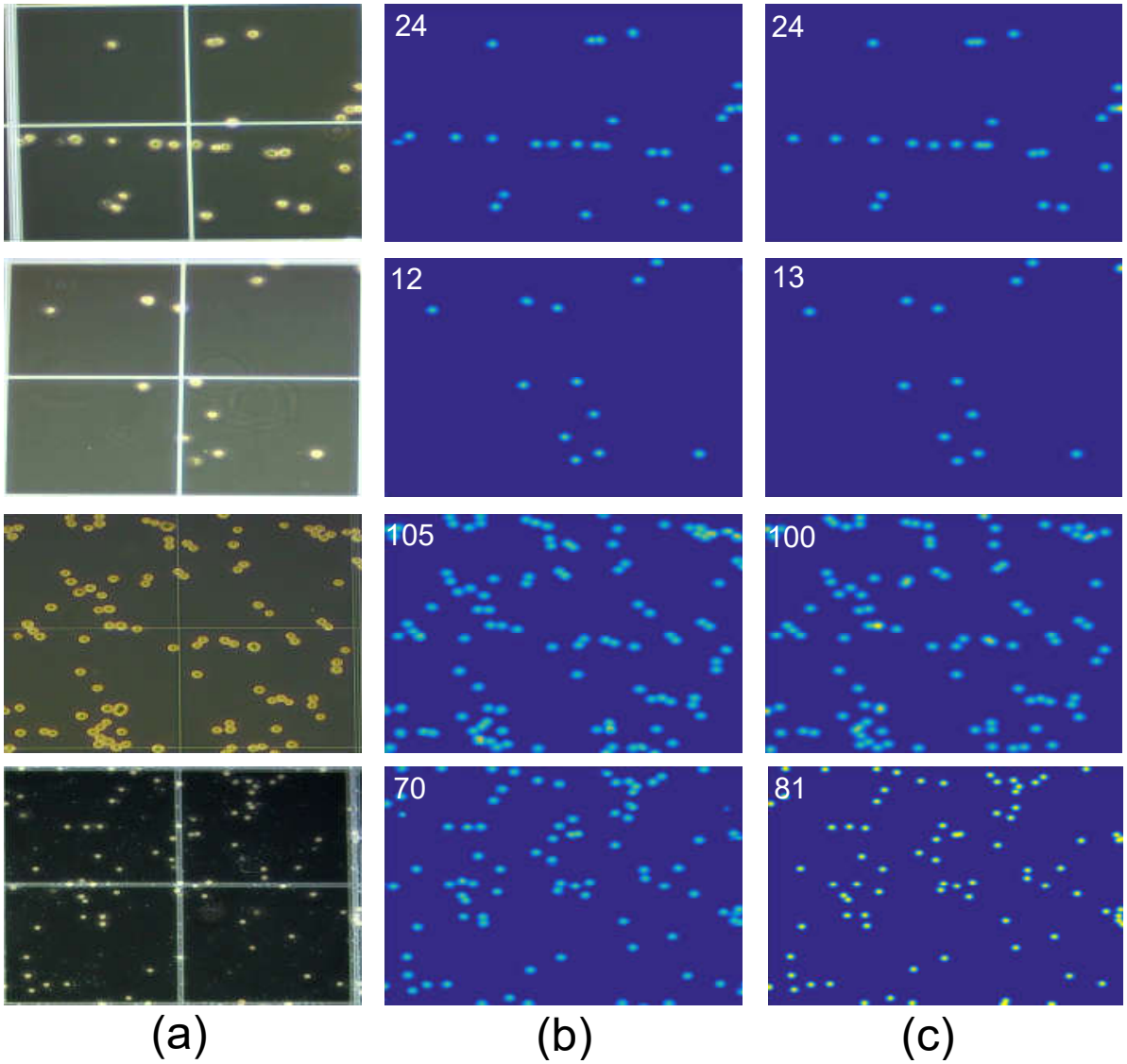

Figure 7: **Qualitative results for DCC.** (a) Sample images, (b) predicted density map, and (c) the ground truth.

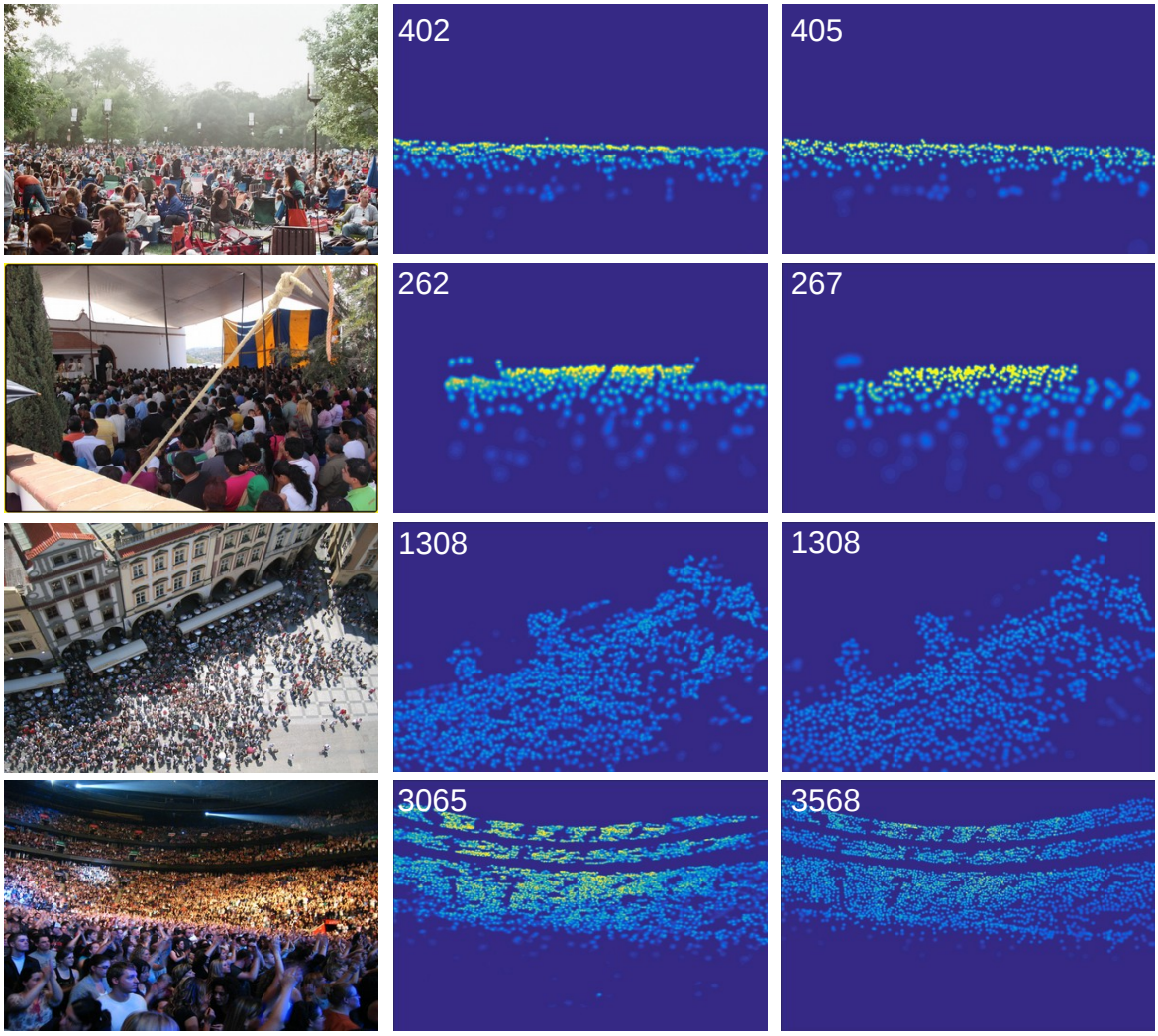

Figure 8: **Qualitative results for UCF-QNRF.** (a) Sample images, (b) predicted density map, and (c) the ground truth.

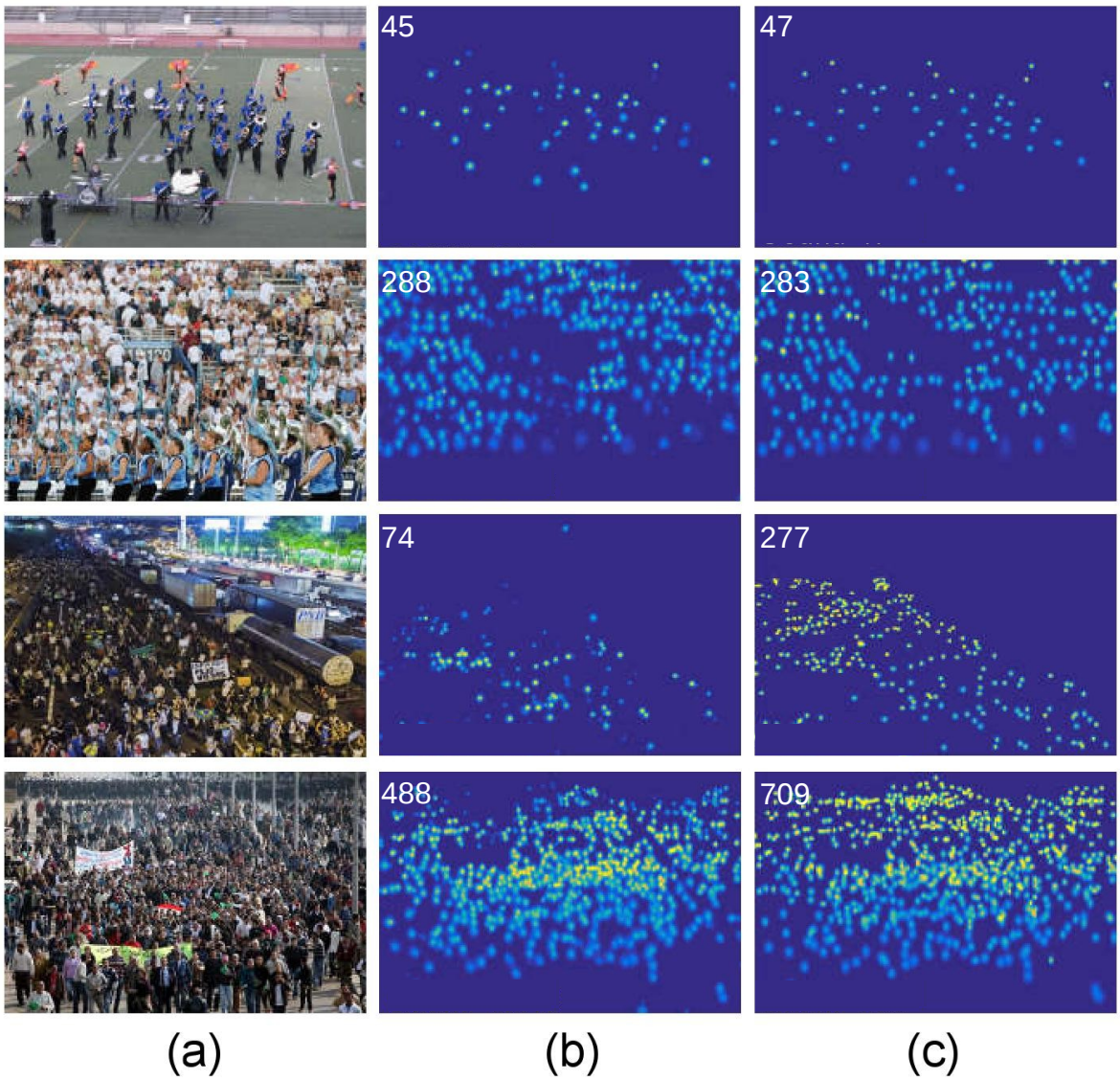

Figure 9: **Qualitative results for WIDER FACE.** (a) Sample images, (b) predicted density map, and (c) the ground truth.