Supplementary Materials for Perspective-Guided Convolution Networks for Crowd Counting

Zhaoyi Yan[1†], Yuchen Yuan[2], Wangmeng Zuo[1,3*], Xiao Tan[2], Yezhen Wang[1], Shilei Wen[2], Errui Ding[2]
[1]Harbin Institute of Technology, [2]Department of Computer Vision Technology (VIS), Baidu Inc., [3]Peng Cheng Laboratory, Shenzhen
yanzhaoyi@outlook.com, wmzuo@hit.edu.cn
{tanxchong, yezhen.wang0305}@gmail.com
{yuanyuchen02, wenshilei, dingerrui}@baidu.com

The following items are included in the supplementary materials:

- Comparisons with single/multi-feature concatenation strategies on ShanghaiTech Dataset.
- The architecture of PENet and the details of three training phases of PENet.
- The visualization of estimated perspective maps in each phase of training PENet.
- Reliability of the prediction of PENet.
- More density maps predicted by the proposed PGCNet.

A. More Comparisons with Single/Multi-feature Concatenation Strategies

To verify the effectiveness of our single-column architecture, we compare it with two kinds of feature concatenations, *i.e*. single- and multi-feature concatenations. Denoted by Θ_k^i the parameters of i-th convolution layer with kernel size k in our network. $F^{i-1}(X)$ denotes the feature of $(i-1)$-th convolution layer with the input X. Then, $\Theta_k^i(F^{i-1}(X))$ is the feature generated by Θ_k^i given $F^{i-1}(X)$ as input. The single-feature concatenation ($k \times k$) indicates the concatenation of $\Theta_k^i(F^{i-1}(X))$ and $F^{i-1}(X)$; on the other hand, the multi-feature concatenation ($1 \otimes K$) indicates the concatenation of $\Theta_1^i(F^{i-1}(X)), \Theta_3^i(F^{i-1}(X)), \cdots, \Theta_K^i(F^{i-1}(X))$ and $F^{i-1}(X)$. The comparisons are conducted on ShanghaiTech Part A, and the results are shown in Table 1. It is observed that our PGC block noticeably outperforms both single- and multi-feature concatenation strategies, which indicates its feasibility over conventional feature concatenations with fixed kernel size. Besides, tables 2, 3, 4, 5, 6 show the results when we adopt 2, 3, 4, 5, 6 feature concatenation blocks, respectively. In summary, our PGCNet can still outperform traditional feature concatenation by a large margin when stacking multiple PGC blocks.

Single feature concatenation			Multi-feature concatenation		
architecture	MAE	MSE	architecture	MAE	MSE
1×1	68.0	105.8	$1 \otimes 1$	68.0	105.8
3×3	67.1	104.0	$1 \otimes 3$	67.8	103.7
5×5	66.4	104.5	$1 \otimes 5$	67.0	104.9
7×7	66.9	105.2	$1 \otimes 7$	66.8	104.8
Ours	**65.8**	**98.0**	-	-	-

Table 1. Comparison between the PGC block and different feature concatenation strategies on ShanghaiTech Part A.

Single feature concatenation			Multi-feature concatenation		
ShanghaiTech Part A					
architecture	MAE	MSE	architecture	MAE	MSE
1×1	68.11	105.95	$1 \otimes 1$	68.11	105.95
3×3	68.36	105.30	$1 \otimes 3$	66.48	103.71
5×5	67.86	106.11	$1 \otimes 5$	67.03	105.04
7×7	68.39	105.80	$1 \otimes 7$	67.55	106.66
Ours	**64.46**	**96.61**	-	-	-
ShanghaiTech Part B					
1×1	10.14	15.70	$1 \otimes 1$	10.14	15.70
3×3	10.10	15.87	$1 \otimes 3$	10.34	16.02
5×5	10.04	16.17	$1 \otimes 5$	10.43	16.03
7×7	9.96	16.36	$1 \otimes 7$	10.48	15.94
Ours	**9.62**	**15.39**	-	-	-

Table 2. Comparison between PGC and different feature concatenation strategies when adopting 2 PGC blocks.

B. The Architecture of PENet and the Training Details

We adopt *Convolution-LeakyReLU* as the basic pattern of the encoder E_p, with each block scaling down the en-

[†]This work was done when Zhaoyi Yan was a research intern at Baidu
[*]Corresponding author

Single feature concatenation			Multi-feature concatenation		
ShanghaiTech Part A					
architecture	MAE	MSE	architecture	MAE	MSE
1×1	68.78	103.68	$1 \otimes 1$	68.78	103.68
3×3	68.74	106.03	$1 \otimes 3$	67.84	104.90
5×5	69.92	105.07	$1 \otimes 5$	67.91	106.59
7×7	67.79	106.66	$1 \otimes 7$	67.34	104.71
Ours	**60.94**	**95.23**	-	-	-
ShanghaiTech Part B					
1×1	9.97	15.68	$1 \otimes 1$	9.97	15.68
3×3	9.78	15.64	$1 \otimes 3$	9.99	15.59
5×5	9.91	15.87	$1 \otimes 5$	10.11	15.71
7×7	9.94	15.96	$1 \otimes 7$	10.37	16.20
Ours	**9.21**	**14.85**	-	-	-

Table 3. Comparison between PGC and different feature concatenation strategies when adopting 3 PGC blocks.

Single feature concatenation			Multi-feature concatenation		
ShanghaiTech Part A					
Single feature concatenation			Multi-feature concatenation		
architecture	MAE	MSE	architecture	MAE	MSE
1×1	69.03	105.01	$1 \otimes 1$	69.03	105.01
3×3	70.87	108.97	$1 \otimes 3$	77.67	121.12
5×5	69.51	101.59	$1 \otimes 5$	70.36	107.60
7×7	70.14	105.04	$1 \otimes 7$	70.32	106.48
Ours	**58.52**	**89.50**	-	-	-
ShanghaiTech Part B					
1×1	9.92	16.22	$1 \otimes 1$	9.92	16.22
3×3	9.82	15.80	$1 \otimes 3$	10.18	15.81
5×5	9.65	15.40	$1 \otimes 5$	10.10	15.88
7×7	9.79	15.72	$1 \otimes 7$	10.15	16.05
Ours	**9.10**	**14.43**	-	-	-

Table 4. Comparison between PGC and different feature concatenation strategies when adopting 4 PGC blocks.

Single feature concatenation			Multi-feature concatenation		
ShanghaiTech Part A					
architecture	MAE	MSE	architecture	MAE	MSE
1×1	67.69	104.48	$1 \otimes 1$	67.69	104.48
3×3	70.70	110.53	$1 \otimes 3$	77.47	119.94
5×5	68.17	103.19	$1 \otimes 5$	69.98	108.13
7×7	70.25	106.31	$1 \otimes 7$	69.51	107.02
Ours	**56.98**	**86.02**	-	-	-
ShanghaiTech Part B					
1×1	9.76	15.67	$1 \otimes 1$	9.76	15.67
3×3	9.91	16.02	$1 \otimes 3$	10.64	16.87
5×5	9.97	15.73	$1 \otimes 5$	10.03	15.94
7×7	10.21	16.26	$1 \otimes 7$	10.38	16.45
Ours	**8.84**	**13.66**	-	-	-

Table 5. Comparison between PGC and different feature concatenation strategies when adopting 5 PGC blocks.

Single feature concatenation			Multi-feature concatenation		
ShanghaiTech Part A					
architecture	MAE	MSE	architecture	MAE	MSE
1×1	69.03	105.01	$1 \otimes 1$	69.03	105.01
3×3	70.87	108.97	$1 \otimes 3$	77.24	116.90
5×5	71.34	108.98	$1 \otimes 5$	73.71	119.42
7×7	72.48	110.42	$1 \otimes 7$	73.83	112.81
Ours	**58.26**	**90.18**	-	-	-
ShanghaiTech Part B					
1×1	10.03	16.19	$1 \otimes 1$	10.03	16.19
3×3	10.20	16.25	$1 \otimes 3$	11.34	18.52
5×5	10.44	17.03	$1 \otimes 5$	10.56	17.30
7×7	10.65	16.93	$1 \otimes 7$	10.80	17.58
Ours	**9.01**	**14.20**	-	-	-

Table 6. Comparison between PGC and different feature concatenation strategies when adopting 6 PGC blocks.

coder feature by ratio 2. While the decoder enlarges the resolution of encoded feature by the combination of *UpConv-ReLU*. Details of the architecture of PENet are demonstrated in Table 8.

In the first phase, PENet is trained to reconstruct the input when given certain perspective map. As PGCNet adopts CSRNet as the baseline, the resolution of perspective needed by PGC block is only $1/8$ size of the original input. Therefore, we do not need to predict the full resolution of perspective map. We downsample the original image to make the resized image be only $1/8$ resolution of the original image and then train PENet as an identity mapping of perspective maps. We get 0.020 MAE and 0.031 MSE in this phase.

In the second stage, we fix the parameters of D_p trained in the first phase, and only train E_p, aiming at constructing the perspective map from its corresponding RGB image. We get 0.101 MAE and 0.142 MSE in the second training phase. And finally in the third stage, *Ours A* denotes directly adopting the estimated perspective map of PENet as the ground-truth, while *Ours B* represents the PENet is embedded as a perspective estimation branch and the whole network can be trained end-to-end. Quantitative results of *Ours A* and *Ours B* have been demonstrated in Sec. 5.3.

C. The Visualization of Estimated Perspective Maps in Each Phase of Training PENet

Beyond the quantitative results given in Sec. B, we also demonstrate visualizations of perspective maps in these three phases. Fig. 1(a)(b) show the input / ground-truth and estimated perspective map, respectively. It can be seen that our PENet performs well in reconstructing the input in the first phase. Fig. 2 demonstrates two examples of estimated maps predicted by PENet. PENet generally produces roughly accurate perspective maps, showing its robustness

in dealing with different scenes. Taking into account the quantitative results, it is obvious that PENet is capable of predicting a meaningful perspective map quantitatively and qualitatively in the first two stages.

For the third phase, Fig. 3 shows the estimated perspective maps of *Ours A* and *Ours B*, respectively shown in the second and third columns. Comparing these two images in Fig. 3(a) vertically, it can be seen that the visual angle of the image in the second row is relatively larger than that of the image in the first row. This observation accords with the directly estimated maps in the Fig. 3(b), which can been seen that the second image contains more larger values comparing with the first image does. When we train the whole network end-to-end, the perspective estimation branch can still predict generally satisfying perspective maps (*i.e.*, Fig. 3(c)). It is seen that larger perspective values move from the right to the left, which is visually explanatory.

Therefore, our PENet works well either in directly predicting perspective maps or in functioning as the perspective map estimator of the end-to-end architecture.

D. Reliability of the Prediction of PENet

PENet is designed as a compromise of the situation that perspective annotations are unavailable, in which the reliability of PENet is essential. Therefore, we conduct an experiment to confirm the feasibility of PENet. Table 7 demonstrates the comparisons of adopting the ground-truth or the estimated perspective map as the guidance of spatially variant smoothing on ShanghaiTech Part A/B and WorldExpo'10. MAEs are respectively 58.1, 9.0 and 8.3, with a small decrease of 1.1, 0.2 and 0.2, respectively. This indicates that PENet is competent to a reasonable perspective map estimator.

Perspective Map	ShanghaiTech Part A/B	WorldExpo'10
Estimated	58.1/9.0	8.3
Ground-truth	**57.0/8.8**	**8.1**

Table 7. Different guidances of PGC on ShanghaiTech Part A/B and WorldExpo'10.

E. More Density Maps Predicted by the Proposed PGCNet

Figs. 4 and 5 demonstrate more density maps predicted by PGCNet as well as by CSRNet. From the visualization, it can be seen that our PGCNet shows its superiority to CSRNet in estimating a more accurate number of pedestrians in either sparse or congested scenes. The quantitative results have been shown in Sec. 5.

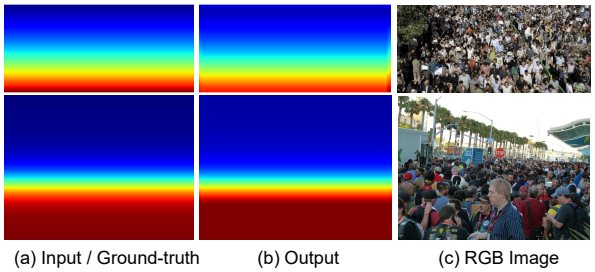


Figure 1. Results of the first phase of training PENet. Given (a) the input of PENet, (b) is the reconstructing output, and (c) denotes the corresponding RGB image.

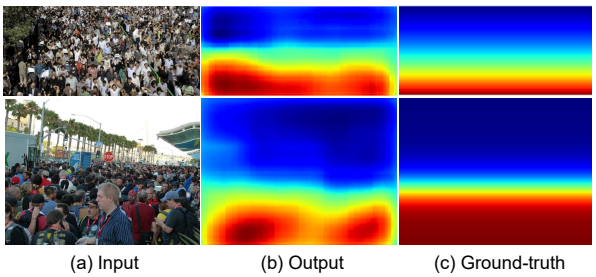


Figure 2. Results of second phase of training PENet. Given (a) the input of PENet, (b) is the output of PENet, and (c) represents the ground-truth.

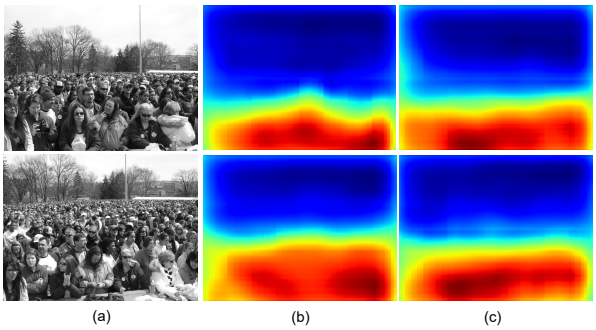
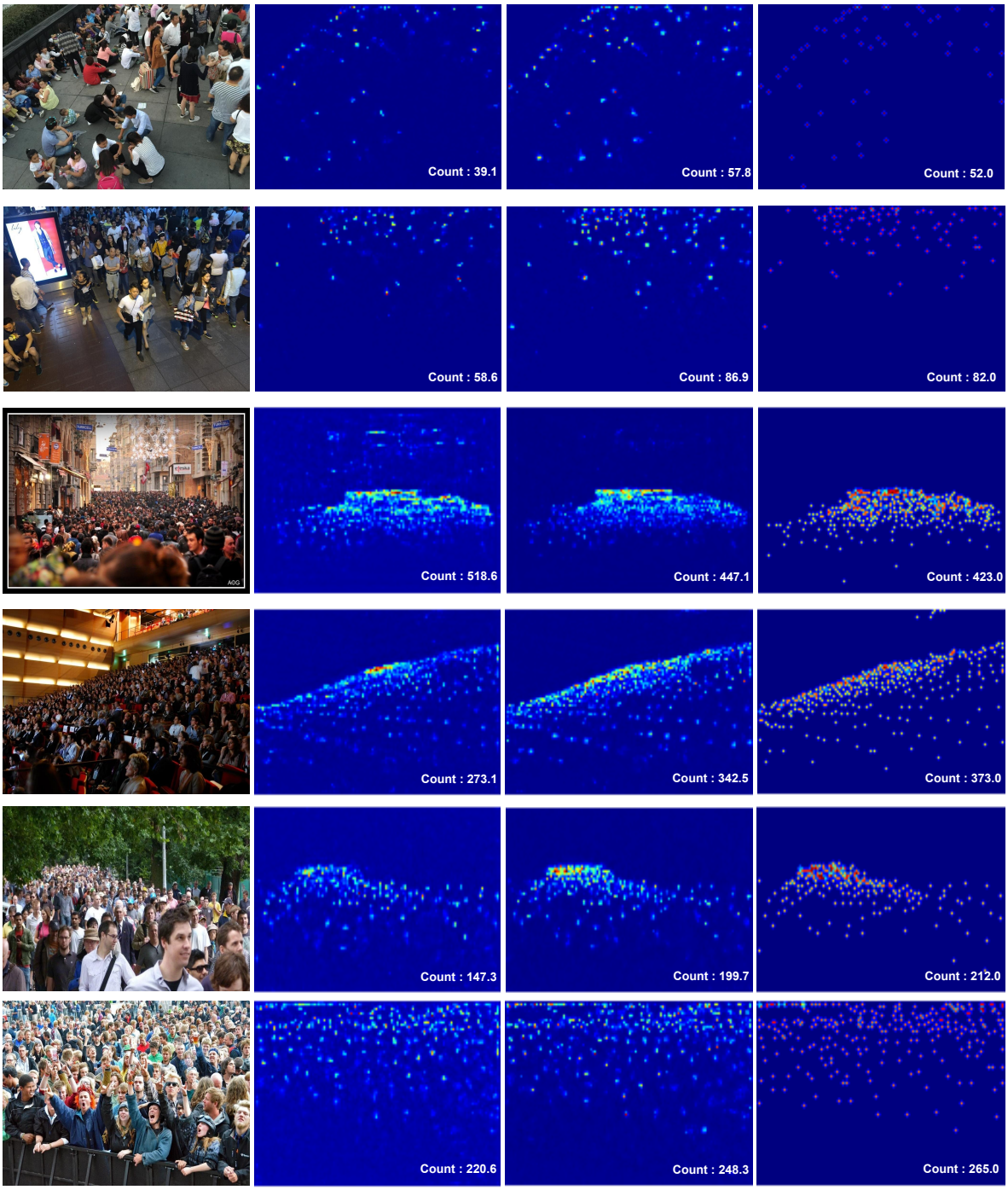


Figure 3. Results of third phase of training PENet. Given (a) the original RGB image, (b) is the corresponding perspective map directly predicted by PENet, and (c) represents perspective map when training end-to-end.

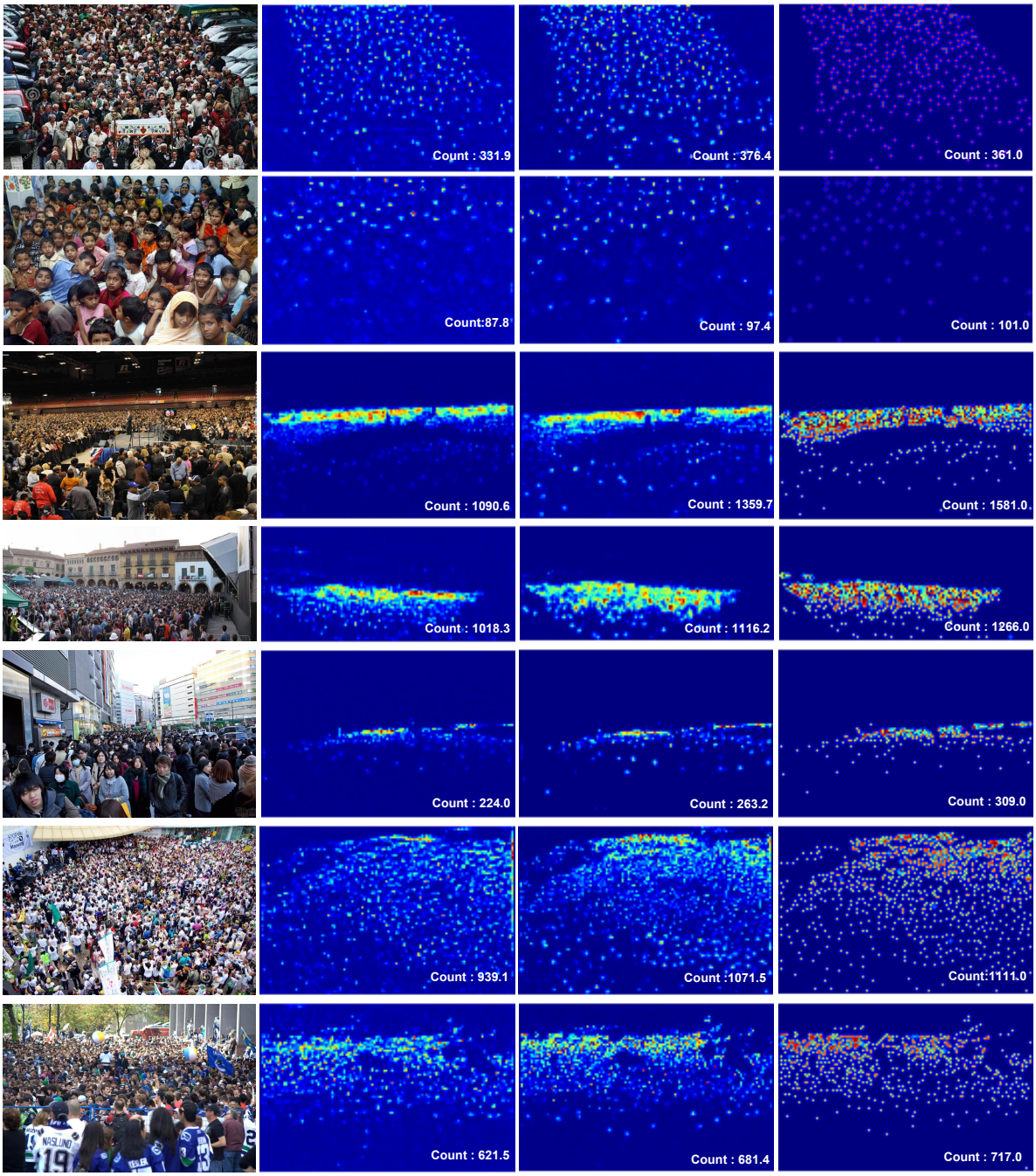
The architecture of PENet
Conv. (3, 3, 64), stride=2; LReLU
Conv. (3, 3, 128), stride=2; LReLU
Conv. (3, 3, 256), stride=2; LReLU
Conv. (3, 3, 512), stride=2; LReLU
UpConv. (3, 3, 256), stride=2; ReLU
UpConv. (3, 3,128), stride=2; ReLU
UpConv. (3, 3, 64), stride=2; ReLU
UpConv. (3, 3, 1), stride=2; ReLU

Table 8. The architecture of PENet. "LReLU" denotes leaky ReLU with the slope of 0.2.



(a) Input (b) CSRNet (c) Ours (d) Ground-truth

Figure 4. Results of density map estimation of CSRNet and our PGCNet.



(a) Input (b) CSRNet (c) Ours (d) Ground-truth

Figure 5. Results of density map estimation of CSRNet and our PGCNet.