Spatial Uncertainty-Aware Semi-Supervised Crowd Counting

Yanda Meng[1], Hongrun Zhang[1], Yitian Zhao[2], Xiaoyun Yang[3], Xuesheng Qian[4], Xiaowei Huang[5], Yalin Zheng[1] ✉
yalin.zheng@liverpool.ac.uk
[1] Department of Eye and Vision Science, University of Liverpool, Liverpool, United Kingdom
[2] Cixi Institute of Biomedical Engineering, Chinese Academy of Sciences, Ningbo, China
[3] Remark AI UK Limited, London, United Kingdom
[4] China Science IntelliCloud Technology Co., Ltd, Shanghai, China
[5] Department of Computer Science, University of Liverpool, Liverpool, United Kingdom

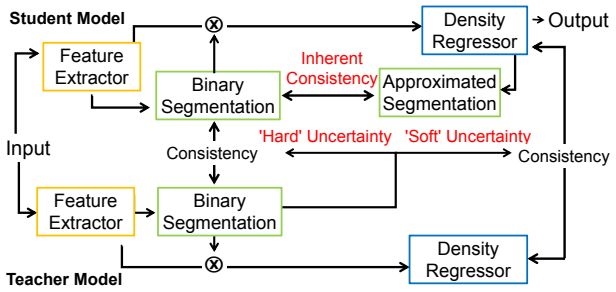

Figure 1. Overview of the proposed model. ⊗ represents the Hadamard Product.

1. Detailed Model Structure

Fig. 1 shows our model's overview, where it contains the student model and teacher model. The teacher model shares the same structure as the student model with Feature Extractor, Binary Segmentation, Density Regressor, except the Approximated Segmentation module. We will elaborate on each of them as follows.

We adopt the first 13 layers of the VGG-16 [9] as the Feature Extractor (backbone), which is the same as [5, 4, 7, 12, 6]. The structure of the student model is shown in Fig. 2. The lower stage of the backbone model retains the high-resolution structure features, and the higher stage features keep rich semantic information. By multi-level aggregating features from the backbone, the binary segmentation and density regressor can obtain affluent semantic and spatial information, which is essential for density map regression and binary segmentation. Previous methods such as [15, 8, 1], have proved that binary segmentation can supply the spatial information into the density regressor branch. We perform Hadamard Product from the binary segmentation's logits to the density regressor's intermediate feature maps.

2. More Results Comparison

Making semi-supervised learning for crowd counting practical is essential. In the manuscript, instead of presenting a relatively impractical performance with 5 %, 10 % or 20 % labeled training data, we report the crowd counting results with 50 % labeled training data, where our model under a semi-supervised manner achieves comparable performance with previous state-of-the-art methods under fully-supervised manner. However, to make a comprehensive comparison, we compare our model with previous semi-supervised approaches [7, 10], and the Baseline model [12] under a different number of training labeled data settings on two datasets. Because they did not provide crowd counting results on the JHU-Crowd [11] and the NWPU-Crowd [13] datasets, which are the largest two public crowd counting datasets containing challenging scenes. We only present the performance comparison on the ShanghaiTech (SHA, SHB) [14] and the QNRF [2] dataset.

To this end, we follow the same training labeled data setting as [7] (10 % or 20 %) and [10] (5 %) on two datasets, respectively. The results of [7, 10] are retrieved from their published paper, and we re-implement the Baseline model [12] through their public code. Note that, [7] adopts the same backbone (VGG-16 [9]) as our model; they build their model based on CSRNet [4], which achieves a comparable performance under fully-supervised manner with ours (*i.e. Ours (Fully)* in the Tab.1 of the manuscript). Furthermore, [10] adopts a more powerful backbone, producing superior performance than *Ours (Fully)* under fully-supervised manner. So the comparison with them in a semi-supervised manner can be seen as straightforward and reasonable.

Specifically, Tab. 1 shows the performance comparison under 10 % of labeled training data for **ShanghaiTech** [14] (SHA and SHB) dataset, and 20 % labeled training data for **QNRF** [2] dataset. Our method achieves 2.0 % and 2.1

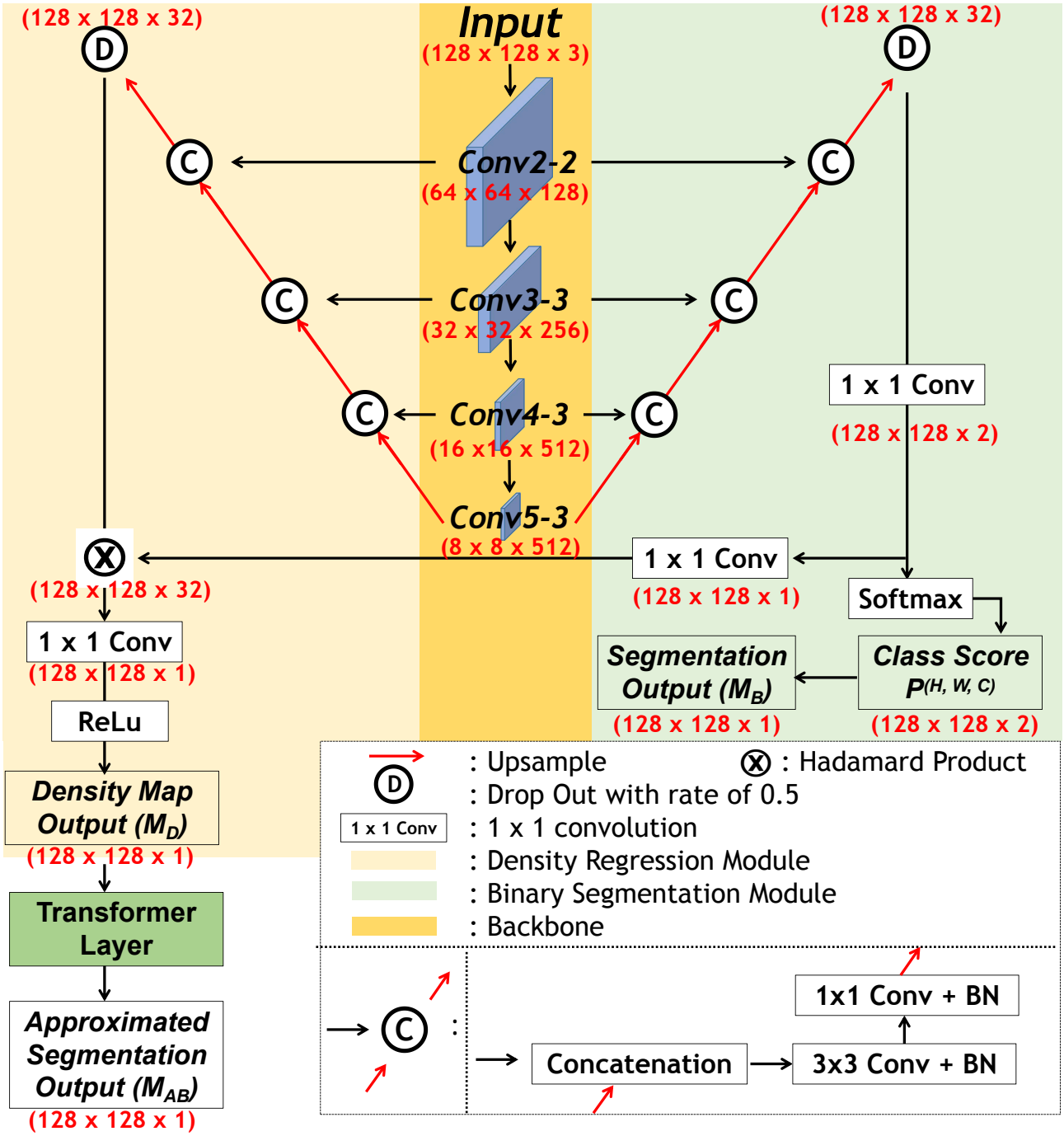

Figure 2. Structure of the student model of the proposed framework. We did not draw the teacher model's structure because it shares the same structure with the Backbone, the Density Regression Module, the Binary Segmentation Module of the student model. The shape of the feature map after each operation is shown in the red bracket as (Height, Width, Channel).

% performance gain in terms of MAE compared with [7] on SHA, QNRF datasets, respectively; we achieve the same MAE but lower RMSE on SHB compared with [7]. Furthermore, Tab 2 demonstrates the results comparison under 5 % labeled training data for two datasets. Our model achieves 10.0 % and 10.1 % lower MAE compared with [10] on SHA, QNRF datasets, respectively; we achieve comparable MAE but lower RMSE on SHB compared with [10].

Methods	SHA		SHB		QNRF	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
Mean-Teacher [12] (Baseline)	100.10	160.7	25.9	41.8	168.7	283.0
Liu *et al.* [7]	86.9	148.9	**14.7**	22.9	135.6	233.4
Ours	**85.1**	**145.2**	**14.7**	**22.5**	**132.8**	**229.3**

Table 1. Results comparison under 10 % of labeled training data on SHA and SHB; 20% of labeled training data on QNRF. Performance is reported with Mean Absolute Error (MAE) and Root Mean Squared Error $(RMSE)$. Our model achieves consistent superior performance over the Baseline [12] and [7] on two datasets in terms of RMSE.

Methods	SHA		SHB		QNRF	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
Mean-Teacher [12] (Baseline)	104.9	167.5	28.1	44.0	171.9	288.8
Sindagi *et al.* [10]	102.0	172.0	15.7	27.9	160.0	275.0
Ours	**91.8**	**148.1**	**15.6**	**25.9**	**143.9**	**253.5**

Table 2. Results comparison under 5 % of labeled training data on SHA, SHB, and QNRF. Performance is reported with Mean Absolute Error (MAE) and Root Mean Squared Error $(RMSE)$. Note that, [10] adopts a more powerful backbone than ours; however, our model still achieves consistent superior performance over the [10] on SHA and QNRF and comparable performance in terms of MAE and RMSE on SHB.

3. More Ablation Studies

We perform extensive experiments to demonstrate that our model is robust to hyper-parameter settings, such as the coefficients of the loss function, *threshold* of 'hard' uncertainty map and weights of 'soft' uncertainty map.

Ablation on Transformation Layer: We perform several experiments to analyse the impact of the proposed transformation layer (Trans). In detail, we remove the transformation layer and inherent consistency loss ($L_{c'}$), and keep the rest components in our model. Then we employ the transformation layer, and $L_{c'}$ upon (1) labeled data only, (2) unlabeled data only, (3) both of the labeled and unlabeled data to demonstrate the performance gain. Tab.3 shows that applying transformation layer only on the unlabeled data can gain close results to ours, and only applying transformation layer on the labeled data results in average 4.4% performance decline via MAE on two datasets compared with ours. The above proves that the performance gain of our model in terms of the proposed differential transformation layer is mainly from the unlabeled data.

Ablation on the coefficients of loss function We perform extensive experiments to demonstrate that our model is robust to hyper-parameter settings, such as the loss function's coefficients, *threshold* of 'hard' uncertainty map and weights of 'soft' uncertainty map.

Ablation on the coefficients of loss function We evaluate the counting performance with different values of the coefficient α, which is used to balance between the main task (density regression) and surrogate task (binary segmentation) in the loss function. In detail, we study the effect of the value of α from 0.001 to 1. Tab.4 shows that our model is robust to this hyper-parameter and achieves the best counting performance in terms of mean absolute error (MAE) and root mean square error (RMSE) on the two datasets when $\alpha = 0.1$.

Methods	SHA		JHU-Crowd	
	MAE	RMSE	MAE	RMSE
w/o Trans	74.8	131.0	89.3	301.2
w/ Trans on Label	73.2	129.5	86.8	296.3
w/ Trans on unlabeled	70.7	123.9	82.1	293.4
w/ Trans on both (ours)	**68.5**	**121.9**	**80.7**	**290.8**

Table 3. Ablation study on the impact of the proposed differential transformation layer. When applying the transformation layer on both the unlabeled and labeled data, ours achieves average 9.1% performance gain than the model without transformation layer via MAE on two datasets.

Ablation on the *threshold* of 'hard' uncertainty map We adopt a time-dependent Gaussian ramp-up paradigm [3] to ramp up the *threshold* of the 'hard' uncertainty U_s from an initial value of 0 to the maximum uncertainty value U_{max} (*i.e.* $\ln 2$ in our work), along with the training process. We conduct several experiments to select the value of U_s on SHA and JHU-crowd datasets. Note that, the initial $threshold$ value U_s cannot be too large (*i.e.* U_{max}) or too small (*i.e.* 0), because the 'hard' uncertainty mechanism hardly filters out uncertain regions or filters out most relatively certain regions during the beginning training period. This will result in the counting performance reduction as the 'hard' uncertainty mechanism barely works, as illustrated in Tab.5. Our model achieves the best counting performance in terms of MAE and RMSE on the two datasets when U_s is equal to 3/4 U_{max}.

Ablation on the weights of 'soft' uncertainty map We perform several experiments to evaluate the counting performance with different weight values M of the 'soft' uncertainty maps. In detail, we change the value of M from 3 to 11 and keep the rest of the components the same in

Methods	SHA		JHU-Crowd	
	MAE	RMSE	MAE	RMSE
α = 0.001	70.3	123.6	83.1	292.5
α = 0.01	69.2	123.1	81.9	292.0
α = 0.5	68.8	122.0	81.1	291.7
α = 1	69.5	122.7	82.3	290.8
α = 0.1 (ours)	**68.5**	**121.9**	**80.7**	**290.8**

Table 4. Performance comparison of different coefficients α in the loss function. The proposed model is robust to this hyper-parameter as the counting performance is very consistent for each of the two datasets.

Methods	SHA		JHU-Crowd	
	MAE	RMSE	MAE	RMSE
U_s = 0	73.0	130.1	85.1	295.6
U_s= 1/4 U_{max}	70.5	125.1	83.7	295.9
U_s = 1/3 U_{max}	69.1	122.3	81.4	291.5
U_s = 1/2 U_{max}	68.8	122.8	82.1	291.7
U_s = 2/3 U_{max}	68.7	122.0	81.2	291.6
U_s = U_{max}	72.8	129.7	84.7	295.4
U_s = 3/4 U_{max} (ours)	**68.5**	**121.9**	**80.7**	**290.8**

Table 5. Ablation study on the effect of the *threshold* in the 'hard' uncertainty map. It shows that the proposed model can achieve comparable counting performance when U_s ranges from 1/3 U_{max} to 3/4 U_{max}. This proves that the proposed model is robust to U_s.

our framework. Tab.6 shows that our model is robust to this hyper-parameter and achieves the best counting performance via MAE and RMSE on the two datasets when $M = 7$.

Methods	SHA		JHU-Crowd	
	MAE	RMSE	MAE	RMSE
M = 3	69.8	122.9	81.8	291.7
M = 5	69.1	122.3	81.0	291.1
M = 9	68.8	122.0	80.7	290.9
M = 11	70.2	123.5	82.3	292.2
M = 7 (ours)	**68.5**	**121.9**	**80.7**	**290.8**

Table 6. The effect of the hyper-parameter weights M on the 'soft' uncertainty map. Our model achieves consistent counting performance with M ranging from 3 to 11 on the two datasets via MAE and RMSE. The best counting performance is achieved with $M = 7$.

4. More Qualitative Results

We present more qualitative results of the binary segmentation predictions (Seg), approximated binary segmentation maps, density map predictions, 'hard' and 'soft' uncertainty maps. Fig.3 shows that our model can produce accurate counting performance compared with the ground truth. The corresponding consistent segmentation predictions and the approximated segmentation predictions demonstrate the effectiveness of the proposed differential transformation layer. The 'hard' and 'soft' uncertainty maps indicate reasonable uncertain regions (*i.e. crowd boundaries*) because the complex backgrounds make it hard for the model to distinguish the crowd boundaries.

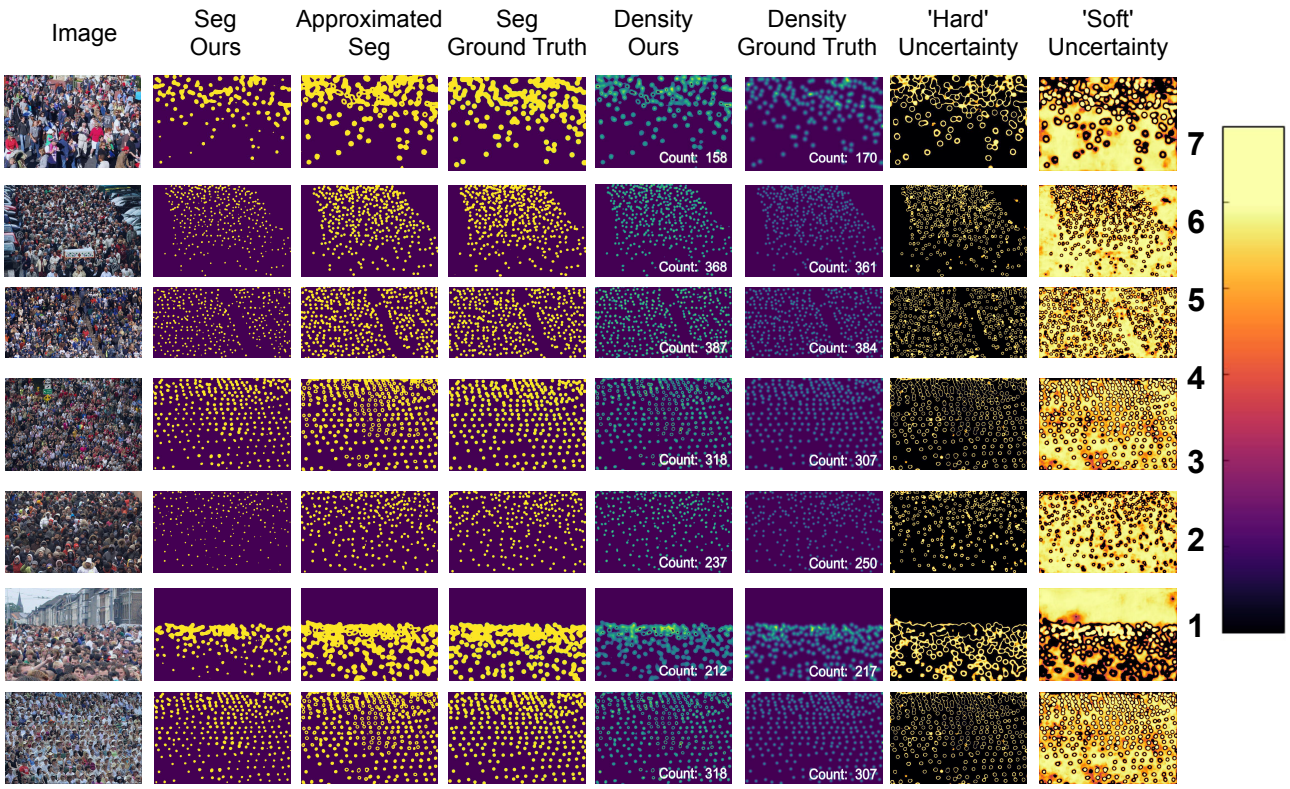

Figure 3. Qualitative results of the binary segmentation predictions, approximated binary segmentation maps, density map predictions, 'hard' and 'soft' uncertainty maps. In the 'hard' uncertainty maps, the yellow pixels represent uncertain regions, and the black pixels are certain regions. In the 'soft' uncertainty maps, the color map shows the confidence of the density predictions in different regions of the image, where the value of 1 represents low confidence while 7 is high confidence.

References

[1] Junyu Gao, Qi Wang, and Xuelong Li. Pcc net: Perspective crowd counting via spatial convolutional network. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019. 1

[2] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 532–546, 2018. 1

[3] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 3

[4] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1091–1100, 2018. 1

[5] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5099–5108, 2019. 1

[6] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Leveraging unlabeled data for crowd counting by learning to rank. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7661–7669, 2018. 1

[7] Yan Liu, Lingqiao Liu, Peng Wang, Pingping Zhang, and Yinjie Lei. Semi-supervised crowd counting via self-training on surrogate tasks. *European Conference on Computer Vision*, 2020. 1, 2, 3

[8] Zenglin Shi, Pascal Mettes, and Cees GM Snoek. Counting with focus for free. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4200–4209, 2019. 1

[9] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 1

[10] Vishwanath A Sindagi, Rajeev Yasarla, Deepak Sam Babu, R Venkatesh Babu, and Vishal M Patel. Learning to count in the crowd from limited labeled data. *European Conference on Computer Vision*, 2020. 1, 2, 3

[11] Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1221–1231, 2019. 1

[12] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural*

information processing systems, pages 1195–1204, 2017. 1, 3

[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] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 589–597, 2016. 1

[15] Muming Zhao, Jian Zhang, Chongyang Zhang, and Wenjun Zhang. Leveraging heterogeneous auxiliary tasks to assist crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12736–12745, 2019. 1