Bipartite Matching for Crowd Counting with Point Supervision

Hao Liu[1,2,3], **Qiang Zhao**[1], **Yike Ma**[1] and **Feng Dai**[1*]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences,
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Artificial Intelligence on Electric Power System Joint Laboratory of SGCC,
Global Energy Interconnection Research Institute Co., Ltd., Beijing, China
{liuhao2018, zhaoqiang, ykma, fdai}@ict.ac.cn

Abstract

For crowd counting task, it has been demonstrated that imposing Gaussians to point annotations hurts generalization performance. Several methods attempt to utilize point annotations as supervision directly. And they have made significant improvement compared with density-map based methods. However, these point based methods ignore the inevitable annotation noises and still suffer from low robustness to noisy annotations. To address the problem, we propose a bipartite matching based method for crowd counting with only point supervision (BM-Count). In BM-Count, we select a subset of most similar pixels from the predicted density map to match annotated pixels via bipartite matching. Then loss functions can be defined based on the matching pairs to alleviate the bad effect caused by those annotated dots with incorrect positions. Under the noisy annotations, our method reduces MAE and RMSE by 9% and 11.2% respectively. Moreover, we propose a novel ranking distribution learning framework to address the imbalanced distribution problem of head counts, which encodes the head counts as classification distribution in the ranking domain and refines the estimated count map in the continuous domain. Extensive experiments on four datasets show that our method achieves state-of-the-art performance and performs better crowd localization.

1 Introduction

Recently, due to its significance in various applications, crowd counting has became an important research problem in deep learning community. Earlier methods estimate crowd counts via the detection of people, bodies or heads [Rabaud and Belongie, 2006; Li *et al.*, 2008; Ge and Collins, 2009] in the image, which may suffer from heavy occlusions. Therefore, these methods require fine-grained annotations for training and have limited applications. Current methods mainly cast crowd counting as a density map estimation problem [Zhang *et al.*, 2016; Li *et al.*, 2018; Xiong *et al.*, 2019; Liu *et al.*, 2019a; Jiang *et al.*, 2020]. Specifically, a crowd density map is regressed by neural networks, whose values are summed to give the total size of the crowd. This type of methods have achieved excellent improvements in terms of the overall error rate compared with earlier methods.

However, most of the crowd density map estimation networks are not directly trained with point supervision. Instead, they are all supervised by pseudo ground truths, where each annotated point in the original ground truths is turned into a Gaussian blob that represents the spatial extent of each person. Unfortunately, it is a non-trivial task to set the correct width for each Gaussian blob, as the spatial extents of persons are various and these informations are not provided by the dataset. Instead of constraining the value at every pixel in the density map, [Ma *et al.*, 2019] proposed a Bayesian loss function adopting a more reliable supervision on the count expectation at each annotated point. Although the new loss makes substantial improvements over the baseline loss, it also requires Gaussian kernels to construct the likelihood functions for annotated points and may lead to density maps that are very different from the ground truths. [Wang *et al.*, 2020a] proposed to use distribution matching for crowd counting (DM-Count) with only point supervision. In this work, the authors use optimal transport to measure the similarity between the normalized predicted density map and the normalized ground truth density map. This measurement can provide valid gradients for network training even if there are no overlap between source and target distribution. However, to stabilize optimal transport computation, total variation loss is incorporated into the model, which may *reduce* the robustness to noisy annotations and ruin the performance. Furthermore, as the size of the input for optimal transport, i.e. the number of the pixels in density map, is much larger than that of crowd, the calculation of optimal transport loss is also *time consuming*.

In this paper, we propose a new crowd counting framework with only point supervision. Our method casts crowd counting as a bipartite matching problem, which selects a subset of most similar pixels from the predicted density map for the annotated pixels in ground truth. As we do not impose any constraints on the coordinates of predicted values and their supervision points, our method is more *robust* to noisy annotations. Unlike DM-Count, which treats all the pixels of ground truth density map equally, our method only considers

*Corresponding author

the sparse pixels with point annotations. Therefore, the size of the input of our method is much smaller, and our network can be trained more *efficiently*. Based on the matching pairs computed by bipartite matching, we present a new ranking distribution loss, which can deal with the imbalanced distribution problem of head count commonly appeared in public datasets. Extensive experiments show that our method can achieve the best performance compared with previous methods. In summary, the contributions of our work are threefold:

- We propose a new bipartite matching based crowd counting method with only point supervision (BM-Count). Our method is more robust to noisy annotations and can be trained more efficiently.
- We present a new ranking distribution loss for performance improvement, which can be used to deal with the imbalanced distribution of head count in crowd counting problem.
- Our method gives new state-of-the-art performance with light weight backbone on four public crowd counting datasets. Our predicted density maps are interpretable and are more useful in real world applications.

2 Related Work

The existing crowd counting methods can be divided into three categories: traditional counting methods, density-map based methods and point based methods.

2.1 Traditional Methods

Most of the early traditional works focus on detection-based methods [Rabaud and Belongie, 2006; Li *et al.*, 2008; Ge and Collins, 2009] via the detection of people, bodies or heads in the crowd image. However, severe occlusions of highly congested scenes limit the performance of these methods. To overcome the problem, regression-based methods [Victor and Andrew, 2010; Idrees *et al.*, 2013; Viet-Quoc *et al.*, 2015] are proposed to learn a mapping from the extracted feature to the count number directly. But their results are less interpretable and the dot annotations are underutilized.

2.2 Density-Map Based Methods

Recent methods mainly conduct crowd counting via density map estimation, which have surpassed the traditional methods by a large margin. [Zhang *et al.*, 2016; Cao *et al.*, 2018; Jiang *et al.*, 2019; Liu *et al.*, 2019a; Dai *et al.*, 2021] mainly focus on the scale variation problem by enlarging the diversity of receptive field and fusing multi-scale features effectively. Instead, many methods attempt to utilize auxiliary information to boost the performance, such as segmentation and semantic priors [Zhao *et al.*, 2019; Wan *et al.*, 2019], attention [Jiang *et al.*, 2020], perspective map [Yang *et al.*, 2020], context information [Liu *et al.*, 2019b] and adaptive density maps [Wan *et al.*, 2020]. Moreover, several approaches address the density pattern shift problem by learning rich representations covering adequate density levels [Xiong *et al.*, 2019]. However, all of these methods are based on density maps whose quality limits the upper bound of network performance. Because it is not trivial to set suitable Gaussian widths in the process of turning annotated dots into blobs.

2.3 Point Based Methods

Most recently, several methods have made use of annotated points directly to alleviate the noises caused by Gaussian smoothed operation. [Ma *et al.*, 2019] propose a Bayesian loss that adopts a more reliable supervision on the count expectation at each annotated point. But it still requires Gaussian kernels and may predict incorrect density maps because the loss is underdetermined. Hence, [Wang *et al.*, 2020a] makes use of Optimal Transport to perform distribution matching under points supervision. Specifically, it can provide valid gradients to train count networks by measuring the similarity between sparse density map and ground truth. However, the time complexity of solving Optimal Transport problem is still not efficient and robust enough. And the method ignores the natural ranking of head counts per pixel that provides richer information.

3 Our Method: BM-Count

In this section, we introduce our BM-Count that considers crowd counting as bipartite matching problem. Given predicted density map, BM-Count first selects a subset of pixels from it that are most similar to the pixels of ground truth with point annotations, then it optimizes the predicted density map with the supervision of matched pixel pairs. Besides of generally used counting loss and regression loss, we additionally introduce a ranking distribution loss for further performance improvement. As we do not make any assumption on the architecture of the network, BM-Count can be applied to all existing counting networks and is end-to-end trainable.

3.1 Crowd Counting as Bipartite Matching

To get rid of Gaussian smoothing ground truth annotations, DM-Count uses optimal transport loss to measure the similarity between the prediction and the ground truth. To increase the stabilization, DM-Count further incorporates a total variation loss. However, for each pixel value of prediction to be optimized, the total variation loss assumes that its supervision has the same coordinates on the ground truth. This makes DM-Count fragile to noises, which is inevitable in crowd annotation. Actually, there is also the same problem for methods supervised with Gaussian smoothed ground truth. However, as each annotated point is turned into a Gaussian blob, the problem is not obvious. To train our network with only point supervision, we select a subset of pixels from predicted density map to make them most compatible with ground truth annotations. As no constraints are imposed on the coordinates of predicted values and their supervision, our method is more robust to noisy annotations.

Given a predictive density map and its corresponding ground truth, let $\mathcal{P} = \{p_i\}_{i=1}^{n}$ denote the flattened points of prediction, where n is the number of pixels. Let $\mathcal{G} = \{g_i\}_{i=1}^{v}$ denote the annotated points on ground truth, where v is the number of point annotations. As the annotations in crowd counting dataset are very sparse, v is much smaller than n. Our goal is to find an optimal subset from $\mathcal{P}$ such that this subset is most compatible with $\mathcal{G}$. To be specific, we try to

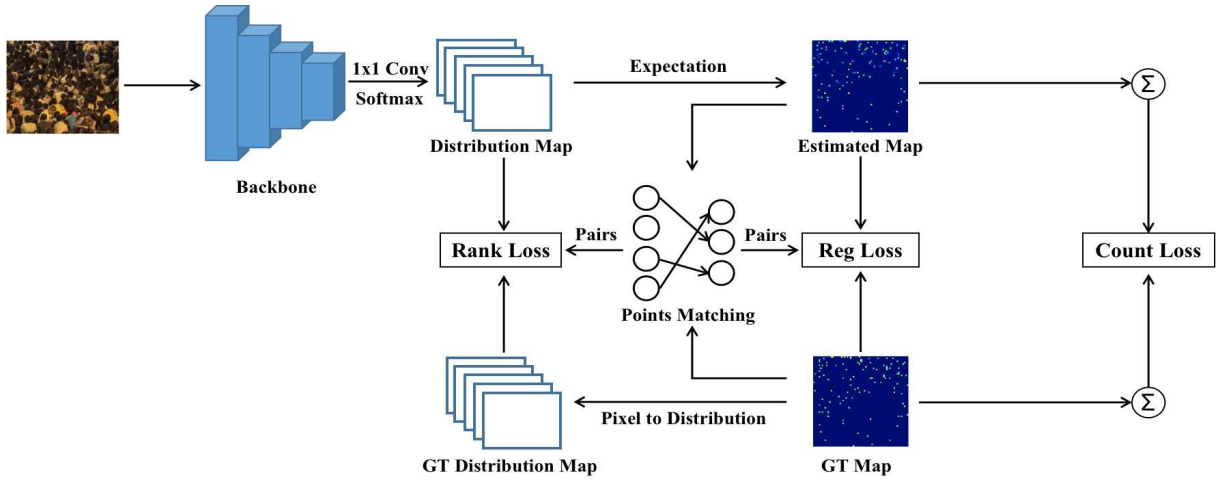

Figure 1: The illustration of the proposed BM-Count, where a novel ranking distribution learning framework uses the results of bipartite matching between estimated map and GT map to optimize Ranking Distribution Loss, Regression Loss and Count Loss.

find a set $\mathcal{S} = \{s_i\}_{i=1}^{v}$ of size v, such that

$$\mathcal{S} = \underset{\hat{\mathcal{S}} \subset \{1,\cdots,n\}}{\arg\min} \sum_{i}^{v} Cost(p_{s_i}, g_i), \tag{1}$$

where $Cost(p_{s_i}, g_j)$ is the matching cost between estimated point p_{s_i} and ground truth point g_i. The matching cost is defined as

$$Cost(p_{s_i}, g_j) = ||\hat{c}_{s_i} - c_j||_1 + \lambda||\hat{\mathrm{x}}_{s_i} - \mathrm{x}_j||_2 \tag{2}$$

where $\hat{c}_{s_i}$ and c_j is the count number for estimated point p_{s_i} and ground truth point g_j respectively, $\hat{\mathrm{x}}_{s_i}$ and x_j are the coordinates of p_{s_i} and g_j respectively, and λ is the weight factor to balance the importance of these two items. In the formula, the former item measures the value difference between p_{s_i} and g_i, while the latter item computes the spatial distance between the points.

As the two points sets $\mathcal{P}$ and $\mathcal{G}$ satisfy $\mathcal{P} \cap \mathcal{G} = \emptyset$, the above minimization problem can be formulated as a bipartite matching problem and can be solved using Hungarian algorithm. Then we can get the matched pixel pairs between predictive density map and ground truth $\mathcal{M} = \{(p_{s_i}, g_i)|i = 1, \cdots, v\}$. The unmatched points in the predictive density map are denoted as $\mathcal{U}$.

Discussion. The time complexity of Sinkhorn algorithm used to compute the optimal transport loss in DM-Count is $O(n^2 \log n/\epsilon^2)$, where ϵ is the desired optimality gap. The time complexity of Hungarian algorithm used in our method is $O(n^2 v)$. Although it seems to be that the former one is faster, the size of the input of our method is much smaller than that of DM-Count, i.e. $v \ll \log n/\epsilon^2$, thus our method can be trained more quickly as shown in experimental result section.

3.2 Ranking Distribution Learning

In this section, we propose a novel supervision way to substitute for original regression form for crowd counting. As shown in Figure 2, the original supervision only comes from the regression loss and count loss, which models the crowd counting as a regression problem. Unlike previous methods, our proposed supervision way models the task as a distribution learning problem to address the imbalanced distribution problem of head counts by using ranking relation of head counts. And the overall loss consists of ranking distribution loss, regression loss and count loss.

Ranking Distribution Loss. As discussed above, previous methods make use of 1×1 convolution layer to regress continuous values for each pixel in estimated map. DM-Count [Wang *et al.*, 2020a] also estimates the count map under points supervision in the same way. Because the ground truth maps are reshaped to $1/8$ of input size. So the head counts of pixels are summed values of non-overlapped local patches in original ground truth maps, not only **1**s and **0**s. However, according to our statistics, there exists severe imbalanced distribution of head counts among all datasets. [Chen *et al.*, 2013] has demonstrated that conventional regression models are difficult to estimate accurately under such distribution.

Notably, head counts are strongly correlated and neighboring values have closer similarities than those further apart, e.g. the count number of 2 is more similar to that of 3 than that of 6. To leverage the observation, we encode the implicit ranking relation into ranking distribution. It reduces the bad effect of imbalanced distribution because additional supervision can be provided implicitly based on the plenty of neighboring head counts. Moreover, compared with one-hot labels for classification, the proposed ranking distribution not only keeps the characteristic of classification but also provides richer ranking information.

For a specific pixel in the ground truth map, its count number c is encoded as a distribution Z of $K+1$ dimension, with its element z_j expressed as follows:

$$z_j = \frac{e^{-(j-y)^2}}{\sum_{i=0}^{K} e^{-(i-y)^2}} \tag{3}$$

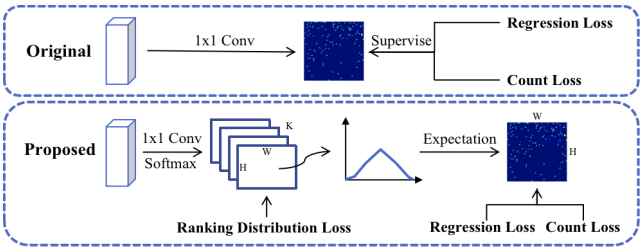

Figure 2: Comparisons of learning framework for crowd counting between existing works (top) and ours (bottom).

where K denotes the maximum count number of specific dataset. According to our statistics, most of popular datasets almost share the same number. Hence, the element z_j denotes the probability that count number is j. By doing so, the probability distribution is formulated by the Euclidean distance of its inter-class and numbers closer to the ground truth number have higher probabilities, as shown in Figure 3.

To measure the difference between prediction and target distribution, we use Kullback-Leibler (KL) divergence. So the proposed ranking distribution loss is defined as follows:

$$L_{rank} = \frac{1}{B}\sum_{i=1}^{B}\Big(\sum_{(p,g)\in\mathcal{M}} KL(\hat{z}_p, z_g) + \sum_{p\in\mathcal{U}} KL(\hat{z}_p, z_0)\Big) \tag{4}$$

where B is the number of images in the batch, $\hat{z}_p$ is the estimated distributions, z_g and z_0 are the ground truth distributions with and without head counts respectively.

Regression Loss. By computing the expectation of predicted distribution for each pixel, the estimated count map can be acquired in the ranking domain. Apart from this, we also leverage the original regression loss to further refine the estimated count map in the continuous domain, which is expressed as follows:

$$L_{reg} = \frac{1}{B}\sum_{i=1}^{B}\Big(\sum_{(p,g)\in\mathcal{M}} |\hat{c}_p - c_g| + \sum_{p\in\mathcal{U}} |\hat{c}_p - 0|\Big) \tag{5}$$

where $\hat{c}_p$ and c_g are the estimated and ground truth numbers respectively.

Count Loss. The goal of crowd counting is to make the overall estimated count as close as possible to ground truth number. So it is essential to utilize the count loss to optimize network auxiliary, and the count loss is defined as the absolute difference between them:

$$L_{count} = \frac{1}{B}\sum_{i=1}^{B}|\hat{C}_i - C_i| \tag{6}$$

where $\hat{C}_i$ and C_i denotes the overall estimated and ground truth counts for i_{th} image respectively.

Overall Loss. By weighting the above three loss functions, the counting network is trained using the following objective function:

$$L = L_{count} + \alpha L_{rank} + \beta L_{reg} \tag{7}$$

where α and β are the weights to balance the ranking distribution loss, regression loss and count loss.

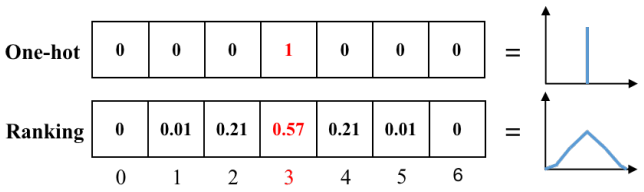

Figure 3: Example for ranking distribution with $c = 3$ and $K = 6$.

4 Experiments

In this section, we first describe the details of experiment settings. Then we compare our proposed method with recent state-of-the-art methods on four public challenging datasets. Finally, ablation studies are further conducted to demonstrate the effectiveness of each component of our method.

4.1 Experiment Settings

Learning Settings. For a fair comparison, we use the same network(VGG-19) as [Ma *et al.*, 2019] and [Wang *et al.*, 2020a] that utilize point supervision. Adam optimizer is applied with fixed learning rate at 1e-5 and weight decay of 1e-4. And the network is trained with batch size of 10 following DM-Count on an NVIDIA 2080Ti GPU. We evaluate our method on four crowd counting datasets: UCF-QNRF [Idrees *et al.*, 2018], NWPU [Wang *et al.*, 2020b], ShanghaiTech [Zhang *et al.*, 2016] and JHU-CROWD++ [Sindagi *et al.*, 2020]. Moreover, due to large sizes of images in UCF-QNRF and NWPU datasets, we limit the shorter size of image within 2048 and 1920 respectively. Also, random crops are taken for training and crop sizes are based on the datasets. Specifically, 256 for ShanghaiTech Part A, 512 for ShanghaiTech Part B and UCF-QNRF, 384 for NWPU and JHU-CROWD++.

Evaluation metrics. Following the existing works [Zhang *et al.*, 2016; Ma *et al.*, 2019; Wang *et al.*, 2020b; Wang *et al.*, 2020a], we adopt the mean absolute error (MAE) and root mean squared error (RMSE) as metrics to evaluate the accuracy of crowd counting estimation, which are defined as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{C}_i - C_i|, RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{C}_i - C_i)^2} \tag{8}$$

where N is the total number of testing images, $\hat{C}_i$ and C_i denotes the overall estimated and ground truth counts for i_{th} image respectively.

4.2 Comparisons with State-of-the-art

Quantitative Results. The experimental results are shown in Table 1. Previous state-of-the-art methods are based on density-map supervision or point supervision. Specifically, density-map based methods pay attention to the design of complex networks by utilizing multi-scale features or auxiliary information to boost performance, leading to more parameters and slow inference. Contrarily, point based methods concentrate on the effective ways of point supervision while ignore the network architectures.

Method	Reference	UCF-QNRF MAE	UCF-QNRF RMSE	NWPU MAE	NWPU RMSE	Shanghai A MAE	Shanghai A RMSE	Shanghai B MAE	Shanghai B RMSE	JHU++ MAE	JHU++ RMSE
density-map:											
CSRNet [Li *et al.*, 2018]	CVPR18	110.6	190.1	121.3	387.8	68.2	115.0	10.6	16.0	85.9	309.2
SANet [Cao *et al.*, 2018]	ECCV18	-	-	190.6	491.4	67.0	104.5	8.4	13.6	91.1	320.4
CAN [Liu *et al.*, 2019b]	CVPR19	107.0	183.0	106.3	386.5	62.3	100.0	7.8	12.2	100.1	314.0
SFCN [Wang *et al.*, 2019]	CVPR19	102.0	171.0	105.7	424.1	64.8	107.5	7.6	13.0	77.5	297.6
S-DCNet [Xiong *et al.*, 2019]	ICCV19	104.4	176.1	**90.2**	**370.5**	58.3	95.0	**6.7**	10.7	-	-
DSSINet [Liu *et al.*, 2019a]	ICCV19	99.1	159.2	-	-	60.6	96.0	6.8	**10.3**	133.5	416.5
PaDNet [Tian *et al.*, 2020]	TIP20	96.5	170.2	-	-	59.2	98.1	8.1	12.2	-	-
CG-DRC [Sindagi *et al.*, 2020]	PAMI20	95.5	164.3	-	-	60.2	94.0	7.5	12.1	71.0	278.6
KDMG [Wan *et al.*, 2020]	PAMI20	99.5	173	100.5	415.5	63.8	99.2	7.8	12.7	69.7	268.3
RPNet [Yang *et al.*, 2020]	CVPR20	-	-	-	-	61.2	96.9	8.1	11.6	-	-
ASNet [Jiang *et al.*, 2020]	CVPR20	91.6	159.7	-	-	**57.8**	**90.2**	-	-	-	-
AMRNet [Liu *et al.*, 2020]	ECCV20	86.6	152.2	-	-	61.6	98.4	7.0	11.0	-	-
NoisyCC [Wan and Chan, 2020]	NeurIPS20	**85.8**	**150.6**	96.9	534.2	61.9	99.6	7.4	11.3	**67.7**	**258.5**
point:											
Pixel-wise Loss [Ma *et al.*, 2019]	ICCV19	106.8	183.7	-	-	68.6	110.1	8.5	13.9	-	-
Bayesian Loss [Ma *et al.*, 2019]	ICCV19	88.7	154.8	105.4	454.2	62.8	101.8	7.7	12.7	75.0	299.9
DM-Count [Wang *et al.*, 2020a]	NeurIPS20	85.6	148.3	88.4	388.6	59.7	95.7	7.4	11.8	63.9	268.7
BM-Count(ours)	-	**81.2**	**138.6**	**83.4**	**358.4**	**57.3**	**90.7**	**7.3**	**11.4**	**61.5**	**263**

Table 1: Comparisons between state-of-the-art methods based on density-map supervision and point supervision on four datasets.

For point based methods, our proposed method achieves the state-of-the-art performance on all datasets, when used in the same network architecture and training procedure. Specifically, on the UCF-QNRF dataset, BM-Count reduces the MAE and RMSE of the DM-Count from 85.6 to 81.2 and from 148.3 to 138.6 respectively. On the NWPU dataset, our method improves 5.65% in MAE and 7.77% in RMSE. On the ShanghaiTech dataset, proposed method reduces the MAE of Part A and Part B from 59.7 to 57.3 and from 7.4 to 7.3 respectively. Moreover, for the JHU++ dataset, BM-Count reduces the MAE from 63.9 to 61.5 and the RMSE from 268.7 to 263.

Compared with density-map based methods, BM-Count still achieves overall best performance on all datasets without using complex network architectures. Specifically, on the UCF-QNRF dataset, BM-Count improves 5.36% in MAE and 7.97% in RMSE compared with NoisyCC. Notably, on the NWPU dataset, our method improve the performance significantly compared with S-DCNet, 7.54% in MAE and 3.27% in RMSE. Moreover, BM-Count achieves comparable MAE and RMSE on the ShanghaiTech dataset compared with ASNet and S-DCNet. Also, on the JHU++ dataset, proposed method improves 9.15% in MAE significantly and achieves comparable RMSE compared with NoisyCC.

Qualitative Results. As shown in Table 2, our proposed method produces much higher PSNRs and SSIMs compared with DM-Count on all datasets, which demonstrates that our estimated maps are more close to the ground truths. Also, Figure 4 presents the estimated density maps that perform better than DM-Count in both sparse and dense areas.

Dataset	DM-Count PSNR	DM-Count SSIM	BM-Count PSNR	BM-Count SSIM
UCF-QNRF	40.68	0.6966	45.59	0.7093
NWPU	46.39	0.8641	50.37	0.8658
Shanghai A	34.43	0.4463	39.54	0.5123
Shanghai B	44.09	0.8533	47.08	0.8537
JHU++	41.85	0.6951	46.95	0.7034

Table 2: Qualitative results on different datasets.

4.3 Ablation Studies

Hyper-parameter study. The matching cost of bipartite matching introduces the weight factor λ to control the importance between value difference and spatial distance. As can be seen from the curves in Figure 5, varying λ between 1 and 4, the results are robust and comparable, with the best performance at $\lambda = 2$. Also, the proposed three losses utilize α and β to balance these items. And our method acquires the best performance with $\alpha = 2$ and $\beta = 3$. Thus, we adopt these configurations on all the datasets.

Contribution of each component. To validate the effectiveness of our proposed losses, we train the model with five different combinations: 1) Count Loss; 2) Count, Reg and Rank Loss; 3) Count and Reg+ Loss; 4) Count and Rank+ Loss; 5) Count, Reg+ and Rank+ Loss. As shown in Table 3, the proposed Reg+ Loss with matching pairs improves the MAE from 65.3 to 59.8 and RMSE from 106.5 to 96.2, compared with the Count Loss. With Rank+ Loss, it improves the MAE from 65.3 to 58.9 and RMSE from 106.5 to 93.7 when models crowd counting as distribution learning problem, which proves the superiority of the novel learning framework. And the combination of all three losses achieves the best performance, 57.3 in MAE and 90.7 in RMSE. Compared with the combination without bipartite matching, it improves 8.2% in MAE and 9.6% in RMSE significantly.

Training time study. To validate the efficiency of our proposed method, we also conduct comparison experiments of

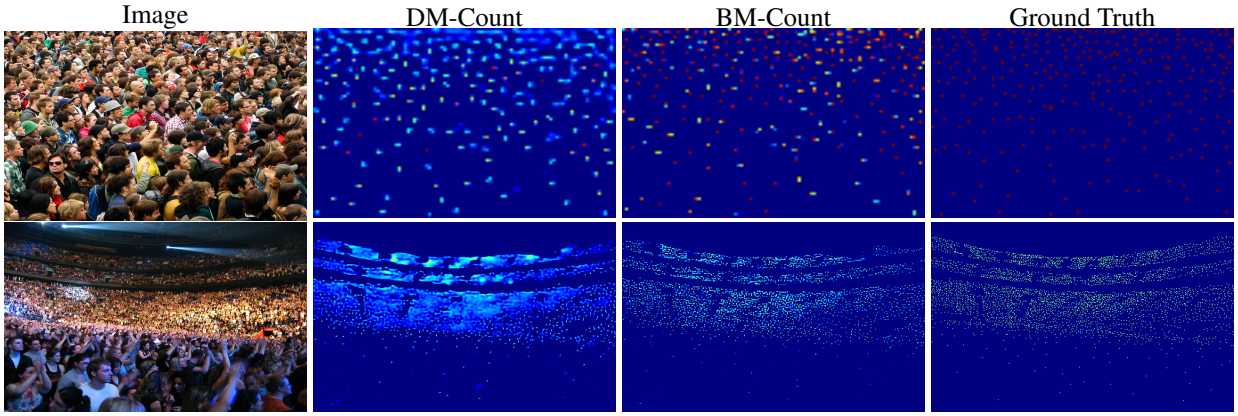

Figure 4: Visualization of density maps. BM-Count localizes people better in both sparse and dense areas compared with DM-Count.

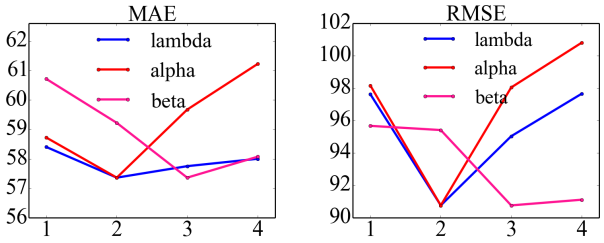

Figure 5: The effect of λ, α and β on ShanghaiTech A dataset

Component			Combinations		
Count Loss	✓	✓	✓	✓	✓
Reg Loss		✓			
Rank Loss		✓			
Reg Loss+			✓		✓
Rank Loss+				✓	✓
MAE	65.3	62.4	59.8	58.9	57.3
RMSE	106.5	100.3	96.2	93.7	90.7

Table 3: Effect of each component on ShanghaiTech A dataset. The symbol + denotes the methods using bipartite matching.

training time between BM-Count and DM-Count. Results are listed in Table 4. As seen in the table, BM-Count reduces the average training time for one epoch by a large margin on all datasets. Specifically, 1.5× on the UCF-QNRF dataset, 2× on the NWPU and JHU++ datasets, 3× on the ShanghaiTech A dataset. The experiment has demonstrated the efficiency of bipartite matching compared with optimal transport.

Robustness to noisy annotations. It is essential to measure the robustness of proposed method based on point supervision to noisy annotation. Because the process of points annotation for crowd is ambiguous and could lead to inevitable noises. Following DM-Count, we also add the same uniform random noise to the original annotation. As shown in Table 5, our method outperforms all previous point based methods and reduces the MAE and RMSE of DM-Count by a large margin, from 105.6 to 96.1 in MAE and from 181.6 to 161.3 in RMSE. Compared with original annotation, DM-Count reduces 23.4% in MAE and 22.5% in RMSE while BM-Count only reduces 18.3% in MAE and 16.4% in RMSE, which demonstrates the stable robustness to noisy annotations.

Method	QNRF	NWPU	JHU++	ShanghaiA
DM-Count	92.2s	179.4s	155.8s	19.0s
BM-Count	60.2s	84.5s	75.5s	6.1s

Table 4: Comparisons of average training time for one epoch on different datasets.

Method	Pixel	Bayesian	DM-Count	BM-Count
MAE	144.1	108.4	105.6	96.1
RMSE	232.5	187.2	181.6	161.3

Table 5: Robustness to noisy annotations on UCF-QNRF dataset.

5 Conclusion

In this paper, we propose a bipartite matching based method called BM-Count for crowd counting to alleviate the effect of inevitable annotation noises. Moreover, a novel ranking distribution learning framework that leverages ranking relation of head counts is proposed to address the imbalanced distribution problem of head counts. Extensive experiments have demonstrated the advantage of our proposed method in terms of accuracy, efficiency and robustness. And the current framework is fairly general and can be easily incorporated with existed networks to further improve performance. However, BM-Count is still not able to locate the position of each person with original resolution. And we will extend our method to crowd localization task in the future.

Acknowledgments

This work was supported by National Natural Science Foundation of China (62072438, 61771458).

References

[Cao *et al.*, 2018] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. Scale aggregation network for accurate and efficient crowd counting. In *ECCV*, pages 757–773, 2018.

[Chen *et al.*, 2013] Ke Chen, Shaogang Gong, Tao Xiang, and Chen Change Loy. Cumulative attribute space for age and crowd density estimation. In *CVPR*, pages 2467–2474, 2013.

[Dai *et al.*, 2021] Feng Dai, Hao Liu, Yike Ma, Xi Zhang, and Qiang Zhao. Dense scale network for crowd counting. In *ICMR*, 2021.

[Ge and Collins, 2009] Weina Ge and Robert T. Collins. Marked point processes for crowd counting. In *CVPR*, pages 2913–2920, 2009.

[Idrees *et al.*, 2013] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *CVPR*, pages 2547–2554, 2013.

[Idrees *et al.*, 2018] 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 *ECCV*, pages 544–559, 2018.

[Jiang *et al.*, 2019] Xiaolong Jiang, Zehao Xiao, Baochang Zhang, Xiantong Zhen, Xianbin Cao, David Doermann, and Ling Shao. Crowd counting and density estimation by trellis encoder-decoder networks. In *CVPR*, pages 6126–6135, 2019.

[Jiang *et al.*, 2020] Xiaoheng Jiang, Li Zhang, Mingliang Xu, Tianzhu Zhang, Pei Lv, Bing Zhou, Xin Yang, and Yanwei Pang. Attention scaling for crowd counting. In *CVPR*, pages 4706–4715, June 2020.

[Li *et al.*, 2008] Min Li, Zhaoxiang Zhang, Kaiqi Huang, and Tieniu Tan. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In *ICPR*, pages 1–4, 2008.

[Li *et al.*, 2018] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *CVPR*, pages 1091–1100, 2018.

[Liu *et al.*, 2019a] Lingbo Liu, Zhilin Qiu, Guanbin Li, Shufan Liu, Wanli Ouyang, and Liang Lin. Crowd counting with deep structured scale integration network. In *ICCV*, 2019.

[Liu *et al.*, 2019b] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In *CVPR*, June 2019.

[Liu *et al.*, 2020] Xiyang Liu, Jie Yang, Tieqiang Wang, and Wenrui Ding. Adaptive mixture regression network with local counting map for crowd counting. In *ECCV*, 2020.

[Ma *et al.*, 2019] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *ICCV*, pages 6142–6151, 2019.

[Rabaud and Belongie, 2006] Vincent Rabaud and Serge Belongie. Counting crowded moving objects. In *CVPR*, pages 705–711, 2006.

[Sindagi *et al.*, 2020] Vishwanath A. Sindagi, Rajeev Yasarla, and Vishal M. Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *TPAMI*, 2020.

[Tian *et al.*, 2020] Yukun Tian, Yiming Lei, Junping Zhang, and James Z. Wang. Padnet: Pan-density crowd counting. *TIP*, 29:2714–2727, 2020.

[Victor and Andrew, 2010] Lempitsky Victor and Zisserman Andrew. Learning to count objects in images. In *NeurIPS*, page 1324–1332, 2010.

[Viet-Quoc *et al.*, 2015] Pham Viet-Quoc, Kozakaya Tatsuo, Yamaguchi Osamu, and Okada Ryuzo. Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation. In *ICCV*, page 3253–3261, 2015.

[Wan and Chan, 2020] Jia Wan and Antoni B. Chan. Modeling noisy annotations for crowd counting. *NeurIPS*, 33, 2020.

[Wan *et al.*, 2019] Jia Wan, Wenhan Luo, Baoyuan Wu, Antoni B. Chan, and Wei Liu. Residual regression with semantic prior for crowd counting. In *CVPR*, pages 4036–4045, 2019.

[Wan *et al.*, 2020] Jia Wan, Qingzhong Wang, and Antoni B. Chan. Kernel-based density map generation for dense object counting. *TPAMI*, 2020.

[Wang *et al.*, 2019] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In *CVPR*, pages 8198–8207, 2019.

[Wang *et al.*, 2020a] Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai. Distribution matching for crowd counting. In *NeurIPS*, 2020.

[Wang *et al.*, 2020b] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *TPAMI*, 2020.

[Xiong *et al.*, 2019] Haipeng Xiong, Hao Lu, Chengxin Liu, Liang Liu, Zhiguo Cao, and Chunhua Shen. From open set to closed set: Counting objects by spatial divide-and-conquer. In *ICCV*, pages 8362–8371, 2019.

[Yang *et al.*, 2020] Yifan Yang, Guorong Li, Zhe Wu, Li Su, Qingming Huang, and Nicu Sebe. Reverse perspective network for perspective-aware object counting. In *CVPR*, pages 4373–4382, 06 2020.

[Zhang *et al.*, 2016] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *CVPR*, pages 589–597, 2016.

[Zhao *et al.*, 2019] Muming Zhao, Jian Zhang, Chongyang Zhang, and Wenjun Zhang. Leveraging heterogeneous auxiliary tasks to assist crowd counting. In *CVPR*, pages 12728–12737, 2019.