Reciprocal Distance Transform Maps for Crowd Counting and People Localization in Dense Crowd

Dingkang Liang¹, Wei Xu², Yingying Zhu¹, Yu Zhou¹
¹Huazhong University of Science and Technology
²Beijing University of Posts and Telecommunications

Abstract

In this paper, we propose a novel map for dense crowd counting and people localization. Most crowd counting methods utilize convolution neural networks (CNN) to regress a density map, achieving significant progress recently. However, these regression-based methods are often unable to provide a precise location for each person, attributed to two crucial reasons: 1) the density map consists of a series of blurry Gaussian blobs, 2) severe overlaps exist in the dense region of the density map. To tackle this issue, we propose a novel Reciprocal Distance Transform (R-DT) map for crowd counting. Compared with the density maps, the R-DT maps accurately describe the people’s location, without overlap between nearby heads in dense regions. We simultaneously implement crowd counting and people localization with a simple network by replacing density maps with R-DT maps. Extensive experiments demonstrate that the proposed method outperforms state-of-the-art localization-based methods in both counting and people localization tasks, achieving very competitive performance compared with the regression-based methods in counting tasks. In addition, the proposed method achieves a good generalization performance under cross-dataset validation, which further verifies the effectiveness of the R-DT map. The code and models are available at [https://github.com/dk-liang/RDTM](https://github.com/dk-liang/RDTM).

1. Introduction

Crowd counting is a popular topic in the computer vision community, playing an essential role in video surveillance, public safety, and city management [34]. The task aims to count the number of people appearing in the image or the video sequence, which is quite difficult due to the heavy occlusions, scale variations, and perspective distortions. Moreover, it is even more challenging to simultaneously predict each person’s location and size in a crowd scene.

The deep-learning-based detectors [11, 28] count the crowd by detecting each instance, encountering difficulties under highly congested scenes, as illustrated in Fig. 1. In general, annotating the bounding box for each person in the dense crowd is expensive and laborious, so most current crowd counting datasets [14, 48] only provide point-level annotations, making the detectors [11, 28] untrainable. Current regression-based methods [17, 32, 43, 48] regard the task as regress a density map and output the count by integrating over the density map, achieving remarkable progress. However, these regression-based methods can not provide individual location and size, mainly because the Gaussian blobs of nearby people in dense regions severely overlap each other, making the local maxima unequal to the individual locations. Whereas, the location and size information also play an essential role in many high-level applications, such as dense pedestrian tracking and crowd analysis. To tackle the problem, PSDDN [24] and LSC-CNN [31] utilize similar nearest-neighbor head distances to initialize the pseudo ground truth bounding boxes in a detection-like model. Essentially, both of them use bounding boxes for the training phase, still applying a complex detection framework (e.g., Faster R-CNN) [24] or cumbersome feature fusion module [31] to enhance the detector performance.

Alternatively, some methods focus on designing an appropriate map to cope with counting and localization, such as trimap [1] and distance label map [42]. However, both
Figure 2: The localization advantages of the R-DT map. (a) Input image, which has heavy occlusions and cluttered backgrounds. (b) The image only provides point-level annotations. (c) A series of Gaussian blobs represent the density map and usually accumulate density values in the dense region, making the person location indistinguishable. (d) R-DT map uses the nearest neighbor distance information to represent each person’s location, and nearby heads remain distinguishable even in dense regions.

trimap [1] and distance label map [42] need to set the threshold of distance and number of the label in handcraft, which is empirical. Trimap is divided into positive, negative, and ignore regions by thresholding. The distance label map is obtained by dividing the distance transform map into eleven classes. Besides, trimap and distance label maps are regarded as semantic segmentation task, which is difficult to replace the density map in most current regression-based methods directly.

In view of the issues mentioned above, we propose a new map named Reciprocal Distance Transform (R-DT) map, which provides precise location information for each person. It is well known that the widely used density maps show blur and indistinguishable due to each head annotation filtered with a Gaussian kernel, as shown in Fig. 2(c). However, the proposed R-DT map is discriminative without any overlaps between nearby heads, even in extremely dense crowds, and the head locations correspond to local maxima, as shown in Fig. 2(d). We extract the coordinates of the heads (local maxima) from the predicted R-DT map and generate a bounding box for each head by a K-nearest neighbors strategy, although the real head size is not annotated. Note that PSDNN [24] and LSC-CNN [31] utilize the ground truth dot annotations to generate pseudo bounding boxes, which are used to supervise the model in the training phase. We obtain the bounding boxes through the predicted positions, which means our method is more flexible due to the proposed method without any anchor setting and IoU-based non-maxima suppression (NMS) processing.

To verify the effectiveness of the proposed R-DT map, we utilize a simple yet effective pipeline to regress the R-DT map instead of the widely used density map. Extensive experiments on five challenging datasets show that our method achieves superior performance to the localization-based methods, obtaining comparable performance among the density map regression-based methods. Moreover, the proposed method achieves an impressive generalization performance under cross-dataset validation.

2. Related Works

2.1. Localization-based crowd counting.

Recently, deep-learning-based detectors [22, 27, 28] achieve appealing performance. However, these methods rely on bounding box annotations, which means impractical in the dense crowd due to expensive annotation costs. To solve the bounding box annotation cost problem, PSDNN [24] and LSC-CNN [31] propose an impressive mechanism to generate bounding boxes only using point-level annotations. PSDDN [24] utilizes the nearest neighbor distance to generate the initial pseudo boxes, and the ground truth boxes will be updated during the training phase. LSC-CNN [31] locates, sizes each person of the dense crowd, and counts crowd person since the number of predicted bounding boxes is equal to count. Besides, CL [14] finds the local peaks on of predicted density map with a small Gaussian kernel as the center locations of heads.

2.2. Regression-based crowd counting.

The mainstream of crowd counting is regression-based methods since the use of robust CNN features and density maps. Before that, traditional regression-based [3, 4, 13] methods directly regress a global crowd count. The current regression-based methods regress a density map for the crowd, and the count is obtained by integrating the density map. Some methods focus on multi-layer or multi-scale feature fusion [2, 15, 16, 30, 32, 35, 48]. MBTTBF [35] proposes a multilevel bottom-top and top-bottom fusion mechanism to combine features between low-level and high-level layers. ZoomCount [30] proposes a zoom-in strategy and zoom-out strategy, mapping the sparse region and dense region, respectively. L2SM [43] learn a scale factor for the dense region and rescales all dense regions into a similar density level to reduce pattern shifts. In addition, some methods [21, 46, 47] incorporate the attention mechanism into the framework. To be specific, RANet [46] proposes local self-attention (LSA) and global self-attention (GSA) to capture short-range and long-range information, respectively. ADCrowdnet [21] utilizes a binary classification network to obtain the crowd region. Besides the mentioned methods above, using perspective information to diminish
the scale variations is effective [7, 33, 44, 45]. PACNN [33] proposes a novel generating ground truth perspective maps strategy and predicts both the perspective maps and density maps at the test phase. Yang et al. [45] propose a reverse perspective network to estimate the perspective factor of the input image and then warp the image.

2.3. Distance transform.

Distance transform [29] is a classical image processing operator, which is applied in many deep-learning-based algorithms recently [1, 10, 40, 41]. Specifically, Hayder et al. [10] introduce a novel segment representation based on the distance transform. Wang et al. [40] present the Deep Distance Transform (DDT) for accurate tubular structure segmentation. In the counting area, Arteta et al. [1] and Xu et al. [42] propose the trimap and distance label map, both discretizing a distance transform map by setting distance threshold empirically.

3. Our Method

The overview of our method is shown in Fig. 3. It contains a base network, regression branch, and segmentation branch. The initial estimated R-DT map is given by the base network and regression branch, achieving high localization accuracy in the sparse region. However, it may be highly underestimated or overestimated in dense regions due to a mass of tiny head and diversity density. Inspired by recently progressive methods using zoom and refinement mechanism [18, 25, 30, 42, 43], we utilize a segmentation branch to select the dense region, then zoomed uniformly by 2x to generate a high-quality R-DT map. Each functional part of the proposed method is described in the following sections.

3.1. Reciprocal Distance Transform Map.

Here, we illustrate the formulation of the Euclidean distance transform map first, which is defined as:

\[
P(x, y) = \min_{(x', y') \in B} \sqrt{(x-x')^2 + (y-y')^2}
\]

where \(B\) represents the set of all annotations. For an arbitrary pixel \((x, y)\), Eq. (1) means that the pixel value \(P(x, y)\) represents the distance between the pixel and its nearest head position (annotation). It is difficult to directly regress the distance transform map, mainly due to the large distance variations, so we improve it to suit crowd counting by using the Reciprocal function to refrain the distance variations. Specifically, we propose the R-DT map, as defined below:

\[
I = \frac{1}{P(x, y) + C}
\]

where \(I\) is the R-DT map, and \(C\) is an additional constant to avoid being divided by zero, as the range of distance transform map is \([0, +\infty)\). Compared to the widely used density map, the R-DT map can accurately represent the individual locations, which correspond to the local maxima, as shown in Fig. 2(d).

3.2. The proposed network architecture.

To verify the effectiveness of the proposed R-DT maps, we use a straightforward yet effective pipeline to regress the R-DT maps, which consists of the base network, segmentation branch, and regression branch, the overview of the proposed pipeline shown in Fig. 3. The trained VGG-16 is used as the base network, discarding the fully connected
layers and the max-pooling layer between stage4 and stage5 to preserve sufficient spatial information. We demonstrate the architecture details of the regression branch and segmentation branch in supplemental materials because of the limited space.

The base network with the regression branch provides an initial estimated R-DT map and relatively accurate in the sparse region. However, it is difficult to locate persons of the dense region on account of the heads usually close to each other. To solve this problem, we utilize a segmentation branch to divide the image into the sparse region and dense region, as shown in Fig. 3. The coordinates of the dense region can be obtained by the attention map, rescaling the selected dense region for refining count. To simplicity, the chosen dense region is zoomed uniformly by 2x, following ZoomCount [30] and RAZ Net [18]. We utilize the average nearby distance to measure the patches’ density level, following [42]. It is a more intuitive and clear density indicator. Specifically, we firstly generate a position map, while the value of each head position in the position map is equal to the distance from the nearest head. Secondly, we use a $M \times M$ sliding window to scan the ground truth position map to obtain all local distance counts. Thirdly, we calculate the average distance $D$ for each $M \times M$ local patch, where $D$ smaller means the patch denser. Finally, we use a distance threshold $J$ to divide the local patches into two levels: sparse region and dense region.

We train the segmentation branch to produce the dense region attention map based on the Cross-Entropy loss. As shown in Fig. 4, the segmentation branch segments the crowd into the dense region and sparse region. The corresponding dense region image patch will be zoomed 2x to generate the high-quality R-DT map by the parameter-shared base network with the regression branch for better counting local maxima. The final estimated R-DT map consists of an initial map and a refinement map, containing the sparse region and dense region, respectively.

3.3. From R-DT map to bounding boxes.

Although the real individual sizes are not provided, we can generate pseudo individual sizes from the estimated R-DT map. Once we get the estimated R-DT map, we first extract the coordinates of local maxima, which can be implemented efficiently using a $3 \times 3$ max-pooling operation, the details are listed in Algorithm 1, and the location map is shown in Fig. 3. Then, we calculate the instance size, using the K-nearest neighbors distance, which is defined as:

$$s(x,y) = \min \left\{ \bar{d} = f \times \frac{1}{k} \sum_{k=1}^{k} d^k_{(x,y)}, \min (\text{img}_w, \text{img}_h) \times 0.05 \right\}$$

where $s(x,y)$ means the size of the instance, which locate in $(x,y)$, and $P$ is the set of predicted head position. $\bar{d}$ is the average distance, which is calculated between point $P_{(x,y)}$ and its K-nearest neighbors, and a scalar factor $f$ is used to restrain the size. In very sparse regions, the $\bar{d}$ may be bigger than the real size of people, so we choose a threshold related to image size to restrain the object size, as described in Eq. 3.

**Algorithm 1 Extract the local maxima (head positions)**

1. **Input:** Predicted R-DT map, threshold $\lambda$ (set as 100 for all dataset)
2. **Output:** The count of the crowd and the coordinates of the persons
3. function `EXTRACT_POSITIOS(input, threshold $\lambda$)
4.    pos_ind = maxpooling(input, size = (3, 3))
5.    pos_ind = (pos_ind == input)
6.    matrix = pos_ind * input
7.    matrix[matrix < $\lambda$/255.0 * max(matrix)] = 0
8.    matrix[matrix > 0] = 1
9.    count = sum(matrix)
10.   coordinates = nonzero(matrix)
11. return count, coordinates
12. end function

3.4. Loss function.

We use the Euclidean loss to measure the difference between the estimated R-DT map and ground truth, as defined by:

$$L_E = \frac{1}{N} \|E - G\|_2^2$$

where $N$ is the number of pixels, $E$ and $G$ stand for the estimated map and corresponding ground truth map, respectively. The final training objective $L$ is defined as bellow:

$$L = L_E + L_{CE}$$

where $L_E$ and $L_{CE}$ refer to the Euclidean loss and Cross-Entropy loss, respectively.

4. Experiments

4.1. Implementation details.

We augment the training data by horizontal. The additional constant $C$ is set to 1. The sliding window size $M$
Table 1: Comparison of the counting performance on the NWPU-Crowd. In the ranking strategy, the Overall MAE is the primary key. $S0 \sim S4$ respectively indicate five categories according to the different number range: 0, (0, 100], (100, 500], (500, 5000], > 5000.

Table 2: Results on the JHU-CROWD+ dataset (“Val set” and “Test set”). Low, Medium, and High respectively indicate three categories according to the different range: [0, 50], (50, 500), and > 500.

4.2. Evaluation Metrics.

Counting Metrics. We use the Mean Absolute Error (MAE) and Mean Square Error (MSE) as counting metrics, defined as: $MAE = \frac{1}{N} \sum_{i=1}^{N} |P_i - G_i|$, $MSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (P_i - G_i)^2}$, where $N$ is the number of testing images, $P_i$ and $G_i$ are the estimated and ground truth count of the $i$-th image, respectively. Besides, we adopt the Grid Average Mean Absolute Error (GAME) [9] as the evaluation metric for vehicle counting, defined as: $GAME(L) = \frac{1}{L} \sum_{i=1}^{L} (|P_i - G_i|)$, which splits an input image into $4^L$ sub-regions.

Localization Metrics. For the localization task, box-level Precision, Recall, and F1-measure are adopted to evaluate the performance on the NWPU-Crowd dataset, defined by [38]. When the distance between the given predicted point $P_f$ and ground truth point $P_g$ is less than a distance threshold $\sigma$, it means the $P_f$ and $P_g$ are successfully matched. The $\sigma$ is related to the real head size. Specifically, Wang et al. [38] give two threshold $\sigma_r = \min(w, h)/2$ and $\sigma_r = \sqrt{w^2 + h^2}/2$, and the former is a stricter criterion than the latter. Besides, for the ShanghaiTech PartB dataset, Mean Localization Error (MLE) and Average precision (AP) are adopted to evaluate the performance, which is proposed by PSDDN [24] and LSC-CNN [31], respectively. MLE is a distance indicator, computing the distance between the predicted head center location to its ground.
truth, averaged over the test set. A one-to-one matching mechanism is used in MLE, and the absent or spurious detections are given a fixed penalty of 16 pixels. AP defines the distance between the true positive head location and its nearest ground truth smaller than a threshold of 20 pixels. The results of other methods are from [38].

4.3. Dataset.

We evaluate our method on five public challenging datasets, each being elaborated below.

NWPU-Crowd [38] consists of 5,109 images, elaborately annotating 2,133,375 instances. The dataset provides 351 negative samples, testing the robustness of the model. The results are obtained by an online evaluation benchmark website.

JHU-CROWD++ [37] contains 2,722 training images, 500 validation images, and 1600 test images, collected from diverse scenarios. The total number of people in each image ranges from 0 to 25,791, which is a large-scale and challenging dataset.

UCF-QNRF [14] contains 1,535 images and about one million annotations. It has a count range of 49 to 12,865, with an average count of 815.4.

ShanghaiTech PartB [48] contains consists of 400 training images and 316 testing images captured from busy streets in Shanghai, which is a sparse dataset.

TRANCOS [9] contains 1,244 images captured in traffic congestion situations with 46,796 annotations, providing a region of interest (ROI) for each image.

5. Results and Analysis

The comparative experiments include crowd counting and people localization.

5.1. Crowd counting.

Table 1, Table 2, and Table 3 show the quantitative results of our method and state-of-the-art methods.

**Compared with the localization-based methods**, which can provide the position information, our method significantly outperforms the state-of-the-art localization-based method LSC-CNN [31] on the two large-scale and challenging datasets (JHU-Crowd++ and UCF-QNRF).

Specifically, on the test set of JHU-Crowd++, our method achieves 71.9 MAE and 297.2 MSE, which is lower than LSC-CNN [31] by a significant margin of 40.8 MAE and 157.2 MSE. Our method also obtains the best performance on ShanghaiTech PartB, which is a sparse dataset. It indicates that the proposed method can cope with both sparse crowd scenes and dense crowd scenes.

**Compared with the regression-based methods**, which only provide the count without position information, it can be observed that the proposed method outperforms all existing methods on NWPU-Crowd and JHU-Crowd++ while obtaining comparable performance on UCF-QNRF and PartB dataset. To intuitionistically demonstrate the difference between R-DT maps and density maps, we choose the UCF-QNRF dataset for estimated R-DT maps and density maps visualization (the density maps are trained with the same network), as shown in Fig. 5. We can see that the estimated density maps lose the position information and show severe overlaps in dense regions. However, the estimated R-DT maps provide an almost accurate location for each individual, even in the extremely dense scene.

5.2. People localization.

Table 4 and Table 5 compare the localization performance of the proposed method against the localization-based methods, which can provide the position information. The results of other methods are from an online benchmark website, making sure to evaluate the localization performance on QNRF dataset.

We evaluate our method on five public challenging datasets, each being elaborated below.

**NWPU-Crowd** [38] consists of 5,109 images, elaborately annotating 2,133,375 instances. The dataset provides 351 negative samples, testing the robustness of the model. The results are obtained by an online evaluation benchmark website.

**JHU-CROWD++** [37] contains 2,722 training images, 500 validation images, and 1600 test images, collected from diverse scenarios. The total number of people in each image ranges from 0 to 25,791, which is a large-scale and challenging dataset.

**UCF-QNRF** [14] contains 1,535 images and about one million annotations. It has a count range of 49 to 12,865, with an average count of 815.4.

**ShanghaiTech PartB** [48] contains 400 training images and 316 testing images captured from busy streets in Shanghai, which is a sparse dataset.

**TRANCOS** [9] contains 1,244 images captured in traffic congestion situations with 46,796 annotations, providing a region of interest (ROI) for each image.

5. Results and Analysis

The comparative experiments include crowd counting and people localization.

5.1. Crowd counting.

Table 1, Table 2, and Table 3 show the quantitative results of our method and state-of-the-art methods.

**Compared with the localization-based methods**, which can provide the position information, our method significantly outperforms the state-of-the-art localization-based method LSC-CNN [31] on the two large-scale and challenging datasets (JHU-Crowd++ and UCF-QNRF).

Specifically, on the test set of JHU-Crowd++, our method achieves 71.9 MAE and 297.2 MSE, which is lower than LSC-CNN [31] by a significant margin of 40.8 MAE and 157.2 MSE. Our method also obtains the best performance on ShanghaiTech PartB, which is a sparse dataset. It indicates that the proposed method can cope with both sparse crowd scenes and dense crowd scenes.

**Compared with the regression-based methods**, which only provide the count without position information, it can be observed that the proposed method outperforms all existing methods on NWPU-Crowd and JHU-Crowd++ while obtaining comparable performance on UCF-QNRF and PartB dataset. To intuitionistically demonstrate the difference between R-DT maps and density maps, we choose the UCF-QNRF dataset for estimated R-DT maps and density maps visualization (the density maps are trained with the same network), as shown in Fig. 5. We can see that the estimated density maps lose the position information and show severe overlaps in dense regions. However, the estimated R-DT maps provide an almost accurate location for each individual, even in the extremely dense scene.

5.2. People localization.

Table 4 and Table 5 compare the localization performance of the proposed method against the localization-based methods, which can provide the position information. The results of other methods are from an online benchmark website, making sure to evaluate the localization performance on QNRF dataset.

The comparative experiments include crowd counting and people localization.

5.1. Crowd counting.

Table 1, Table 2, and Table 3 show the quantitative results of our method and state-of-the-art methods.

**Compared with the localization-based methods**, which can provide the position information, our method significantly outperforms the state-of-the-art localization-based method LSC-CNN [31] on the two large-scale and challenging datasets (JHU-Crowd++ and UCF-QNRF). Specifically, on the test set of JHU-Crowd++, our method achieves 71.9 MAE and 297.2 MSE, which is lower than LSC-CNN [31] by a significant margin of 40.8 MAE and 157.2 MSE. Our method also obtains the best performance on ShanghaiTech PartB, which is a sparse dataset. It indicates that the proposed method can cope with both sparse crowd scenes and dense crowd scenes.

**Compared with the regression-based methods**, which only provide the count without position information, it can be observed that the proposed method outperforms all existing methods on NWPU-Crowd and JHU-Crowd++ while obtaining comparable performance on UCF-QNRF and PartB dataset. To intuitionistically demonstrate the difference between R-DT maps and density maps, we choose the UCF-QNRF dataset for estimated R-DT maps and density maps visualization (the density maps are trained with the same network), as shown in Fig. 5. We can see that the estimated density maps lose the position information and show severe overlaps in dense regions. However, the estimated R-DT maps provide an almost accurate location for each individual, even in the extremely dense scene.
Ground truth
Estimated R-DT map
Estimated density map
3970 3967 4176
1018 1082 1089

Figure 5: Examples from the UCF-QNRF dataset. From left to right, there are images, ground truth maps, estimated R-DT maps, and estimated density maps (trained with the same network). It can be seen that the heads of the estimated R-DT maps are distinguishable, and the estimated density maps show severe overlaps.

Figure 6: Qualitative visualization of detected person locations by proposed method. We use the proposed K-nearest neighbors strategy to generate bounding boxes (green boxes), compared with LSC-CNN [31].

fairly. The Faster-RCNN [28] and TinyFaces [11] are trained with box-level (this dataset provides bounding-box annotations), and the rest methods are trained with point-level. Specifically, our method achieves 69.9% F1-measure under $\sigma_I$ and 64.1% F1-measure under $\sigma_s$, a significant margin improvement compared with other methods. Additionally, our method produces superior counting results on the NWPU-Crowd dataset. For the ShanghaiTech PartB dataset, containing mass sparse crowd scenes, we follow PSDDN [24] and LSC-CNN [31] to utilize the MLE and AP as the metric. In particular, our method has a 36.7% MLE and a 17.0% AP improvement compared with the second-best, LSC-CNN [31]. We qualitatively evaluate the proposed method by visualizing the bounding boxes on the various crowd scenes in Fig. 6, which gives competitive bounding boxes compared with LSC-CNN [31].

6. Ablation Study

The ablation studies are conducted on the UCF-QNRF dataset. In the following, we mainly study the annotation transform functions and the cross-domain capability.

6.1. Annotation transform function.

We perform a detailed ablation study on annotation transform function to understand the localization effectiveness of the R-DT map. We conduct four annotation transform maps, including Distance Transform (DT) map, Min-Max Normalization Distance Transform (MM-DT) map, L2 Normalization Distance Transform (L2-DT) map, and Reciprocal Distance Transform (R-DT) map, as shown in Fig. 7. During the testing phase, the count of the R-DT map is equal to the number of local maxima, and the count of rest transform maps is equal to the local minima. The results are shown in Tab 6. As expected, the counting error reaches an unacceptable level based on the DT map,
we present the F1-measure-MAE curve and Recall-MAE curve on the UCF-QNRF dataset, a fixed $\sigma$ adopted to evaluate since the dataset does not provide the real head size. For the density map, the smaller the kernel used, the better the localization performance is. The counting performance achieves best when the Gaussian kernel is set as 8, but it has a bad localization performance. As expected, the proposed R-DT map achieves the ideal trade-off between counting and localization.

### 6.2. Cross-dataset evaluation.

We conduct cross-dataset experiments on the NWPU-Crowd, JHU-Crowd++, UCF-QNRF, and ShanghaiTech PartB datasets to explore the transferability of the proposed method. In the Cross-dataset evaluation, models are trained on the source dataset and tested on the target dataset without further fine-tuning. Quantitative results are shown in Tab 7. Our method outperforms the state-of-the-art methods on all datasets, which shows remarkable transferability. This inspires us that can use the proposed method to annotate the unlabeled data for semi-supervised crowd counting tasks due to the excellent localization and transferring ability in the future.

![Figure 7: The different transform functions.](image)

![Figure 8: F1-measure-MAE curve and Recall-MAE curve under a fixed $\sigma$ (set as 20) on the UCF-QNRF dataset.](image)

Table 7: Experimental results on the transferability of different methods under cross-dataset evaluation. * means the localization-based methods.

### 7. Conclusion

In this paper, we present a novel map named R-DT map, which is designed to cope with both crowd counting and people localization. The proposed method simultaneously implement counting, localization, and size, which only rely on the point-level annotations. Extensive experiments demonstrate that the proposed method achieves state-of-the-art performance compared with the localization-based methods on all metrics. We hope that the community would switch from the density map regression to R-DT map regression for more practical.

**References**

[1] Carlos Arteta, Victor Lempitsky, and Andrew Zisserman. Counting in the wild. In *European conference on computer vision*, pages 483–498. Springer, 2016.

[2] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. Scale aggregation network for accurate and efficient crowd count-
ing. In Proceedings of the European Conference on Computer Vision (ECCV), pages 734–750, 2018. 2, 5
[3] Antoni B Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In 2008 IEEE Conference on Computer Vision and Pattern Recognition, pages 1–7. IEEE, 2008. 2
[4] Antoni B Chan and Nuno Vasconcelos. Bayesian poisson regression for crowd counting. In 2009 Proceedings 12th international conference on computer vision, pages 545–551. IEEE, 2009. 2
[5] Junyu Gao, Tao Han, Qi Wang, and Yuan Yuan. Domain-adaptive crowd counting via inter-domain features segregation and gaussian-prior reconstruction. arXiv preprint arXiv:1912.03677, 2019. 6
[6] Junyu Gao, Wei Lin, Bin Zhao, Dong Wang, Chenyu Gao, and Jun Wen. C∗3 framework: An open-source pytorch code for crowd counting. arXiv preprint arXiv:1907.02724, 2019. 5
[7] 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. 3, 5
[8] Junyu Gao, Qi Wang, and Yuan Yuan. Scar: Spatial-/channel-wise attention regression networks for crowd counting. Neurocomputing, 363:1–8, 2019. 5, 8
[9] Ricardo Guerrero-Gómez-Olmedo, Beatriz Torre-Jiménez, Roberto López-Sastre, Saturnino Maldonado-Bascón, and Daniel Onoro-Rubio. Extremely overlapping vehicle counting. In Iberian Conference on Pattern Recognition and Image Analysis, pages 423–431. Springer, 2015. 5, 6
[10] Zeeshan Hayden, Xuming He, and Mathieu Salzmann. Boundary-aware instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5696–5704, 2017. 3
[11] Peiyun Hu and Deva Ramanan. Finding tiny faces. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 951–957, 2017. 1, 6, 7
[12] Yutao Hu, Xiaolong Jiang, Xuhui Liu, Baochang Zhang, Jun-gong Han, Xianbin Cao, and David Doermann. Nas-count: Counting-by-density with neural architecture search. In Proceedings of the European Conference on Computer Vision, 2020. 5
[13] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2547–2554, 2013. 2
[14] 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, 2018. 1, 2, 5, 6
[15] Xiaolong Jiang, Zehao Xiao, Baochang Zhang, Xiantong Chen, Xianbin Cao, David Doermann, and Ling Shao. Crowd counting and density estimation by trellis encoder-decoder networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6133–6142, 2019. 2, 5
[16] Xiaoheng Jiang, Li Zhang, Mingliang Xu, Tianzhu Zhang, Pei Lv, Bing Zhou, Xin Yang, and Yanwei Pang. Attention scaling for crowd counting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4706–4715, 2020. 2, 5
[17] Yuhong Li, Xiaofan Zhang, and Deming Chen. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. In IEEE Conf. Comput. Vis. Pattern Recog., pages 1091–1100, 2018. 1, 5, 8
[18] Chenchen Liu, Xinyu Weng, and Yadong Mu. Recurrent attentive zooming for joint crowd counting and precise localization. In CVPR, pages 1217–1226. IEEE, 2019. 3, 4, 5, 6
[19] Liang Liu, Hao Lu, Hongwei Zou, Haipeng Xiong, Zhiguo Cao, and Chunhua Shen. Weighing counts: Segmental crowd counting by reinforcement learning. In Proceedings of the European Conference on Computer Vision, 2020. 5, 8
[20] Lingbo Liu, Zhilin Qian, Guanbin Li, Shufen Liu, Wanli Ouyang, and Liang Lin. Crowd counting with deep structured scale integration network. In Proceedings of the IEEE International Conference on Computer Vision, pages 1774–1783, 2019. 5
[21] Ning Liu, Yongchao Long, Changqing Zou, Qun Niu, Li Pan, and Hefeng Wu. Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding. In IEEE Conf. Comput. Vis. Pattern Recog., pages 3225–3234, 2019. 2
[22] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In European conference on computer vision, pages 21–37. Springer, 2016. 2
[23] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In IEEE Conf. Comput. Vis. Pattern Recog., pages 5099–5108, 2019. 5
[24] Yuting Liu, Miaojing Shi, Qijun Zhao, and Xiaofang Wang. Point in, box out: Beyond counting persons in crowds. In IEEE Conf. Comput. Vis. Pattern Recog., pages 6469–6478, 2019. 1, 2, 5, 6, 7
[25] Yongtuo Liu, Qiang Wen, Haoxin Chen, Wenxi Liu, Jing Qin, Guoqiang Han, and Shengfeng He. Crowd counting via cross-stage refinement networks. IEEE Transactions on Image Processing, 2020. 3
[26] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In Proceedings of the IEEE International Conference on Computer Vision, pages 6142–6151, 2019. 5, 8
[27] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788, 2016. 2
[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In Adv. Neural Inform. Process. Syst., pages 91–99, 2015. 1, 2, 6, 7
[29] Azriel Rosenfeld and John L Pfaltz. Distance functions on digital pictures. Pattern recognition, 1(1):33–61, 1968. 3
[30] Osman Sajid, Hasan Sajid, Hongcheng Wang, and Guanghui Wang. Zoomcount: A zooming mechanism for crowd counting in static images. IEEE Transactions on Circuits and Sys-
[31] Deepak Babu Sam, Skand Vishwanath Peri, Mukuntha Narayanan Sundararaman, Amogh Kamath, and Venkatesh Babu Radhakrishnan. Locate, size and count: Accurately resolving people in dense crowds via detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 2, 3, 4

[32] Deepak Babu Sam, Shiv Surya, and R Venkatesh Babu. Switching convolutional neural network for crowd counting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, volume 1, page 6, 2017. 1, 2

[33] Miaojing Shi, Zhaohui Yang, Chao Xu, and Qijun Chen. Revisiting perspective information for efficient crowd counting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7279–7288, 2019. 3

[34] Vishwanath A Sindagi and Vishal M Patel. A survey of recent advances in cnn-based single image crowd counting and density estimation. *Pattern Recognition Letters*, 107:3–16, 2018. 1

[35] Vishwanath A Sindagi and Vishal M Patel. Multi-level bottom-top and top-bottom feature fusion for crowd counting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1002–1012, 2019. 2, 5

[36] 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. 5

[37] Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 6

[38] 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. 5, 6

[39] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8198–8207, 2019. 5, 8

[40] Yan Wang, Xu Wei, Fengze Liu, Jieneng Chen, Yuyin Zhou, Wei Shen, Elliot K Fishman, and Alan L Yuille. Deep distance transform for tubular structure segmentation in ct scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3833–3842, 2020. 3

[41] Yukang Wang, Yongchao Xu, Stavros Tsogkas, Xiang Bai, Sven Dickinson, and Kaleem Siddiqi. Deeplflux for skeletons in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5287–5296, 2019. 3

[42] Chenfeng Xu, Dingkang Liang, Yongchao Xu, Song Bai, Wei Zhan, Masayoshi Tomizuka, and Xiang Bai. Autoscale: Learning to scale for crowd counting. *arXiv preprint arXiv:1912.09632*, 2019. 1, 2, 3, 4, 5, 6

[43] Chenfeng Xu, Kai Qiu, Jianlong Fu, Song Bai, Yongchao Xu, and Xiang Bai. Learn to scale: Generating multipolar normalized density map for crowd counting. *Int. Conf. Comput. Vis.*, 2019. 1, 2, 3, 5