

Focal Inverse Distance Transform Maps for Crowd Localization

Dingkang Liang, Wei Xu, Yingying Zhu✉, and Yu Zhou✉

Abstract—In this paper, we focus on the crowd localization task, a crucial topic of crowd analysis. Most regression-based methods utilize convolution neural networks (CNN) to regress a density map, which can not accurately locate the instance in the extremely dense scene, 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 Focal Inverse Distance Transform (FIDT) map for the crowd localization task. Compared with the density maps, the FIDT maps accurately describe the persons' locations without overlapping in dense regions. Based on the FIDT maps, a Local-Maxima-Detection-Strategy (LMDS) is derived to effectively extract the center point for each individual. Furthermore, we introduce an Independent SSIM (I-SSIM) loss to make the model tend to learn the local structural information, better recognizing local maxima. Extensive experiments demonstrate that the proposed method reports state-of-the-art localization performance on six crowd datasets and one vehicle dataset. Additionally, we find that the proposed method shows superior robustness on the negative and extremely dense scenes, which further verifies the effectiveness of the FIDT maps. The code and model will be available at <https://github.com/dk-liang/FIDTM>.

Index Terms—Crowd localization, Crowd counting, Crowd analysis, Distance transform, FIDT map

I. INTRODUCTION

Crowd analysis contains many sub-tasks, such as crowd detection [65], crowd counting [23], [19], and crowd localization [11], [41]. Specifically, the crowd detection task is to detect persons based on bounding boxes, an expensive way of labeling. The crowd counting aims to estimate a density map and give the total count of a crowd scene based on point-level annotations. In this paper, we focus on crowd localization, predicting a point for each person's head only based on point-level annotations, which is a more complex task compared with crowd detection and crowd counting.

The deep-learning-based detectors [38], [13] predict the bounding box for each instance, encountering difficulties under highly congested scenes [51]. In general, annotating the bounding box for each person in the dense crowd is expensive and laborious, so most current crowd datasets [66],

This research was supported by the National Key Research and Development Program of China under Grant No. 2018AAA0100400, the National Natural Science Foundation of China (62176098 and 61703049) and the Natural Science Foundation of Hubei Province of China under Grant 2019CFA022. (✉ Corresponding authors)

Dingkang Liang, Yingying Zhu, Yu Zhou are with Huazhong University of Science and Technology; (email: dkliang@hust.edu.cn, yyzhu@hust.edu.cn, yuzhou@hust.edu.cn)

Wei Xu is with Beijing University of Posts and Telecommunications; (email: xuwei2020@bupt.edu.cn)

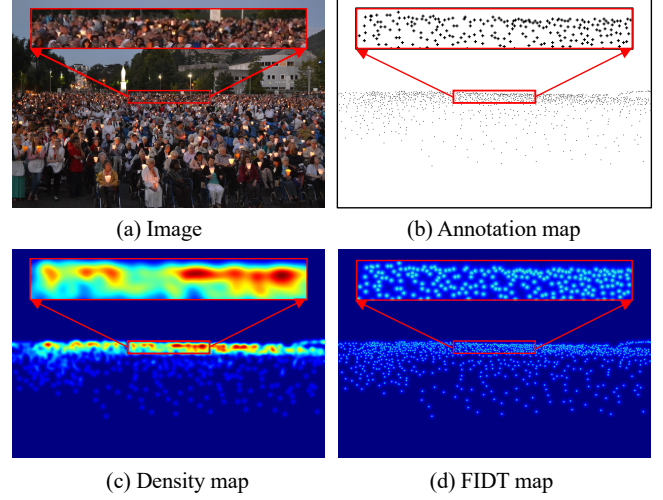


Fig. 1: The localization advantages of the FIDT map. (a) Input image, which has heavy occlusions and cluttered backgrounds. (b) The image only provides point-level annotations. (c) A series of Gaussian blobs represent the density map and usually accumulate density values in the dense region, making the person's location indistinguishable. (d) FIDT map uses the nearest neighbor distance information to represent each person's location, and nearby heads remain distinguishable even in dense regions.

[16] only provide point-level annotations (Fig. 1(b)), making the detectors [38], [13] untrainable. Current regression-based methods [66], [3], [23], [19] regress a density map and output the count by integrating over the density map. However, these regression-based methods can not provide individual location and size, mainly because the Gaussian blobs of the widely used density map overlap in dense regions, making the local maxima unequal to the head locations. However, the location and size information also play an essential role in many high-level applications, such as pedestrian tracking [21] and crowd analysis [22], [64]. To tackle the problem, PSDDN [32] and LSC-CNN [41] utilize similar nearest-neighbor head distances to initialize the pseudo ground truth (GT) 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). Actually, the pseudo GT boxes do not reflect the real head size well, leading to poor performance.

Alternatively, some methods focus on designing an appropriate map to cope with crowd localization, such as binary-

like maps [1], [25], [7] and segmentation-like maps [2], [57]. Among them, both trimap [2] and distance label map [57] need to set the threshold of distance and number of the label in handcraft, which is empirical. Topological maps [1] and IIM [7] still apply box-level annotations in some challenge datasets (*e.g.*, JHU-Crowd++ and NWPU-Crowd), which limits its application to the real world. During the testing phase, these methods regard the connected components of predicted maps as the head location. However, they easily fail in the congested scenes because the adjacent connected components may be linked together in dense regions. In other words, it is possible to incorrectly predict many heads as one.

Different from the above methods, this paper proposes a novel label named Focal Inverse Distance Transform (FIDT) map for the localization task, which provides precise location information for each person. It is well known that the widely used density map is blurry and indistinguishable due to each head annotation filtered with a Gaussian kernel, as shown in Fig. 1(c). In contrast, the proposed FIDT map is discriminative without any overlaps between nearby heads, even in extremely dense crowds, as shown in Fig. 1(d). In the proposed FIDT map, the closer pixels are to the head center, the higher responses they will have, which means the local maxima are equal to the head centers. Accordingly, the counting result is equal to the number of local maxima.

In FIDT maps, a local maximum represents an individual instance, so detailed local structural information can help locate the FIDT maps' local maxima. A straightforward way is to utilize the SSIM loss to improve the similarity between the predicted FIDT map and the ground truth map. However, in the FIDT map, the background's pixel value is close to 0, without structure information. The traditional SSIM loss may cause high responses for the background, which may produce false local maxima. Thus, we introduce the Independent SSIM (I-SSIM) loss to further improve the model's ability to enhance the structure information of local maxima and reduce the false local maxima in background regions.

As we mentioned above, for a given FIDT map, we can obtain the heads' position by localizing the local maxima, so the final key step is how to extract the local maxima of FIDT maps. In this paper, we propose simple yet effective post-processing named Local-Maxima-Detection-Strategy (LMDS), implemented by a simple max-pooling layer with an adaptive threshold. Furthermore, the proposed LMDS can help to classify the negative samples (*e.g.*, Terra-Cotta Warriors images).

In summary, this work contributes to the following:

- 1) In order to effectively cope with the crowd localization task in dense scenes, we propose the FIDT maps. The local maxima of FIDT maps represent exact persons' locations.
- 2) We introduce the I-SSIM loss to make the model focus on the independent regions, enhancing the model's ability to handle the local maxima and background regions.
- 3) Based on the FIDT map, we design a Local-Maxima-Detection-Strategy, LMDS, which can effectively locate the predicted local maxima (head centers).
- 4) Extensive experiments demonstrate that the proposed method achieves state-of-the-art localization perfor-

mance. Additionally, our method is robust for the negative and extremely dense scenes.

II. RELATED WORKS

A. Crowd analysis

Current crowd analysis methods mainly focus on the counting task, which usually adopts CNN to regress the density maps. And the total count is obtained by integrating the density maps. Some methods work on multi-layer or multi-scale feature fusion [43], [17], [18] to improve the quality of predicted density maps. Some methods [62], [63], [28] incorporate the attention mechanism into the framework, which effectively attends to the foreground regions. Multi-head layers [34] are useful that can effectively aggregate features from the conv-backbone. Using different density map representations [42], [48] is also an essential procedure in the training phase, which can promote the model's ability. Some methods make efforts to minimize the expensive labeling work in a semi-supervised [35], [31], [59] or weakly-supervised [24], [61] paradigm. Unfortunately, these counting methods only give the total count or coarse density map, which can not provide the precise position of each head, limiting the application in the real-world.

Recently, crowd localization, aiming to predict the precise position of each person's head, has been a hot topic in crowd analysis. The deep-learning-based detectors [38], [37], [29] rely on bounding box annotations, which is impractical in the dense crowd due to expensive labeling costs. To address this problem, some approaches [32], [41], [53] attempt to initialize the pseudo GT bounding boxes from the point-level annotations, applying the two-stage detection framework. However, the generated pseudo GT bounding boxes do not reflect the actual head sizes well, leading to unsatisfactory performance. CL [16] finds the local maxima of the predicted density map with a small Gaussian kernel. Several methods attempt to predict a location map as a binary-like [1], [25], [7] or segmentation-like map [57], [20]. Specifically, Xu *et al.* [57] propose the distance label map, which formulates the problem as a segmentation task. Shahira *et al.* [1] generate the binary mask by thresholding the topological map. A recent work [7] proposes Independent Instance Maps (IIM), and a differentiable Binarization Module is used to learn adaptive thresholds for different heads. Both topological map [1] and IIM [7] still need box-level annotations when facing challenging datasets (*e.g.*, NWPU-Crowd [51]). In general, these methods usually obtain the head position by detecting the connected components of predicted maps. However, in dense regions, the connected component may be linked together, which is possible to incorrectly predict many heads as one.

Unlike the above localization methods, we propose a new label named FIDT map. This non-overlap map utilizes the local maxima to represent persons' heads, *i.e.*, the closer pixels are to the head center, the higher responses they will have.

B. Loss function

Most crowd counting methods apply MSE as the loss function. However, only using MSE loss will cause blur, and

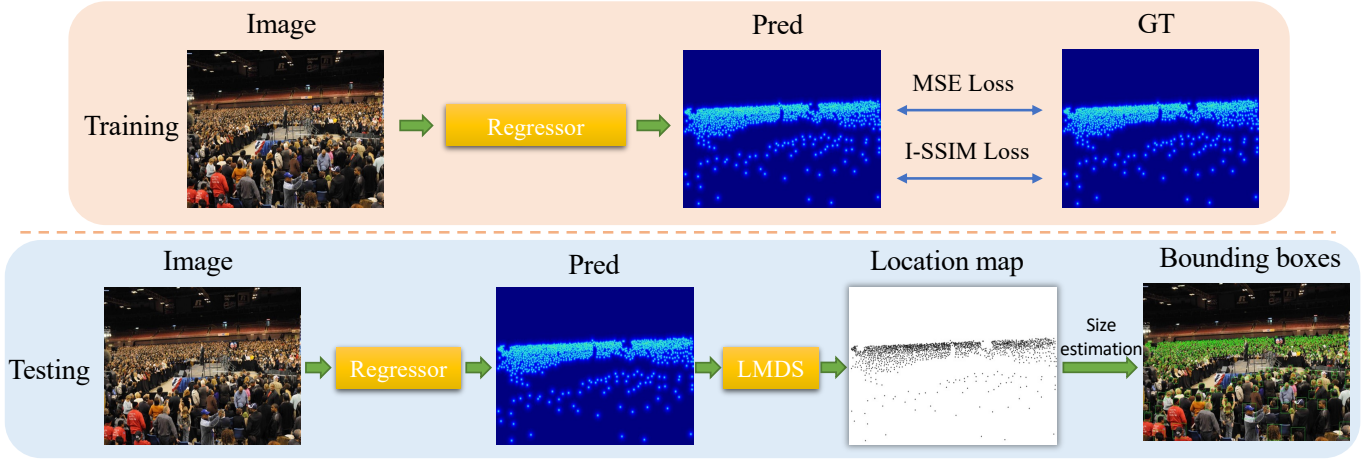


Fig. 2: The pipeline of our method. During the training phase, MSE loss and the proposed I-SSIM loss are adopted. During the testing phase, each person’s location can be obtained by the LMDS, and the final count is equal to the number of local maxima. Additionally, the bounding boxes can be obtained through the size estimation step.

lose the local structure information [4]. To this end, some methods focus on designing an appropriate loss function to promote the model’s ability. Specifically, BL [33] regards the density map as a probability map, calculating the expected count of pixels. SPANet [5] proposes Maximum Excess over Pixels (MEP) loss by finding the pixel-level subregion with the highest discrepancy with ground truth. DM-Count [49] uses Optimal Transport (OT) to measure the similarity between the normalized predicted density map and the normalized ground truth density map. DSSINet [27] propose the DMS-SSIM loss to measure the structural similarity between the multiscale regions centered at the given pixel on an estimated density map and the corresponding regions on ground-truth.

The above loss functions usually calculate the loss on the global level. In contrast, the proposed loss focuses on the structure information of local regions (instance level), which benefits the model better detecting the local maxima (*i.e.*, head centers).

C. Distance transform

Distance transform [40] is a classical image processing operator applied in many deep-learning-based algorithms recently [12], [55], [56], [2]. Specifically, Hayder *et al.* [12] introduce a novel segment representation based on the distance transform. Wang *et al.* [55] present the Deep Distance Transform (DDT) for accurate tubular structure segmentation. In the crowd analysis, Arteta *et al.* [2] and Xu [57] *et al.* propose semantic-like maps, discretizing a distance transform map by setting distance thresholds and transforming the localization task into a semantic task. To the best of our knowledge, we are the first to leverage the local maxima of such distance transform maps for regression-based crowd localization.

III. OUR METHOD

The overview of our method is shown in Fig. 2. At the training stage, a regressor is used to generate the predicted FIDT map. The MSE loss and the proposed I-SSIM loss are

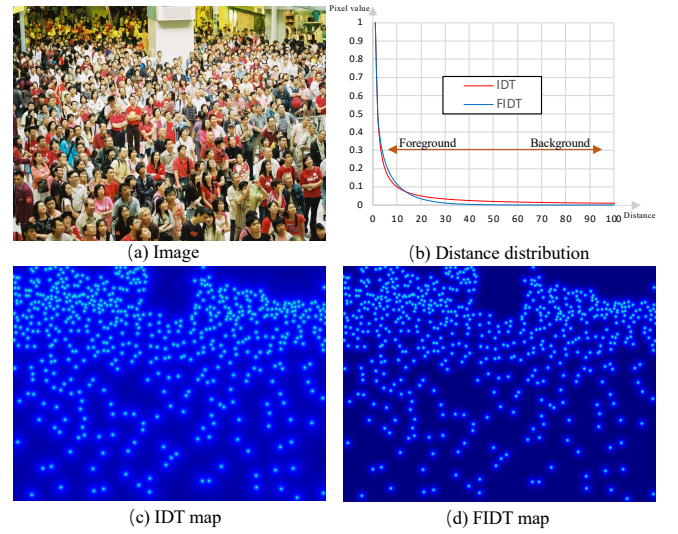


Fig. 3: (a) Original image. (b) The distance distribution of IDT map and FIDT map. (c) IDT map shows faster response decay from the head center and keeps high response in the background. (d) FIDT map shows slower response decay from the head center and keeps low response (close to 0) in the background.

used to measure the difference between prediction results and ground truth. At the testing stage, a predicted FIDT map is generated, and the location map is obtained by the proposed Local-Maxima-Detection-Strategy (LMDS). Furthermore, we can obtain the bounding boxes for better visualization by a simple KNN strategy.

A. Focal Inverse Distance Transform Map

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

$$P(x, y) = \min_{(x', y') \in B} \sqrt{(x - x')^2 + (y - y')^2}, \quad (1)$$

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 (range from 0 to the length of the image). A way is to use the inverse function to refrain from the distance variations. Specifically, the Inverse Distance Transform (IDT) map is generated, defined as:

$$I' = \frac{1}{P(x, y) + C}, \quad (2)$$

where I' is the IDT map, and C is an additional constant (set $C = 1$) to avoid being divided by zero, as the range of distance transform map is $[0, +\infty)$. The IDT map is a special form of the iKNN map [36] (when the K is set as 1). It is noteworthy that [36] only uses the iKNN map for the counting task instead of the localization task. Compared to the widely used density map, the IDT (i1NN [36]) map can accurately represent the individual locations, which correspond to the local maxima. However, the IDT map presents a faster response decay away from the head center and slower response decay in the background, as shown in the distance distribution curve in Fig. 3(b). Ideally, the decay should be slower away from the head, and the response of the background should be quickly close to 0, which means the model should focus on the foregrounds (head regions). Thus, we propose the Focal Inverse Distance Transform (FIDT) map, defined as:

$$I = \frac{1}{P(x, y)^{(\alpha \times P(x, y) + \beta)} + C}, \quad (3)$$

where I is the FIDT map we proposed, α and β set as 0.02 and 0.75, respectively. As shown in Fig. 3(b), the curve examples of the IDT and FIDT map are illustrated. Compared with the IDT map, the FIDT map shows slower response decay away from the head center, and the response of background is close to 0, as shown in Fig. 3(c) and Fig. 3(d). It is noteworthy that the proposed FIDT map is totally different from the density map that uses small Gaussian kernels. The latter still presents overlap in extremely dense scenes, and a recent method [57] has demonstrated that small Gaussian kernels can not report satisfying localization and counting performance.

B. Localization framework

1) *Regressor*: To verify the effectiveness of the proposed FIDT maps, we use a straightforward base network to regress the FIDT maps. The corresponding individual center of the FIDT map is equal to the local maximum, so high-resolution representations are essential. Here, following IIM [7], we use HRNET [50] as the base network, and we add one convolution and two transposed convolution layers as the representation head based on the HRNET [50]. Note that the regressor can be replaced by any crowd regressor, such as CSRNET [23], BL [33].

2) *Local Maxima Detection Strategy*: Given a predicted FIDT map, we can obtain the persons' positions by localizing the local maxima. We call this process as Local-Maxima-Detection-Strategy (LMDS), as illustrated in Algorithm 1.

Algorithm 1 Local Maxima Detection Strategy (LMDS)

```

1: Input: Predicted FIDT map
2: Output: The coordinates of the persons and the total count
3: function EXTRACT_POSITION(input)
4:   pos_ind = maxpooling(input, size = (3, 3))
5:   pos_ind = (pos_ind == input)
6:   matrix = pos_ind × input
7:   if max(matrix) <  $T_f$  then
8:     count = 0
9:     coordinates = None
10:  else
11:     $T_a = 100/255.0 \times \max(\text{matrix})$ 
12:    matrix[matrix <  $T_a$ ] = 0
13:    matrix[matrix > 0] = 1
14:    count = sum(matrix)
15:    coordinates = nonzeros(matrix)
16:  end if
17:  return count, coordinates
18: end function

```

Specifically, we first utilize a 3×3 max-pooling to obtain all local maxima (candidate points). However, these candidate points may contain some false positives from the background. We observe that the pixel values of true positives are much larger than the pixel values of false positives, which means a local maximum is likely to be a person if its pixel value is large enough. This inspires us to utilize an adaptive threshold T_a to filter the false positives. Thus, given a series of candidate points M , the final selected points are those whose values are no less than T_a , which is equal to $100/255.0$ times the maximum of M . Recent dataset [51] provides some negative samples, which consists of some scenes without persons and is similar to crowd scenes. We can not judge whether the original images contain persons based on the predicted density maps. However, given a predicted FIDT map, if the maximum of M is smaller than a tiny fixed threshold T_f (set as 0.10), this means the input image is a negative sample, and the LMDS will set the counting result as 0. An example of obtained location map is shown in Fig. 2.

Although the real individual sizes are not provided, we can generate pseudo individual bounding boxes from the predicted FIDT map. Here, the bounding box is a square surrounding a head. Once we get the predicted FIDT map, we first extract the coordinates of the head centers, which can be implemented efficiently using the proposed LMDS. Then, we estimate the instance size, using the K -nearest neighbours distance, which is defined as:

$$s_{(x,y) \in P} = \min \begin{cases} \bar{d} = f \times \frac{1}{k} \sum_{i=1}^k d_{(x,y)}^k \\ \min(\text{img_w}, \text{img_h}) \times 0.05 \end{cases} \quad (4)$$

where $s_{(x,y)}$ means the size of the instance, which locate in (x, y) , and P is the set of predicted head positions. \bar{d} is the average distance, which is calculated between point $P_{(x,y)}$ and its K -nearest neighbours, and a scalar factor f is used

TABLE I: Quantitative comparison of the localization performance on the NWPU-Crowd dataset. The results of other methods are from the online benchmark website [51]. F, P, and R refer to the F-measure, precision and recall, respectively.

Method	Training Labels	Validation set						Test set					
		σ_l			σ_s			σ_l			σ_s		
		F (%)	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	P (%)	R (%)
Faster RCNN [38]	Box	7.3	96.4	3.8	6.8	90.0	3.5	6.7	95.8	3.5	6.3	89.4	3.3
TinyFaces [13]	Box	59.8	54.3	66.6	55.3	50.2	61.7	56.7	52.9	61.1	52.6	49.1	56.6
TopoCount [1]	Box	-	-	-	-	-	-	69.1	69.5	68.7	60.1	60.5	59.8
VGG+GPR [6]	Point	56.3	61.0	52.2	46.0	49.9	42.7	52.5	55.8	49.6	42.6	45.3	40.2
RAZ_Loc [25]	Point	62.5	69.2	56.9	54.5	60.5	49.6	59.8	66.6	54.3	51.7	57.6	47.0
Crowd-SDNet [53]	Point	-	-	-	-	-	-	63.7	65.1	62.4	-	-	-
AutoScale* [57]	Point	66.8	70.1	63.8	60.0	62.9	57.3	62.0	67.3	57.4	54.4	59.1	50.4
GL [47]	Point	-	-	-	-	-	-	66.0	80.0	56.2	-	-	-
SCALNet [54]	Point	72.4	73.5	71.4	66.9	67.9	65.9	69.1	69.2	63.6	63.6	63.7	63.6
Ours	Point	78.9	82.2	75.9	73.7	76.7	70.9	75.5	79.7	71.7	70.5	74.4	66.9

to restrain the size. In very sparse regions, the \bar{d} may be bigger than the real size of persons, so we choose a threshold related to image size to restrain the object size, as described in Eq. 4. Note that Eq. 4 is only used in the testing phase for visualizations, and the size of bounding boxes does not influence the localization performance.

C. Independent SSIM Loss

Just using MSE loss to supervise the training phase will cause some negative impacts, such as blur effect and losing local structure information [4]. Some methods [4], [27] have proved that SSIM loss can improve the quality of the predicted map. SSIM is defined as:

$$SSIM(E, G) = \frac{(2\mu_E\mu_G + \lambda_1)(2\sigma_{EG} + \lambda_2)}{(\mu_E^2 + \mu_G^2 + \lambda_1)(\sigma_E^2 + \sigma_G^2 + \lambda_2)}, \quad (5)$$

where E and G represent the estimated map and ground-truth map, respectively. The μ and σ are the mean and variance. λ_1 and λ_2 are set to 0.0001 and 0.0009 to avoid being divided by zero. The value range of $SSIM$ is $[-1, 1]$, and $SSIM = 1$ means the estimated map is the same as the ground truth, so the SSIM loss is defined as:

$$L_S(E, G) = 1 - SSIM(E, G). \quad (6)$$

In general, the SSIM loss utilizes a sliding window to scan the whole predicted map without distinguishing the foreground (head region) and background. However, for the localization task, relying on detecting the local maxima, the model should focus on local maxima. The global SSIM loss may generate high responses, causing some false local maxima in the background. Thus, we propose the Independent SSIM (I-SSIM) loss, defined as:

$$L_{I-S} = \frac{1}{N} \sum_{n=1}^N L_S(E_n, G_n), \quad (7)$$

where N means the total number of persons, E_n and G_n mean the estimated and ground truth for the n -th independent instance region, and the region size of each instance is set as 30×30 for all datasets, mainly because we observe that

TABLE II: Quantitative evaluation of localization-based methods on the UCF-QNRF dataset. We report the average precision, average Recall, and average F-measure at different distance thresholds (1, 2, 3, ..., 100).

Method	Av.Precision	Av.Recall	F-measure
MCNN [66]	59.93%	63.50%	61.66%
CL [16]	75.80%	59.75%	66.82%
LCFCN [20]	77.89%	52.40%	62.65%
Method in [39]	75.46%	49.87%	60.05%
LSC-CNN[41]	74.62%	73.50%	74.06%
GL [47]	78.20%	74.80%	76.30%
TopoCount[1]	81.77%	78.96%	80.34%
Ours	84.49%	80.10%	82.23%

TABLE III: Quantitative evaluation of localization-based methods on the JHU-Crowd++ dataset using Precision (P), Recall (R), and F-measure (F).

Method	$\sigma = 4$			$\sigma = 8$		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
TopoCount [1]	31.5%	28.8%	30.1%	63.6%	58.3%	60.8%
Ours	38.9%	38.7%	38.8%	62.5%	62.4%	62.4%

this size can contain the entire head region without redundant background for most independent instance. The final training objective L is defined as below:

$$L = L_{MSE} + L_{I-S}, \quad (8)$$

where L_{MSE} and L_{I-S} refer to the MSE loss and the proposed I-SSIM loss, respectively.

IV. IMPLEMENT DETAILS

We augment the training data using random cropping and horizontal flipping. The crop size is 256×256 for Part A and Part B and 512×512 for other datasets. We set k as 4 and f as 0.1 to generate the bounding boxes. The α and β set as 0.02 and 0.75 respectively. We use Adam to optimize the model with the learning rate of $1e-4$, and the weight decay is set as $5e-4$. We set the size of the training batch to 16. We resize the images to make sure that the longer side is smaller



Fig. 4: Qualitative visualization of detected persons' locations by the proposed method. We use the proposed KNN strategy to generate bounding boxes (green boxes), compared with LSC-CNN [41].

than 2048 for NWPU-Crowd [51], JHU-Crowd++ [45], and UCF-QNRF [16] datasets.

A. Evaluation metrics

Localization Metrics. Precision, Recall, and F-measure are adopted to evaluate the performance on the NWPU-Crowd dataset, defined by [51]. When the distance between the given predicted point P_p and ground truth point P_g is less than a distance threshold σ , it means the P_p and P_g are successfully matched. The σ is related to the real head size (this dataset provides box-level annotation). Specifically, Wang *et al.* [51] give two thresholds:

$$\sigma_s = \min(w, h)/2, \quad (9)$$

$$\sigma_l = \sqrt{w^2 + h^2}/2, \quad (10)$$

and the former is a stricter criterion than the latter. For the UCF-QNRF dataset, similar to CL [16], we calculate the Precision, Recall, and F-measure at various thresholds (1, 2, 3, . . . , 100 pixels). For JHU-Crowd++, ShanghaiTech Part A, Part B, and UCF_CC_50 datasets, we choose two fixed thresholds ($\sigma = 4, 8$) for evaluation.

Counting Metrics. We use the Mean Absolute Error (MAE) and Mean Square Error (MSE) as the counting metrics, defined as:

$$MAE = \frac{1}{M} \sum_{i=1}^M |P_i - G_i|, \quad (11)$$

$$MSE = \sqrt{\frac{1}{M} \sum_{i=1}^M |P_i - G_i|^2}, \quad (12)$$

where M is the number of testing images, P_i and G_i are the predicted and ground truth count of the i -th image, respectively.

B. Dataset

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

NWPU-Crowd [51], a large-scale and challenging dataset, 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 from an online evaluation benchmark website.

JHU-CROWD++ [45] contains 2,722 training images, 500 validation images, and 1,600 test images, collected from diverse scenarios. The total number of persons in each image ranges from 0 to 25,791.

UCF-QNRF [16] 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 [66] consists of Part A and Part B with a total count of 1,198 images. In particular, Part A contains 300 training images and 182 testing images, and Part B consists of 400 training images and 316 testing images.

UCF_CC_50 [15] contains 50 gray images captured in extremely congested scenes. The number of crowd counts varies from 96 to 4,633. It is a challenging dataset due to the heavy background noise and the limited number of images.

TRANCOS [10] contains 1,244 images captured in traffic congestion situations with 46,796 annotations, providing a region of interest (ROI) for each image.

V. RESULTS AND ANALYSIS

A. Crowd localization

Tab. I, II, III, IV, and V compare the localization performance of the proposed method against the state-of-the-art methods. The results of other methods [41], [1], [57] are from the official code and model, and we directly utilize their predicted coordinates for evaluation. The evaluated localization code is provided by [51].

The results of the NWPU-Crowd dataset are from an online benchmark website, making sure to evaluate the localization performance fairly. As shown in Tab. I, we can observe that the proposed method outperforms the popular detectors, including Faster RCNN [38] and TinyFaces [13], by a significant margin. Compared with SCALNet [54] and TopoCount [1], our method outperforms them by at least 6.4% for σ_l (6.9% for σ_s) F-measure. Note that TopoCount [1] still applies the box-level annotations for training on the NWPU-Crowd dataset, while our method just utilizes the point-level.

For the dense dataset UCF-QNRF, as shown in Tab. II, the proposed method reports the highest Precision and Recall. For the JHU-Crowd++ dataset, as depicted in Tab. III, the proposed method improves the state-of-the-art method TopoCount [1] by 8.7% F-measure for the very strict setting $\sigma = 4$.

For the two sparse datasets, ShanghaiTech Part A and Part B, as depicted in Tab. IV, the proposed method improves the TopoCount [1] by 17.5% F-measure for the stricter setting $\sigma = 4$ on part A, and 1.5% F-measure on part B. It indicates that the proposed method can effectively cope with dense and sparse scenes.

For the gray images, UCF_CC_50 dataset (Tab. V), our method surpasses the other localization methods by a significant margin, *i.e.*, more than 7% F-measure improvement on the $\sigma = 4$. This impressive result demonstrates that our method is robust to the degraded images.

Additionally, we qualitatively evaluate the proposed method by visualizing the bounding boxes on the various crowd

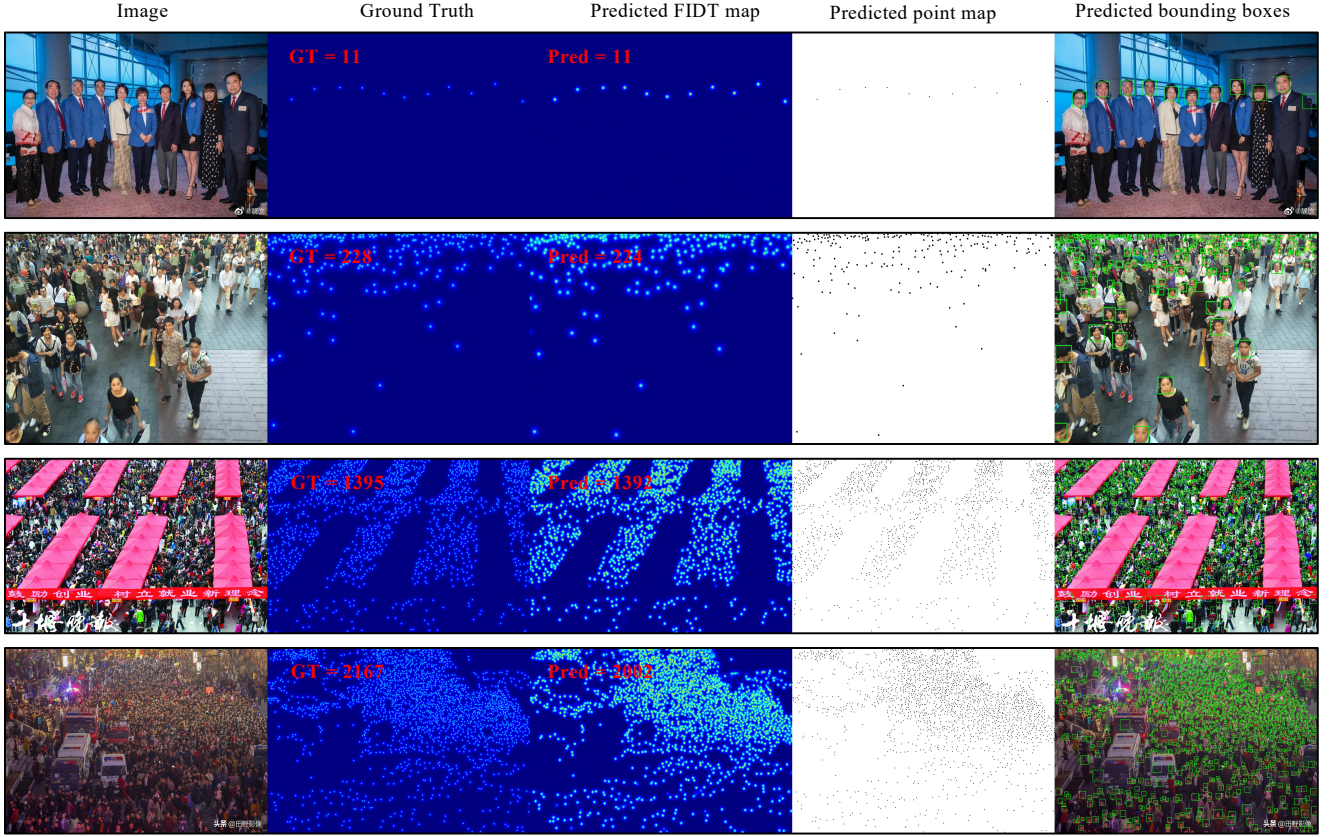


Fig. 5: Qualitative visualizations of our method. From left to right, there are testing images, ground truth maps, predicted FIDT maps, predicted point maps, and predicted bounding boxes.

TABLE IV: Comparison of the localization performance on the ShanghaiTech Part A [66] and ShanghaiTech Part B [66] datasets using Precision (P), Recall (R), and F-measure (F).

Method	Part A						Part B					
	$\sigma = 4$			$\sigma = 8$			$\sigma = 4$			$\sigma = 8$		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
LCFCN[20]	43.3%	26.0%	32.5%	75.1%	45.1%	56.3%	-	-	-	-	-	-
Method in [39]	34.9%	20.7%	25.9%	67.7%	44.8%	53.9%	-	-	-	-	-	-
LSC-CNN [41]	33.4%	31.9%	32.6%	63.9%	61.0%	62.4%	29.7%	29.2%	29.5%	57.5%	56.7%	57.0%
TopoCount [1]	41.7%	40.6%	41.1%	74.6%	72.7%	73.6%	63.4%	63.1%	63.2%	82.3%	81.8%	82.0%
Ours	59.1%	58.2%	58.6%	78.2%	77.0%	77.6%	64.9%	64.5%	64.7%	83.9%	83.2%	83.5%

TABLE V: Quantitative evaluation of localization-based methods on the UCF_CC_50 dataset using Precision (P), Recall (R), and F-measure (F). † represents that the networks are trained by ourselves.

Method	$\sigma = 4$			$\sigma = 8$		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
LSC-CNN† [41]	37.7%	39.5%	38.6%	57.8%	61.1%	59.4%
AutoScale† [57]	37.8%	40.5%	39.1%	59.0%	62.3%	60.6%
TopoCount† [1]	39.5%	42.0%	40.7%	62.5%	66.9%	64.6%
Ours	46.5%	49.0%	47.7%	67.0%	70.6%	68.7%

scenes in Fig. 4 and Fig. 5. The proposed method gives competitive bounding boxes compared with LSC-CNN [41]

and achieves impressive localization performance under various crowd scenes. It is noteworthy that the bounding boxes are only used for visualizations during the testing phase, and the size of bounding boxes does not affect the localization performance.

B. Crowd counting

In this work, we mainly focus on the crowd localization task, while the counting result can also be easily obtained since the total count is equal to the number of local maxima. Tab. VI, and Tab. VII show the quantitative counting 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

TABLE VI: Comparison of the counting performance on the NWPU-Crowd. $S0 \sim S4$ respectively indicate five categories according to the different number range: 0, (0, 100], (100, 500], (500, 5000], >5000. * means the localization-based methods, which can provide the position information.

Method	Output Position Information	validation set		Test set			
		Overall		Overall		Scene Level (only MAE)	
		MAE	MSE	MAE	MSE	Avg.	$S0 \sim S4$
C3F-VGG [8]	✗	105.79	504.39	127.0	439.6	666.9	140.9/26.5/58.0/307.1/2801.8
CSRNet [23]	✗	104.89	433.48	121.3	387.8	522.7	176.0/35.8/59.8/285.8/2055.8
CAN [30]	✗	93.58	489.90	106.3	386.5	612.2	82.6/14.7/46.6/269.7/2647.0
SCAR [9]	✗	81.57	397.92	110.0	495.3	718.3	122.9/16.7/46.0/241.7/3164.3
BL [33]	✗	93.64	470.38	105.4	454.2	750.5	66.5/8.7/41.2/249.9/3386.4
SFCN [52]	✗	95.46	608.32	105.7	424.1	712.7	54.2/14.8/44.4/249.6/3200.5
KDMG [48]	✗	-	-	100.5	415.5	632.7	77.3/10.3/38.5/259.4/2777.9
NoisyCC [46]	✗	-	-	96.9	534.2	608.1	218.7/10.7/35.2/203.2/2572.8
DM-Count [49]	✗	70.5	357.6	88.4	388.6	498.0	146.6/7.6/31.2/228.7/2075.8
RAZ_loc* [25]	✓	128.7	665.4	151.4	634.6	1166.0	60.6/17.1/48.3/364.7/5339.0
AutoScale* [57]	✓	97.3	571.2	123.9	515.5	871.0	42.3/18.8/46.1/301.7/3947.0
TopoCount* [1]	✓	-	-	107.8	438.5	-	-
SCALNet* [54]	✓	64.4	251.1	86.8	339.9	429.5	92.0/11.2/41.1/227.7/1775.3
Ours*	✓	51.4	107.6	86.0	312.5	390.6	21.6/13.7/55.6/217.1/1645.4

TABLE VII: Comparison of the counting performance on the JHU-Crowd++, UCF-QNRF, ShanghaiTech Part A, Part B and UCF_CC_50 datasets. * means the localization-based methods, which can provide the position information.

Method	Output Position Information	JHU		QNRF		Part A		Part B		UCF_CC_50	
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
CSRNet [23]	✗	85.9	309.2	-	-	68.2	115.0	10.6	16.0	266.1	397.5
SFCN [52]	✗	77.5	297.6	102.0	171.4	64.8	107.5	7.6	13.0	214.2	318.2
L2SM [58]	✗	-	-	104.7	173.6	64.2	98.4	7.2	11.1	188.4	315.3
CG-DRCN [44]	✗	82.3	328.0	112.2	176.3	64.0	98.4	8.5	14.4	-	-
MUD-iKNN [36]	✗	-	-	104.0	172.0	68.0	117.7	13.4	21.4	237.7	305.7
DSSI-Net [27]	✗	133.5	416.5	99.1	159.2	60.6	96.0	6.9	10.3	216.9	302.4
MBTTBF [43]	✗	81.8	299.1	97.5	165.2	60.2	94.1	8.0	15.5	233.1	300.9
BL [33]	✗	75.0	299.9	88.7	154.8	62.8	101.8	7.7	12.7	229.3	308.2
RPNNet [60]	✗	-	-	-	-	61.2	96.9	8.1	11.6	-	-
ASNet [18]	✗	-	-	91.6	159.7	57.8	90.1	-	-	174.8	251.6
AMSNNet [14]	✗	-	-	101.8	163.2	56.7	93.4	6.7	10.2	208.6	296.3
LibraNet [26]	✗	-	-	88.1	143.7	55.9	97.1	7.3	11.3	181.2	262.2
KDMG [48]	✗	69.7	268.3	99.5	173.0	63.8	99.2	7.8	12.7	-	-
NoisyCC [46]	✗	67.7	258.5	85.8	150.6	61.9	99.6	7.4	11.3	-	-
DM-Count [49]	✗	-	-	85.6	148.3	59.7	95.7	7.4	11.8	211.0	291.5
RAZ_loc+* [25]	✓	-	-	118.0	198.0	71.6	120.1	9.9	15.6	-	-
PSDDN* [32]	✓	-	-	-	-	65.9	112.3	9.1	14.2	359.4	514.8
LSC-CNN* [41]	✓	112.7	454.4	120.5	218.2	66.4	117.0	8.1	12.7	225.6	302.7
Crowd-SDNet* [53]	✓	-	-	-	-	65.1	104.4	7.8	12.6	-	-
AutoScale* [57]	✓	85.6	356.1	104.4	174.2	65.8	112.1	8.6	13.9	210.5	287.4
TopoCount* [1]	✓	60.9	267.4	89.0	159.0	61.2	104.6	7.8	13.7	184.1	258.3
Ours*	✓	66.6	253.6	89.0	153.5	57.0	103.4	6.9	11.8	171.4	233.1

SCALNet [54] on the NWPU-Crowd (test set) by a significant margin of 27.4 MSE. Our method also obtains the best performance on UCF-QNRF, ShanghaiTech Part A, Part B, and UCF_CC_50 datasets. For the JHU-Crowd++ dataset, the proposed method achieves SOTA performance in MSE and comparable performance in MAE. It indicates that the proposed method can cope with both sparse crowd scenes and dense crowd scenes.

Compared with the regression-based methods. Although it is not fair to compare localization-based count-

ing methods and density-map regression-based counting methods, our method still outperforms all density map regression-based methods on NWPU-Crowd, JHU-Crowd++, and UCF_CC_50 datasets. Meanwhile, the proposed method achieves comparable performance on UCF-QNRF, ShanghaiTech Part A, and Part B datasets. To intuitively demonstrate the difference between FIDT maps and density maps, we provide the predicted FIDT maps and density maps visualization (the density maps are trained with the same network), as shown in Fig. 6 (row 1 and row 2). We can see that the predicted

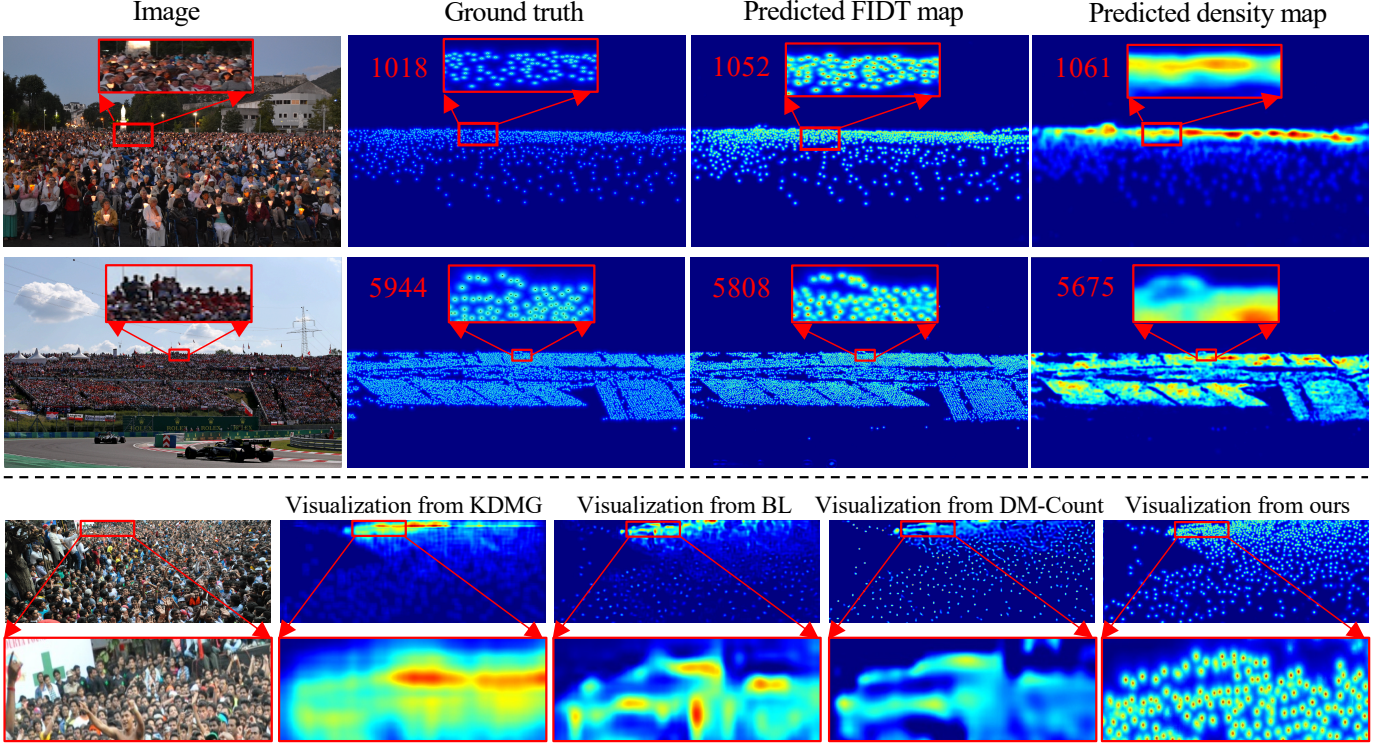


Fig. 6: Row 1 and row 2, from left to right, there are images, ground truth, predicted FIDT maps, and predicted density maps (trained with the same network). It can be seen that the heads of the predicted FIDT maps are distinguishable, and the predicted density maps show severe overlaps. Row 3 and row 4 compare the predicted visualizations from KDMG [48], BL [33], DM-Count [49], and ours. Our method provides precise location information in the dense region (red box).

density maps lose the position information and show severe overlaps in dense regions. However, the predicted FIDT maps provide an almost accurate location for each individual, even in the extremely dense scene. On row 3 and row 4 of Fig. 6, it compares the predicted visualizations from KDMG [48], BL [33], DM-Count [49], and ours. We can see that the DM-Count [49] and our method provide clear position information in the sparse region, while only our method can provide precise location information in the dense region (red box).

C. Evaluation on vehicle dataset

A robustness algorithm should easily generalize to similar tasks (e.g., vehicle localization and counting). Thus, following previous localization methods [32], [41], we evaluate the generalization capability of the proposed method on the TRANCOS [10] dataset for vehicle counting. We adopt the Grid Average Mean Absolute Error (GAME) [10] as the evaluation metric for vehicle counting, defined as:

$$GAME(L) = \frac{1}{N} \sum_{i=1}^N \left(\sum_{l=1}^{4^L} |P_i^l - G_i^l| \right), \quad (13)$$

which splits an input image into 4^L non-overlapping sub-regions. N is the number of the testing images, P_i and G_i are the predicted and ground truth count of the i -th image, respectively. Tab. VIII compares the GAME metric of the proposed method and the state-of-the-art localization-based methods. Specifically, the proposed method achieves the best

performance on GAME(0), GAME(1), and GAME(2) and obtains comparable performance on GAME(3). It means that the proposed method not only achieves accurate global predictions but also has well localization performance.

D. Ablation Study

Analysis of the FIDT map. To understand the FIDT map better, we analyze the distribution of the FIDT map by using different α and β . Only changing the α (resp. β), as shown in Fig. 7, as α (resp. β) increasing (resp. decreasing), the response of FIDT map shows faster (resp. slower) decay in both foreground and background. As discussed in Sec. III-A, the decay should be slower away from the head, and the response of the background should quickly close to 0. Thus, we set $\alpha = 0.02$ and $\beta = 0.75$ in all experiments, and we also report the various α and β settings experiments in Tab. IX. The following ablation study will provide experiments of choosing 0.02 and 0.75 in Eq. 3.

Effectiveness of I-SSIM loss. In this section, we explore the advantage of the proposed I-SSIM loss. Based on Tab. X, we make the following observations: (1) Adding the traditional global SSIM loss can bring improvement. (2) The proposed I-SSIM loss achieves further improvement in terms of localization and counting, mainly because the I-SSIM loss can further optimize the structure information of the predicted FIDT map to find local maxima better and repress the false local maxima in the background.

TABLE VIII: Quantitative comparison of vehicles counting on the TRANCOS [10] dataset. † represents that the networks are trained by ourselves.

Method	GAME(0)	GAME(1)	GAME(2)	GAME(3)
PSDDN [32]	4.79	5.43	6.68	8.40
LSC-CNN [41]	4.60	5.40	6.90	8.30
AutoScale [57]	2.88	4.97	6.64	9.73
Crowd-SDNet† [53]	3.82	5.27	7.72	10.11
TopoCount† [1]	3.42	4.76	6.51	8.55
Ours	2.25	3.91	5.66	8.36

TABLE IX: The results of different α and β setting on Part A. It is noteworthy that IDT map (Eq. 2) means the $\alpha = 0$ and $\beta = 1$.

Method	α	β	Localization ($\sigma = 8$)			Counting	
			P(%)	R(%)	F(%)	MAE	MSE
IDT	0.00	1.00	75.6%	74.6%	75.1%	61.8	109.6
FIDT	0.01	0.65	76.6%	76.0%	76.3%	60.5	107.4
FIDT	0.02	0.75	78.2%	77.0%	77.6%	57.0	103.4
FIDT	0.03	0.85	77.0%	76.8%	76.9%	58.3	106.6

We further ablate the influence of the independent instance region sizes of I-SSIM loss, as shown in Tab. XI. Larger region sizes may contain too much background (without structure information), leading to excess false local maxima in the background. Smaller region sizes may not involve the entire independent head, which can not effectively enhance the structure information of the local maxima (head region). Based on the experiments, we use 30×30 as the region size for all datasets, and it works well.

Analysis of T_a . On the proposed post-processing, LMDS, the T_a is used to choose the positive points. Its value is adaptive, which is set to $\frac{100}{255} \times \max(M)$, where $\max(M)$ is the max value of the predicted FIDT map. As shown in Tab. XII, using fixed thresholds is worse than the adaptive threshold since the local-maxima pixel values of different predicted FIDT maps are not the same. This inspires us to

TABLE X: The effectiveness of the proposed I-SSIM loss on Part A.

Method	Localization ($\sigma = 8$)			Counting	
	P(%)	R(%)	F(%)	MAE	MSE
L2	73.5%	76.3%	74.9%	62.1	108.8
L2 + SSIM	76.8%	76.6%	76.7%	59.3	106.5
L2 + I-SSIM (ours)	78.2%	77.0%	77.6%	57.0	103.4

TABLE XI: The influence of the independent instance region size of I-SSIM loss on Part A dataset.

Region Size	Localization ($\sigma = 8$)			Counting	
	P(%)	R(%)	F(%)	MAE	MSE
20×20	78.0%	76.4%	77.2%	58.1	104.4
30×30	78.2%	77.0%	77.6%	57.0	103.4
40×40	77.4%	76.4%	76.9%	58.6	105.9

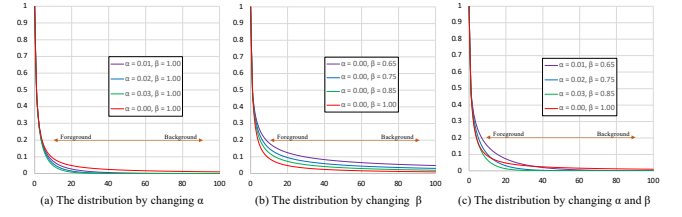


Fig. 7: The effect of changing α and β on the distribution of FIDT map.

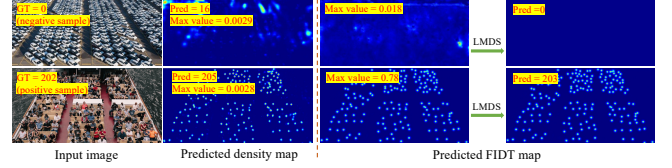


Fig. 8: Row 1 and row 2 are negative and positive samples, respectively. In the density maps, the max pixel value between positive and negative samples is very similar. In the FIDT maps, the negative and positive samples present a significant difference in max pixel value.

utilize adaptive threshold. Using large adaptive thresholds will filter out the true-positive points, and small will reserve some false-positive points. Hence, we choose the $\frac{100}{255} \times \max(M)$ as an adaptive threshold for all datasets.

Robustness on negative and dense scenes. The S0 of NWPU-Crowd consists of some “dense fake humans” (e.g., Terra-Cotta Warriors), called negative samples. In contrast, S4 means the extremely dense crowd scenes, containing more than 5,000 persons. Thus, the S0 and S4 are usually adopted to evaluate the model’s robustness [51]. Tab. XIII lists the results of some popular methods on the NWPU-Crowd’s

TABLE XII: The ablation study on the threshold T_a .

Threshold value	Adaptive	MAE	MSE
50/255	✗	95.6	169.9
70/255	✗	92.1	167.0
90/255	✗	80.6	152.3
100/255	✗	115.3	206.4
110/255	✗	122.6	233.5
$90/255 \times \max(M)$	✓	60.4	105.3
$100/255 \times \max(M)$	✓	57.0	103.4
$110/255 \times \max(M)$	✓	58.1	107.1

TABLE XIII: The results of S0 and S4 on NWPU-Crowd test set.

Method	S0-level		S4-level	
	MAE	MSE	MAE	MSE
KDMG [48]	77.3	303.0	2777.9	3521.8
DM-Count [49]	146.7	736.1	2075.8	2895.2
NoisyCC [46]	218.7	1415.6	2572.5	3414.9
SCALNet [54]	92.0	479.3	1775.3	2676.4
Ours	21.6	129.3	1645.4	2288.2

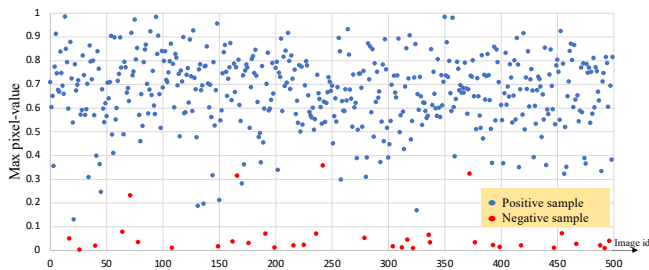


Fig. 9: The max pixel-value distribution of predicted FIDT maps (NWPU-Crowd validation set).

TABLE XIV: The results of different regressors on the NWPU-Crowd [51] (validation set) dataset.

Method	Density map (counting by integration)		FIDT map (counting by localization)				
	MAE	MSE	MAE	MSE	P(%)	R(%)	F(%)
CSRNET [23]	104.9	433.5	100.6	464.3	68.8%	66.2%	67.5%
CSRNET [23] + FPN	95.5	450.8	70.6	369.7	73.9%	69.8%	71.8%
BL [33]	93.6	470.4	89.7	446.6	69.9%	68.9%	69.4%
BL [33] + FPN	85.4	412.9	65.7	259.2	77.8%	70.0%	73.7%

negative samples and extremely dense scenes. As expected, the proposed method achieves the lowest counting error, reporting superior robustness. As shown in Fig. 8, the maximum pixel value between negative and positive samples is similar for density maps while it is distinct for the FIDT maps. Given a predicted FIDT map, if its maximum pixel value is smaller than threshold T_f , the LMDS will regard the input as a negative sample and set the counting result as 0, as illustrated in Algorithm 1.

We set the T_f as 0.1 according to the statistics. Specifically, we give the max pixel-value of each predicted FIDT map based on NWPU-Crowd (validation set, including 500 images), as shown in Fig. 9. We can observe that all positive samples' value is much bigger than 0.1, and most negative samples are smaller than 0.1. Thus, the threshold T_f set as 0.1 is reasonable.

Generalization on different regressors. In this section, to demonstrate the proposed FIDT map can be generalized to different regressors, we implement the CSRNET [23], BL [33] with the FIDT maps on the NWPU-Crowd dataset (validation set). Besides, we add a FPN into the CSRNET [23] and BL [33] to capture rich spatial context. The quantitative results are listed in Tab. XIV, where we can observe that using the FIDT map can realize the localization task, and the counting performance is competitive compared with the density map. Notability, the experiments of FIDT maps only adopt the L2 loss, and the image scaling strategy is the same as [51]. The results indicate that the FIDT map is suitable for the crowd localization task.

E. Limitation

The main limitation is that the proposed method inference will be slower than some real-time methods [47]. As shown in Tab. XV, we report the Multiply-Accumulate Operations

TABLE XV: The comparisons of complexity. The F-measure is from the NWPU-Crowd benchmark (test set).

Method	MACs (G)	Inference speed	F-measure
LSC-CNN [41]	1244.3	2.6 FPS	-
AutoScale [57]	1074.6	5.7 FPS	62.0%
Crowd-SDNet ¹ [53]	-	0.8 FPS	63.7%
GL [47]	324.6	20.3 FPS	66.0%
TopoCount [49]	797.2	9.4 FPS	69.1%
Ours	426.7	7.1 FPS	75.5%

(MACs) and Frames Per Second (FPS) to analyze the complexity. All methods are evaluated on the official code with a size of 768×1024 image, and the GPU device is NVIDIA RTX 3090. Although our method achieves the second MACs and the third FPS, there is still a lot of room for improvement. In the future, we are interested in extending our method for real-time.

VI. CONCLUSION

In this paper, we present a novel label named FIDT map, designed to cope with the crowd localization task. The proposed FIDT map is a non-overlap map, which utilizes local maxima to represent the head's center. To extract the corresponding individual center, a Local-Maxima-Detection-Strategy (LMDS) is proposed. Besides, we introduce a novel I-SSIM loss to make the model tend to focus on the foreground regions, improving the structure information of local maxima. By performing experiments on six publicly available datasets, we demonstrate that the proposed method achieves state-of-the-art localization performance and shows superior robustness for the negative samples and extreme scenes. We hope the community switches from the density map regression to FIDT map regression for more practical.

REFERENCES

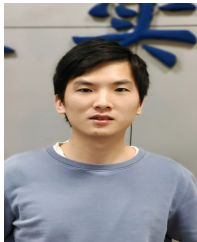
- [1] Shahira Abousamra, Minh Hoai, Dimitris Samaras, and Chao Chen. Localization in the crowd with topological constraints. In *Proc. of the AAAI Conf. on Artificial Intelligence*, 2021.
- [2] Carlos Arteta, Victor Lempitsky, and Andrew Zisserman. Counting in the wild. In *Proc. of European Conference on Computer Vision*, pages 483–498, 2016.
- [3] Deepak Babu Sam, Shiv Surya, and R Venkatesh Babu. Switching convolutional neural network for crowd counting. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 5744–5752, 2017.
- [4] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. Scale aggregation network for accurate and efficient crowd counting. In *Proc. of European Conference on Computer Vision*, pages 734–750, 2018.
- [5] Zhi-Qi Cheng, Jun-Xiu Li, Qi Dai, Xiao Wu, and Alexander G Hauptmann. Learning spatial awareness to improve crowd counting. In *Proc. of IEEE Intl. Conf. on Computer Vision*, pages 6152–6161, 2019.
- [6] 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.
- [7] Junyu Gao, Tao Han, Yuan Yuan, and Qi Wang. Learning independent instance maps for crowd localization. *arXiv preprint arXiv:2012.04164*, 2020.
- [8] Junyu Gao, Wei Lin, Bin Zhao, Dong Wang, Chenyu Gao, and Jun Wen. C³ framework: An open-source pytorch code for crowd counting. *arXiv preprint arXiv:1907.02724*, 2019.
- [9] Junyu Gao, Qi Wang, and Yuan Yuan. Scar: Spatial/channel-wise attention regression networks for crowd counting. *Neurocomputing*, 363:1–8, 2019.

¹We try our best to calculate the MACs of Crowd-SDNet, but the official code relies on the old version Keras, which is hard to obtain the MACs.

- [10] Ricardo Guerrero-Gómez-Olmedo, Beatriz Torre-Jiménez, Roberto López-Sastre, Saturnino Maldonado-Bascón, and Daniel Onoro-Rubio. Extremely overlapping vehicle counting. In *IbPRIA*, pages 423–431, 2015.
- [11] Tao Han, Junyu Gao, Yuan Yuan, Xuelong Li, et al. Ldc-net: A unified framework for localization, detection and counting in dense crowds. *arXiv preprint arXiv:2110.04727*, 2021.
- [12] Zeeshan Hayder, Xuming He, and Mathieu Salzmann. Boundary-aware instance segmentation. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 5696–5704, 2017.
- [13] Peiyun Hu and Deva Ramanan. Finding tiny faces. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 951–959, 2017.
- [14] Yutao Hu, Xiaolong Jiang, Xuhui Liu, Baochang Zhang, Jungong Han, Xianbin Cao, and David Doermann. Nas-count: Counting-by-density with neural architecture search. In *Proc. of European Conference on Computer Vision*, pages 747–766, 2020.
- [15] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2013.
- [16] Haroon Idrees, Muhammad 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 *Proc. of European Conference on Computer Vision*, pages 532–546, 2018.
- [17] 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 *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 6133–6142, 2019.
- [18] Xiaoheng Jiang, Li Zhang, Mingliang Xu, Tianzhu Zhang, Pei Lv, Bing Zhou, Xin Yang, and Yanwei Pang. Attention scaling for crowd counting. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 4706–4715, 2020.
- [19] Xiaoheng Jiang, Li Zhang, Tianzhu Zhang, Pei Lv, Bing Zhou, Yanwei Pang, Mingliang Xu, and Changsheng Xu. Density-aware multi-task learning for crowd counting. *IEEE Transactions on Multimedia*, 23:443–453, 2020.
- [20] Issam H Laradji, Negar Rostamzadeh, Pedro O Pinheiro, David Vazquez, and Mark Schmidt. Where are the blobs: Counting by localization with point supervision. In *Proc. of European Conference on Computer Vision*, pages 547–562, 2018.
- [21] Jing Li, Lison Wei, Fangbing Zhang, Tao Yang, and Zhaoyang Lu. Joint deep and depth for object-level segmentation and stereo tracking in crowds. *IEEE Transactions on Multimedia*, 21(10):2531–2544, 2019.
- [22] Yuke Li. A deep spatiotemporal perspective for understanding crowd behavior. *IEEE Transactions on Multimedia*, 20(12):3289–3297, 2018.
- [23] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 1091–1100, 2018.
- [24] Dingkan Liang, Xiwu Chen, Wei Xu, Yu Zhou, and Xiang Bai. Transcrowd: weakly-supervised crowd counting with transformers. *Science China Information Sciences*, 65(6):1–14, 2022.
- [25] Chenchen Liu, Xinyu Weng, and Yadong Mu. Recurrent attentive zooming for joint crowd counting and precise localization. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 1217–1226, 2019.
- [26] Liang Liu, Hao Lu, Hongwei Zou, Haipeng Xiong, Zhiguo Cao, and Chunhua Shen. Weighing counts: Sequential crowd counting by reinforcement learning. In *Proc. of European Conference on Computer Vision*, pages 164–181, 2020.
- [27] Lingbo Liu, Zhilin Qiu, Guanbin Li, Shufan Liu, Wanli Ouyang, and Liang Lin. Crowd counting with deep structured scale integration network. In *Proc. of IEEE Intl. Conf. on Computer Vision*, pages 1774–1783, 2019.
- [28] Ning Liu, Yongchao Long, Changqing Zou, Qun Niu, Li Pan, and Hefeng Wu. Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 3225–3234, 2019.
- [29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Proc. of European Conference on Computer Vision*, pages 21–37, 2016.
- [30] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 5099–5108, 2019.
- [31] Yan Liu, Lingqiao Liu, Peng Wang, Pingping Zhang, and Yinjie Lei. Semi-supervised crowd counting via self-training on surrogate tasks. In *Proc. of European Conference on Computer Vision*, pages 242–259. Springer, 2020.
- [32] Yuting Liu, Miaoqing Shi, Qijun Zhao, and Xiaofang Wang. Point in, box out: Beyond counting persons in crowds. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 6469–6478, 2019.
- [33] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *Proc. of IEEE Intl. Conf. on Computer Vision*, pages 6142–6151, 2019.
- [34] Pier Luigi Mazzeo, Riccardo Contino, Paolo Spagnolo, Cosimo Distanti, Ettore Stella, Massimiliano Nitti, and Vito Renò. Mh-metronet—a multi-head cnn for passenger-crowd attendance estimation. *Journal of Imaging*, 6(7):62, 2020.
- [35] Yanda Meng, Hongrun Zhang, Yitian Zhao, Xiaoyun Yang, Xuesheng Qian, Xiaowei Huang, and Yalin Zheng. Spatial uncertainty-aware semi-supervised crowd counting. In *Proc. of IEEE Intl. Conf. on Computer Vision*, pages 15549–15559, 2021.
- [36] Greg Olmschenk, Hao Tang, and Zhigang Zhu. Improving dense crowd counting convolutional neural networks using inverse k-nearest neighbor maps and multiscale upsampling. In *15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, volume 5, 2020.
- [37] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 779–788, 2016.
- [38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proc. of Advances in Neural Information Processing Systems*, volume 28, pages 91–99, 2015.
- [39] Javier Ribera, David Güera, Yuhao Chen, and Edward J. Delp. Locating objects without bounding boxes. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 6479–6489, 2019.
- [40] Azriel Rosenfeld and John L Pfaltz. Distance functions on digital pictures. *Pattern recognition*, 1(1):33–61, 1968.
- [41] 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*, pages 2739–2751, 2020.
- [42] Miaoqing Shi, Zhaohui Yang, Chao Xu, and Qijun Chen. Revisiting perspective information for efficient crowd counting. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 7279–7288, 2019.
- [43] Vishwanath A Sindagi and Vishal M Patel. Multi-level bottom-top and top-bottom feature fusion for crowd counting. In *Proc. of IEEE Intl. Conf. on Computer Vision*, pages 1002–1012, 2019.
- [44] Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method. In *Proc. of IEEE Intl. Conf. on Computer Vision*, pages 1221–1231, 2019.
- [45] 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*, pages 1–1, 2020.
- [46] Jia Wan and Antoni Chan. Modeling noisy annotations for crowd counting. In *Proc. of Advances in Neural Information Processing Systems*, pages 3386–3396, 2020.
- [47] Jia Wan, Ziquan Liu, and Antoni B Chan. A generalized loss function for crowd counting and localization. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 1974–1983, 2021.
- [48] Jia Wan, Qingzhong Wang, and Antoni B Chan. Kernel-based density map generation for dense object counting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020.
- [49] Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai. Distribution matching for crowd counting. In *Proc. of Advances in Neural Information Processing Systems*, 2020.
- [50] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Minghui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020.
- [51] 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*, 43(6):2141–2149, 2020.
- [52] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 8198–8207, 2019.
- [53] Yi Wang, Junhui Hou, Xinyu Hou, and Lap-Pui Chau. A self-training

approach for point-supervised object detection and counting in crowds. *IEEE Transactions on Image Processing*, 30:2876–2887, 2021.

- [54] Yi Wang, Xinyu Hou, and Lap-Pui Chau. Dense point prediction: A simple baseline for crowd counting and localization. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2021.
- [55] 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 *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 3833–3842, 2020.
- [56] Yukang Wang, Yongchao Xu, Stavros Tsogkas, Xiang Bai, Sven Dickinson, and Kaleem Siddiqi. Deepflux for skeletons in the wild. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 5287–5296, 2019.
- [57] Chenfeng Xu, Dingkan Liang, Yongchao Xu, Song Bai, Wei Zhan, Xiang Bai, and Masayoshi Tomizuka. Autoscale: Learning to scale for crowd counting. *International Journal of Computer Vision*, pages 1–30, 2022.
- [58] Chenfeng Xu, Kai Qiu, Jianlong Fu, Song Bai, Yongchao Xu, and Xiang Bai. Learn to scale: Generating multipolar normalized density maps for crowd counting. In *Proc. of IEEE Intl. Conf. on Computer Vision*, pages 8382–8390, 2019.
- [59] Yanyu Xu, Ziming Zhong, Dongze Lian, Jing Li, Zhengxin Li, Xinling Xu, and Shenghua Gao. Crowd counting with partial annotations in an image. In *Proc. of IEEE Intl. Conf. on Computer Vision*, pages 15570–15579, 2021.
- [60] Yifan Yang, Guorong Li, Zhe Wu, Li Su, Qingming Huang, and Nicu Sebe. Reverse perspective network for perspective-aware object counting. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 4374–4383, 2020.
- [61] Yifan Yang, Guorong Li, Zhe Wu, Li Su, Qingming Huang, and Nicu Sebe. Weakly-supervised crowd counting learns from sorting rather than locations. In *Proc. of European Conference on Computer Vision*, pages 1–17, 2020.
- [62] Anran Zhang, Jiayi Shen, Zehao Xiao, Fan Zhu, Xiantong Zhen, Xianbin Cao, and Ling Shao. Relational attention network for crowd counting. In *Proc. of IEEE Intl. Conf. on Computer Vision*, pages 6788–6797, 2019.
- [63] Anran Zhang, Lei Yue, Jiayi Shen, Fan Zhu, Xiantong Zhen, Xianbin Cao, and Ling Shao. Attentional neural fields for crowd counting. In *Proc. of IEEE Intl. Conf. on Computer Vision*, pages 5714–5723, 2019.
- [64] Cong Zhang, Kai Kang, Hongsheng Li, Xiaogang Wang, Rong Xie, and Xiaokang Yang. Data-driven crowd understanding: A baseline for a large-scale crowd dataset. *IEEE Transactions on Multimedia*, 18(6):1048–1061, 2016.
- [65] Jialiang Zhang, Lixiang Lin, Jianke Zhu, Yang Li, Yun-chen Chen, Yao Hu, and Steven CH Hoi. Attribute-aware pedestrian detection in a crowd. *IEEE Transactions on Multimedia*, 23:3085–3097, 2020.
- [66] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 589–597, 2016.



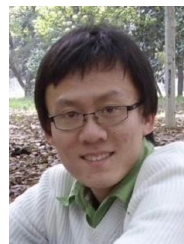
Dingkan Liang is currently working towards the Ph.D. degree in School of Artificial Intelligence and Automation from the Huazhong University of Science and Technology, Wuhan, China. He has served as a reviewer for several top journals and conferences such as TPAMI, TIP, CVPR, ICCV, and ECCV. His research interests include computer vision, especially for crowd analysis and 3D object detection.



Wei Xu is currently working towards the M.S. degree in information and communication engineering from Beijing University of Posts and Telecommunications, Beijing, China. His research interests include crowd analysis, face reconstruction, and object detection.



Yingying Zhu received the bachelor's degree and Ph.D. degree in Electronics and Information Engineering from Huazhong University of Science and Technology (HUST), Wuhan, P.R. China in 2018. She joined the Huazhong University of Science and Technology (HUST), Wuhan, as an Engineer from 2021 to now. Her research interests include computer vision and machine learning.



automatic drive.

Yu Zhou received the M.S. and Ph.D. degrees both in Electronics and Information Engineering from Huazhong University of Science and Technology (HUST), Wuhan, P.R. China in 2010, and 2014, respectively. In 2014, he joined the Beijing University of Posts and Telecommunications (BUPT), Beijing, as a Postdoctoral Researcher from 2014 to 2016, an Assistant Professor from 2016 to 2018. He is currently an Associate Professor with the School of Electronic Information and Communications, HUST. His research interests include computer vision and