Focal Inverse Distance Transform Maps for Crowd Localization

Dingkang Liang, Wei Xu, Yingying Zhu, and Yu Zhou, Member, IEEE

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.

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 [19], [23], 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 [13], [38] 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 [16], [66] only provide point-level annotations (Fig. 1(b)), making the detectors [13], [38] untrainable. Current regression-based methods [3], [19], [23], [66] 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.

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.
Alternatively, some methods focus on designing an appropriate map to cope with crowd localization, such as binary-like maps [1, 7, 25] 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 performance. 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 [17], [18], [43] to improve the quality of predicted density maps. Some methods [28], [62], [63] 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 [31], [35], [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 [29], [37], [38] 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, 7, 25] or segmentation-like map [20], [57]. 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 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 [2], [12], [55], [56]. 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 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.
defined as:

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

where $I$ is the FIDT map we proposed, $\alpha$ and $\beta$ 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. Specifically, we first utilize a $3 \times 3$ 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:

$$s(x,y) \in P = \min \left\{ \frac{d}{f \times \frac{1}{k} \sum_{1}^{k} \delta(x,y)} \min (img \_ w, img \_ h) \times 0.05 \right\}$$

where $s(x,y)$ means the size of the instance, which locate in $(x,y)$, and $P$ is the set of predicted head positions. $d$ is the average distance, which is calculated between point $P(x,y)$ and its K-nearest neighbours, and a scalar factor $f$ is used to restrain the size. In very sparse regions, the $d$ 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 (4). Note that (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:

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

where \(E\) and \(G\) represent the estimated map and ground-truth map, respectively. The \(\mu\) and \(\sigma\) are the mean and variance, \(\lambda_1\) and \(\lambda_2\) 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:

\[
L_S(E, G) = 1 - SSIM(E, G).
\]

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:

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

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 \(30 \times 30\) for all datasets, mainly because we observe that this size can contain the entire head region without redundant background for most independent instance. The final training objective \(L\) is defined as below:

\[
L = L_{MSE} + L_{I-S},
\]

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 \(256 \times 256\) for Part A and Part B and \(512 \times 512\) for other datasets. We set \(k\) as 4 and \(f\) as 0.1 to generate the bounding boxes. The \(\alpha\) and \(\beta\) 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 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 \(P_g\) and ground truth point \(P_p\) is less than a distance threshold \(\sigma\), it means the \(P_g\) and \(P_p\) are successfully matched. The \(\sigma\) is related to the real head size (this dataset provides box-level annotation). Specifically, Wang et al. [51] give two thresholds:

\[
\sigma_s = \min(w, h)/2,
\]

\[
\sigma_l = \sqrt{w^2 + h^2}/2,
\]

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:

\[
MAE = \frac{1}{M} \sum_{i=1}^{M} |P_i - G_i|,
\]

\[
MSE = \sqrt{\frac{1}{M} \sum_{i=1}^{M} |P_i - G_i|^2},
\]

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 Gy 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.
TABLE I

| Method       | Training Labels | Validation set | Test set |
|--------------|-----------------|----------------|----------|
|              |                 | $\sigma_l$ | $\sigma_s$ | $\sigma_l$ | $\sigma_s$ |
|              |                 | F (%) | P (%) | R (%) | F (%) | P (%) | R (%) |
| Faster RCNN [38] | Box              | 7.3 | 96.4 | 3.8 | 6.8 | 90.0 | 3.5 |
| TinyFaces [13]   | Box              | 59.8 | 54.3 | 66.6 | 55.3 | 50.2 | 61.7 |
| TopoCount [1]    | Box              | -   | -   | -   | -   | -   | -   |
| VGG+GPR [6]      | Point            | 56.3 | 61.0 | 52.2 | 46.0 | 49.9 | 42.7 |
| RAZ_Loc [25]     | Point            | 62.5 | 69.2 | 56.9 | 54.5 | 60.5 | 49.6 |
| Crowd-3DNet [53] | Point            | -   | -   | -   | -   | -   | -   |
| AutoScale* [57]  | Point            | 66.8 | 70.1 | 63.8 | 60.0 | 62.9 | 57.3 |
| GL [47]          | Point            | -   | -   | -   | -   | -   | -   |
| SCALNet [54]     | Point            | 72.4 | 73.5 | 71.4 | 66.9 | 67.9 | 65.9 |
| Ours             | Point            | 78.9 | 82.2 | 75.9 | 73.7 | 76.7 | 70.9 |

TABLE II

| 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

| 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% |

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

Table 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 [1], [41], [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 Table 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 $\sigma_l$ (6.9% for $\sigma_s$) 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 Table II, the proposed method reports the highest Precision and Recall. For the JHU-Crowd++ dataset, as depicted in Table 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 Table 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 (Table 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 scenes in Fig. 4 and Fig. 5. The proposed method gives competitive
TABLE IV
COMPARISON OF THE LOCALIZATION PERFORMANCE ON THE SHANGHAI TECH PART A [66] AND SHANGHAI TECH PART B [66] DATASETS USING PRECISION (P), RECALL (R), AND F-MEASURE (F)

| Method       | Part A |          | Part B |          |
|--------------|--------|----------|--------|----------|
|              | σ = 4  | σ = 8    | σ = 4  | σ = 8    |
|              | 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%  |
| TopoCount [1]| 41.7%  | 40.6%    | 41.1%  | 74.6%    | 72.7%    | 73.6%  |
| Ours         | 59.1%  | 58.2%    | 58.6%  | 78.2%    | 77.0%    | 77.6%  |

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 V
QUANTITATIVE EVALUATION OF LOCALIZATION-BASED METHODS ON THE UCF_CC_50 DATASET USING PRECISION (P), RECALL (R), AND F-MEASURE (F)

| Method       | σ = 4  |          | σ = 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%  |

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. Table VI, and Table 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 SCAL-Net [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

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
TABLE VI
Comparison of the Counting Performance on the NWPU-Crowd. S0 ~ S4 respectively indicate five categories according to the different number range: 0, (0, 100], (100, 500], (500, 5000], > 5000

| Method          | Output Position Information | validation set | Test set | Scene Level (only MAE) |
|-----------------|-----------------------------|----------------|----------|------------------------|
|                 |                             | Overall ME    | Overall MAE | Avg. |
|                 |                             | MAE | MSE |       | MAE | MSE |       | S0 | S1 | S2 | S3 | 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 | 2058.5 |
| 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 | 1383.4 |
| BL [33]         | ×                           | 93.64 | 470.38 | 105.4 | 454.2 | 750.5 | 68.5 | 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 | 652.7 | 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 | 6.7 | 31.2 | 228.7 | 2075.8 |

RZLoc* [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 |

* means the localization-based methods, which can provide the position information.

TABLE VII
Comparison of the Counting Performance on the JHU-Crowd++, UCF-QNRF, ShanghaiTech Part A, Part B and UCF_CC_50 Datasets

| Method          | Output Position Information | JHU ME | QNRF ME | Part A ME | Part B ME | UCF_CC_50 ME |
|-----------------|-----------------------------|--------|---------|-----------|-----------|--------------|
| CSRNet [23]     | ×                           | 85.9   | 309.2   | -         | -         | -            |
| SFCN [52]       | ×                           | 77.5   | 297.6   | 102.0     | 171.4     | -            |
| L2SM [38]       | ×                           | -      | -       | 104.7     | 173.6     | 64.2         |
| CG-DRCN [44]    | ×                           | 82.3   | 328.0   | 112.2     | 176.3     | 64.0         |
| MUD-kNN [36]    | ×                           | -      | -       | 104.0     | 172.0     | 68.0         |
| DSSNet [27]     | ×                           | 133.5  | 416.5   | 99.1      | 159.2     | 60.6         |
| MBTBF [43]      | ×                           | 81.8   | 299.1   | 97.5      | 165.2     | 60.2         |
| BL [33]         | ×                           | 75.0   | 299.9   | 88.7      | 154.8     | 62.8         |
| RNN [60]        | ×                           | -      | -       | 88.1      | 143.7     | 55.9         |
| ASNet [18]      | ×                           | -      | -       | 91.6      | 159.7     | 57.8         |
| AMSSNet [14]    | ×                           | -      | -       | 101.8     | 163.2     | 56.7         |
| LibraNet [26]   | ×                           | -      | -       | 88.1      | 143.7     | 55.9         |
| KDMG [48]       | ×                           | 69.7   | 268.3   | 99.5      | 173.0     | 63.8         |
| NoisyCC [46]    | ×                           | 67.7   | 258.5   | 85.8      | 150.6     | 61.9         |
| DM-Count [49]   | ×                           | -      | -       | 85.6      | 148.3     | 59.7         |
| RZLoc* [25]     | ✔                           | -      | -       | 118.0     | 198.0     | 71.6         |
| PSDDN* [32]     | ✔                           | -      | -       | -         | -         | 65.9         |
| LSC-CNN* [41]   | ✔                           | 112.7  | 454.4   | 120.5     | 218.2     | 66.4         |
| Crowd-SDNet* [53] | ✔                         | -      | -       | -         | -         | 65.1         |
| AutoScale* [57] | ✔                           | 85.6   | 356.1   | 104.4     | 174.2     | 65.8         |
| TopoCount* [1]  | ✔                           | 60.9   | 267.4   | 89.0      | 159.0     | 61.2         |
| Ours*           | ✔                           | 66.6   | 253.6   | 89.0      | 153.5     | 57.0         |

* means the localization-based methods, which can provide the position information.

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 counting 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).
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).

We can see that the predicted 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 TRAN-COS [10] dataset for vehicle counting. We adopt the Grid Average Mean Absolute Error (GAME) [10] as the evaluation metric for vehicle counting, defined as:

\[
GAME(L) = \frac{1}{N} \sum_{i=1}^{N} \left( \sum_{l=1}^{4^L} |P_l^i - G_l^i| \right),
\]

which splits an input image into \(4^L\) non-overlapping sub-regions. \(N\) is the number of the testing images, \(P_l^i\) and \(G_l^i\) are the predicted and ground truth count of the \(i\)-th image, respectively. Table 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 \(\alpha\) and \(\beta\). Only changing the \(\alpha\) (resp. \(\beta\)), as shown in Fig. 7, as \(\alpha\) (resp. \(\beta\)) increasing (resp. decreasing), the response of FIDT map shows faster (resp. slower) decay in both foreground and background. As discussed in Section 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 \(\alpha\) and \(\beta\) results in Table VIII.
The effect of changing $\alpha$ and $\beta$ on the distribution of FIDT map.

Table IX

| Method | $\alpha$ | $\beta$ | 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 |
|        | 0.02     | 0.75    | 78.2% | 77.0% | 77.6% | 57.0 | 103.4 |
|        | 0.03     | 0.85    | 77.0% | 76.8% | 76.9% | 58.3 | 106.6 |

Table X

| Method | Localization ($\sigma = 8$) | Counting |
|--------|----------------------------|----------|
|        | P(%) | R(%) | F(%) | MAE | MSE |
| L2     | 73.5% | 76.3% | 74.9% | 62.1 | 108.8 |
| L2 + SSIM | 73.8% | 76.6% | 76.7% | 59.3 | 106.5 |
| L2 + I-SSIM (ours) | 78.2% | 77.0% | 77.6% | 57.0 | 103.4 |

Table XI

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

and $\beta$ settings experiments in Table IX. The following ablation study will provide experiments of choosing 0.02 and 0.75 in (3).

Effectiveness of I-SSIM loss: In this section, we explore the advantage of the proposed I-SSIM loss. Based on Table 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.

We further ablate the influence of the independent instance region sizes of I-SSIM loss on Part A. The following ablation study will provide experiments of choosing 0.02 and 0.75 in (3).

Table XII

| Threshold value | MAE | MSE |
|-----------------|-----|-----|
| 50/255          | X   | 95.6| 169.9|
| 70/255          | X   | 92.1| 167.0|
| 90/255          | X   | 80.6| 152.3|
| 100/255         | X   | 115.3| 206.4|
| 110/255         | X   | 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

| 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 | 2073.8 | 2895.2 |
| NoiseCC [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. 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.

experiments, we use $30 \times 30$ 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 $\frac{100}{255} \times \text{max}(M)$, where $\text{max}(M)$ is the max value of the predicted FIDT map. As shown in Table 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 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 $\frac{100}{255} \times \text{max}(M)$ 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]. Table XIII lists the results of some popular methods on the NWPU-Crowd’s 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 Table 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. Notably, 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 Table XV, we report the Multiply-Accumulate Operations (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] S. Abousamra, M. Hoai, D. Samaras, and C. Chen, “Localization in the crowd with topological constraints,” in Proc. AAAI Conf. Artif. Intell., 2021, pp. 872–881.

[2] C. Arteta, V. Lempiäski, and A. Zisserman, “Counting in the wild,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 483–498.

[3] D. B. Sam, S. Surya, and R. Venkatesh Babu, “Switching convolutional neural network for crowd counting,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., 2017, pp. 5744–5752.

[4] X. Cao, Z. Wang, Y. Zhao, and F. Su, “Scale aggregation network for accurate and efficient crowd counting,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 734–750.

[5] Z.-Q. Cheng, J.-X. Li, Q. Dai, X. Wu, and A. G. Hauptmann, “Learning spatial awareness to improve crowd counting,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 6152–6161.

[6] J. Gao, T. Han, Q. Wang, and Y. Yuan, “Domain-adaptive crowd counting via high-quality image translation and density reconstruction,” IEEE Trans. Neural Netw. Learn. Syst., pp. 1–13, 2021.

[7] J. Gao, T. Han, Y. Yuan, and Q. Wang, “Learning independent instance maps for crowd localization,” 2020, arXiv:2012.04164.

[8] J. Gao et al., “C³ framework: An open-source pytorch code for crowd counting,” 2019, arXiv:1907.02724.

[9] J. Gao, Q. Wang, and Y. Yuan, “Sscar: Spatial- channel-wise attention regression networks for crowd counting,” Neurocomputing, vol. 363, pp. 1–8, 2019.

[10] R. Guerrero-Gómez-Olmedo, B. Torre-Jiménez, R. López-Sastre, S. Maldonado-Bascón, and D. Oneru-Rubio, “Extremely overlapping vehicle counting,” in Proc. Iberian Conf. Pattern Recognit. Image Anal., 2015, pp. 423–431.

[11] T. Han et al., “LDC-Net: A unified framework for localization, detection and counting in dense crowds,” 2021, arXiv:2110.04727.

[12] Z. Hayder, X. He, and M. Salzmann, “Boundary-aware instance segmentation,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., 2017, pp. 5696–5704.

**TABLE XIV**

| Method     | Density map (creating by integration) | FIDT map (creating by localization) |
|------------|---------------------------------------|-------------------------------------|
|            | MAE | MSE | MAE | MSE | P(%) | R(%) | P(%) |
| CSRNET [23]| 104.9 | 433.5 | 100.6 | 464.3 | 68.8% | 66.2% | 67.5% |
| CSRNET [23] + PPN | 95.6 | 470.4 | 91.2 | 472.3 | 73.2% | 69.6% | 71.8% |
| BL [33]    | 87.8 | 412.9 | 85.4 | 411.9 | 73.8% | 70.8% | 73.7% |

**TABLE XV**

| 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† [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%     |

We 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.
Dingkang Liang is currently working toward the Ph.D. degree with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China. He was a Reviewer of several top journals and conferences, such as TPAMI, TIP, CVPR, ICCV, and ECCV. His research focuses on computer vision, especially for crowd analysis and 3D object detection.

Wei Xu is currently working toward the M.S. degree in information and communication engineering from the 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 the Huazhong University of Science and Technology, Wuhan, China, in 2018. In 2021, she joined the Huazhong University of Science and Technology, as an Engineer. Her research interests include computer vision and machine learning.

Yu Zhou (Member, IEEE) received the M.S. and Ph.D. degrees in electronics and information engineering from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2010 and 2014, respectively. From 2014 to 2016, he was a Postdoctoral Researcher with the Beijing University of Posts and Telecommunications, Beijing, China, where he was 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 automatic drive.