# Localization in Crowds with Loosen-structured Pair-of-points Loss

ID: 931556

***Abstract*—Phantom dots are common spatial semantic errors in crowd localization and counting. We attribute these semantic errors to the absence of bounding boxes and deal with them through adjacent nodes. One approach to solving this problem involves constraining the point and its neighbor point set simultaneously. However, learning two point sets that satisfy the surjective relation is challenging because predicting a point set is an open-set problem. In this paper, we propose a loosen-structured pair-of-points loss, making predicting a point set and its neighbor point set simultaneously possible. On the other hand, to suppress the probability of points far away from the target points, we compose the connected component map, which is free with head size. Without bells and whistles, experiment results on several public datasets reach the state-of-the-art.**

***Index Terms*—crowd counting and localization, computer vision, phantom dots, loosen-structured pair-of-points loss, connected component map**


Fig. 1. Some examples of phantom dots. Red and green dots are predicted and ground-truth points, respectively.

## I. Introduction

Crowd counting aims at counting the number of the target objects in the image. Counting approaches f categorized into regression-based [1]–[6] and detection-based [7]–[12] according to the count way. Regression-based methods predict a density map and sum it up as the total number without considering each object's detailed information, such as the localization. The detection-based method offers the localization of each object, which is crucial in many crowd analysis tasks, such as pedestrian detection [13], pedestrian behavior analysis [14], etc. Unfortunately, the detect-then-count approach needs the object bounding boxes labels and is challenging to cope with the severe crowd occlusion scene.

Lempitsky et al. [15] set a Gaussian kernel on each annotated person dot to generate an intermediate representation, the density map. The density map has been widely adopted in the field of crowd counting [1]–[4], [6], [16]–[21] for its excellent performance in the dense crowd. A well-known problem with density-map regression-based methods is that it does not take into account the density map's sensitiveness to head size. Li et al. [2] produce pseudo head size through neighbor distance for a head size adaptive density map. The works of wan et al. [22], [23] show the importance of intermediate representation, and an adaptive density map generator is proposed. Wang et al. [24] theoretically prove the gaussian blurred dot map will hurt the generalization performance of counting. Besides, another primary defect of density map regression method is it failure to offer detailed information of each object, such as localization, bounding box, etc.

Another line of work, detect-then-count, has caused increasing attention because it can offer crucial information, accurate individual localization, etc. The foremost problems of these works are the fact that the absence of bounding box labels [16], [25], [26]. Approaches of [4], [12], [18] introduce detection framework into crowd counting through generating pseudo bounding boxes. Wan et al. [27] and Abousamra et al. [28] achieve localization by post-processing to avoid the need of bounding box labels. For example, [27] set a maximum pooling on the predicted density map and utilize a threshold to lock the person's location. However, this kind of post-processing is sensitive to the size of pooling size and the threshold. Carion et al. [29] utilize the Hungarian algorithm to transform the bounding box regression problem into an open-set issue, directly predicting a bounding box set. Continuing with the idea of [29], Song et al, [30] predict a set of head point set and achieve the best performance for both counting and localization. However, duplicate points maybe positioned to the same object, called phantom dots, because of the constraint absence of the bounding box, as displayed in Fig. 1.

This paper aims to develop a feasibility method for suppressing these phantom dots. Intuitively, we can constrain both the target point and its neighbors. However, since predicting a point set is an open set problem, it is challenging to predict two point sets whose association relation is surjective. Therefore, we relax the set of neighbor points formed by definite positions into an infinite set of relative bearings and relative distances. In addition, we use the nearest neighbor distance to construct the connection component, which is used to suppress points far away from the target position. The key contribution of this work are presented as follows:

- We design a loosen-structured pair-of-points loss (Sec. II-B) to constrain each point and its nearest neighbor

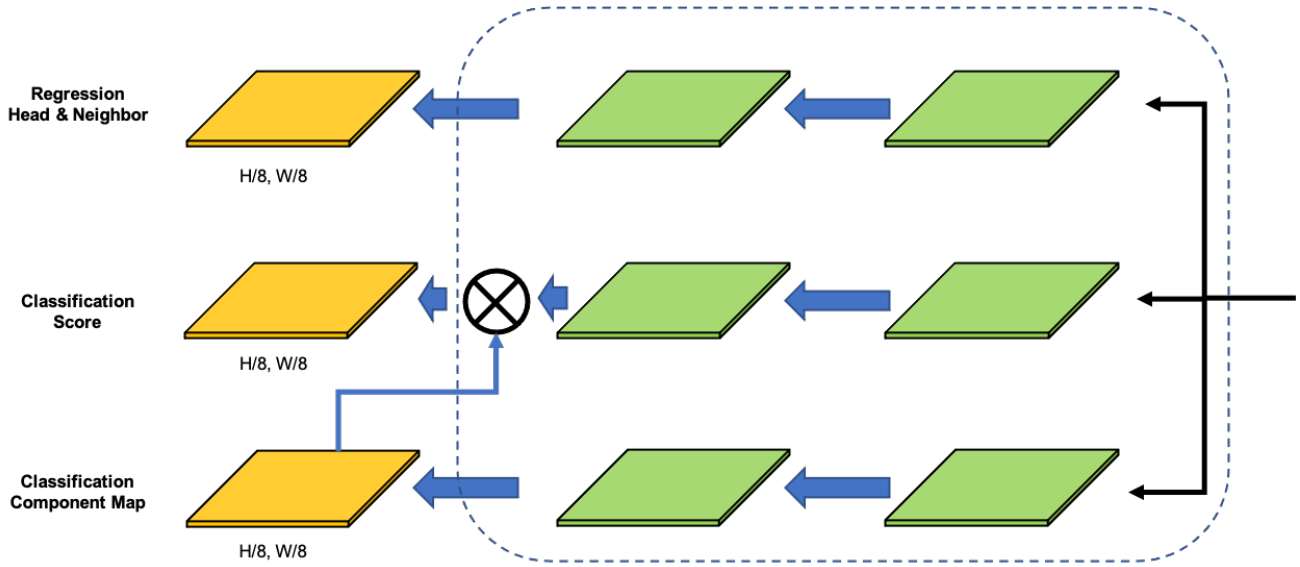



Fig. 2. The architecture of the decoder. The input is the learned feature maps, going through three branches to regress head points, classification score, and component map, respectively.

simultaneously to suppress the generation of phantom dots.

- We utilize the connected component map (Sec. II-C) to suppress the probability of points far away from the target points, which is free with head size.

Experiments on several public datasets show the effectiveness of our method.

## II. Our Approach

In this section, we introduce our method in detail. The following part includes the framework, the loosen-structured pair-of-points loss, and the connected component.

### A. Framework

Our work is based on P2Pnet [30], which is an encoder-decoder framework. It first extracts features from VGG16 and uses a pyramid structure to enrich the representation of the feature map. Then the learned feature map is used to regress head localizations and classification scores. Taking the input in resolution of $H \times W$, the feature map $\frac{1}{8}H \times \frac{1}{8}W$, the number of the predicted head locations is $M = k \times \frac{1}{8}H \times \frac{1}{8}W$. The k is a pre-defined foreign parameter to ensure the predicted number of people is more than the actual number T. The Hungarian algorithm is then adopted to choose the top T predicted head locations and constrain them through a $||\cdot||_2$ distance loss and a cross-entropy score loss.

To highlight our contributions, we only display the difference part from P2Pnet in Fig. 2. The differences include convolution layers in the decoder, a new branch classification component, the output component multiplied with the classification score, and the new output of neighbor point set in the regression branch. We utilize the predicted component to decrease the probability of the point without objects. The neighbor point set combines the point set to generate the loosen-structured pair point set representation. The convolutional layers in the decoder are replaced with layers with a big kernel, according to the big kernel's good performance as displayed in [31]. The green blocks in Fig. 2 are with different convolutional layer in [30], and the detail is shown in Fig. 3. Each convolution with a big kernel assisted with a convolution with a small kernel for better performance, as recommended by [31]. Notable, only the kernel size is changed.

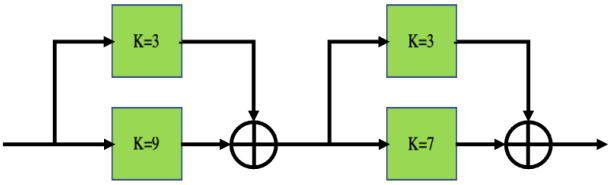



Fig. 3. Structure of the two layers of green blocks shown in Fig. 2. Text on each block represents its kernel size.

### B. Loosen-structured Pair-of-points Loss

The main objective is to investigate methods to suppress phantom dots. The intuitive idea is to simultaneously predicting two point sets which satisfies the surjective relation. However, considering it is an open-set problem to predict a point set, it is arduous to simultaneously predict two point sets with strict relations. Therefore, we loosen the pair of points' relation by representing the neighbor point with the relative orientation and relative distance.

Fig. 4 geometry illustrates the probability change with different loss. Green points are target location and other colored points are predicted points. We use color to indicate each predicted point's probability to be considered as a person's localization. The point has a higher possibility when its color tends to be red. As (a) displayed, $p_1$ and $p_2$ have a higher score

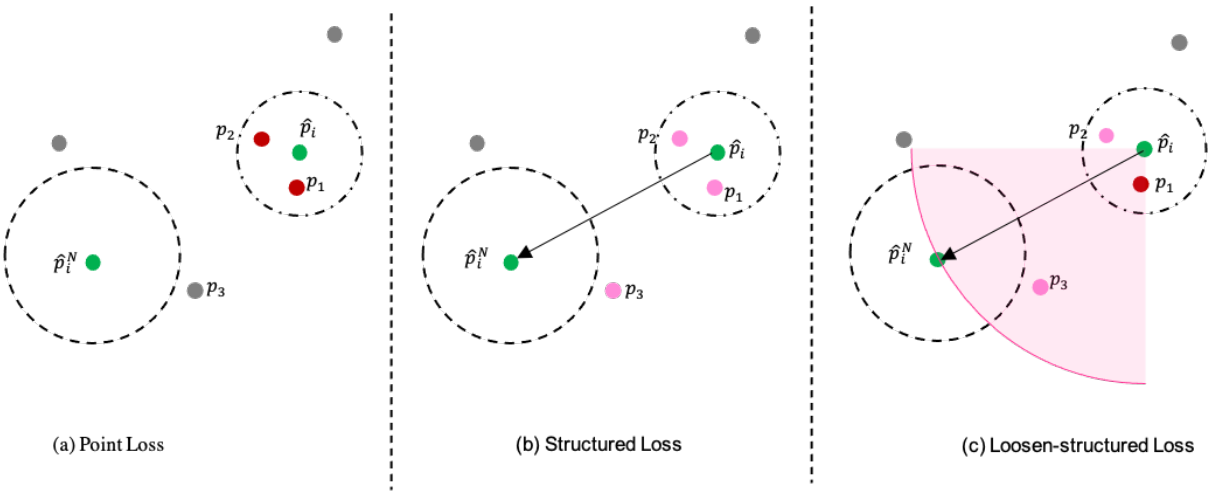



Fig. 4. Geometry illustration of points probability change with different losses. Dots of crossed lines are ground truth points, others are predicted points. Green line directs to nearest point.

for their tiny distance error to $p_i$, which cause the phantom dots. While these two points' score decrease because the their neighbor point's huge distance error. Besides, the score of $p_3$ in (b) is improved when considering its neighbor's cost at the same time. Thus, the structured loss function is not conducive to the convergence of the model. When adopt loosen-structured loss, $p_3$'s score is enhancement, and $p_1$'s score is higher than $p_2$. Because, the neighbor's relative orientation loss of $p_1$ is smaller than $p_2$. The decrease of $p_2$, along with the increase of $p_3$, shows the loosen-structured loss has a positive effect on suppressing phantom dots.

Given an image with T individuals in an image, let $\hat{\mathcal{P}} = \{\hat{p}_i | i \in \{1, 2, ..., \mathrm{T}\}\}$ represents the annotated head center point set. The model outputs point set $\mathcal{P} = \{p_i | i \in \{1, 2, ..., \mathrm{M}\}\}$. We use Euler distance $||\,\boldsymbol{.}\,||_2$ to measure the difference between $\hat{P}$ and $P$. We represent point sets difference with point loss,

$$\mathcal{L}_{\text{point}} = ||\mathcal{P}, \hat{\mathcal{P}}||_2, \tag{1}$$

with structured loss,

$$\mathcal{L}_{\text{point}} = ||\mathcal{P}, \hat{\mathcal{P}}||_2 + ||\mathcal{P}^{\mathrm{N}}, \hat{\mathcal{P}}^{\mathrm{N}}||_2, \tag{2}$$

and with loosen-structured loss,

$$\mathcal{L}_{\text{point}} = ||\mathcal{P}, \hat{\mathcal{P}}||_2 + C(\mathcal{R}, \hat{\mathcal{R}}) + C(\mathbb{I}(\overrightarrow{\mathcal{P}\mathcal{P}^{\mathrm{N}}}), \mathbb{I}(\overrightarrow{\hat{\mathcal{P}}\hat{\mathcal{P}}^{\mathrm{N}}})), \tag{3}$$

where $C \in \mathbb{R}^{\mathrm{M}\times\mathrm{T}}$, and $\mathcal{P}^{\mathrm{N}}$ and $\hat{\mathcal{P}}^{\mathrm{N}}$ indicate the point set consist of the neighbor of point set $\mathcal{P}$ and $\hat{\mathcal{P}}$, respectively. It is a remarkable fact that we normalize $C(\mathcal{P}, \hat{\mathcal{P}})$ with $\frac{\hat{\mathcal{R}}}{2}$,

$$\hat{\mathcal{R}} = ||\hat{\mathcal{P}}, \hat{\mathcal{P}}^{\mathrm{N}}||_2, \tag{4}$$

to decrease to impact of various head size. We set $\mathcal{R}$ and $\hat{\mathcal{R}}$ to the distance of predicted pair points and ground-truth pair points, respectively. The term of $C(\mathcal{R}, \hat{\mathcal{R}})$ means the cost of relative distance, and the $C(\mathbb{I}(\overrightarrow{\mathcal{P}\mathcal{P}^{\mathrm{N}}}), \mathbb{I}(\overrightarrow{\hat{\mathcal{P}}\hat{\mathcal{P}}^{\mathrm{N}}}))$ indicates the cost of relative orientation. We also use the $||\,\boldsymbol{.}\,||_2$ to calculate the distance between point and its neighbor. Moreover, to ensure the dominance of the predicted point's accuracy, we normalize the relative distance,

$$C(\mathcal{R}, \hat{\mathcal{R}}) = \frac{||\mathcal{R}, \hat{\mathcal{R}}||_2}{\hat{\mathcal{R}}}, \tag{5}$$

and pixel-wise weight each relative orientation,

$$C(\mathbb{I}(\overrightarrow{\mathcal{P}\mathcal{P}^{\mathrm{N}}}), \mathbb{I}(\overrightarrow{\hat{\mathcal{P}}\hat{\mathcal{P}}^{\mathrm{N}}})) = \mathcal{W} \times (\mathbb{I}(\overrightarrow{\mathcal{P}\mathcal{P}^{\mathrm{N}}}) \oplus \mathbb{I}(\overrightarrow{\hat{\mathcal{P}}\hat{\mathcal{P}}^{\mathrm{N}}})), \tag{6}$$

where $\mathbb{I}$ represents the sign function, and $\oplus$ means the logical relation XOR. The weight $\mathcal{W}$ is pixel-wise and is related to $||\mathcal{P}\hat{\mathcal{P}}||_2$. We distinguish each predicted point according to the average cost of $C(\mathcal{P}, \hat{\mathcal{P}})$,

$$\mathcal{U} = \frac{1}{\mathrm{T} \times \mathrm{M}} \sum_{\mathrm{i}=1}^{\mathrm{M}} \sum_{j=1}^{\mathrm{T}} c(p_i, \hat{p}_j), \tag{7}$$

where $c(p_i, \hat{p}_j)$ represents the Euler distance between point $p_i$ and point $\hat{p}_j$. When the $c(x_i, y_i)$ is small than $\mathcal{U}$, we punishment its neighbor's relative orientation with a small weight, and vice versa. The weight $w$ can be represented as

$$w = \begin{cases} 0.1, & c(p_i, \hat{p}_j) \leqslant \mathcal{U} \\ 1. & c(p_i, \hat{p}_j) > \mathcal{U} \end{cases} \tag{8}$$

### C. Connected Component Map

We borrow the idea of the connected component map from [28], which is free with head size. The aims here are to distinguish each component's sphere from each other and pre-judgment whether the point contains a person or not. Some examples of the component map are shown in Fig. 5. The location in the non-component map area will not be considered as a position containing people. The area of each component is a square with a radius of $r$,

$$r = min(\lceil \hat{R}/2 \rceil, 7), \tag{9}$$

where 7 is considered as the upper bound according to [28]. We constrain the component map with MSELoss,

$$\mathcal{L}_{\text{component}} = ||\mathrm{M} - \hat{\mathrm{M}}||_2, \tag{10}$$

where M and $\hat{\mathrm{M}}$ means the predicted and ground truth of the connected component map , respectively.

### D. Total Loss

Same as [30], this work utilize Entropy loss $\mathcal{L}_{\text{cls}}$,

$$\mathcal{L}_{\text{cls}} = -\frac{1}{\mathrm{M}}\{\sum_{\mathrm{i=1}}^{\mathrm{T}} log\hat{c}_i + \lambda_1 \sum_{\mathrm{i=T+1}}^{\mathrm{M}} log(1-\hat{c}_i)\} \tag{11}$$

to supervise the proposal classification scores. We adopt the grid layout to ensure the number of the reference point is more than the number of the target point and set k=4. In equation 11, $\lambda_1$ is a re-weight factor for negative proposals. The total loss function $\mathcal{L}$ is the summation of the the above three losses, which is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_2(\mathcal{L}_{\text{loc}} + \mathcal{L}_{\text{component}}), \tag{12}$$

where $\lambda_2$ is a weight term to balance the effect of neighbor points and connected component.

## III. Experiments

### A. Implementation Details

**Dataset.** We mainly tested the effectiveness of the proposed approach by carrying thorough experiments on existing public crowd counting datasets. Specifically, extensive experiments are conducted on four challenging datasets, including ShanghaiTech [16], ShanghaiTechRGBD [32], and $\text{UCF}_{\text{CC\_50}}$ [25]. SHA and SHB are two parts of ShanghaiTech. Generally speaking, the crowd density of the three datasets in descending order is $\text{UCF}_{\text{CC\_50}} >$ SHA $>$ SHB. We show an example from SHA, SHB, and $\text{UCF}_{\text{CC\_50}}$ in Fig. 5. Our component map is related to the neighbor points distance which is sensitive to the density. Therefore, ShanghaiTechRGBD with bounding box labels is introduced to verify the effectiveness of the component map.

**Data Augmentations & Hyper-parameters.** This work is based on P2Pnet [30]. For a fair comparison, we used the exact same data augmentations and hyper-parameters as [30]. The only difference is that we increase the initial learning rate to 1e-4. Besides, we fixed the random seed to ensure the reproducibility of the experiment.

### B. Performance Evaluation

The output of our model includes two parts, the predicted number of people in the image and the location of each individual. We use Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) to evaluate counting and apply nAP, proposed by [30] to evaluate localization. The nAP takes the circle as the target domain, centered at the annotated person and radius at a distance to its neighbors. When the predicted point falls in the target domain, it is recorded as a true positive point (TP); otherwise, a false positive point (FP). Then the nAP is calculated following the common practice in [33].

TABLE I
COMPARISON WITH STATE-OF-THE-ART CROWD COUNTING METHODS ON SHANGHAITECH AND $\text{UCF}_{\text{CC\_50}}$.

| method | SHA | | SHB | | $\text{UCF}_{\text{CC\_50}}$ | |
|---|---|---|---|---|---|---|
| | MAE | RMSE | MAE | RMSE | MAE | RMSE |
| SDANet [34] | 63.6 | 101.8 | 7.8 | 10.2 | 227.6 | 316.4 |
| ADSCNet [35] | 55.4 | 97.7 | 6.4 | 11.3 | 198.4 | 267.3 |
| ASNet [36] | 57.78 | 90.13 | - | - | 174.84 | 251.63 |
| AMRNet [5] | 61.59 | 98.36 | 7.02 | 11.00 | 184.0 | 265.8 |
| AMSNet [37] | 56.7 | 93.4 | 6.7 | 10.2 | 208.4 | 297.3 |
| DM-Count [24] | 59.7 | 95.7 | 7.4 | 11.8 | 211.0 | 291.5 |
| TopoCount [28] | 61.2 | 104.6 | 7.8 | 13.7 | 184.1 | 258.3 |
| Resnet-DC [9] | 73.51 | 118.1 | 13.3 | 22.5 | 254.78 | 326.16 |
| P2PNet [30] | 52.74 | 85.06 | 6.25 | 9.9 | **172.72** | **256.18** |
| ours | **50.82** | **82.52** | **5.96** | **9.78** | 175.94 | 288.32 |

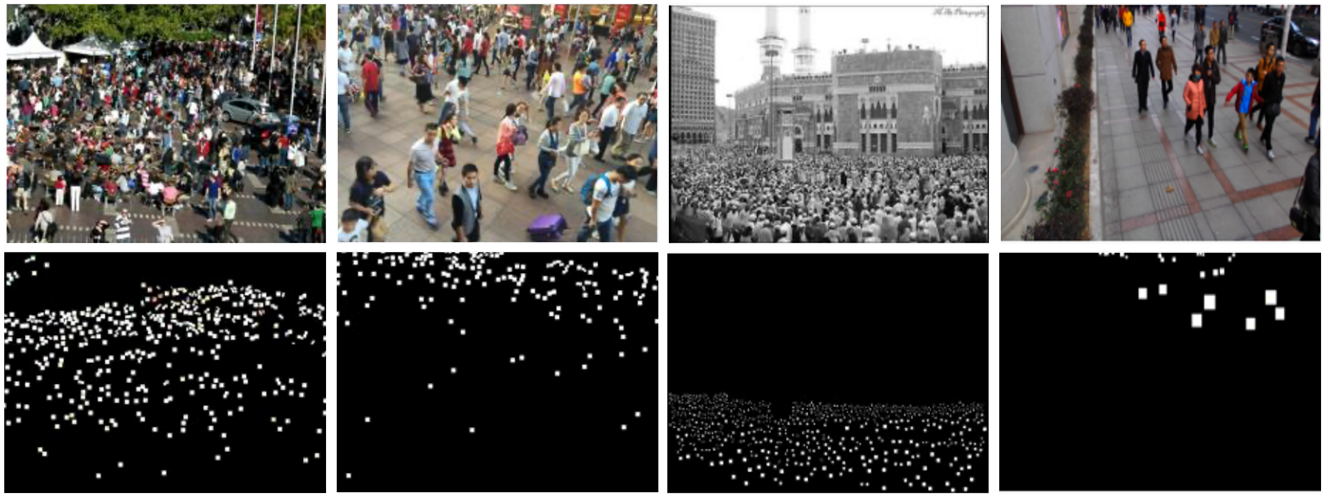
TABLE II
COMPARISON WITH STATE-OF-THE-ART CROWD COUNTING METHODS ON SHANGHAITECHRGBD.

| method | ShanghaiTechRGBD | |
|---|---|---|
| | MAE | RMSE |
| CSR [2] | 4.91 | 7.11 |
| RDNet [38] | 4.96 | 7.22 |
| AGD [18] | 4.18 | 6.75 |
| CSR+IDAM [39] | 4.38 | 7.06 |
| ours | **3.95** | **5.72** |

### C. Results

In Tabel I and Tabel II, we compare our results to those of the method that returns the best results for each one of the 4 public datasets, as currently reported in the literature. They are those of [30], [30], [30], and [18], respectively. Our method ranking first on both ShanghaiTech and ShanghaiTechRGBD. Take SHA as an example, our methods are superior than [30] and decrease 5.3% and 5.27% in terms of MAE and RMSE, respectively. However, on $\text{UCF}_{\text{CC\_50}}$, the counting error of our model is higher than that of P2Pnet. We consider that this is caused by the difficulty of convergence of the pair-of-points loss due to the high crowd density in $\text{UCF}_{\text{CC\_50}}$. For example, when the crowd is too dense, the relative positions of the pair of points will fluctuate significantly due to slight deviations from the predicted points.

Tabel III prints the nAP of different models on SHA. Identical to [30], we use the average distance of the three nearest neighbors to the center point $d$ as the target area. Besides, the $\delta$ is introduced to control the distance error range. The smaller the $\delta$, the higher the precision, and vice versa. When the distance of the predicted point to the target points is smaller than $d * \delta$, the precited point is considered as a true positive point; otherwise false positive point. The column of P2Pnet is calculated according to the released trained checkpoint of [30]. Compared with P2Pnet, the nAP of our method is improved by 75.1%, 14.1%, and 1.9%, respectively, when $\delta$ increases from 0.25 to 0.5.



(a) SHA (b) SHB (c) $\text{UCF}_{\text{CC\_50}}$ (d) RGBD

Fig. 5. Some samples from SHA, SHB, $\text{UCF}_{\text{CC\_50}}$, and ShanghaiTechRGBD. The upper row are images and the lower row are correspondence component map.

TABLE III
THE NAP OF DIFFERENT MODELS ON SHA.

| $\text{nAP}_\delta$ | P2Pnet | Base | Loosen-structured Loss | Component Map | Ours |
|---|---|---|---|---|---|
| $\delta = 0.05$ | 3.46% | 6.15% | 6.72% | 4.97% | 6.06% |
| $\delta = 0.25$ | 54.52% | 61.89% | 62.38% | 59.45% | 62.19% |
| $\delta = 0.5$ | 86.25% | 87.39% | 87.76% | 87.73% | 87.95% |
| $\delta = \{0.05 : 0.05 : 0.50\}$ | 52.67% | 57.54% | 58.18% | 55.97% | 57.95% |

TABLE IV
RESULTS OF MODEL COMBINED WITH COMPONENT MAP OR LOOSEN-STRUCTURED PAIR POINTS LOSS.

| component | loosen-structure | SHA | |
|---|---|---|---|
| | | MAE | RMSE |
| | | 53.9 | 86.71 |
| | √ | 51.64 | 83.63 |
| √ | | 51.14 | 81 |
| √ | √ | **49.94** | **80.57** |

TABLE V
THE NAP OF DIFFERENT MODELS ON RGBD.

| $\text{nAP}_\delta$ | Base | Component Map |
|---|---|---|
| $\delta = 0.05$ | 6.29% | 13% |
| $\delta = 0.25$ | 63.59% | 85.31% |
| $\delta = 0.5$ | 91.7% | 97.41% |
| $\delta = \{0.05 : 0.05 : 0.50\}$ | 59.82% | 74.88% |

## IV. ANALYSIS AND ABLATIONS

We validate the effectiveness of the loosen-structured pair-of points loss and component map on SHA. According to [30], $\lambda_1$ and $\lambda_2$ are set to 0.0002.

### A. Effectiveness of Loosen-structured Pair-of-Points Loss

Table III and Table IV display the localization precision and counting error of models combining with or without loosen-structured pair-of-points loss or component map, respectively. The Base column in Table III means the model consists of convolutions with big kernel sizes. The subsequent two columns represent the base model combined with the Loosen-structured Loss and the Component Map. When introducing the loosen-structured loss, the the overall average precision $\text{nAP}_{\{0.05:0.05:0.5\}}$ is improved about $1.11\%$ and counting error (MAE) is reduced about 5.12%.

### B. Effectiveness of Component Map

Tabel III $\sim$ IV indicates that when combined with the component map, the localization precision ($\text{nAP}_{\{\delta=0.05\}}$) reduced round 19.18%, but the counting accuracy is improved about 5.12%. We assume this is mainly caused by the low-fidelity component map on SHA, as shown in (a) of Fig. 5. Since the component map is related to the distance of the adjacent head and has an upper limit of 7, the head component map close to the camera is smaller than the actual map in most cases. The component map reduces the background noise while suppressing some points falling into the head area beyond the component map. In particular, we compare the localization precision of models with and without the component map on the RGBD dataset, which offers the head bounding boxes. Results in Table V show that when the component map approximates the head bounding boxes is favorable to the localization precision.

## V. Conclusion

This work dedicates to dealing with the phantom dots when localizing each individual in the crowd. We proposed a loosen-structured pair-of-points loss to constrain the target point and its nearest neighbor. Without bells and whistles, the proposed loosen-structured pair-of-points loss improves the counting accuracy and localization precision simultaneously. Besides, our work demonstrates the positive effect on counting accuracy but the reverse effect on localization accuracy of the component map when free with head size. Especially when replacing the component map to the head mask with actual head size, it brings a noticeable improvement in localization.

## Acknowledgment

## References


[1] D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 4031–4039. 1

[2] Y. Li, X. Zhang, and D. Chen, "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1091–1100. 1, 4

[3] L. Zhu, Z. Zhao, C. Lu, Y. Lin, Y. Peng, and T. Yao, "Dual path multi-scale fusion networks with attention for crowd counting," *arXiv preprint arXiv:1902.01115*, 2019. 1

[4] D. Guo, K. Li, Z.-J. Zha, and M. Wang, "Dadnet: Dilated-attention-deformable convnet for crowd counting," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1823–1832. 1

[5] X. Liu, J. Yang, W. Ding, T. Wang, Z. Wang, and J. Xiong, "Adaptive mixture regression network with local counting map for crowd counting," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*. Springer, 2020, pp. 241–257. 1, 4

[6] W. Wang, Q. Liu, and W. Wang, "Pyramid-dilated deep convolutional neural network for crowd counting," *Applied Intelligence*, pp. 1–13, 2021. 1

[7] D. Kong, D. Gray, and H. Tao, "A viewpoint invariant approach for crowd counting," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 3. IEEE, 2006, pp. 1187–1190. 1

[8] H. Fu, H. Ma, and H. Xiao, "Real-time accurate crowd counting based on rgb-d information," in *2012 19th IEEE International Conference on Image Processing*. IEEE, 2012, pp. 2685–2688. 1

[9] J. Zhang, S. Chen, S. Tian, W. Gong, G. Cai, and Y. Wang, "A crowd counting framework combining with crowd location," *Journal of advanced transportation*, vol. 2021, 2021. 1, 4

[10] W. Ge and R. T. Collins, "Marked point processes for crowd counting," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 2913–2920. 1

[11] Q. Wang, J. Gao, W. Lin, and X. Li, "Nwpu-crowd: A large-scale benchmark for crowd counting and localization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 6, pp. 2141–2149, 2020. 1

[12] D. B. Sam, S. V. Peri, M. N. Sundararaman, A. Kamath, and R. V. Babu, "Locate, size, and count: accurately resolving people in dense crowds via detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 8, pp. 2739–2751, 2020. 1

[13] E. Foxlin, "Pedestrian tracking with shoe-mounted inertial sensors," *IEEE Computer graphics and applications*, vol. 25, no. 6, pp. 38–46, 2005. 1

[14] E. Papadimitriou, G. Yannis, and J. Golias, "A critical assessment of pedestrian behaviour models," *Transportation research part F: traffic psychology and behaviour*, vol. 12, no. 3, pp. 242–255, 2009. 1

[15] V. Lempitsky and A. Zisserman, "Learning to count objects in images," *Advances in neural information processing systems*, vol. 23, 2010. 1

[16] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 589–597. 1, 4

[17] M. Liu, J. Jiang, Z. Guo, Z. Wang, and Y. Liu, "Crowd counting with fully convolutional neural network," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 953–957. 1

[18] X. Pan, H. Mo, Z. Zhou, and W. Wu, "Attention guided region division for crowd counting," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 2568–2572. 1, 4

[19] H. Mo, W. Ren, Y. Xiong, X. Pan, Z. Zhou, X. Cao, and W. Wu, "Background noise filtering and distribution dividing for crowd counting," *IEEE Transactions on Image Processing*, vol. 29, pp. 8199–8212, 2020. 1

[20] Y.-B. Liu, R.-S. Jia, Q.-M. Liu, X.-L. Zhang, and H.-M. Sun, "Crowd counting method based on the self-attention residual network," *Applied Intelligence*, vol. 51, no. 1, pp. 427–440, 2021. 1

[21] B. Chen, Z. Yan, K. Li, P. Li, B. Wang, W. Zuo, and L. Zhang, "Variational attention: Propagating domain-specific knowledge for multi-domain learning in crowd counting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 16 065–16 075. 1

[22] J. Wan and A. Chan, "Adaptive density map generation for crowd counting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1130–1139. 1

[23] J. Wan, Q. Wang, and A. B. Chan, "Kernel-based density map generation for dense object counting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1

[24] B. Wang, H. Liu, D. Samaras, and M. H. Nguyen, "Distribution matching for crowd counting," *Advances in neural information processing systems*, vol. 33, pp. 1595–1607, 2020. 1, 4

[25] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2547–2554. 1, 4

[26] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, and M. Shah, "Composition loss for counting, density map estimation and localization in dense crowds," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 532–546. 1

[27] J. Wan, Z. Liu, and A. B. Chan, "A generalized loss function for crowd counting and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1974–1983. 1

[28] S. Abousamra, M. Hoai, D. Samaras, and C. Chen, "Localization in the crowd with topological constraints," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 872–881. 1, 3, 4

[29] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229. 1

[30] Q. Song, C. Wang, Z. Jiang, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and Y. Wu, "Rethinking counting and localization in crowds: A purely point-based framework," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3365–3374. 1, 2, 4, 5

[31] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31x31: Revisiting large kernel design in cnns," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 963–11 975. 2

[32] D. Lian, J. Li, J. Zheng, W. Luo, and S. Gao, "Density map regression guided detection network for rgb-d crowd counting and localization," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 4

[33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755. 4

[34] Y. Miao, Z. Lin, G. Ding, and J. Han, "Shallow feature based dense attention network for crowd counting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 11 765–11 772. 4

[35] S. Bai, Z. He, Y. Qiao, H. Hu, W. Wu, and J. Yan, "Adaptive dilated network with self-correction supervision for counting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4594–4603. 4

[36] X. Jiang, L. Zhang, M. Xu, T. Zhang, P. Lv, B. Zhou, X. Yang, and Y. Pang, "Attention scaling for crowd counting," in *Proceedings of the*

*IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4706–4715. 4

[37] Y. Hu, X. Jiang, X. Liu, B. Zhang, J. Han, X. Cao, and D. Doermann, "Nas-count: Counting-by-density with neural architecture search," in *European Conference on Computer Vision*. Springer, 2020, pp. 747–766. 4

[38] D. Lian, J. Li, J. Zheng, W. Luo, and S. Gao, "Density map regression guided detection network for rgb-d crowd counting and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1821–1830. 4

[39] L. Liu, J. Chen, H. Wu, G. Li, C. Li, and L. Lin, "Cross-modal collaborative representation learning and a large-scale rgbt benchmark for crowd counting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4823–4833. 4