

Mahalanobis Distance-based Multi-view Optimal Transport for Multi-view Crowd Localization

Qi Zhang¹, Kaiyi Zhang^{2,1}, Antoni B. Chan³, and Hui Huang^{1*}

¹ Visual Computing Research Center, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China

² Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), China

³ Department of Computer Science, City University of Hong Kong, HK SAR, China
qi.zhang.opt@gmail.com, zhangkaiyi2022@email.szu.edu.cn,
abchan@cityu.edu.hk, hhzhiyan@gmail.com

Abstract. Multi-view crowd localization predicts the ground locations of all people in the scene. Typical methods usually estimate the crowd density maps on the ground plane first, and then obtain the crowd locations. However, existing methods’ performances are limited by the ambiguity of the density maps in crowded areas, where local peaks can be smoothed away. To mitigate the weakness of density map supervision, optimal transport-based point supervision methods have been proposed in the single-image crowd localization tasks, but have not been explored for multi-view crowd localization yet. Thus, in this paper, we propose a novel Mahalanobis distance-based multi-view optimal transport (M-MVOT) loss specifically designed for multi-view crowd localization. First, we replace the Euclidean-based transport cost with the Mahalanobis distance, which defines elliptical iso-contours in the cost function whose long-axis and short-axis directions are guided by the view ray direction. Second, the object-to-camera distance in each view is used to adjust the optimal transport cost of each location further, where the wrong predictions far away from the camera are more heavily penalized. Finally, we propose a strategy to consider all the input camera views in the model loss (M-MVOT) by computing the optimal transport cost for each ground-truth point based on its closest camera. Experiments demonstrate the advantage of the proposed method over density map-based or common Euclidean distance-based optimal transport loss on several multi-view crowd localization datasets.

Keywords: Multi-view · Optimal transport · Crowd localization

1 Introduction

Multi-view crowd localization [10, 11, 26, 31] has been proposed to predict the people’s locations on the ground of the scene, which could be used for applications like crowd analysis, autonomous driving, public traffic management, *etc.*

* Corresponding author

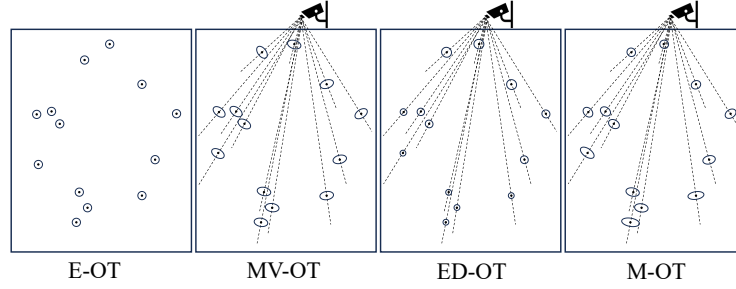


Fig. 1: The comparison of different optimal transport (OT) losses on the ground plane. The original OT uses Euclidean distance cost (E-OT) and treats all deviations from the ground-truth equally. MV-OT uses the view ray direction to the camera to change the cost function, ED-OT considers the camera distance in E-OT, while M-OT considers both the influence of the view ray direction and the distance to the camera.

It fuses multi-camera information via feature extraction and projection of each camera view into a common ground plane, followed by multi-view fusion and decoding. Current methods mainly rely on fixed-size Gaussian-kernel density maps as supervision to train the multi-view crowd localization models. The people locations are further extracted after a local non-maximum suppression (NMS) as a post-processing step on the estimated crowd density maps, where the pixel-wise mean square error (MSE) loss between the predicted and ground-truth density maps is adopted to train the whole model. However, these methods' localization performance is limited by the ambiguity of the density maps in crowded areas, where there may not be clear density peaks over each person for localization.

To mitigate the problems of Gaussian density map supervision in the single-image crowd localization task, point-supervision methods based on optimal transport (OT) [23, 34] have been proposed and achieved significant gains in localization performance compared to Gaussian density map methods trained with MSE loss. OT loss directly uses point annotations as supervision and generates more compact density maps [34]. However, point supervision for multi-view crowd localization methods has not been explored yet.

In this paper, we explore using point supervision for multi-view crowd localization and propose a novel Mahalanobis distance-based multi-view optimal transport (M-MVOT) loss. In the M-MVOT loss, the transport cost is defined using the Mahalanobis distance, which adjusts the cost based on the view-ray direction between the ground-truth point and the camera, and the distance to the camera. Specifically, the Mahalanobis distance defines elliptical iso-contours of the cost function for a ground-truth annotation, where the ellipse's long- and short-axes are guided by the view-ray direction of the camera, and the axes extents are guided by the distance of the point to the camera. The projection step of the multi-view crowd localization framework causes streaking artifacts of the features along the view ray on the ground plane, which could smear the

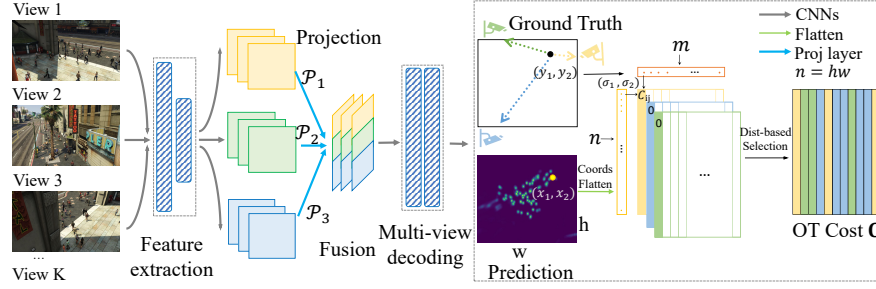


Fig. 2: The model architecture and the proposed Mahalanobis distance-based multi-view optimal transport (M-MVOT) loss for multi-view crowd localization. The model consists of feature extraction, projection, and multi-view fusion and decoding. In the proposed M-MVOT, each point’s transport cost \mathbf{C} is calculated via the Mahalanobis distance instead of the common Euclidean distance **under the closest camera**, which is **directed by the view ray and adjusted by the object-to-camera distance**.

predicted density map making localization less accurate. In order to counteract this effect, we penalize predictions that deviate from the ground-truth position more along the view ray. This can be implemented by assigning larger weights in the Mahalanobis distance along the view ray direction. In addition, the object-to-camera distance in each view is used to further adjust the cost (denoted as M-OT), where the locations far away from the camera are more heavily penalized in order to overcome the more severe projection errors there, compared to locations near to the camera.

Finally, to consider all the input camera views in the model loss (M-MVOT), we integrate the transport costs over multiple views, by setting each point’s optimal transport cost to be calculated with the closest camera. In Fig. 1, we compare different optimal transport loss together, where E-OT uses Euclidean distance and considers each point equally regardless of camera locations, MV-OT uses Mahalanobis distance based on view ray direction, ED-OT considers the camera distance in E-OT, and M-OT considers both the view ray direction and camera distance. M-MVOT considers all camera views together based on M-OT (see Fig. 3 right). Fig. 2 shows the brief architecture of the multi-view crowd localization model and the proposed M-MVOT loss. The model consists of single-view feature extraction, projection, and multi-view fusion and decoding steps. In the proposed M-MVOT loss, the cost for transporting pixels from the predicted density map to the ground-truth points is measured with the Mahalanobis distance to obtain the transport cost \mathbf{C} for the optimal transport problem. The main contributions of this work are as follows.

- To the best of our knowledge, this is the first study on using point supervision optimal transport (OT) for training multi-view crowd localization.
- We propose a new OT loss specifically designed for multi-view crowd localization, which uses view-ray direction and object-to-camera distance to adjust

the Mahalanobis distance transport cost and create a multi-view version of OT based on calculating the cost under the closest camera.

- The experiments validate that the proposed M-MVOT loss achieves better localization performance than density map-based MSE and Euclidean distance-based OT losses.

2 Related Work

2.1 Multi-view crowd localization

Traditional methods. Early traditional multi-view crowd localization methods first detect each person in the images and then associate the camera view results via correspondence matching with hand-crafted features [8, 27, 33, 36]. W. Ge et al. [9] presented a Bayesian people detection and counting approach by sampling from a posterior distribution over crowd configurations. Traditional methods rely on hand-crafted features and background subtraction preprocessing, which limit their performance and application scenarios under heavy occlusions.

Deep learning methods. With the development of deep learning frameworks and more multi-view crowd datasets, deep learning-based multi-view crowd localization methods [3, 5, 31] have seen great progress. The early deep-learning methods did not consider the feature alignment between camera views [5]. MVMS [42] proposed an end-to-end multi-view fusion framework with a feature projection layer for the wide-area crowd-counting task. MVDet [11] also used feature perspective transformations to fuse multi-view information for the multi-view crowd localization. These are the first methods utilizing end-to-end frameworks for multi-view crowd localization tasks, which include single-view feature extraction, projection, and multi-view fusion and decoding. In more recent work, SHOT [31] utilized multi-height projection with an attention-based soft selection module and replaced the perspective transformation with a homography transformation layer in the model for better localization performance. MVDeTr [10] proposed the multi-view deformable transformer network MVDeTr for multi-view fusion by extending the deformable transformer framework in [44]. CVCS [43] has studied multi-view crowd counting methods under the cross-view cross-scene setting, and proposed a large synthetic multi-view crowd dataset CVCS, which could also be used for multi-view crowd localization tasks. 3DROM [26] proposed a data augmentation method to relieve the model overfitting issue by generating random 3D cylinder occlusion masks in the 3D space. 3DROM achieved state-of-the-art performance on 2 small single-scene datasets Multi-viewX and Wildtrack, which have fixed cameras for training and testing. However, it is unclear how well 3DROM performs on larger cross-view cross-scene datasets, such as the CVCS dataset [43].

Overall, the existing multi-view crowd localization methods [10, 11, 26, 31] are trained with fixed-size Gaussian density maps as supervision. However, using Gaussian density maps causes issues for localization in dense crowd regions since when the Gaussian kernels significantly overlap, the local peaks will be smoothed out and the locations unrecoverable. Note that reducing the Gaussian

width could potentially solve this problem, but leads to a more difficult learning problem. Considering that point-based supervision, such as optimal transport loss, achieved better performance on single-image crowd localization task, in this paper, we explore the multi-view optimal transport loss for the multi-view crowd localization task.

2.2 Single-image crowd localization

Single-image crowd localization aims to estimate people’s locations in crowd images with points, which are usually on each person’s head. The task is similar to people detection, which is based on bounding boxes, but crowd localization only uses points to indicate people’s locations. Most single-image counting methods are based on Gaussian density map supervision [2, 6, 7, 13, 14, 16, 21, 24, 28–30, 39–41], and their density map predictions could also be used for crowd localization. However, due to the smoothness of the predicted Gaussian density maps in crowded regions, their crowd localization performance is quite limited.

Thus, point supervision methods [1, 19, 20, 22, 32, 38], blob segmentation [15] and inverse distance maps [17] have been proposed, many of which show better localization performance, especially the **optimal transport**-based point-supervision methods [23, 34, 35]. [22] proposed Bayesian loss, which constructed a density contribution probability model from the point annotations. TopoCount [1] introduced a topological constraint to reason about the spatial arrangement of dots by a persistence loss based on the theory of persistent homology. Z. Ma [23] minimized the unbalanced optimal transport distance between the ground-truth points and predicted density maps, and applied it to crowd counting and localization. Concurrently, GL [34] explored the unbalanced optimal transport (UOT) for crowd counting and proposed a perspective-guided transport cost function to better handle perspective transformations. [34] also proved that the MSE loss with Gaussian ground-truth density maps and the Bayesian loss in [22] are special suboptimal cases of the unbalanced optimal transport loss. P2PNet [32] proposed a pure point framework in a one-to-one matching manner using the Hungarian algorithm. Most previous works perform localization from the density map by first thresholding the density map and then finding local peaks. W. Lin [18] proposed the optimal transport minimization (OT-M) algorithm for crowd localization from density maps, which is an iterative algorithm that finds the set of locations that have minimal OT loss with the density map.

Despite OT’s success in single-view crowd localization, point supervision with optimal transport for multi-view crowd localization is an unexplored area. The multi-view setting introduces new difficulties, such as variations in the camera view, projection errors due to poor camera calibration, and artifacts introduced when projecting features from the camera view to the 2D ground plane, which could hinder model training and localization ability. *To address these issues, our proposed loss considers the camera view direction and distance when computing the transport cost, which is implemented via a Mahalanobis distance function.*

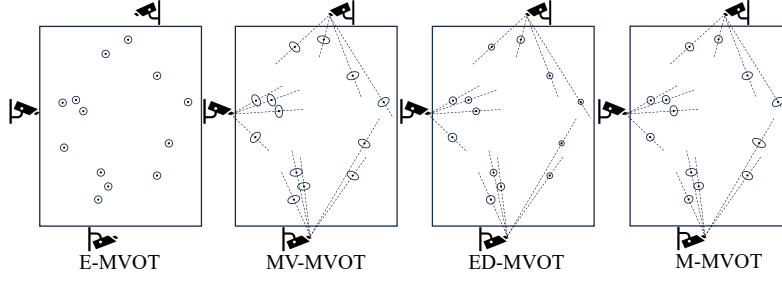


Fig. 3: The comparison of different versions for the multi-view optimal transport, where the cost for each ground-truth point is calculated using the closest camera. E-MVOT considers each point equally regardless of the camera views, which is the same as E-OT for a single camera view. MV-MVOT replaces the Euclidean distance with the Mahalanobis distance and the direction is guided by the view ray direction. ED-MVOT introduces the object-to-camera influence in the optimal transport cost of E-OT. M-MVOT considers both the view-ray direction guidance and the object-to-camera distance adjustment in the optimal transport based on M-OT.

3 Multi-view Optimal Transport

In this section, we first review optimal transport based on Euclidean distance cost (E-OT) for single-image crowd localization. Then, we introduce the proposed Mahalanobis distance optimal transport (M-OT), where the cost can be guided either by the view ray direction of each GT location (MV-OT), or by the object-to-camera distance of each location (ED-OT), or both (M-OT). Finally, the proposed M-OT is applied under multi-views, where each prediction point's optimal transport cost is calculated under the closest camera view via M-OT, denoted as M-MVOT.

3.1 Euclidean distance optimal transport

We follow the loss formulation in [34], which is based on an unbalanced optimal transport formulation with Euclidean distance cost (E-OT). Denote $\mathcal{X} = \{(a_i, \mathbf{x}_i)\}_{i=1}^n$ as the predicted density map, where a_i is pixel \mathbf{x}_i 's predicted density, $\mathbf{a} = [a_i]_i$ is the density of all pixels, and n is the number of pixels. We denote $\mathcal{Y} = \{(b_j, \mathbf{y}_j)\}_{j=1}^m$ as the ground-truth dot map, where $b_j \in \{0, 1\}$ indicates there is a person at location \mathbf{y}_j , $\mathbf{b} = [b_j]_j = \mathbf{1}_m$, and m is the ground-truth number of people. The entropic-regularized unbalanced optimal transport loss is as follows.

$$\begin{aligned} \mathcal{L}_C^\tau = \min_{\mathbf{P} \in \mathcal{R}_+^{n \times m}} & \langle \mathbf{C}, \mathbf{P} \rangle - \epsilon \sum_{ij} P_{ij} \log P_{ij} \\ & + \tau \|\mathbf{P} \mathbf{1}_m - \mathbf{a}\|_2^2 + \tau \|\mathbf{P}^T \mathbf{1}_n - \mathbf{b}\|_1, \end{aligned} \quad (1)$$

where $\mathbf{P} \in \mathcal{R}_+^{n \times m}$ is the transport plan matrix, which assigns each density \mathbf{x}_i of \mathcal{X} to location \mathbf{y}_i of \mathcal{Y} for measuring the cost. $\mathbf{C} \in \mathcal{R}_+^{n \times m}$ is the transport

cost matrix, where C_{ij} measures the cost of moving the predicted density at \mathbf{x}_i to ground-truth dot location \mathbf{y}_j . The cost function is defined as the exponential Euclidean distance between the predicted density of \mathbf{x}_i and the ground-truth dot \mathbf{y}_j :

$$C_{ij} = c(\mathbf{x}_i, \mathbf{y}_j) = \exp(\|\mathbf{x}_i - \mathbf{y}_j\|). \quad (2)$$

The second term $H(P) = \sum_{ij} P_{ij} \log P_{ij}$ is the entropic regularization term. [34] sets $D_1(\mathbf{P}\mathbf{1}_m|\mathbf{a}) = \|\mathbf{P}\mathbf{1}_m - \mathbf{a}\|_2^2$, ensuring that all predicted density values have a corresponding annotation, and $D_2(\mathbf{P}^T\mathbf{1}_n|\mathbf{b}) = \|\mathbf{P}^T\mathbf{1}_n - \mathbf{b}\|_1$, ensuring that all ground-truth points are used in the transport plan. ϵ and τ are weighted factors to adjust the weight of the last three terms. The optimization of (1) can be computed via sinkhorn iteration [25].

3.2 Mahalanobis distance optimal transport

The Euclidean distance in the transport cost matrix \mathbf{C} in (2) considers all directions of deviation from the ground-truth point equally. For multi-view crowd localization, we aim to apply the OT loss on the predicted density map on the ground plane. However, the projection step, which projects features from the camera view to the ground plane introduces streaking artifacts in the feature map due to missing information, *i.e.*, the unknown height of the corresponding object in the 3D world, which may lead to streaking artifacts in the density map that hinder its localization ability. To counteract this effect, we use Mahalanobis distance to replace the Euclidean distance in (2), where the contours of the Mahalanobis distance are guided by the view ray direction at the location, denoted as MV-OT. Besides, the object-to-camera distance could also adjust the calculation of the cost matrix for different regions, denoted as ED-OT. And both the view ray direction and object-to-camera distance can be considered in the Mahalanobis distance optimal transport loss, denoted as M-OT.

View ray-directed transport cost (MV-OT). Using the Mahalanobis distance defines elliptical iso-contours in the transport cost function, where certain directions will incur larger costs than others. Here we set the ellipse shape to counteract the errors and streaking artifacts introduced by the projection step. Specifically, the short axis of the ellipse is set to be aligned with the view ray between the ground-truth point and the camera, and the long axis is set to be perpendicular to the view ray (MV-OT). In this way, the predictions that deviate from the ground-truth along the view ray from the ground-truth points are assigned with larger penalties (see ‘‘MV-OT’’ in Fig. 1, where more penalties are given along the view ray), in order to penalize prediction errors caused by projection distortion. Concretely, we first calculate the covariance matrix for each GT point:

$$\mathbf{S} = \mathbf{R}\mathbf{\Sigma}\mathbf{R}^{-1}, \mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}, \mathbf{R} = \begin{bmatrix} \cos \beta & -\sin \beta \\ \sin \beta & \cos \beta \end{bmatrix}, \quad (3)$$

where \mathbf{R} represents the rotation matrix from the original coordinate to the view-ray coordinate, and β is the rotation angle. Σ is the variance matrix with diagonal entries σ_1^2 and σ_2^2 corresponding to the variances along the view-ray direction and perpendicular to the view-ray direction, respectively. Here we set $\sigma_1^2 < \sigma_2^2$, and thus $1/\sigma_1^2 > 1/\sigma_2^2$, indicating deviating predictions along the view ray direction incur the larger cost and thus are more heavily penalized. Then, the Mahalanobis distance cost in MV-OT can be written as:

$$C_{ij} = c(\mathbf{x}_i, \mathbf{y}_j) = \exp(\sqrt{(\mathbf{x}_i - \mathbf{y}_j)^T \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{y}_j)}). \quad (4)$$

Note when $\sigma_1^2 = \sigma_2^2 = 1$, Mahalanobis distance degenerates into the Euclidean distance. We choose $\sigma_1^2 = 1$ and $\sigma_2^2 = 1.2$ in the experiments for MV-OT.

Distance-adjusted transport cost (ED-OT). The object-to-camera distance also influences the localization accuracy for the points, where faraway points tend to have larger prediction errors. Thus, we use the object-to-camera distance to adjust the cost calculation and give more penalties to faraway regions. We rewrite σ_1^2 and σ_2^2 as follows.

$$\sigma_1^2 = \sigma_2^2 = 1 / \exp(\alpha * \text{MinMaxNorm}(d_{cam})), \quad (5)$$

where d_{cam} denotes the distance from the ground-truth point to the camera, which can be obtained from the camera calibrations and point coordinates, and MinMaxNorm normalizes the object-to-camera distance into $[0, 1]$ through $(d_{cam} - \min(d_{cam})) / (\max(d_{cam}) - \min(d_{cam}))$. Here α is an adaptive factor to adjust the influence of distance to the cost function in (5) and (6). When d_{cam} is larger, σ_1^2 and σ_2^2 decrease, and thus more penalties are assigned. Since $\sigma_1^2 = \sigma_2^2$, the Euclidean distance is actually used in the cost, but the object-to-camera distance adjusts the Euclidean distance for different points, denoted as ED-OT.

View ray-directed and distance-adjusted transport cost (M-OT). To consider both the influence of the view ray direction and object-to-camera distance, we use the object-to-camera distance to adjust the long-axis and short-axis extents of the ellipse iso-contours in the Mahalanobis distance. Specifically, for the GT points that are far away from the camera (d_{cam} is larger), a larger variance is assigned perpendicular to the view ray (σ_2^2), which has the effect of penalizing deviations perpendicular to the view ray less, and penalizing deviations along the view ray (σ_1^2) more:

$$\sigma_1^2 = 1, \quad \sigma_2^2 = \exp(\alpha * \text{MinMaxNorm}(d_{cam})). \quad (6)$$

By combining (3) and (6), we obtain the view ray-directed and distance-adjusted Mahalanobis distance cost function for OT, denoted as M-OT. Note that when $\alpha = 0$, then $\sigma_1^2 = \sigma_2^2 = 1$, and the proposed Mahalanobis distance cost of M-OT degenerates into E-OT in (2).

3.3 Mahalanobis distance multi-view optimal transport (M-MVOT)

We introduce the Mahalanobis distance optimal transport (M-OT) for single views in Sec. 3.2. We further introduce the Mahalanobis distance multi-view

optimal transport (M-MVOT) by combining the proposed Mahalanobis distance optimal transport via a distance-based selection strategy. Specifically, the cost of a point is calculated using the closest camera view’s M-OT defined as follows.

$$C_{ij} = \sum_{k=1}^K \mathbb{1}(d_{cam}^k) \exp(\sqrt{(\mathbf{x}_i - \mathbf{y}_j)^T \mathbf{S}_k^{-1} (\mathbf{x}_i - \mathbf{y}_j)}), \quad (7)$$

where $\mathbb{1}(d_{cam}^k) = \begin{cases} 1, & d_{cam}^k = \min_{p \in \{1, \dots, K\}} d_{cam}^p, \\ 0, & \text{otherwise.} \end{cases}$ is binary and chooses the

closest camera’s M-OT value, K is the camera number, and d_{cam}^k means the distance from the ground-truth point to camera k .

Following this strategy, we extend our E-OT, MV-OT, ED-OT, and M-OT to multi-camera view settings for multi-view crowd localization tasks, denoted as E-MVOT, MV-MVOT, ED-MVOT and M-MVOT, respectively, where the closest camera is used to compute the cost function for a given ground-truth point. See details in Fig. 3.

4 Experiments and Results

4.1 Experiment Settings

Datasets. We evaluate the proposed multi-view optimal transport loss on 3 datasets in the experiments, including CVCS [43], Wildtrack [4] and MultiviewX [11]. **CVCS** is a synthetic multi-view crowd dataset, containing 31 scenes, where 23 are for training and the remaining 8 for testing. Each scene contains 100 multi-view frames with about 60-120 camera views. The size of the scenes in CVCS varies from 10m×20m to 90m×80m. Each scene contains 90-180 people. **Wildtrack** is a real-world single-scene multi-view dataset, whose scene size is 12m×36m. There are 20 people per frame on average in the dataset. The dataset consists of 7 camera views and 400 synchronized multi-view frames, in which the first 360 frames are used for training and the 40 frames are for testing. **MultiviewX** is a synthetic single-scene multi-view dataset, whose scene size is 16m×25m and consists of 6 camera views. The dataset includes 400 multi-view frames and the first 360 are used for training and 40 are for testing. MultiviewX has around 40 synthetic people per frame. The resolution of the images in the 3 datasets is 1920×1080. Note that CVCS is a *cross-scene* dataset, where the same scene and camera layout do not appear in both the training set and test set. In contrast, both Wildtrack and MultiviewX are *single-scene* datasets with the same camera layout for both training and testing. Thus, CVCS is a more challenging dataset, and can better demonstrate the generalization of multi-view localization methods to new scenes and camera layouts.

Comparison methods. We compare the proposed M-MVOT loss based multi-view crowd localization model against the Euclidean distance version (E-MVOT) using the same architecture, as well as against existing methods MVDet

Table 1: The multi-view crowd localization results on the CVCS dataset. The distance threshold is 0.5m in the world for all methods.

Method	MODA \uparrow	MODP \uparrow	Precision \uparrow	Recall \uparrow	F1_score \uparrow
MVDet [11]	14.2	59.3	85.0	17.3	28.7
SHOT [31]	31.7	72.1	<u>94.5</u>	33.6	49.6
MVDeTr [10]	24.9	79.6	98.1	25.4	40.4
3DROM [26]	20.1	74.2	84.1	23.7	37.0
E-MVOT (ours)	<u>43.1</u>	<u>74.3</u>	85.6	<u>51.8</u>	<u>64.5</u>
M-MVOT (ours)	43.5	74.1	85.5	52.3	64.9

[11], SHOT [31], MVDeTr [10], 3DROM [26]. We report the performance of previous methods on Wildtrack and MultiviewX according to the published results, and also apply them on the CVCS dataset for comparison. We also compare with RCNN [37], POM-CNN [8], DeepMCD [5], DeepOcc. [3], and Volumetric [12] on the Wildtrack and MultiviewX datasets.

Implementation details. In the experiments, for CVCS, the input image resolution is 1280×720 . In the training, 5 views are randomly selected for 5 times in each iteration per frame of each scene, and the same number of views (5) is randomly selected for 21 times per frame in the evaluation. For Wildtrack, the input image resolution is resized to 1280×720 , and the scene map resolution is 120×360 , where each pixel represents 0.1m in the real world. For MultiviewX, the input image resolution is also resized to 1280×720 , and the scene map resolution is 160×250 , where each pixel is 0.1m in the real world. The map distance threshold is 0.5m in the real world as in [11] for all methods on the 3 datasets. On the CVCS dataset, the multi-view crowd localization model is implemented based on the existing multi-view crowd localization method MVDet [11]. On the Wildtrack and MultiviewX datasets, the implementation for the multi-view crowd localization is based on MVDeTr [10]. The ground-plane density map prediction losses in [10, 11] are replaced with the proposed Mahalanobis distance multi-view optimal transport loss (M-MVOT). The models are trained with the proposed M-MVOT loss and the auxiliary 2D density map losses in [11] together. For Wildtrack and CVCS, τ is set to 1, and τ is set to 20 for MutiviewX. α is 1 on CVCS and 0.05 on both MultiviewX and Wildtrack. See more model implementation and training details in the supplemental.

Evaluation metrics. We use 5 metrics to evaluate and compare the multi-view crowd localization methods: Multiple Object Detection Accuracy (MODA), Multiple Object Detection Precision (MODP), Precision (P), Recall (R), and F1_score (F1). Specifically, $MODA = 1 - (FP + FN) / (TP + FN)$ measures the detection accuracy. $MODP = (\sum (1 - d[d < t] / t)) / TP$ measures the precision of detection, where d is the distance from a detected person point to its ground truth and t is the distance threshold. Meanwhile, $P = TP / (FP + TP)$, $R = TP / (TP + FN)$, and $F1 = 2P * R / (P + R)$. Here, TP , FP , and FN are the number of true positives, false positives, and false negatives, respectively. Overall, we use MODA and F1_score as the main metrics, since they comprehensively measure the performance of detecting the crowd both completely and precisely.

Table 2: Comparison of the multi-view people detection performance on single-scene datasets Wildtrack and MultiviewX. The proposed method is the best on MultiviewX and comparable on Wildtrack. The distance threshold is 0.5m in the real world [11].

Dataset Method	MultiviewX					Wildtrack				
	MODA	MODP	P.	R.	F1.	MODA	MODP	P.	R.	F1.
RCNN [37]	18.7	46.4	63.5	43.9	51.9	11.3	18.4	68	43	52.7
POM-CNN [8]	-	-	-	-	-	23.2	30.5	75	55	63.5
DeepMCD [5]	70.0	73.0	85.7	83.3	84.5	67.8	64.2	85	82	83.5
DeepOcc. [3]	75.2	54.7	97.8	80.2	88.1	74.1	53.8	95	80	86.9
Volumetric [12]	84.2	80.3	97.5	86.4	91.6	88.6	73.8	95.3	93.2	94.2
MVDet [11]	83.9	79.6	96.8	86.7	91.5	88.2	75.7	94.7	93.6	94.1
SHOT [31]	88.3	82.0	96.6	91.5	94.0	90.2	76.5	96.1	94.0	95.0
MVDeTr [10]	93.7	91.3	99.5	94.2	97.8	91.5	82.1	97.4	94.0	95.7
3DROM [26]	95.0	84.9	<u>99.0</u>	96.1	97.5	93.5	75.9	<u>97.2</u>	96.2	96.7
E-MVOT (ours)	<u>96.3</u>	85.2	98.1	98.1	<u>98.1</u>	91.9	80.9	94.1	98.0	96.0
M-MVOT (ours)	96.7	<u>86.1</u>	98.8	<u>97.9</u>	98.3	<u>92.1</u>	<u>81.3</u>	94.5	<u>97.8</u>	<u>96.1</u>

4.2 Multi-view crowd localization results

CVCS. In Table 1, we compare the proposed Mahalanobis distance multi-view optimal transport loss (M-MVOT) with other multi-view crowd localization methods and the Euclidean distance multi-view optimal transport loss (E-MVOT) on CVCS dataset. Table 1 shows that the proposed M-MVOT loss achieves better overall localization performance than existing multi-view crowd localization methods. Even though MVDeTr achieves the highest MODP, our proposed M-MVOT and E-MVOT are the best in terms of MODA and F1_score, which are significantly increased from previous methods due to M-MVOT and E-MVOT’s higher recall rates. These results validate the advantages of our proposed M-MVOT loss, which considers the camera view variations in the task and can handle the localization ambiguities, as compared to Gaussian density map MSE loss. M-MVOT also outperforms E-MVOT, due to its ability to more flexibly change the cost function based on the relationships between the camera view and ground-truth positions. Note that CVCS is a large multi-scene multi-view dataset, which is more challenging than other single-scene multi-view datasets.

Wildtrack and MultiviewX. In Table 2, we compare our M-MVOT/E-MVOT with other multi-view crowd localization methods on the MultiviewX (left) and Wildtrack datasets (right), respectively. On **MultiviewX**, the proposed M-MVOT and E-MVOT outperform the comparison methods, with the highest MODA and F1_score. M-MVOT also outperforms E-MVOT on the MultiviewX dataset (according to MODA and F1_score), which also demonstrates the advantage of the proposed M-MVOT over the common Euclidean distance optimal transport loss. On **Wildtrack**, 3DROM achieves the best results on the MODA and F1_score metrics, and M-MVOT and E-MVOT achieve the second-best results. The possible reason is that 3DROM is a data augmentation method that could deal with the overfitting issue on small datasets like Wildtrack. But M-MVOT is still better than E-MVOT, demonstrating the effectiveness of the view ray direction and camera distance guidance.

Visualization results. We show the visualization result on the 3 datasets in Fig. 4. Generally, the proposed M-MVOT/E-MVOT shows better localization

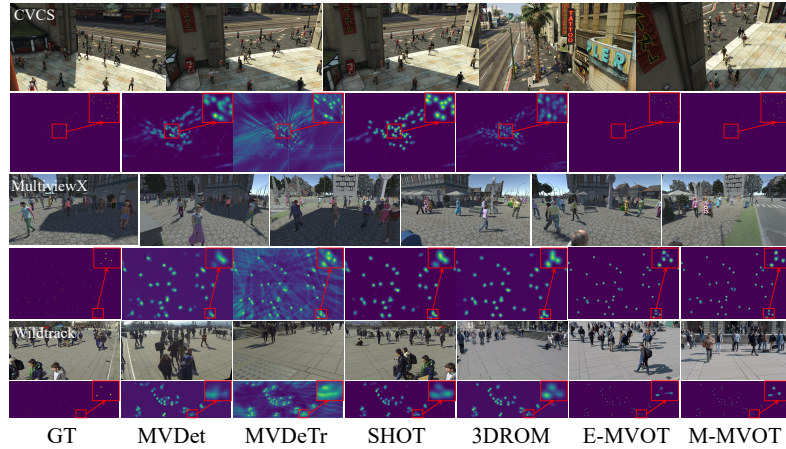


Fig. 4: The predicted crowd occupancy maps of different methods on the 3 datasets CVCS, MultiviewX, and Wildtrack (zoom in for better view). The value of crowd occupancy maps indicates the person probability of each location.

performance under crowded regions (in red boxes) and reduces the projection artifacts of comparison methods on faraway regions. Furthermore, as seen in the middle and the bottom of Fig. 4, M-MVOT predicts fewer false positives than E-MVOT on MultiviewX; M-MVOT also has fewer tail (streaking) artifacts than E-MVOT on Wildtrack, which proves the effectiveness of the proposed the view ray direction and distance guidance.

4.3 Ablation study

Ablation study on the MVOT loss. We conducted an ablation study on the setting of the MVOT loss in Table 3 on the CVCS dataset, comparing the proposed M-MVOT loss with MSE, MVOT with Euclidean cost (E-MVOT), MVOT with Mahalanobis cost based on view ray guidance only (MV-MVOT), and camera distance only (ED-MVOT). From Table 3, M-MVOT achieves better performance than ED-MVOT, MV-MVOT, E-MVOT, and MSE loss, which demonstrates the advantage of the proposed distance guidance and view ray direction in the covariance matrix over other losses. Note that ED-MVOT got lower results, due to more penalties being added to faraway points in all directions, leading to fewer predictions (so it has higher precision). M-MVOT considers directions and adds relaxations to directions perpendicular to the view-ray, achieving better results. As shown in the red boxes of Fig. 5, M-MVOT predicts more accurate results than MV-MVOT and ED-MVOT, which demonstrates the effectiveness of using the view ray direction and object-to-camera distance in the covariance matrix of the MVOT loss.

Ablation study on α and τ in M-MVOT. We conducted an ablation study on the α and τ of M-MVOT in Table 4 on the MultiviewX dataset. α

Table 3: The ablation study on the MVOT loss for multi-view crowd localization on the CVCS dataset with other settings kept the same except for the training loss.

Loss	Dir.	Dis.	MODA	MODP	P	R	F1
MSE	✗	✗	14.2	59.3	85.0	17.3	28.7
E-MVOT	✗	✗	43.1	74.3	85.6	51.8	64.5
MV-MVOT	✓	✗	43.3	74.4	85.9	51.8	64.6
ED-MVOT	✗	✓	42.9	75.0	87.7	49.9	63.3
M-MVOT	✓	✓	43.5	74.1	85.5	52.3	64.9

Table 4: The ablation study on α and τ of M-MVOT loss on the MultiviewX dataset. The best results are achieved by $\alpha = 0.05$ and $\tau = 20$.

α	MODA	MODP	P.	R.	F1.	τ	MODA	MODP	P.	R.	F1.
0	96.3	85.2	98.1	98.1	98.1	1	95.2	90.2	99.1	96.1	97.6
0.05	96.7	86.1	98.8	97.9	98.3	10	96.1	86.5	99.2	96.9	98.0
0.2	96.7	84.3	98.6	98.1	98.3	20	96.7	86.1	98.8	97.9	98.3
0.6	96.3	84.5	98.2	98.1	98.1	30	95.2	83.8	97.9	97.3	97.6

adjusts the object-to-camera distance in the M-MVOT loss in (6). $\alpha = 0$ is equivalent to the E-MVOT loss. When α increases, the effect of the distance change on the model is enlarged, which penalizes the faraway regions for more accurate localization performance. However, when α is too large, the penalties are too large, which may force the model to neglect the near regions. τ adjusts the effect of the last two regularization terms in (1) to ensure that all predicted density values have a corresponding annotation and all ground-truth points are used in the transport plan. When τ increases, MODA and Recall increase while MODP and Precision drop for detecting more points. But when τ is too large, the penalties of the last two terms are too large, harming the optimization of the main term. The best result is achieved at $\alpha = 0.05$, $\tau = 20$.

Ablation study on the multi-view fusion strategy. We conduct an ablation study on the multi-view fusion strategy, where soft weights learned from camera distance replace the binary choice strategy in Eq. (4). Table 5 shows that our proposed M-MVOT with binary multi-view fusion still outperforms the soft choices for fusing multi-cameras. The possible reason is the binary camera choice forces the model to focus on the closest camera (the more reliable one) and reduces the complexity of the covariance matrix for model learning.

Comparison with learnable covariance matrix (LCM). We compare our proposed M-MVOT, which uses manually designed covariance matrices, with a data-driven learnable covariance matrix (LCM) in Table 5, where the covariance matrices are estimated from the ground-plane features using two convolutional layers. Table 5 shows the results of using learnable covariance matrices (LCM), which is not as good as M-MVOT. The reason is that our M-MVOT considers the multi-views’ unique domain knowledge in the covariance matrix and is more suitable for the multi-view task, although exploring adding priors to the learnable covariance matrix could be good future work.

Comparison with the single-image SOTA OT method with a simple multi-view fusion (Multi-UOT). We compare our proposed M-MVOT with

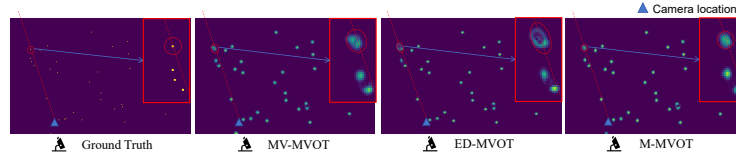


Fig. 5: Comparison of MV-MVOT, ED-MVOT, and M-MVOT. The blue triangle is the camera location. M-MVOT predicts fewer artifacts than MV-MVOT and ED-MVOT (see red boxes), demonstrating the effectiveness of distance adjustment and view-ray direction, respectively.

Table 5: Comparisons with soft choices for fusing multi-cameras (Soft), a learnable covariance matrix (LCM) and multi-view fusion of single-view methods (Multi-UOT).

Comparison	MODA	MODP	P.	R.	F1.
Soft	96.5	86.0	98.7	97.8	98.2
LCM	95.1	81.5	96.7	98.5	97.6
Multi-UOT	23.9	72.0	75.2	35.7	48.4
M-MVOT	96.7	86.1	98.8	97.9	98.3

the single-image SOTA OT method UOT [34] with a simple multi-view fusion. Specifically, we project and sum the 2D results of [34] on the ground and show results as Multi-UOT in Table 5. M-MVOT is much better than Multi-UOT and it demonstrates that single-image methods with simple multi-view fusion cannot fully utilize the multi-view information well, consistent with [11, 42].

5 Discussion and Conclusion

In this paper, we propose a novel Mahalanobis distance-based multi-view optimal transport (M-MVOT) loss, which is specially designed for multi-view crowd localization. As far as we know, this is the first study on point supervision for multi-view crowd localization methods., this is the first study on point supervision for multi-view crowd localization, which advances the research on multi-view crowd localization tasks in the aspect of point supervision training losses. The proposed loss uses view-ray directed and object-to-camera distance-adjusted Mahalanobis distance in the cost calculation in order to counteract the distortions created during the projection step, and then it is applied to multi-camera views via calculating the cost under the closest camera. Overall, the experiments validate that the proposed M-MVOT achieves better localization performance, compared to density map-based mean squared error loss (MSE) and common Euclidean distance-based optimal transport loss. **Ethical considerations:** Multi-view crowd localization could be used for crowd analysis, autonomous driving, public traffic management, etc. For applications, we could encode the images as features, rely on synthetic datasets, or mask out human faces first instead of directly inputting images with human faces to the crowd localization models for higher privacy protection standards.

Acknowledgements

This work was supported in parts by NSFC (62202312, U21B2023, U2001206), Guangdong Basic and Applied Basic Research Foundation (2023B1515120026), DEGP Innovation Team (2022KCXTD025), City University of Hong Kong Strategic Research Grant (7005665), Shenzhen Science and Technology Program (KQTD 20210811090044003, RCJC20200714114435012), Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), and Scientific Development Funds from Shenzhen University.

References

1. Abousamra, S., Hoai, M., Samaras, D., Chen, C.: Localization in the crowd with topological constraints. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 872–881 (2021)
2. Bai, S., He, Z., Qiao, Y., Hu, H., Wu, W., Yan, J.: Adaptive dilated network with self-correction supervision for counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4594–4603 (2020)
3. Baqué, P., Fleuret, F., Fua, P.: Deep occlusion reasoning for multi-camera multi-target detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 271–279 (2017)
4. Chavdarova, T., Baqué, P., Bouquet, S., Maksai, A., Jose, C., Bagautdinov, T., Lettry, L., Fua, P., Van Gool, L., Fleuret, F.: Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5030–5039 (2018)
5. Chavdarova, T., Fleuret, F.: Deep multi-camera people detection. In: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 848–853. IEEE (2017)
6. Cheng, Z.Q., Dai, Q., Li, H., Song, J., Wu, X., Hauptmann, A.G.: Rethinking spatial invariance of convolutional networks for object counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19638–19648 (2022)
7. Cheng, Z.Q., Li, J.X., Dai, Q., Wu, X., Hauptmann, A.G.: Learning spatial awareness to improve crowd counting. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6152–6161 (2019)
8. Fleuret, F., Berclaz, J., Lengagne, R., Fua, P.: Multicamera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **30**(2), 267–282 (2007)
9. Ge, W., Collins, R.T.: Crowd detection with a multiview sampler. In: ECCV. pp. 324–337 (2010)
10. Hou, Y., Zheng, L.: Multiview detection with shadow transformer (and view-coherent data augmentation). In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 1673–1682 (2021)
11. Hou, Y., Zheng, L., Gould, S.: Multiview detection with feature perspective transformation. In: Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16. pp. 1–18. Springer (2020)
12. Isakov, K., Burkov, E., Lempitsky, V., Malkov, Y.: Learnable triangulation of human pose. In: ICCV (2019)

13. Jiang, X., Zhang, L., Xu, M., Zhang, T., Lv, P., Zhou, B., Yang, X., Pang, Y.: Attention scaling for crowd counting. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
14. Kang, D., Chan, A.: Crowd counting by adaptively fusing predictions from an image pyramid. In: BMVC (2018)
15. Laradji, I.H., Rostamzadeh, N., Pinheiro, P.O., Vazquez, D., Schmidt, M.: Where are the blobs: Counting by localization with point supervision. In: Proceedings of the european conference on computer vision (ECCV). pp. 547–562 (2018)
16. Li, Y., Zhang, X., Chen, D.: Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In: CVPR. pp. 1091–1100 (2018)
17. Liang, D., Xu, W., Zhu, Y., Zhou, Y.: Focal inverse distance transform maps for crowd localization. *IEEE Transactions on Multimedia* (2022)
18. Lin, W., Chan, A.B.: Optimal transport minimization: Crowd localization on density maps for semi-supervised counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21663–21673 (June 2023)
19. Liu, C., Lu, H., Cao, Z., Liu, T.: Point-query quadtree for crowd counting, localization, and more. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1676–1685 (2023)
20. Liu, H., Zhao, Q., Ma, Y., Dai, F.: Bipartite matching for crowd counting with point supervision. In: IJCAI. pp. 860–866 (2021)
21. Liu, W., Salzmann, M., Fua, P.: Context-aware crowd counting. In: CVPR. pp. 5099–5108 (2019)
22. Ma, Z., Wei, X., Hong, X., Gong, Y.: Bayesian loss for crowd count estimation with point supervision. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6142–6151 (2019)
23. Ma, Z., Wei, X., Hong, X., Lin, H., Qiu, Y., Gong, Y.: Learning to count via unbalanced optimal transport. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 2319–2327 (2021)
24. Onoro-Rubio, D., López-Sastre, R.J.: Towards perspective-free object counting with deep learning. In: ECCV. pp. 615–629. Springer (2016)
25. Peyré, G., Cuturi, M., et al.: Computational optimal transport. Center for Research in Economics and Statistics Working Papers (2017-86) (2017)
26. Qiu, R., Xu, M., Yan, Y., Smith, J.S., Yang, X.: 3d random occlusion and multi-layer projection for deep multi-camera pedestrian localization. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X. pp. 695–710. Springer (2022)
27. Sabzmeydani, P., Mori, G.: Detecting pedestrians by learning shapelet features. In: CVPR. pp. 1–8. IEEE (2007)
28. Sam, D.B., Surya, S., Babu, R.V.: Switching convolutional neural network for crowd counting. In: CVPR. pp. 4031–4039 (2017)
29. Shi, M., Yang, Z., et al.: Revisiting perspective information for efficient crowd counting. In: CVPR. pp. 7279–7288 (2019)
30. Sindagi, V.A., Patel, V.M.: Generating high-quality crowd density maps using contextual pyramid cnns. In: ICCV. pp. 1879–1888 (2017)
31. Song, L., Wu, J., Yang, M., Zhang, Q., Li, Y., Yuan, J.: Stacked homography transformations for multi-view pedestrian detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6049–6057 (2021)
32. Song, Q., Wang, C., Jiang, Z., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., Wu, Y.: Rethinking counting and localization in crowds: A purely point-based framework. *arXiv preprint arXiv:2107.12746* (2021)

33. Viola, P., Jones, M.J.: Robust real-time face detection. *International Journal of Computer Vision* **57**(2), 137–154 (2004)
34. Wan, J., Liu, Z., Chan, A.B.: A generalized loss function for crowd counting and localization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 1974–1983 (2021)
35. Wang, B., Liu, H., Samaras, D., Nguyen, M.H.: Distribution matching for crowd counting. *Advances in neural information processing systems* **33**, 1595–1607 (2020)
36. Wu, B., Nevatia, R.: Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *International Journal of Computer Vision* **75**(2), 247–266 (2007)
37. Xu, Y., Liu, X., Liu, Y., Zhu, S.C.: Multi-view people tracking via hierarchical trajectory composition. In: *CVPR*. pp. 4256–4265 (2016)
38. Yan, S., Xiong, X., Arnab, A., Lu, Z., Zhang, M., Sun, C., Schmid, C.: Multiview transformers for video recognition. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 3333–3343 (2022)
39. Yan, Z., Yuan, Y., Zuo, W., Tan, X., Wang, Y., Wen, S., Ding, E.: Perspective-guided convolution networks for crowd counting. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 952–961 (2019)
40. Yang, Y., Li, G., Wu, Z., Su, L., Huang, Q., Sebe, N.: Reverse perspective network for perspective-aware object counting. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4374–4383 (2020)
41. Zhang, C., Li, H., et al.: Cross-scene crowd counting via deep convolutional neural networks. In: *CVPR*. pp. 833–841 (2015)
42. Zhang, Q., Chan, A.B.: Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns. In: *CVPR*. pp. 8297–8306 (2019)
43. Zhang, Q., Lin, W., Chan, A.B.: Cross-view cross-scene multi-view crowd counting. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 557–567 (2021)
44. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159* (2020)