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

Qi Zhang[1], Kaiyi Zhang[2,1], Antoni B. Chan[3], 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

Table 1: The overview of the proposed methods. They differ in the variance in the covariance matrix, and with or without the distance-based camera selection.

Method	Variance	Cam Selection
E-OT	$\sigma_1^2 = \sigma_2^2 = 1$	without
MV-OT	$\sigma_1^2 = 1, \sigma_2^2 = 1.2$	without
ED-OT	$\sigma_1^2 = \sigma_2^2 = 1/\exp(\alpha * \text{MinMaxNorm}(d_{cam}))$	without
M-OT	$\sigma_1^2 = 1, \sigma_2^2 = \exp(\alpha * \text{MinMaxNorm}(d_{cam}))$	without
E-MVOT	$\sigma_1^2 = \sigma_2^2 = 1$	with
MV-MVOT	$\sigma_1^2 = 1, \sigma_2^2 = 1.2$	with
ED-MVOT	$\sigma_1^2 = \sigma_2^2 = 1/\exp(\alpha * \text{MinMaxNorm}(d_{cam}))$	with
M-MVOT	$\sigma_1^2 = 1, \sigma_2^2 = \exp(\alpha * \text{MinMaxNorm}(d_{cam}))$	with

1 Overview of the proposed methods

In order for better understanding, we give an overview of the proposed methods (E-OT, MV-OT, ED-OT, M-OT, E-MVOT, MV-MVOT, ED-MVOT and M-MVOT) in Table 1. They differ in the variance defined in the covariance matrix when computing the transport cost matrix and with or without the distance-based camera selection for the fusion of multi-cameras. σ_1^2 is the variance along the camera view ray. σ_2^2 is the variance perpendicular to the camera view ray.

2 More Experiments

We conducted a series of extra experiments to analyze the performance of the proposed Mahalanobis distance-based multi-view optimal transport method.

⋆ Corresponding author

Table 2: The ablation study on $\sigma_1^2 : \sigma_2^2$ of MV-MVOT loss on the MultiviewX dataset. The best results of fixed ratios are achieved by $\sigma_1^2 : \sigma_2^2$ = 1:1.2, but it is still lower than M-MVOT's performance.

$\sigma_1^2 : \sigma_2^2$	MODA	MODP	Precision	Recall	F1_score
1:1	96.3	85.2	98.1	98.1	98.1
1.2:1	96.5	84.6	98.3	**98.2**	98.2
1.5:1	96.2	85.2	98.2	98.0	98.1
2:1	96.3	86.1	98.5	97.8	98.1
1:1.2	96.5	85.6	98.3	**98.2**	98.2
1:1.5	96.5	85.2	98.5	97.9	98.2
1:2	96.3	85.5	98.4	97.9	98.1
M-MVOT	**96.7**	**86.1**	**98.8**	97.9	**98.3**

Table 3: The ablation study on different strategies to fuse M-MVOT loss on the MultiviewX dataset. The best results are achieved by Dist-based Selection.

Strategy	MODA	MODP	Precision	Recall	F1_score
Average	96.5	85.2	98.7	97.8	98.2
Soft	96.5	86.0	98.7	97.8	98.2
Dist-based Selection	**96.7**	**86.1**	**98.8**	**97.9**	**98.3**

2.1 Ablation study on the fixed ratios of σ_1^2 and σ_2^2

We conducted more experiments on MV-MVOT in Table 2 with various fixed ratios of σ_1^2 and σ_2^2 on the MultiviewX dataset. When $\sigma_1^2 = \sigma_2^2 = 1$, it is equal to E-MVOT.

When $\sigma_1^2 > \sigma_2^2$ (1.2:1, 1.5:1, and 2:1), it means we encourage the prediction points to appear along the view-ray direction, which is the opposite of MV-MVOT. 1.2:1 achieves the best performance with MODA=96.5, MODP=84.6, which is lower than the best performance of using $\sigma_1^2 < \sigma_2^2$ (1:1.2), with MODA=96.5, MODP=85.6. The reason is the prediction points already appear along the view-ray direction with a higher probability but with an offset to the ground-truth locations due to projection distortions.

So we enforce more punishments along the view-ray direction to obtain more accurate predicted points, namely σ_1^2 should be smaller than σ_2^2 (1:1.2, 1:1.5, and 1:2). From Table 2, we can see that the best performance on fixed ratio is achieved at $\sigma_1^2 : \sigma_2^2$ = 1:1.2. When the ratio continues to increase, the performance starts to decrease.

Overall, the best performance on fixed ratios is still worse than the proposed M-MVOT. Because M-MVOT considers the object-to-camera distance in the cost calculation and can adjust the cost adaptively and according to the distance to the camera.

2.2 Ablation study on the multi-view fusion strategy

We conduct extra experiments on the multi-view fusion strategy with a very basic form on M-MVOT, that is, **Average**. We sum the OT losses under all

Table 4: The ablation study on the backbone of M-MVOT loss on the MultiviewX dataset. The best results are achieved by MVDeTr.

Backbone	Method	MODA	MODP	Precision	Recall	F1_score
SHOT [4]	MSE	88.3	82.0	96.6	91.5	94.0
	M-MVOT	89.2	85.6	96.4	92.7	94.5
MVDeTr [1]	MSE	93.9	**90.3**	98.0	95.9	96.9
	M-MVOT	**96.7**	86.1	**98.8**	**97.9**	**98.3**

Table 5: Ablation study on M-MVOT loss with view confidence attentions.

Method	MODA	MODP	Precision	Recall	F1_score
MVOT-conf	96.0	85.8	98.3	97.7	98.0
M-MVOT(Ours)	**96.7**	**86.1**	**98.8**	97.9	**98.3**

cameras together and then averaging them. And **Soft** is to replace the binary choice strategy with soft weights learned from camera distance for fusing multi-cameras (as mentioned in the manuscript). The experiment in Table 3 shows that the distance-based selection strategy achieves better results than Average or Soft fusion of multi-cameras, a simple but effective strategy for selecting the most reliable camera. While the average strategy considers all cameras equally without any preference, which is not flexible for multi-view fusion.

2.3 Ablation study on backbones

We conducted experiments on our proposed M-MVOT with two different backbones on the MutiviewX dataset: SHOT [4] and MVDetr [1]. In the MVDeTr backbone, we also add a multi-height selection module proposed by SHOT to reduce the impact of projection errors. The experiment in Table 4 shows that the MVDeTr backbone performs better than the SHOT backbone. More importantly, no matter if it is implemented on which backbone, the proposed M-MVOT always achieves better performance than the density-map-based MSE loss, which indicates the advantage of the proposed M-MVOT loss.

2.4 Ablation study on M-MVOT loss with view confidence attentions

We conduct an extra experiment on M-MVOT loss with extra view confidence attentions in Table 5. Specifically, extra self-confidence attentions for each camera are added in the multi-view OT loss for dealing with the occlusions (denoted as MVOT-conf) on the MultiviewX dataset. The experiments show the M-MVOT is still better than MVOT-conf. The multi-view crowd localization model already can fuse multi-view features for handling occlusions.

2.5 Ablation study on different distance thresholds

In addition to using 0.5m as the distance threshold, we also evaluate and compare the results on different distance thresholds in Table 6. 0.5m is a suitable thresh-

Table 6: Ablation study on different distance thresholds.

Threshold	Method	MODA	MODP	Precision	Recall	F1_score
0.4m	SHOT [4]	89.0	78.0	97.1	91.7	94.3
	MVDeTr [1]	93.4	**89.4**	**99.4**	93.9	96.4
	M-MVOT (Ours)	**96.5**	82.7	98.6	**97.8**	**98.2**
0.5m	SHOT [4]	88.3	82.0	96.6	91.5	94.0
	MVDeTr [1]	93.7	**91.3**	**99.5**	94.2	97.8
	M-MVOT (Ours)	**96.7**	86.1	98.8	**97.9**	**98.3**
0.6m	SHOT [4]	89.5	85.1	97.4	92.0	94.6
	MVDeTr [1]	93.6	**92.9**	**99.6**	94.0	96.7
	M-MVOT (Ours)	**96.7**	88.4	98.8	**97.9**	**98.3**

Table 7: Loss computation cost comparison.

Loss	Memory(GB)	FLOPs(G)	Train(s)	Test(s)
Focal	10.205	923.702	0.416	0.29
E-MVOT	11.159	923.702	0.484	0.29
M-MVOT	11.251	923.702	0.531	0.29

old considering human sizes, and all SOTAs [1–4] evaluate using this threshold. We also tested with {0.4m, 0.6m} thresholds on MultiviewX in Table 6, and our MODA & F1_score are still better than MVDeTr [1] and SHOT [4], demonstrating the advantage of the proposed M-MVOT loss.

2.6 Ablation study on the computational cost of different losses.

We compare the computational cost of different losses in Table 7. The GPU memory usage is M-MVOT > E-MVOT > Focal, and the single-batch training speed rank is the opposite. Our OT losses' training computation is higher than Focal loss, but the test speed and FLOPs are the same since the loss computation is removed at test time.

3 More Implementation and Training Details

On the CVCS dataset, the proposed methods are implemented on MVDet [2] backbone. We first pretrain the 2D encoder and decoder with MSE loss to give a better initialization for 2D feature extraction. After that, we enable the multi-view decoder into the training state with the proposed loss function for 300 epochs. In each epoch, 10 frames are randomly chosen from a total of 100 frames of each scene with 5 times sampling.

On MultiviewX and Wildtrack, the proposed methods are implemented based on the MVDetr backbone with an extra muti-height selection module proposed by SHOT. We use the same loss function proposed by MVDetr but replace the ground plane focal loss with the proposed M-MVOT loss. After pretraining the

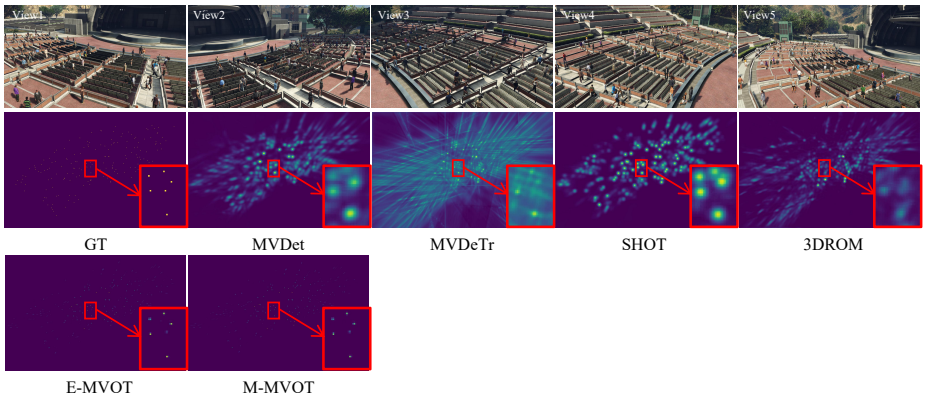

Fig. 1: Visualization results on CVCS.

feature extraction, the 2D encoder and multi-view decoder are trained simultaneously for 120 epochs with the OneCycle scheduler. The maximum and minimum learning rates are 1e-4 and 1e-5 respectively.

4 More Visualization Results

We provide more visualization results for each dataset. figure 1 and 2, figure 3 and 4, figure 5 and 6 give visualizations results on CVCS dataset, MutiviewX dataset and Wildtrack dataset respectively. It is worth noting our methods E-MVOT and M-MVOT get a better visual effect than other methods, especially in areas with dense crowds or closed to the border because they reduce the projection artifacts in these areas.

References

1. 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)
2. 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)
3. 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)
4. 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)

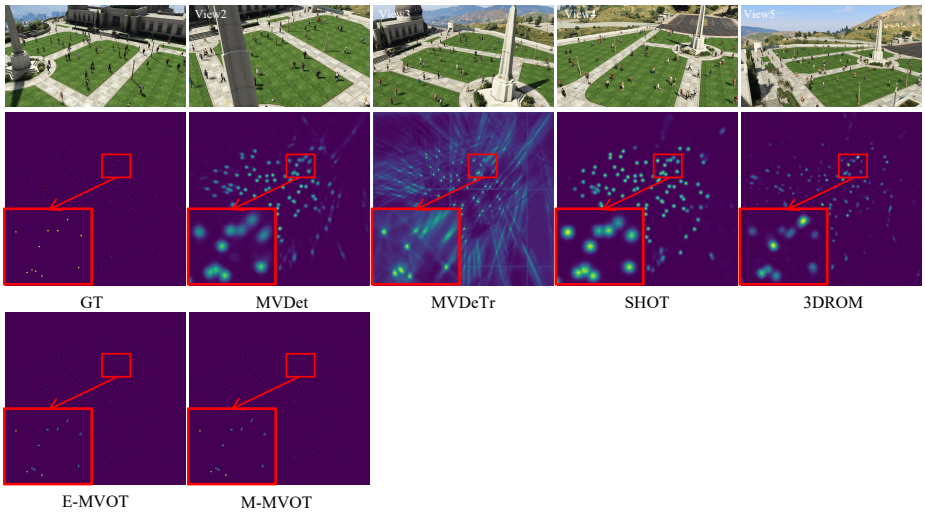

Fig. 2: Visualization results on CVCS.

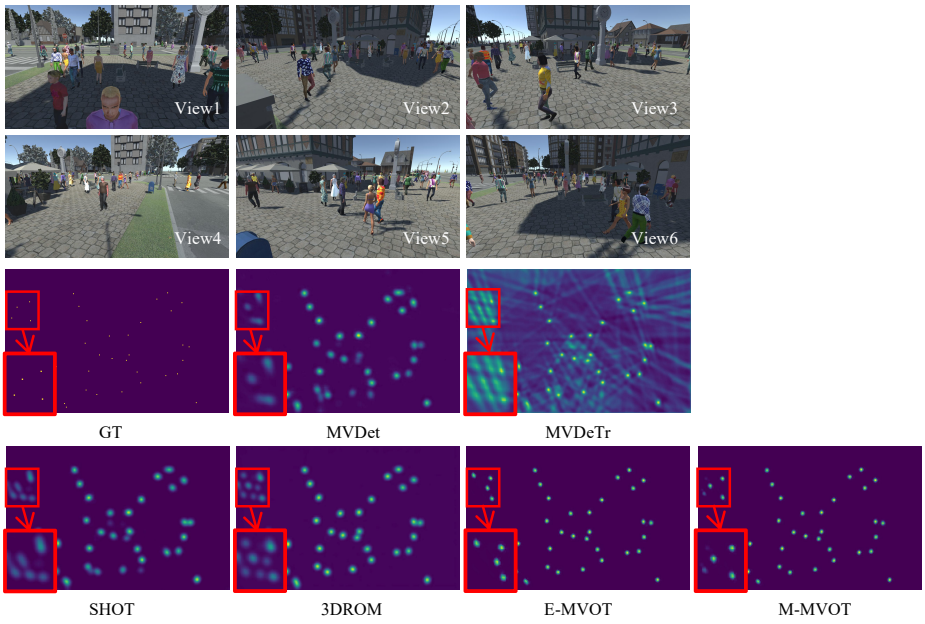

Fig. 3: Visualization results on MutiviewX.

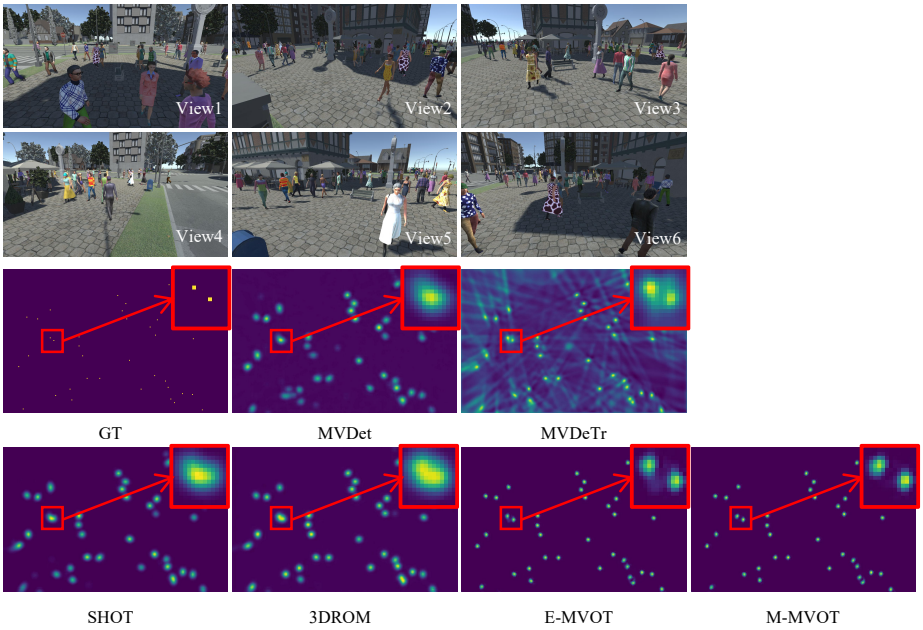

Fig. 4: Visualization results on MutiviewX.

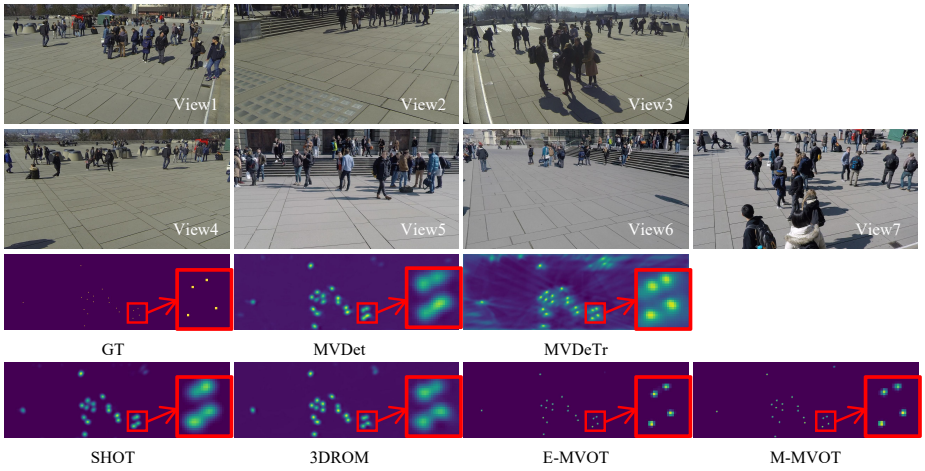

Fig. 5: Visualization results on Wildtrack.

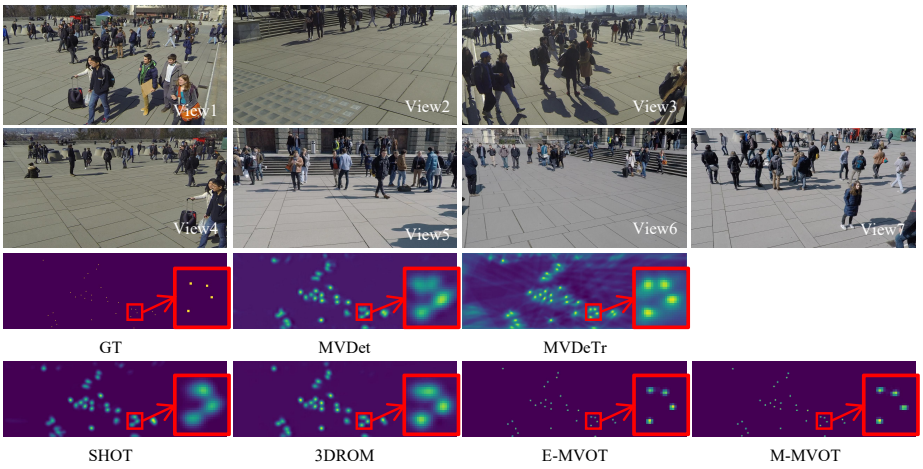

Fig. 6: Visualization results on Wildtrack.