

# Pinpointing Crowd in Bird’s Eye View via Proximal Contexts

## Supplementary Material

### 1 ADDITIONAL ABLATION STUDY

#### 1.1 EVALUATION OF TARGETS UNDER OCCLUSION

Thanks to the annotation of the CityUHK-X-BEV dataset, we select those heavily occluded objects for evaluation. Table 1 shows the localization performance (measured by CD) of different methods for these occluded targets. Experimental results show that our proposed method achieves the best localization performance (1.45) compared with the comparison methods including BEV-Net (1.86), CLTR (1.80), DINO (1.85), and AlignDETR (1.93). This verifies the effectiveness of our method under complex occlusion conditions.

#### 1.2 ANALYSIS OF BEV-TO-IMAGE PROJECTION AND FUSION STRATEGIES FOR COLLABORATIVE DUAL-SPACE LEARNING

To explore different BEV-to-IV projection and feature fusion strategies between two branches, we conduct experiments for analysis, as shown in Table 2. For cross-space projection, we compare MLP against the direct inverse homographic transform. As observed, the inverse homographic transform results in poor localization performance. It reflects that, the BEV-to-IV projection is not simply project BEV features to IV for enhancement, which possibly introduces undesired noises. Instead, MLP enables the projection to leverage the salient BEV contexts to enhances the IV features. For fusion strategies, we evaluate the trivial addition, cross-attention based fusion, and our practice which first concatenates features along the channel dimension and then compresses them. It is observed that both addition and cross-attention operation lead to performance degradation, due to the unaligned features between two spaces.

Table 1: Evaluation for occluded targets on CityUHK-X-BEV.

	BEV-Net	CLTR	DINO	AlignDETR	Ours
Localiz. CD	1.86	1.80	1.85	1.93	<b>1.45</b>

Table 2: We evaluate different projection and fusion strategies for collaborative dual-space learning.

Projection	Fusion	Localization CD ↓	Local Risk %IoU ↑	Global Risk MSE × ↓
Inverse Homography	Concat	0.66	80.56	<b>1.63</b>
MLP	Add	0.65	81.33	2.28
MLP	Cross Atten.	0.69	79.26	2.07
MLP	Concat	<b>0.57</b>	<b>83.64</b>	2.05

Table 3: Analysis on the threshold of proximity-aware suppression.

Threshold	Localization CD × 1 m ↓	Local Risk %IoU ↑	Global Risk MSE × 10 <sup>-4</sup> ↓
0.00010	0.59	82.22	<b>1.95</b>
0.00025	0.58	82.44	2.10
0.00050	<b>0.57</b>	<b>83.64</b>	2.05
0.00100	0.59	81.31	2.06

### 1.3 ANALYSIS ON THE THRESHOLD OF PROXIMITY-AWARE SUPPRESSION $\epsilon$

In the proximity-aware suppression loss, we simply set the threshold  $\epsilon$  of the pairwise distance for determining whether to suppress the certain point pairs,  $\epsilon = 0.0005$ . To verify the threshold is optimal, we select four different values ranging from 0.0001 to 0.001 and calculate the corresponding metrics, as shown in Table 3. As shown, the threshold 0.0005 gives rise to the optimal performance.

Table 4: Analysis on BEV features from different layers for projection.

Block	Localization $CD \times 1 \text{ m} \downarrow$	Local Risk $\%IoU \uparrow$	Global Risk $MSE \times 10^{-4} \downarrow$
0	0.60	81.65	1.91
1	0.60	82.11	<b>1.31</b>
3	0.59	82.44	1.48
6	<b>0.57</b>	<b>83.64</b>	2.05

### 1.4 BEV FEATURES FROM DIFFERENT LAYERS FOR PROJECTION

Table 4 evaluates the results of BEV features collected from the 1st, 3rd, and 6th blocks of the encoder projected to IV. It reveals that the features obtained from the last (i.e., 6th) encoder block are the most effective.

### 1.5 EFFECTS OF THE WEIGHTING FACTOR $\omega$ OF $\mathcal{L}_{PM}$

To verify that the value set by the weighting factor is optimal, we set different values and calculate the corresponding metrics, the results are shown in Table 5.

Table 5: Analysis on the weighting factor of the point matching loss.

Weight	Localization $CD \times 1 \text{ m} \downarrow$	Local Risk $\%IoU \uparrow$	Global Risk $MSE \times 10^{-4} \downarrow$
1.0	0.69	78.96	2.12
2.5	<b>0.57</b>	<b>83.64</b>	2.05
4.0	0.64	82.76	<b>1.63</b>

### 1.6 HEAD VS. FEET PREDICTION

We visually compare the results obtained by predicting head coordinates and feet locations in Fig. 1. As highlighted, the model based on head prediction tends to cause imprecise results deviated from the ground-truth coordinates.

## 2 MORE EXAMPLES

In Fig. 2, we present more multi-person BEV localization results in comparison to BEV-Net (Dai et al., 2021), DINO (Zhang et al., 2022), CLTR (Liang et al., 2022), and Mask-RCNN (He et al., 2017).

In Fig. 3 and Fig. 4, we visualize the prediction results in BEV and IV Spaces with and without proximity-aware suppression loss  $\mathcal{L}_{PM}^{IV+}$  on the SoccerNet (Cioppa et al., 2022) dataset.

In Fig. 1, we visualized the results of predicting head coordinates and predicting foot coordinates on the CityUHK-X-BEV dataset. Where, 'Feet' denotes regressing the foot coordinates and projecting them into BEV space, whereas 'Head' refers to first regressing the head coordinates and then employing a fixed human height of 1.75 meters to project into BEV space.

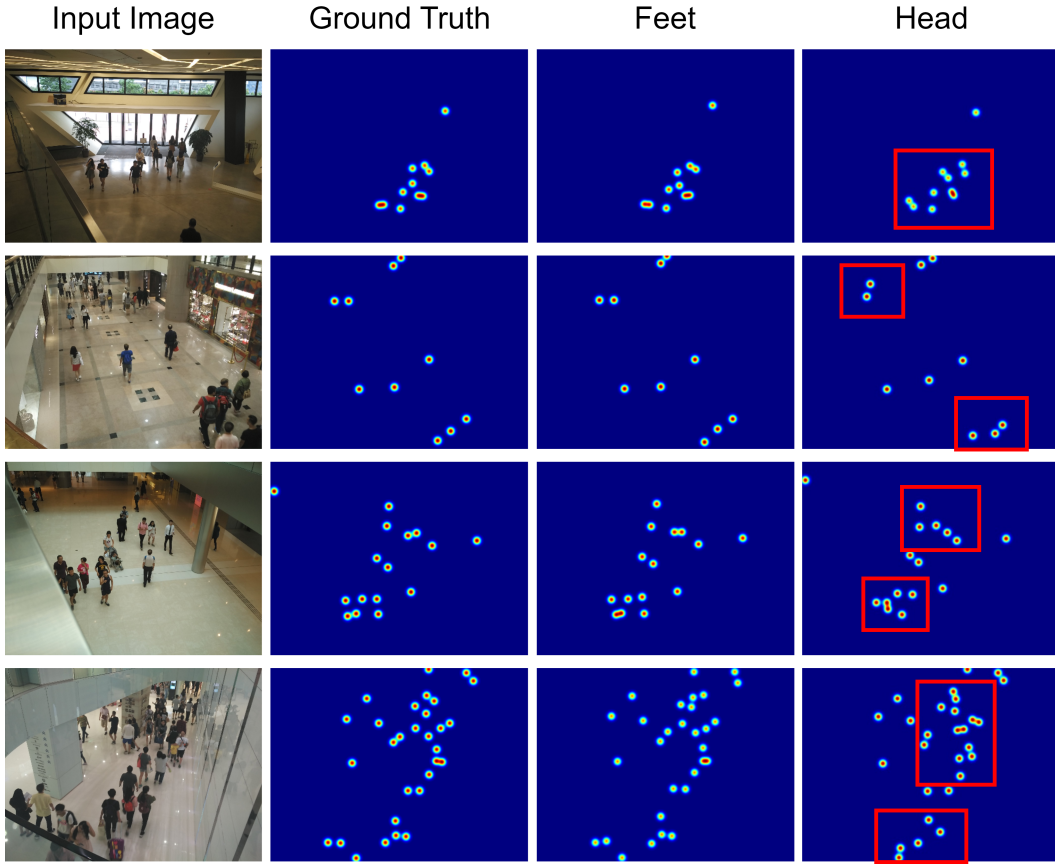


Figure 1: We visualize the results of predicting head coordinates and predicting foot coordinates on the CityUHK-X-BEV dataset.

## REFERENCES

- Anthony Cioppa, Silvio Giancola, Adrien Deliege, Le Kang, Xin Zhou, Zhiyu Cheng, Bernard Ghanem, and Marc Van Droogenbroeck. Soccernet-tracking: Multiple object tracking dataset and benchmark in soccer videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3491–3502, 2022.
- Zhirui Dai, Yuepeng Jiang, Yi Li, Bo Liu, Antoni B Chan, and Nuno Vasconcelos. Bev-net: Assessing social distancing compliance by joint people localization and geometric reasoning. In *Proceedings of the IEEE/CVF international conference on computer Vision*, pp. 5401–5411, 2021.
- Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.
- Dingkang Liang, Wei Xu, and Xiang Bai. An end-to-end transformer model for crowd localization. In *European Conference on Computer Vision*, pp. 38–54. Springer, 2022.
- Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

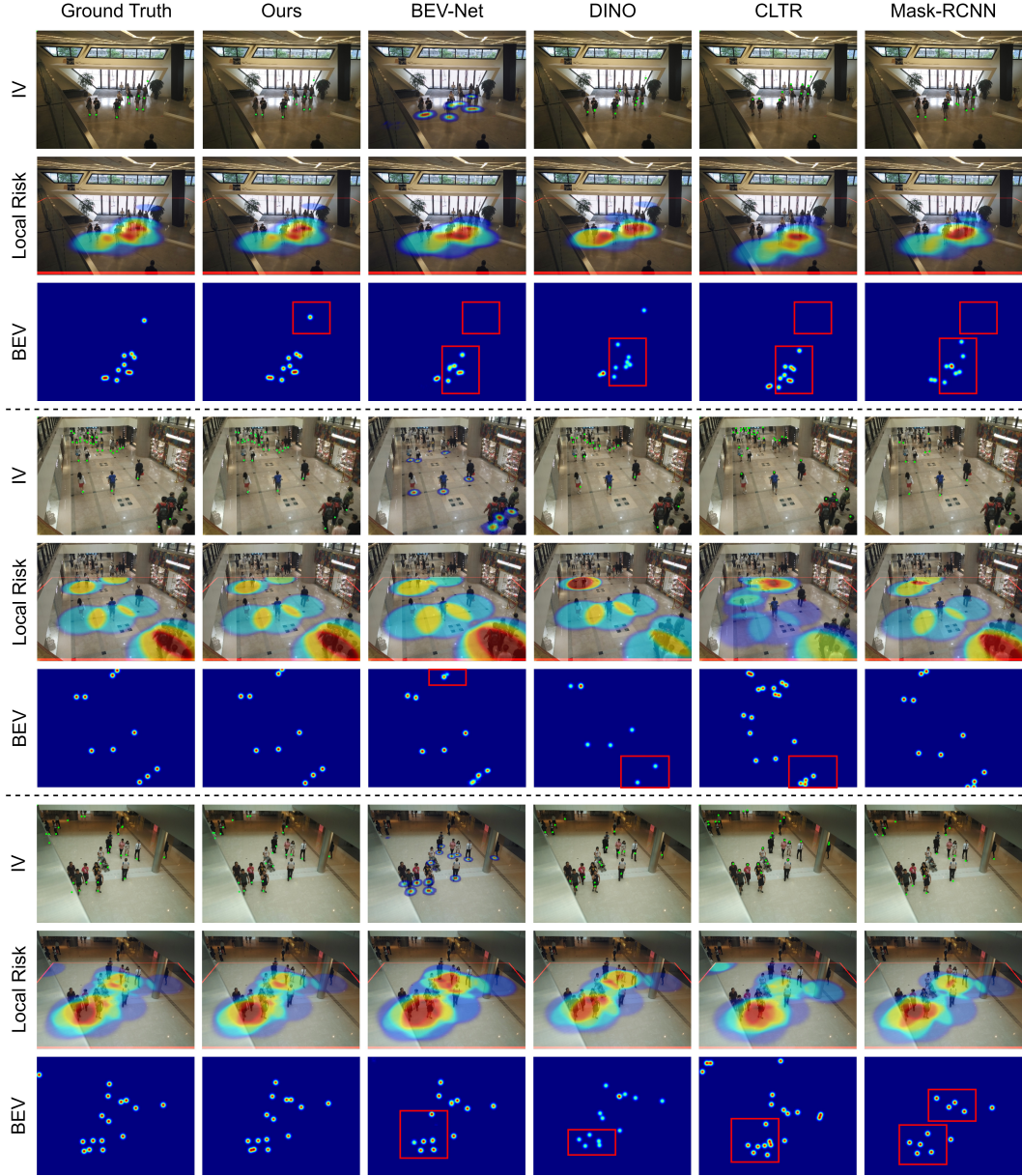


Figure 2: We visualize the multi-human BEV localization results on the CityUHK-X-BEV dataset.





Figure 3: We visualize the prediction results in BEV and IV Spaces with and without the loss  $\mathcal{L}_{PM}^{IV+}$  on the SoccerNet dataset.



Figure 4: We visualize the prediction results in BEV and IV Spaces with and without the loss  $\mathcal{L}_{PM}^{IV+}$  on the SoccerNet dataset.