Multi-view Pedestrian Detection Results and Analysis

Aung Sithu

MS-Student

14.12.23

UST
과학기술연합대학원대학교
UNIVERSITY OF SCIENCE & TECHNOLOGY

KIST 한국과학기술연구원
Korea Institute of Science and Technology

Contents

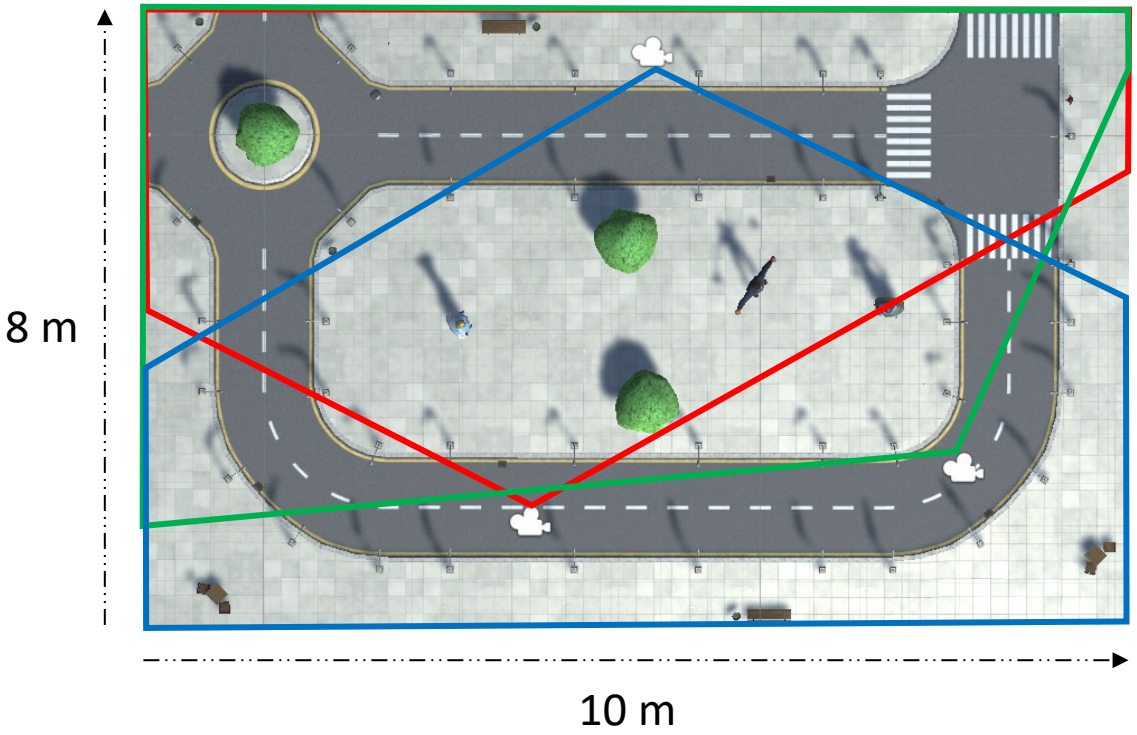
Background

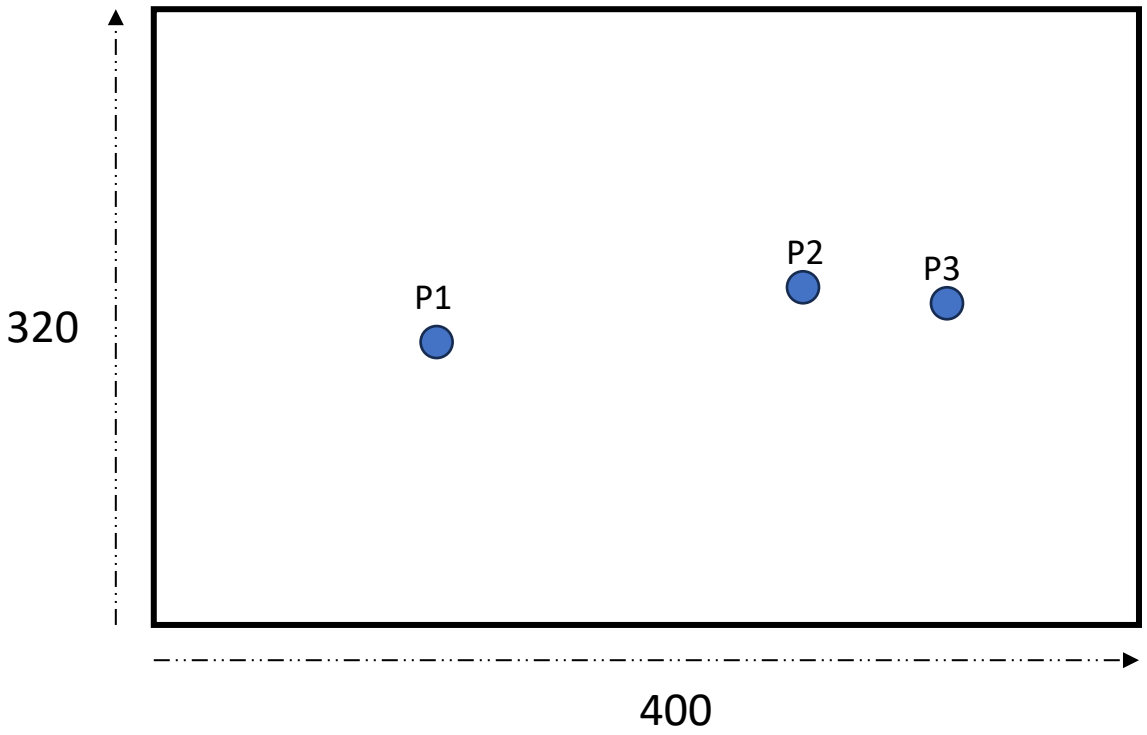
Multi-view Pedestrian Detection




Background

Localization within Ground Plane

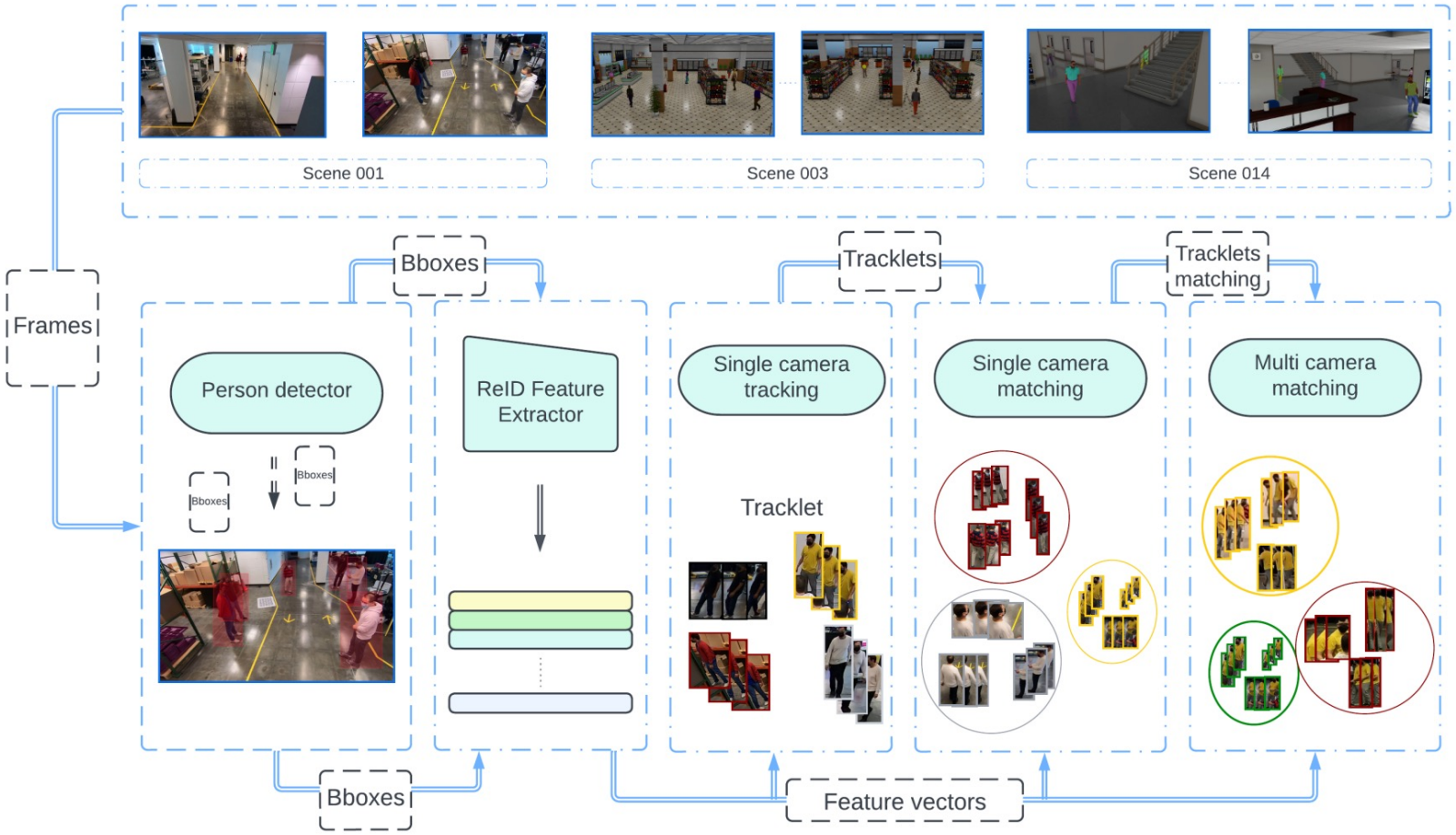




Discretized with 2.5 cm resolution

Background

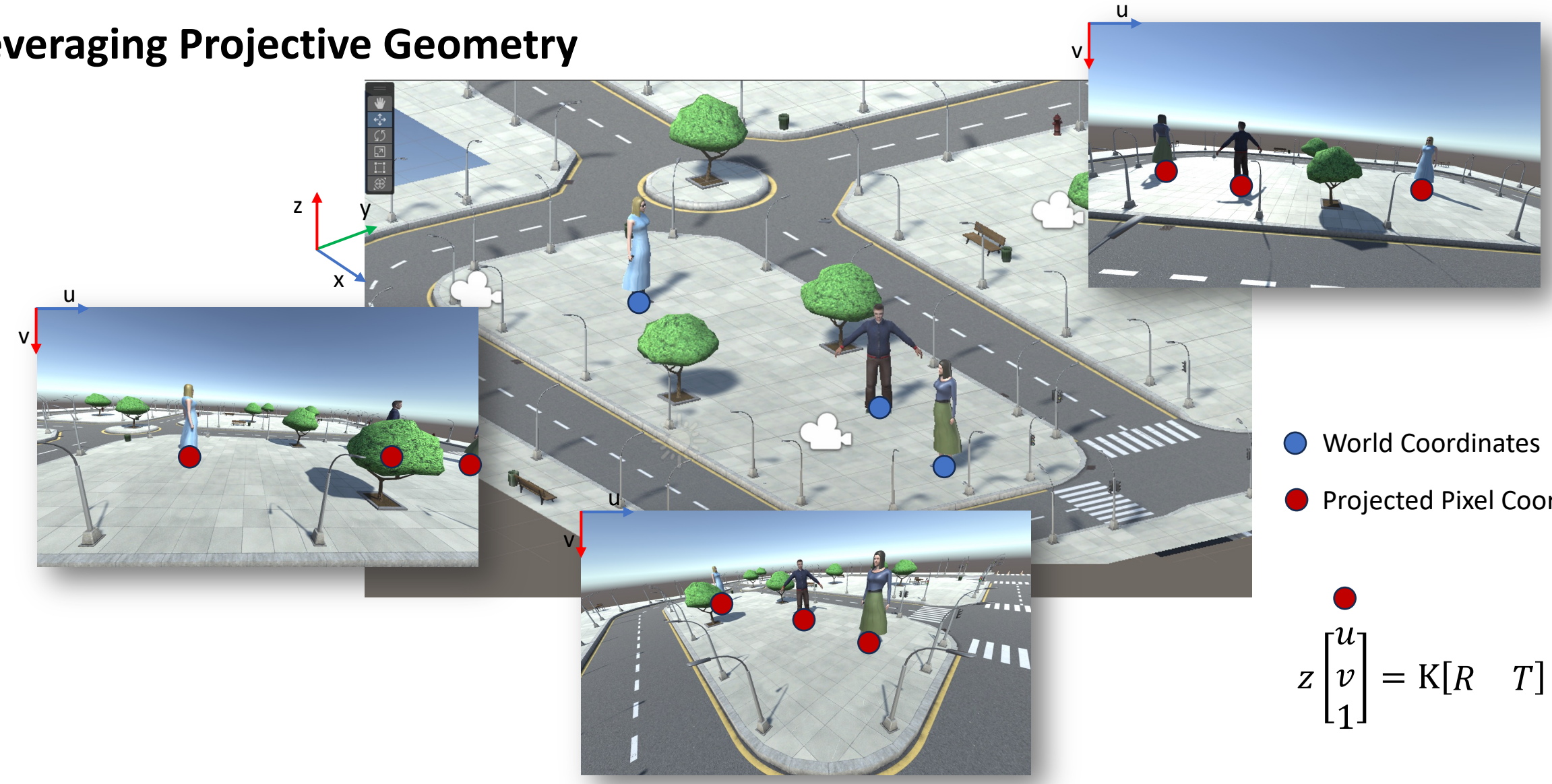
Conventional 2D-based Approaches



Quang Qui-Vinh Nguyen, Huy Dinh-Anh Le, Truc Thi-Thanh Chau, Duc Trung Luu, Nhat Minh Chung, Synh Viet-Uyen Ha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2023, pp. 5496-5506

Background

Leveraging Projective Geometry



- World Coordinates
- Projected Pixel Coordinates

$$z\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathrm{K}[R \quad T]\begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix}$$

Method

Overview

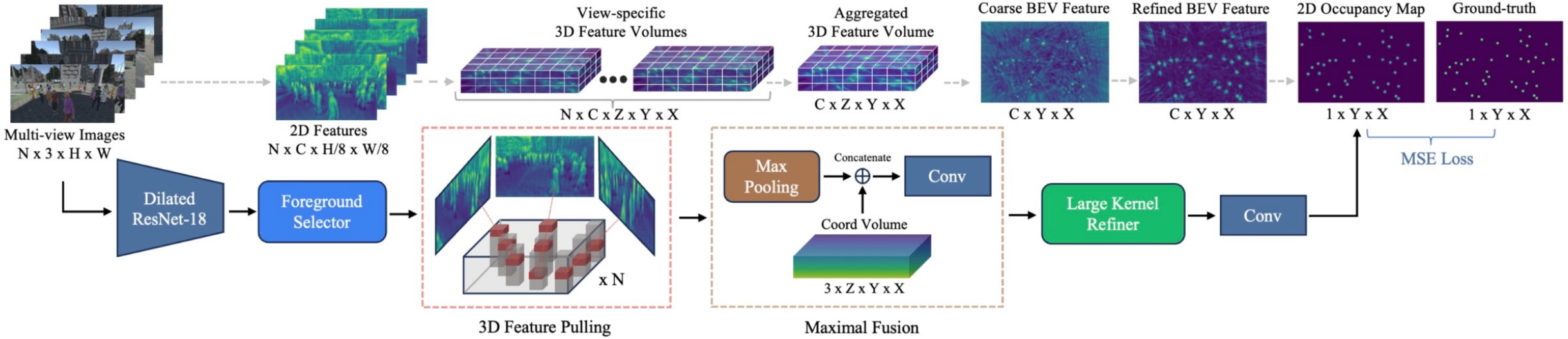


Figure 2. **Overall architecture of the proposed model**. A dilated ResNet-18, coupled with our foreground selector module, is used to extract multi-view features. In the 3D feature-pulling, a sub-pixel 2D feature is pulled for each 3D voxel using projection and bilinear sampling. A maximal fusion module is employed to produce an aggregated 3D feature volume and subsequently, reduce the vertical dimension to create a 2D BEV feature map. Finally, a large kernel refiner module is used to enhance the output, and a 2D occupancy map is predicted. ”Conv” indicates a 1×1 conv. layer.

Experiments

Backbone	**Dilated ResNet-18**
Channel Dimension	256
Voxel Size	10cm x 10cm x 20cm
GPU	4x Nvidia A100 GPUs
Optimizer	AdamW
Weight decay	1e-4
Batch size	4 (1 on each GPU)

Experiments

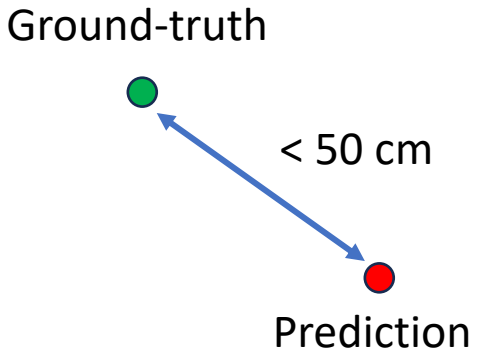
Evaluation Protocols

Experiments

Metrics (2D Localization)

$$MODA = 1 - \frac{FP + FN}{N}$$ ➡ Considers both false positives and false negatives.

$$MODP = \frac{\sum 1 - \frac{d[d < t]}{t}}{TP}$$ ➡ Considers localization precision of true positives.



$$Precision = \frac{TP}{FP + TP}$$ ➡ Considers true positive rate.

$$Recall = \frac{TP}{N}$$ ➡ Considers accurate localization performance.

d = distance from a detection to its ground truth
t = threshold (50 cm)
N = total number of detections in the ground truth

Results

Same-domain Testing (Quantitative Results)

Method	WildTrack				MultiviewX			
	MODA	MODP	Prec.	Recall	MODA	MODP	Prec.	Recall
MVDet [7]	88.2	75.7	94.7	93.6	83.9	79.6	96.8	86.7
GMVD* [20]	86.7	76.2	95.1	91.4	88.2	79.9	96.8	91.2
SHOT [19]	90.2	76.5	96.1	94.0	88.3	82.0	96.6	91.5
MVDeTr† [6]	91.5	**82.1**	**97.4**	94.0	93.7	**91.3**	**99.5**	94.2
MVAug† [3]	93.2	79.8	96.3	97.0	95.3	89.7	99.4	95.9
3DROM† [17]	93.5	75.9	97.2	96.2	95.0	84.9	99.0	96.1
Ours*	**94.1**	78.8	96.4	**97.7**	**95.7**	85.1	98.4	**97.2**

Table 1. **Comparison against state-of-the-art methods**. Same-domain testing on WildTrack and MultiviewX datasets. * shows the method that works with different camera setups while other methods are configured to work on a fixed camera setup as in training. † shows the method that uses additional augmentations. Our method outperforms previous methods with larger MODA and recall scores, accurately identifying the number of individuals involved in the scene.

Results

Same-domain Testing (Qualitative Results)

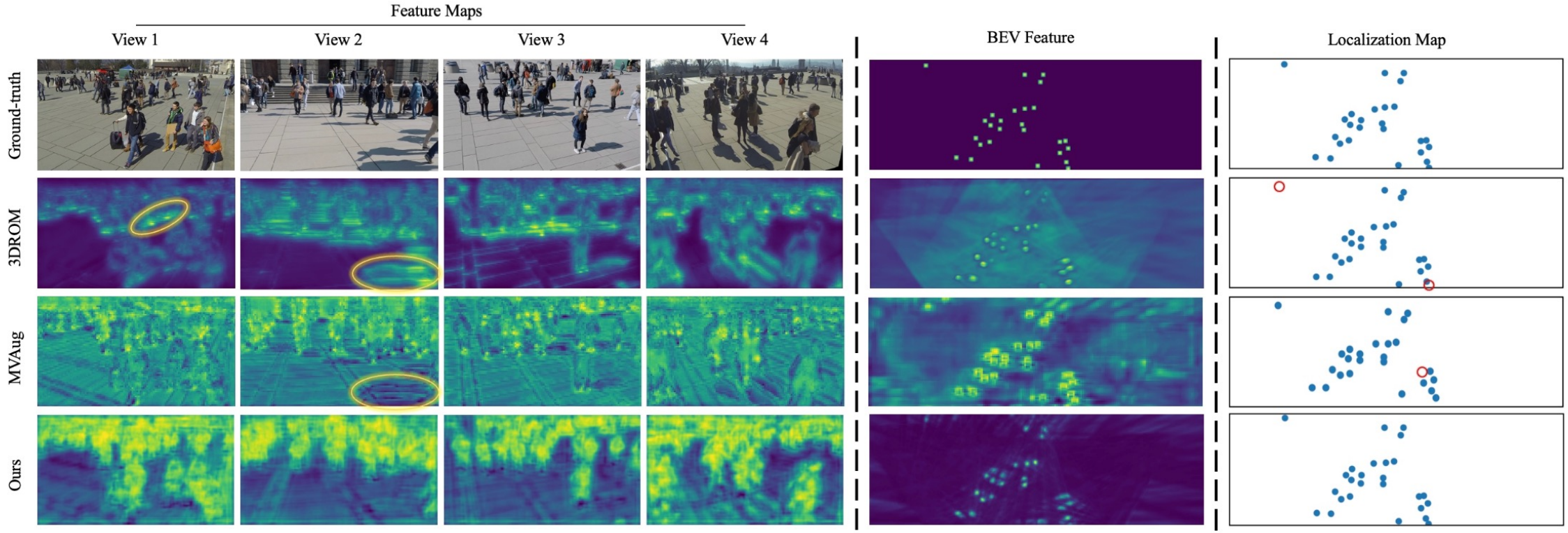


Figure 6. **Qualitative comparison between 3DROM [17], MVAug [3], and our method on the WildTrack dataset**. The initial four columns provide visual representations of example views and the corresponding extracted feature maps. The central column displays the aggregated BEV feature, while the final column illustrates ground plane localizations. Yellow ovals draw attention to the shadow of the person, while red circles emphasize instances of missed detections.

Results

Scene Generalization (Quantitative Results)

Method	NV_t	NV_i	MODA	MODP	Prec	Recall
MVDet [7]	6	6	17.0	65.8	60.5	48.8
MVAug [3]	6	6	26.3	58.0	71.9	50.8
MVDeTr [6]	6	6	50.2	69.1	74.0	77.3
SHOT [19]	6	6	53.6	72.0	75.2	79.8
GMVD [20]	6	6	66.1	72.2	82.0	84.7
3DROM [17]	6	6	67.5	65.6	**94.5**	71.7
Ours	6	6	**76.7**	**74.9**	85.2	**92.8**
GMVD [20]	6	7	70.7	73.8	89.1	80.6
Ours	6	7	**82.6**	**76.2**	**89.6**	**93.4**

Table 2. **Scene generalization evaluation with the MultiviewX**

Method	NV_t	GMVD					WildTrack				
		NV_i	MODA	MODP	Prec.	Recall	NV_i	MODA	MODP	Prec.	Recall
GMVD [20]	3,5,6,7	6,8	68.2	76.3	91.5	75.5	7	80.1	75.6	90.9	89.1
Ours	3,5,6,7	6,8	**73.3**	**76.5**	**93.0**	**79.2**	7	**85.6**	**78.0**	**91.8**	**94.0**

Table 3. **Scene generalization evaluation with the GMVD dataset**. Trained on GMVD train-set and tested on GMVD test-set and real dataset (WildTrack). NV_t and NV_i represent the number of views in training and inference, respectively.

Results

Scene Generalization (Qualitative Results)

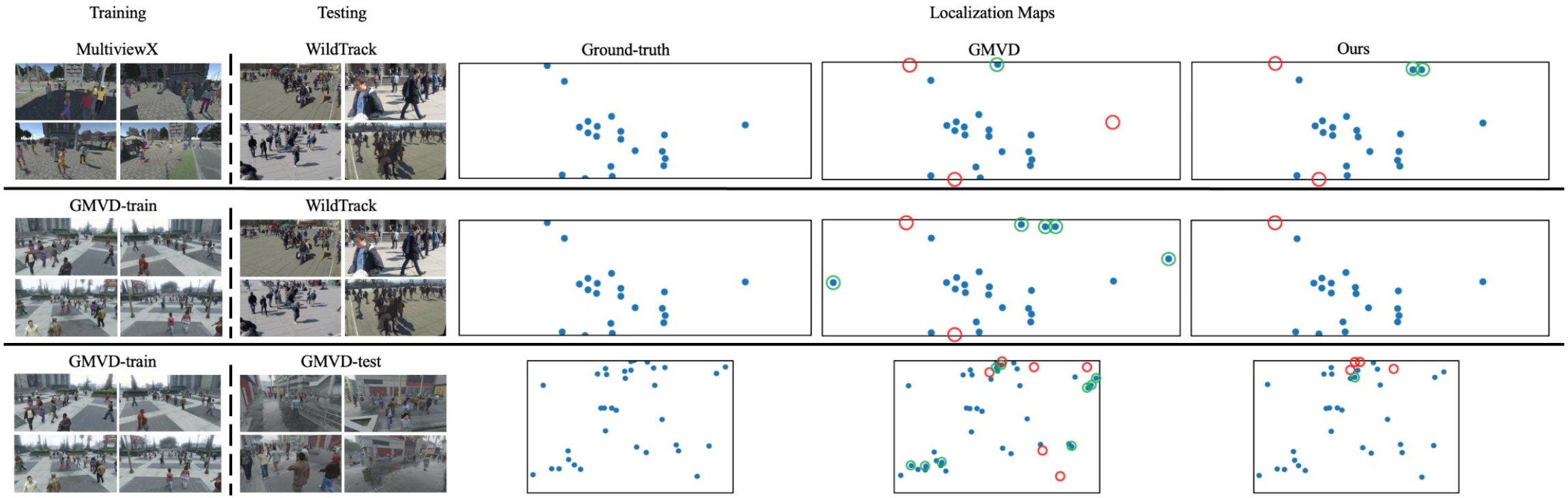
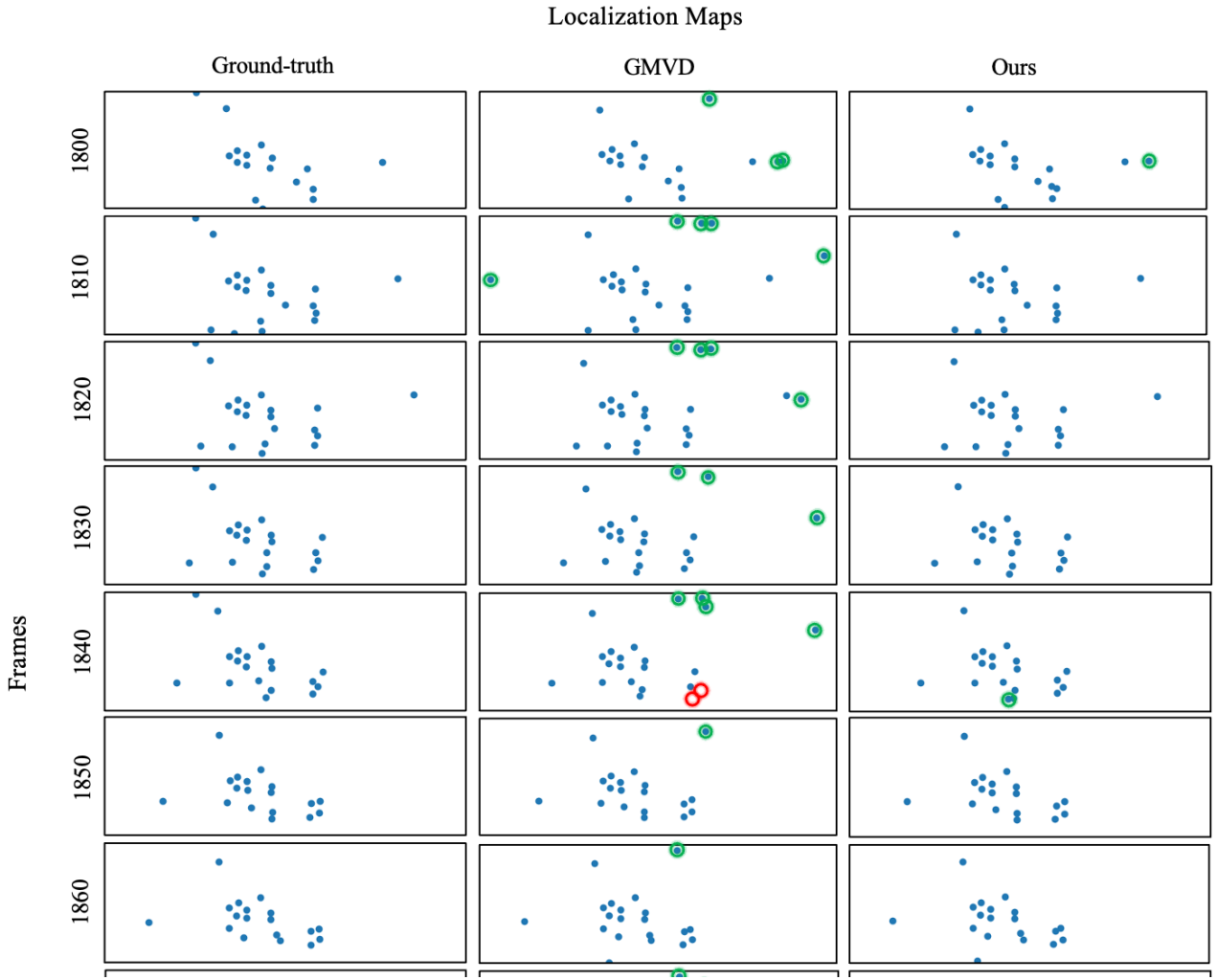


Figure 7. **Qualitative comparison between GMVD [20] and our method on scene generalization**. The first column illustrates samples from the training set, while the second column visualizes samples from the testing set. Subsequent columns depict the ground truth and predicted localization maps. Red circles denote missed detections (false negatives) while green circles denote false positives.

Results

Scene Generalization (Multi-frame Analysis)

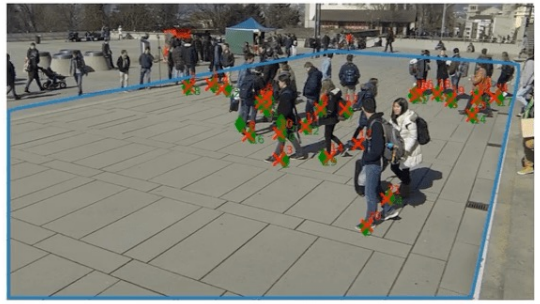


Results

Comparison with Ground-truth

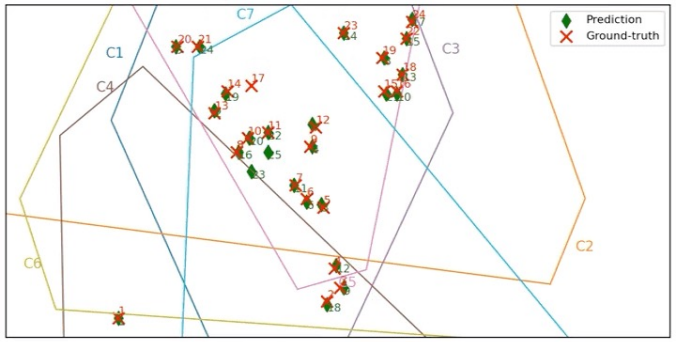
Ours

Camera 1

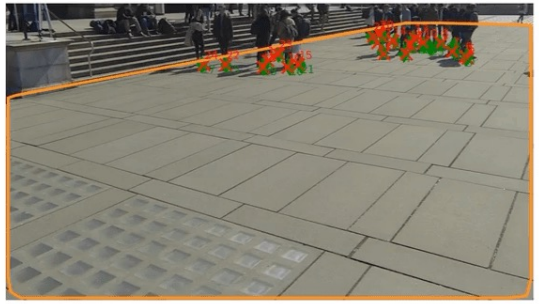


Frame: 1910, Actual: 24, Detected: 25

Ground Plane



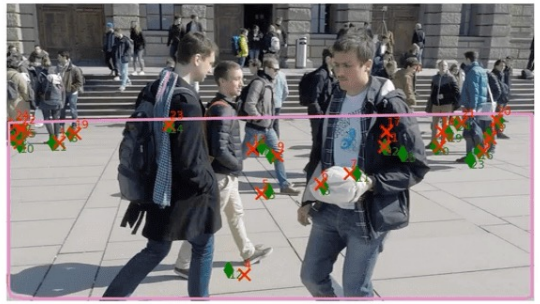
Camera 2



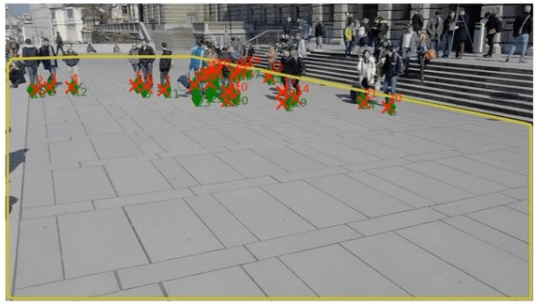
Camera 3

Camera 4

Camera 5



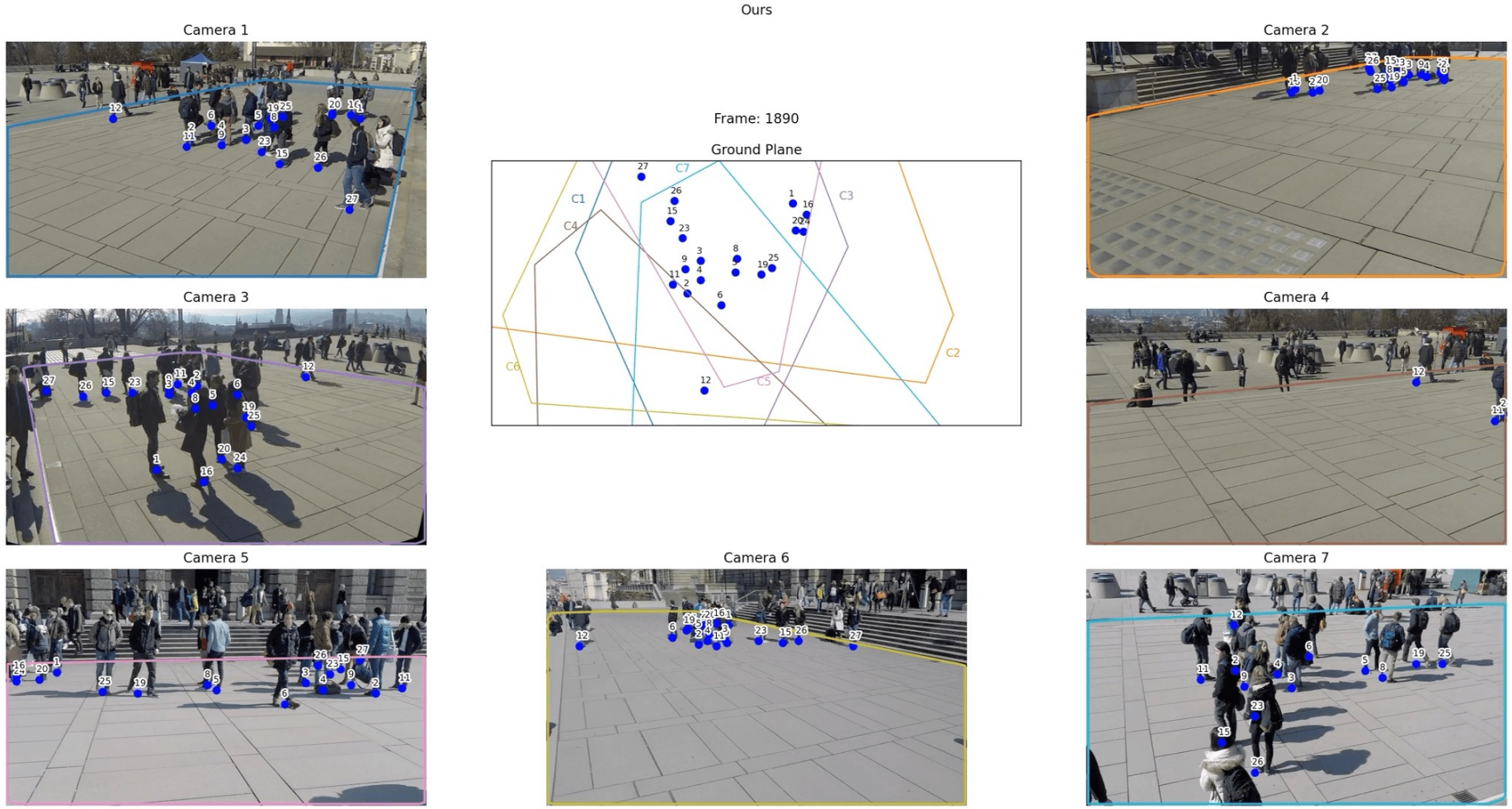
Camera 6



Camera 7

Results

Tracking Results



Notes: The tracked results are shown with a SORT-based online tracking algorithm.

Results

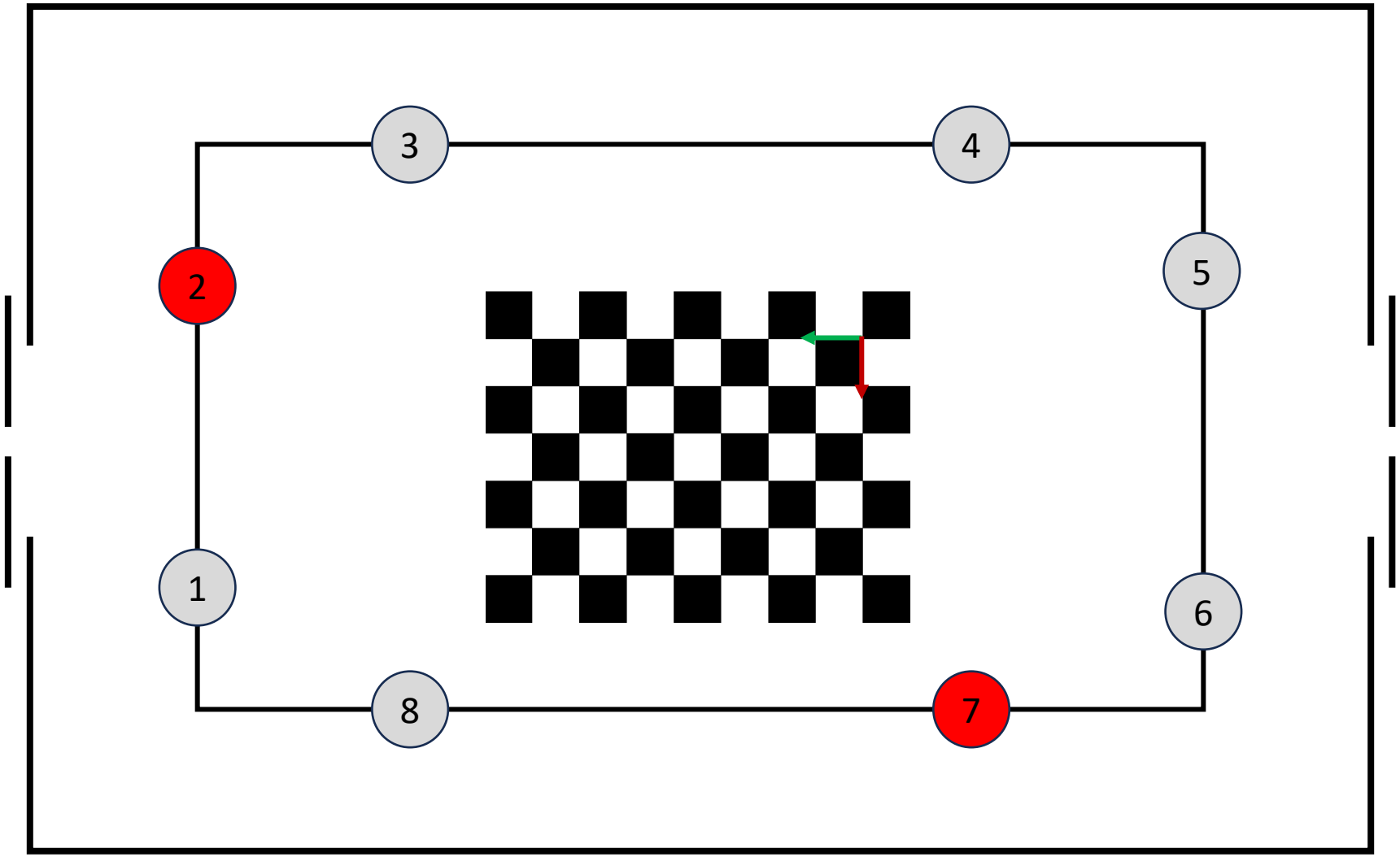
Latency Analysis

GPU	RTX 2080 Ti
Image Size	1280x720
Num. of views	7

Precision\Framework	PyTorch	TorchScript
FP32	4 FPS	5 FPS
FP16	9 FPS	12 FPS
FP16 (960x540)	13 FPS	18 FPS

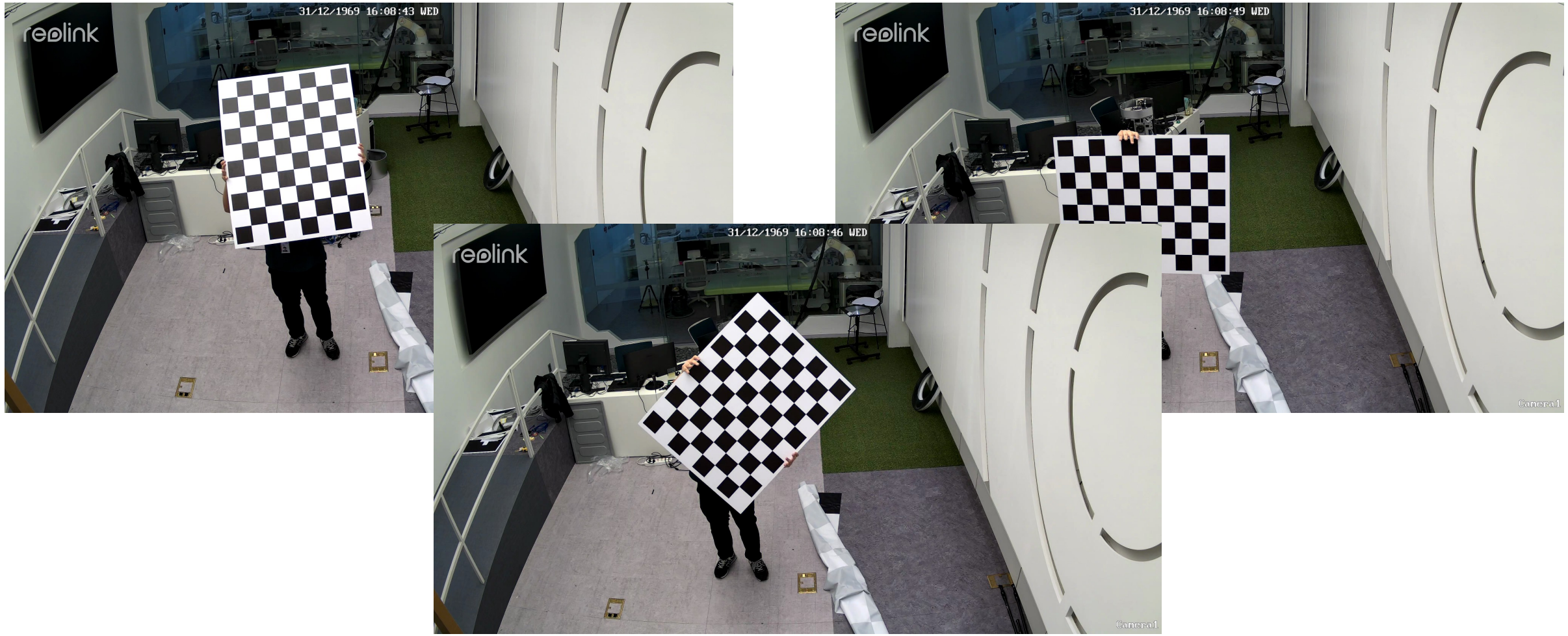
Real-World Testing

Camera Setup



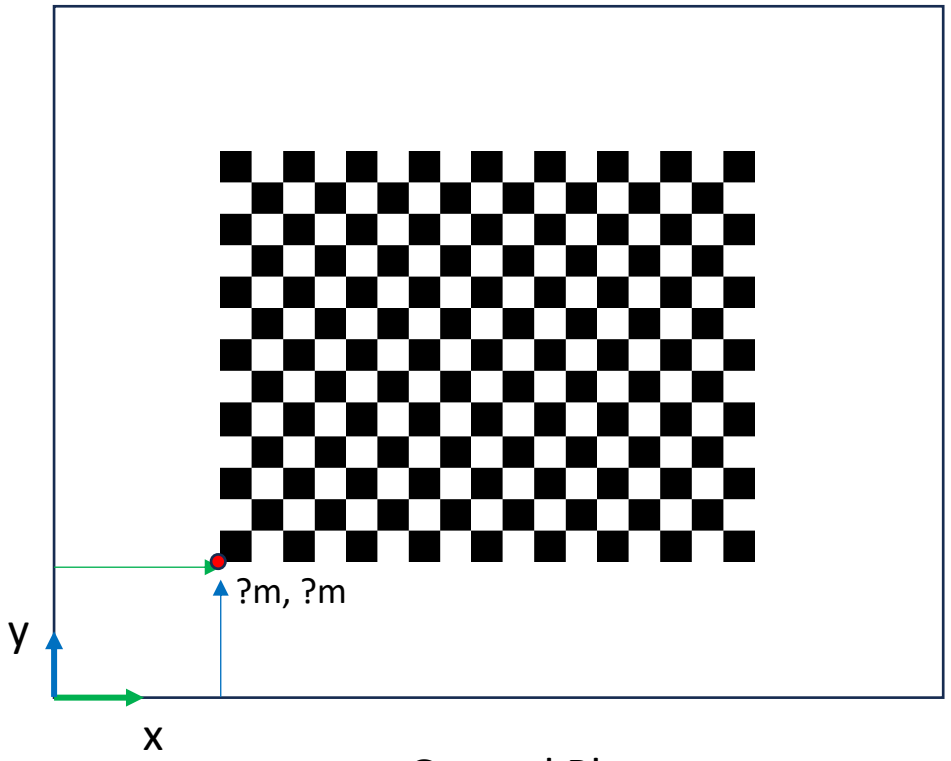
Real-World Testing

Calibration (Intrinsic)



Real-World Testing

Calibration (Extrinsic)



Ground Plane

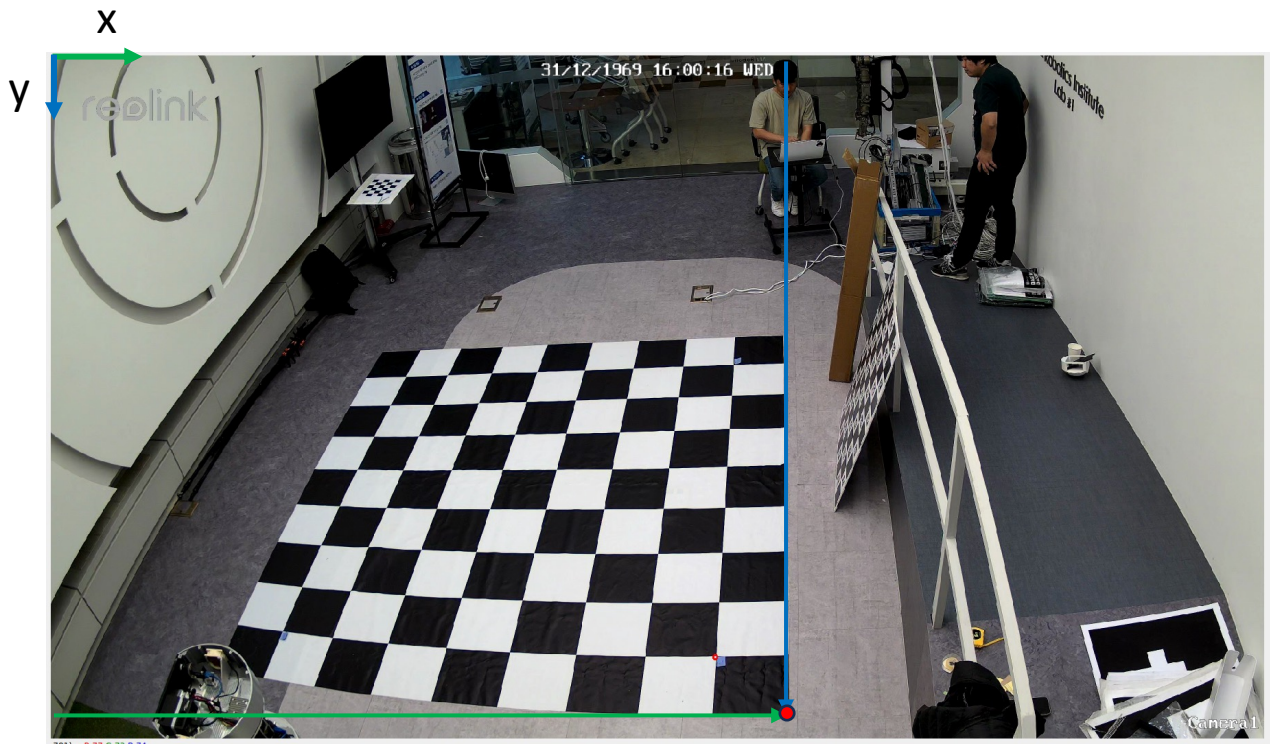


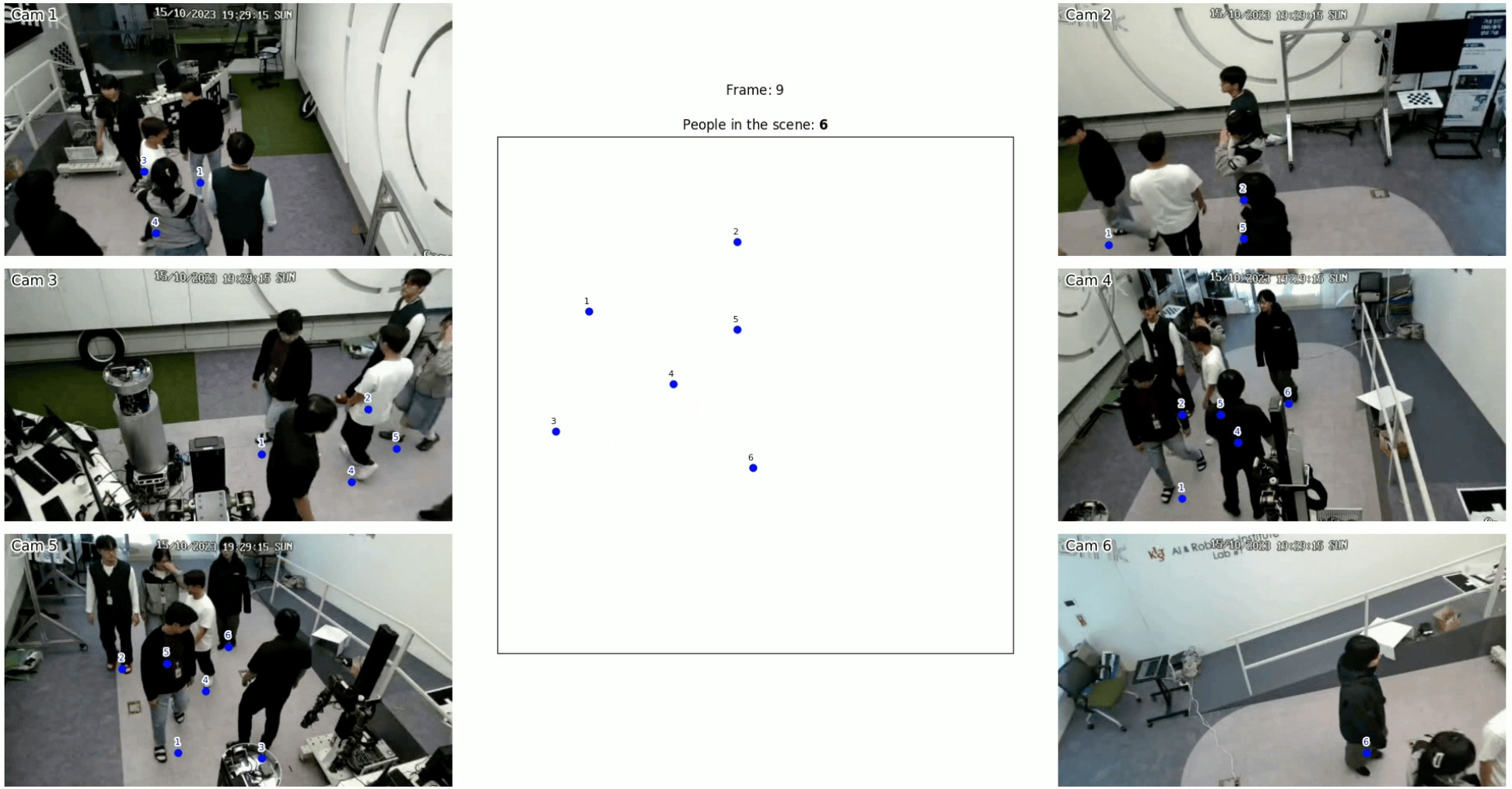
Image View

Real-World Testing

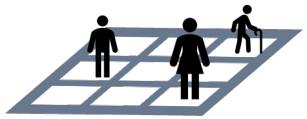
Conventional 2D-based Approach

Real-World Testing

Our 3D-based Approach



Conclusion



- We introduced an **effective multi-view detection framework** with novel modules which **significantly enhances the overall performance**.
- The proposed model showcases **adaptability across diverse scenes** with varying camera setups.
- The proposed model **achieves zero-shot performance** on real-world scenarios while training with synthetic datasets.
- The *performance with larger datasets might not be optimal* since our model was initially designed to excel on small datasets.
- The *model's capacity to capture the heightened complexity present in the larger datasets* could be a potential area for enhancement.