Multi-view Pedestrian Detection

Aung Sithu

MS-Student

12.10.23

UST
과학기술연합대학원대학교
UNIVERSITY OF SCIENCE & TECHNOLOGY

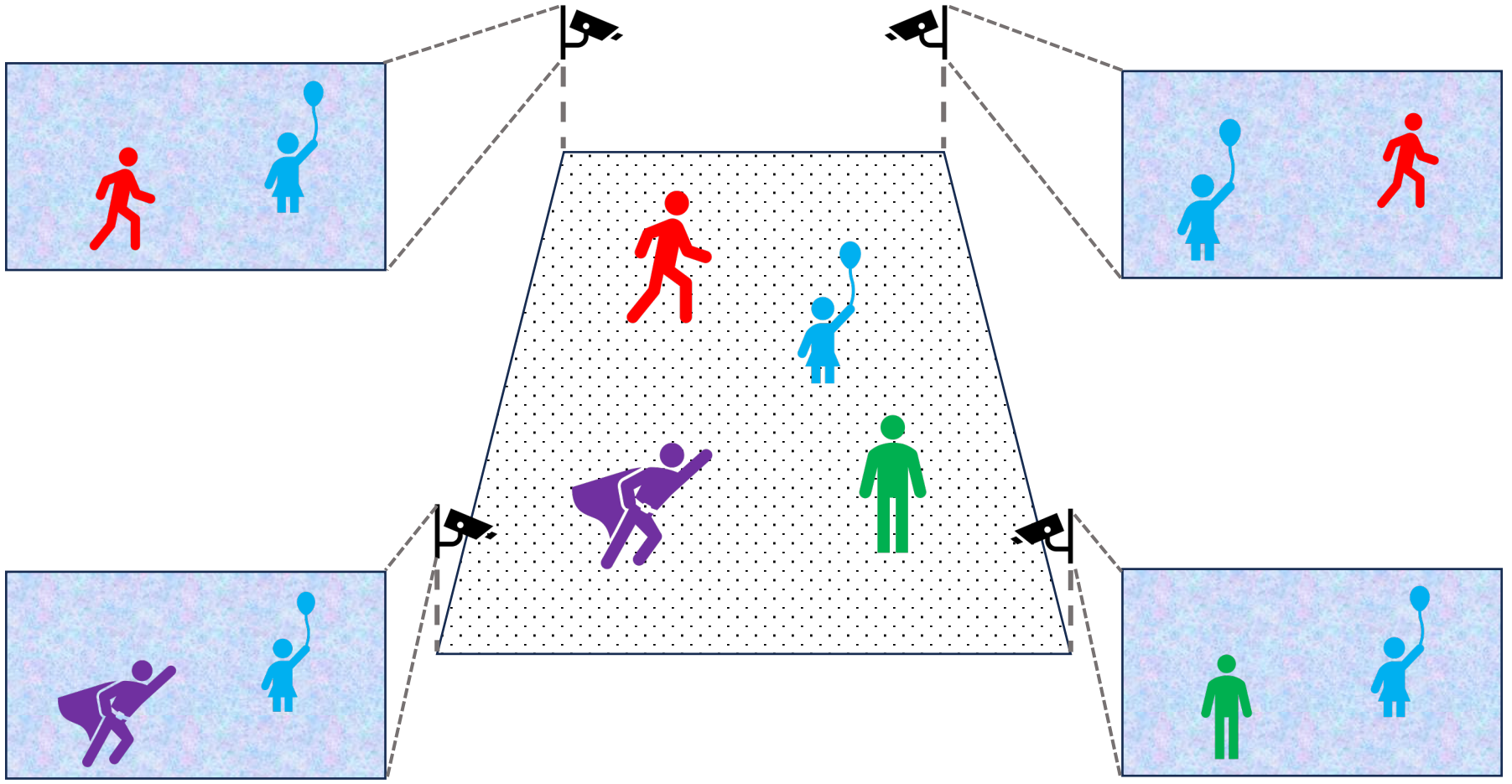
KIST 한국과학기술연구원
Korea Institute of Science and Technology

Contents

- Background
- Preliminaries
- Related Work (2D to BEV Feature Transform)
- Previous Work (Multi-view Detection)

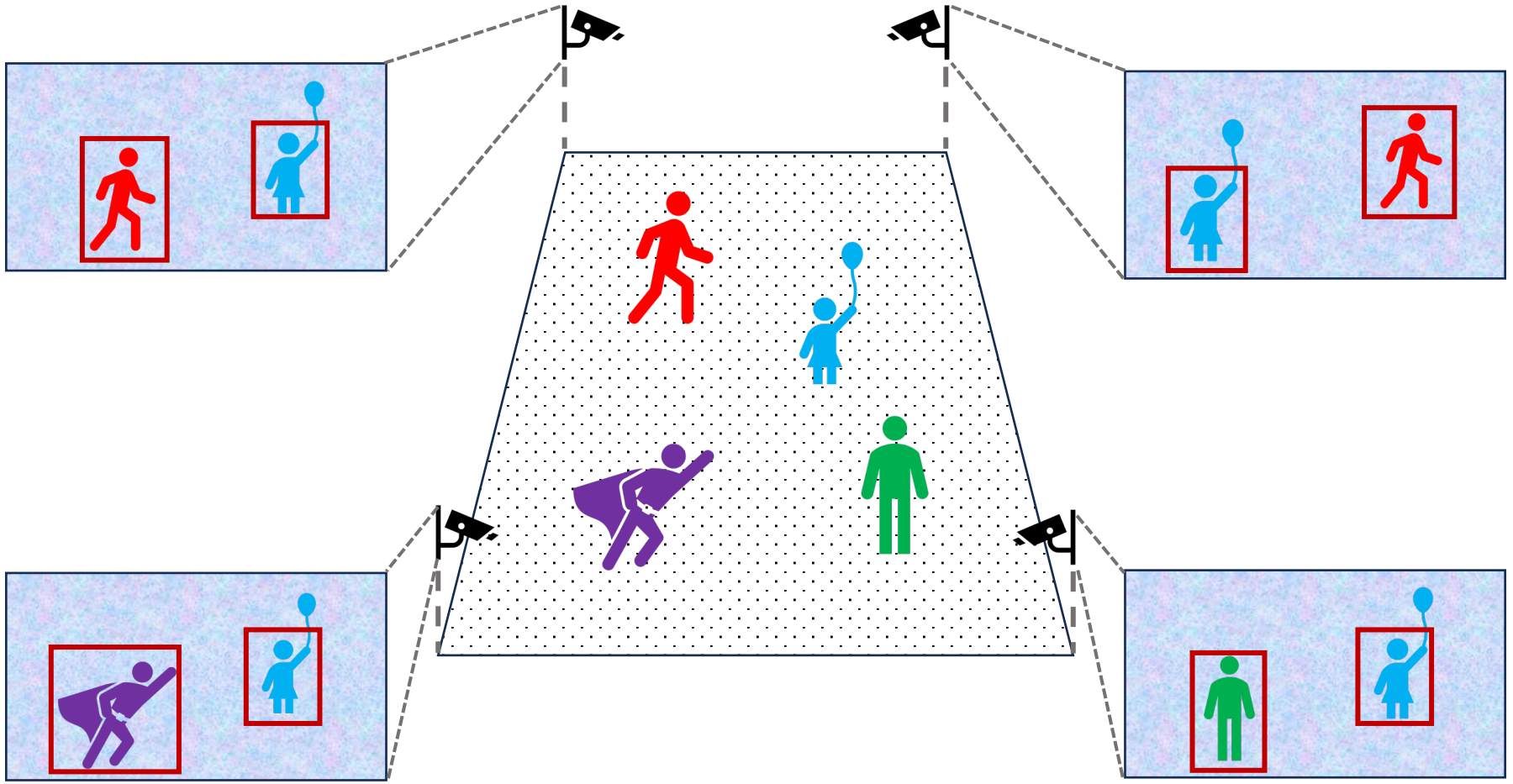
Background

Multi-camera Indoor People Localization



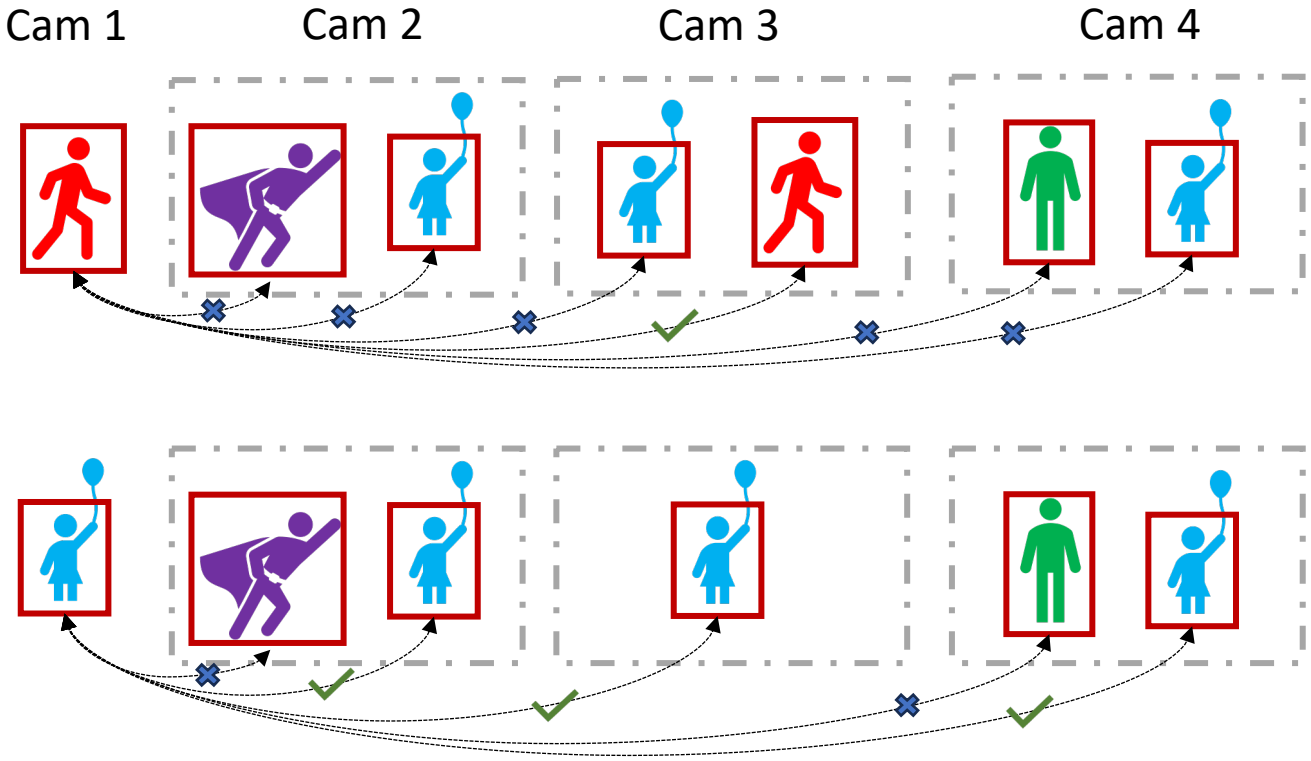
Background

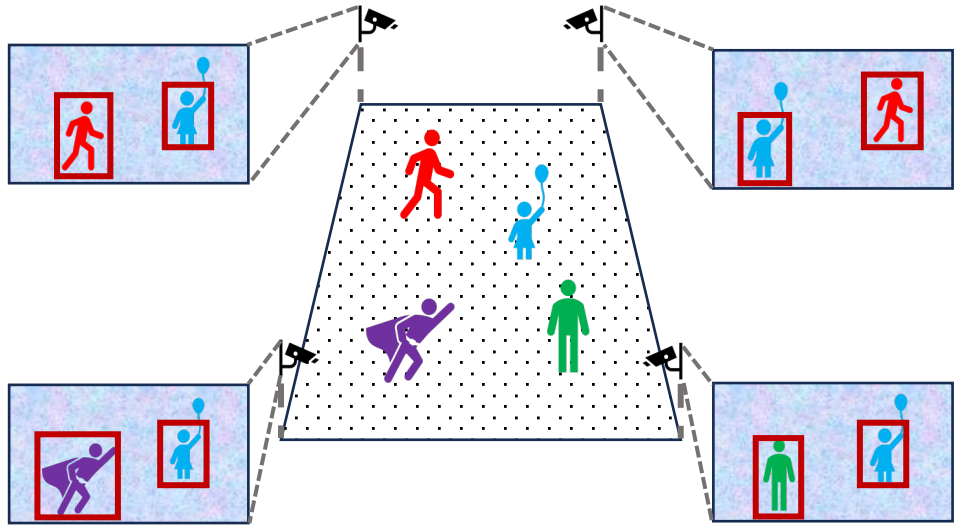
Conventional Approach (2D Detection + Similarity Ranking + MOT)



Background

Conventional Approach (Continued)

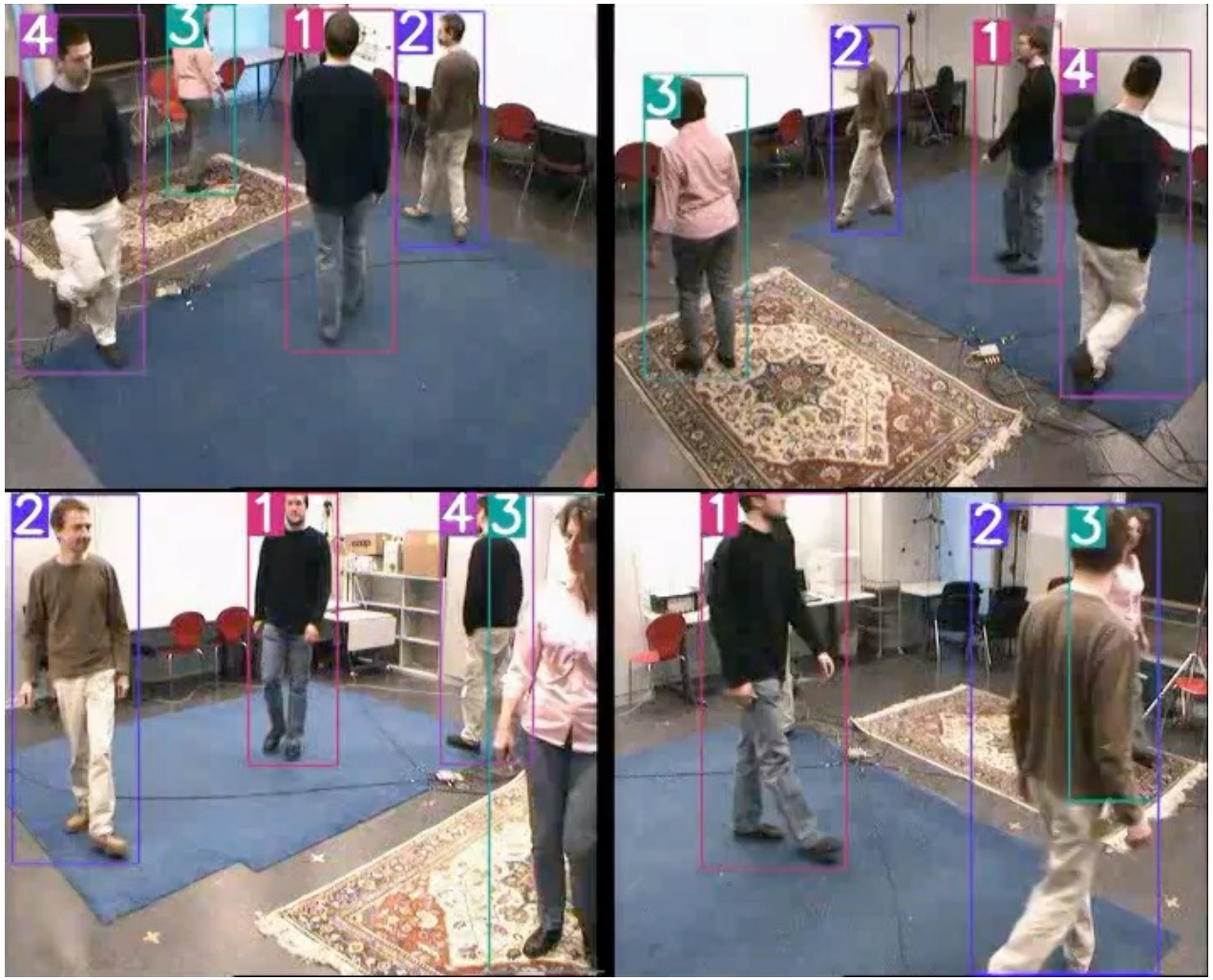




Severe Occlusions

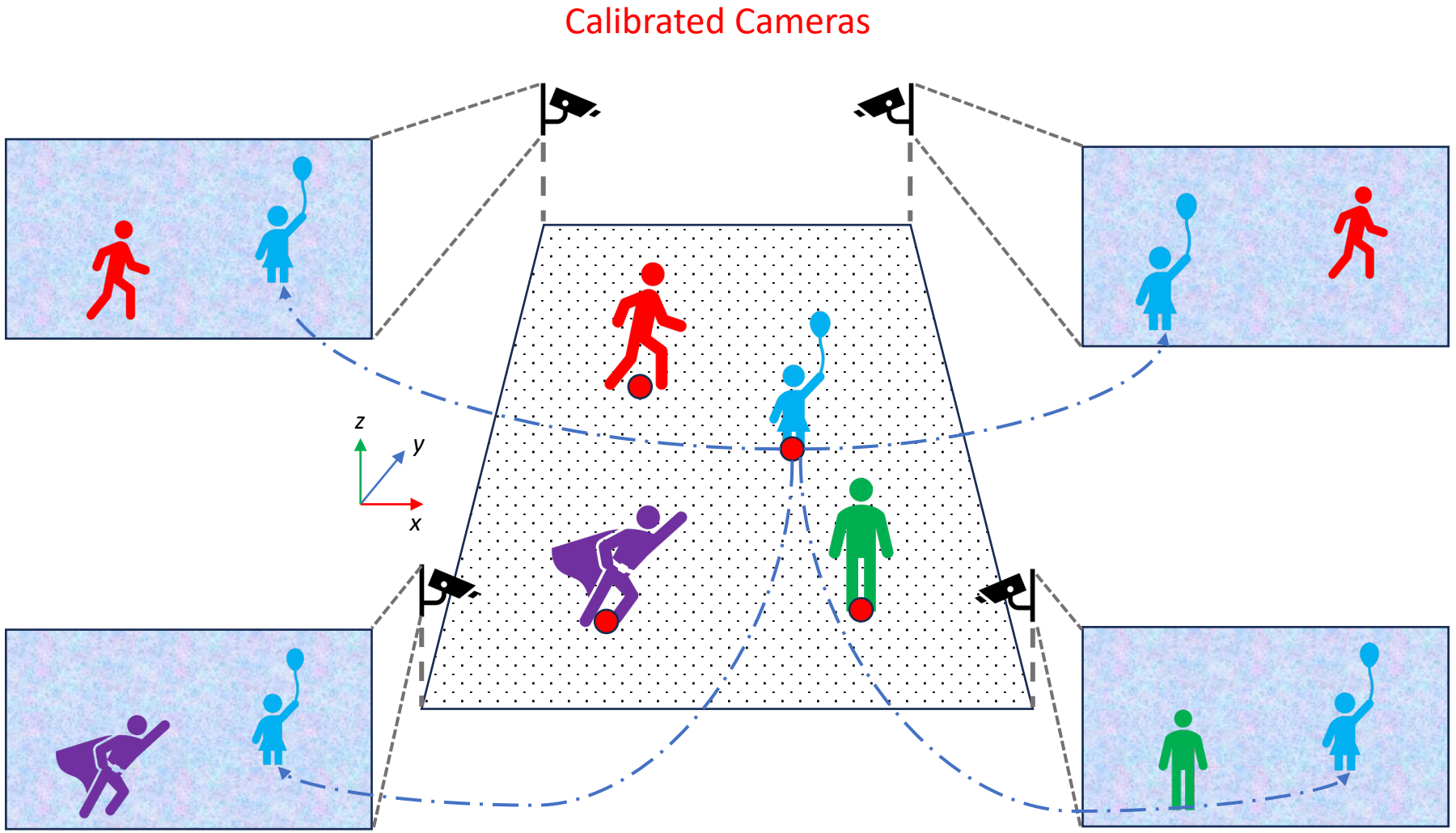
Background

Conventional Approach (Continued)



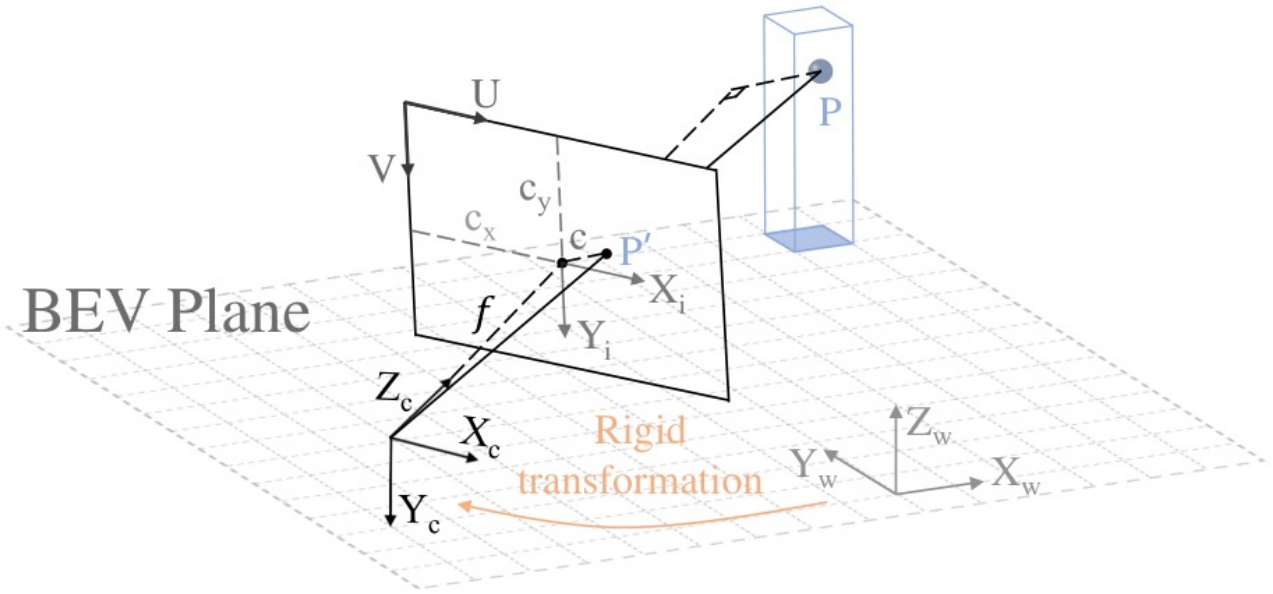
Background

Towards 3D Bird's-Eye-View (BEV) Perception



Preliminaries

Projective Geometry



- $P = (X_w, Y_w, Z_w)$ -> world coordinate
- $P' = (X_c, Y_c, Z_c)$ -> camera coordinate
- $P' = (X_i, Y_i)\ or\ (U, V)$ -> pixel coordinate or image coordinate

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

- K = intrinsic
- $[R \quad T]$ = extrinsic

Given the world coordinate of P

$\mathrm{K}[R \quad T]$ →

Retrieve the pixel coordinates of P’

Preliminaries

Ground Plane AoI (Area of Interest)

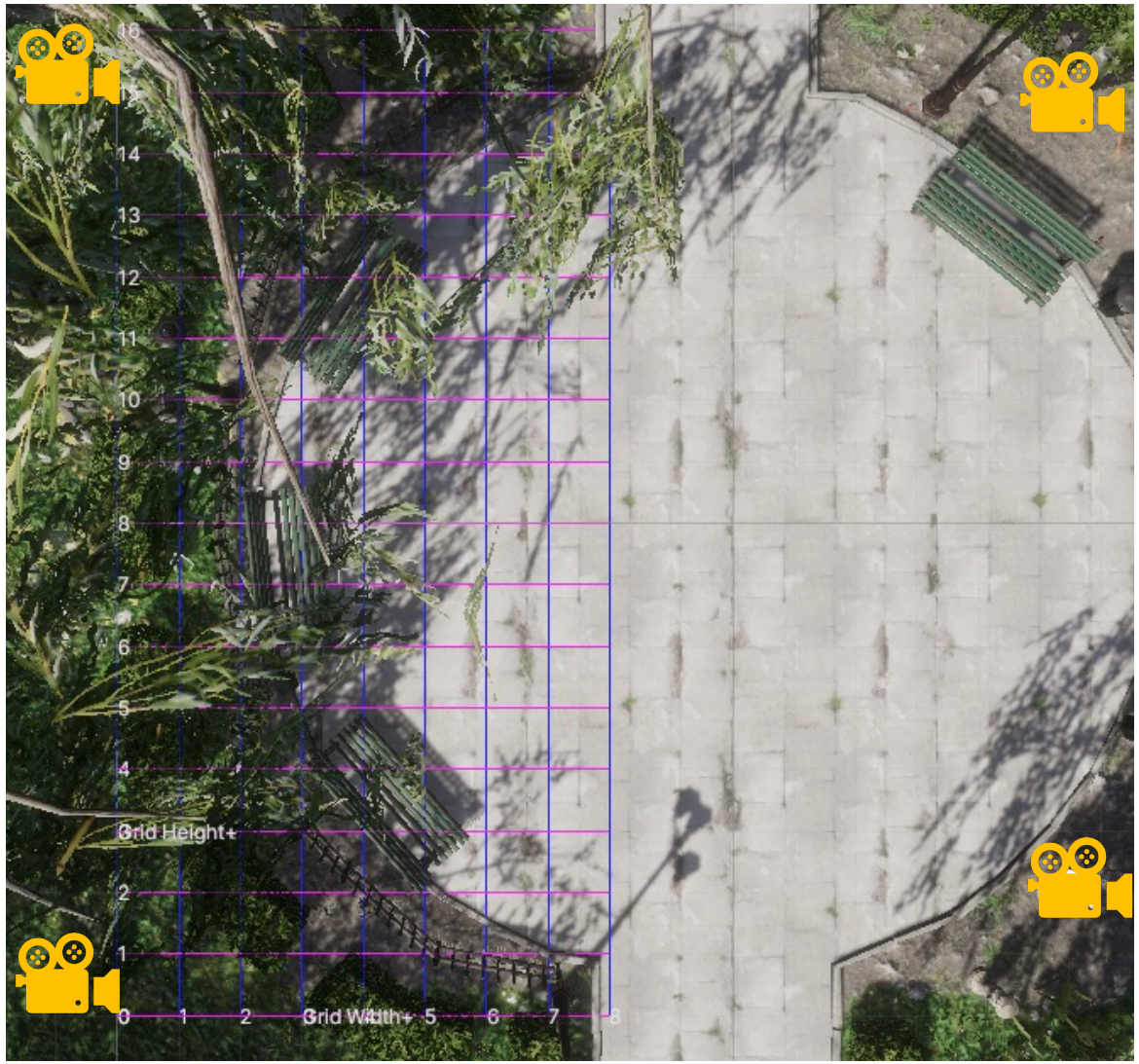
Grid resolution ↕ 0.025 m

Ground plane height = 15 m

Grid height = 600

↑ Ground plane width = 10 m

Grid width = 400

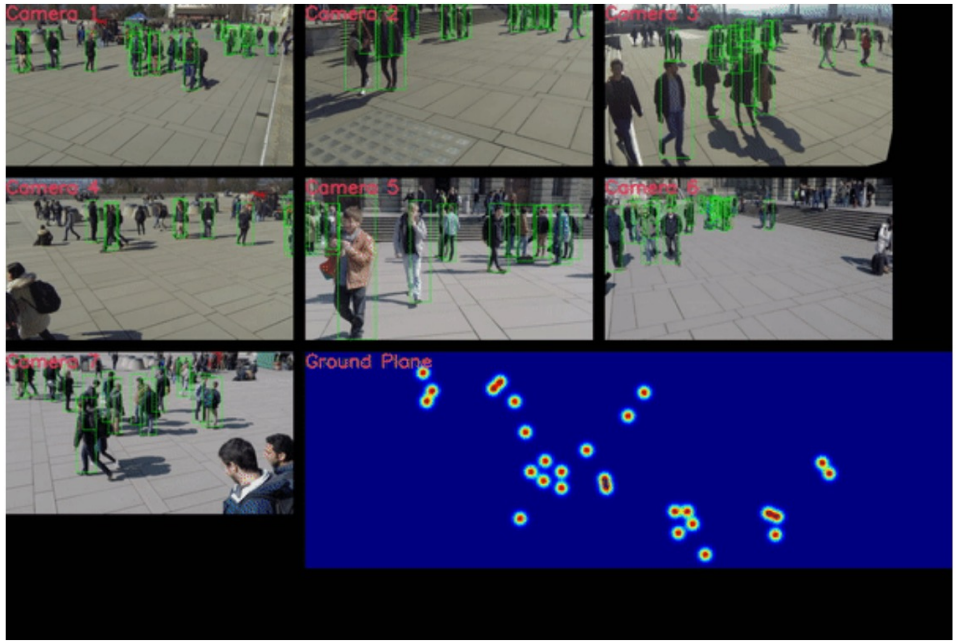


Preliminaries

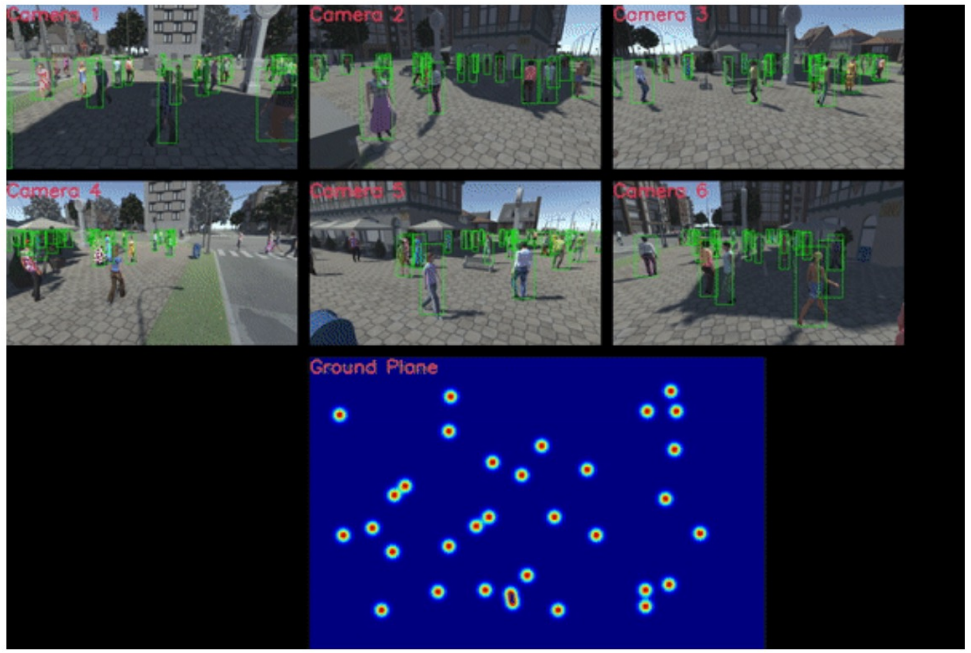
Multi-view Datasets

	#camera	resolution	frames	area	crowdedness	avg. coverage
Wildtrack	7	1080×1920	400	12×36 m^2	20 person/frame	3.74 cameras
MultiviewX	6	1080×1920	400	16×25 m^2	40 person/frame	4.41 cameras

Wildtrack



MultiviewX



Preliminaries

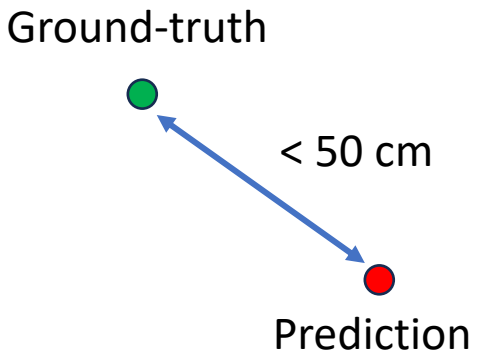
Metrics (2D Localization)

$$MODA = 1 - \frac{FP + FN}{N}$$ ➡ 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

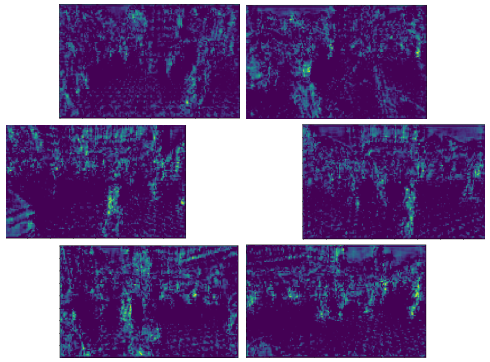
Preliminaries

Overall Framework

Multi-view Images
B x N x 3 x H x W

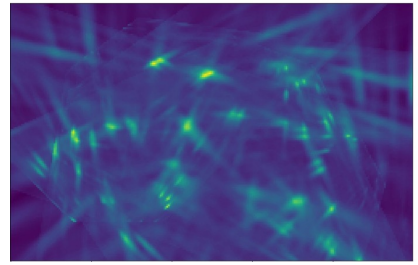
Shared Backbone

ResNet-18



Multi-view Feature Maps
B x N x C x H_f x W_f

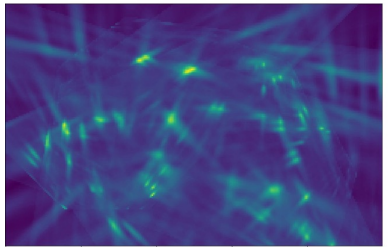
BEV Generator



BEV Feature Maps
B x N x C x H_g x W_g

Fuser

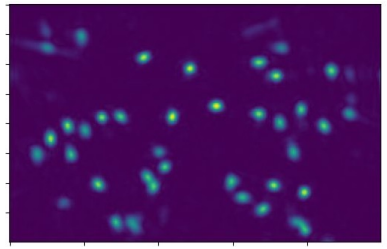
BEV Feature Map



B x C x H_g x W_g

Refiner

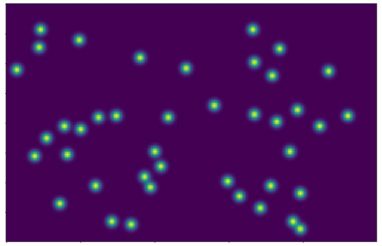
2D Occupancy Map



B x C x H_g x W_g

MSE Loss

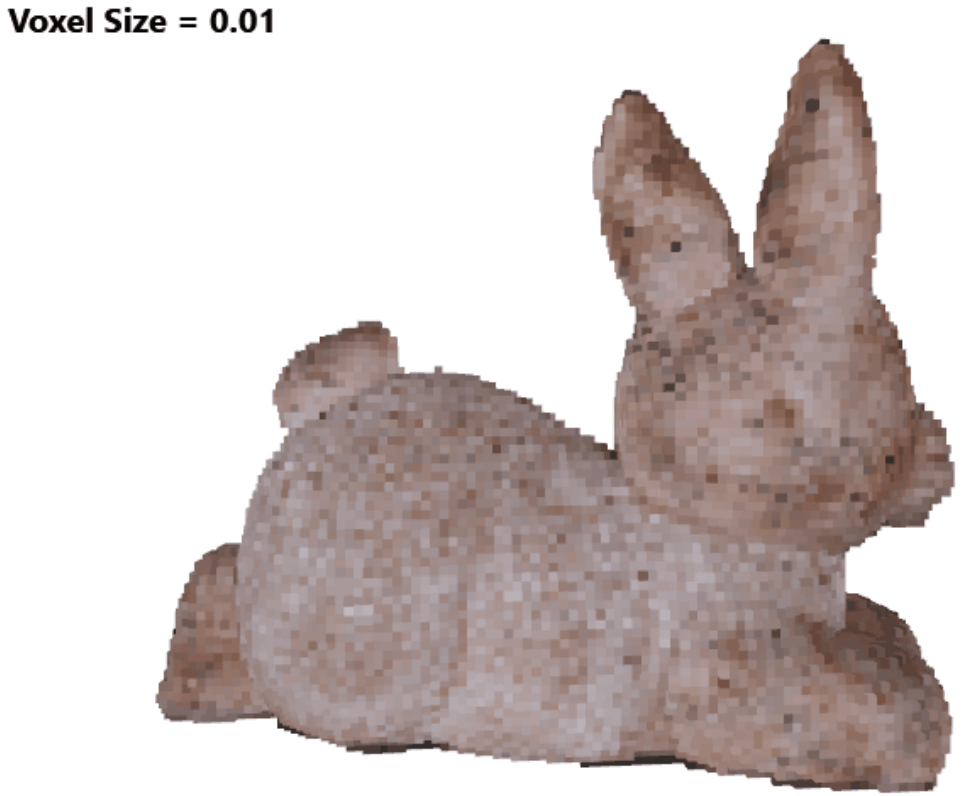
Ground-truth



B x C x H_g x W_g

Preliminaries

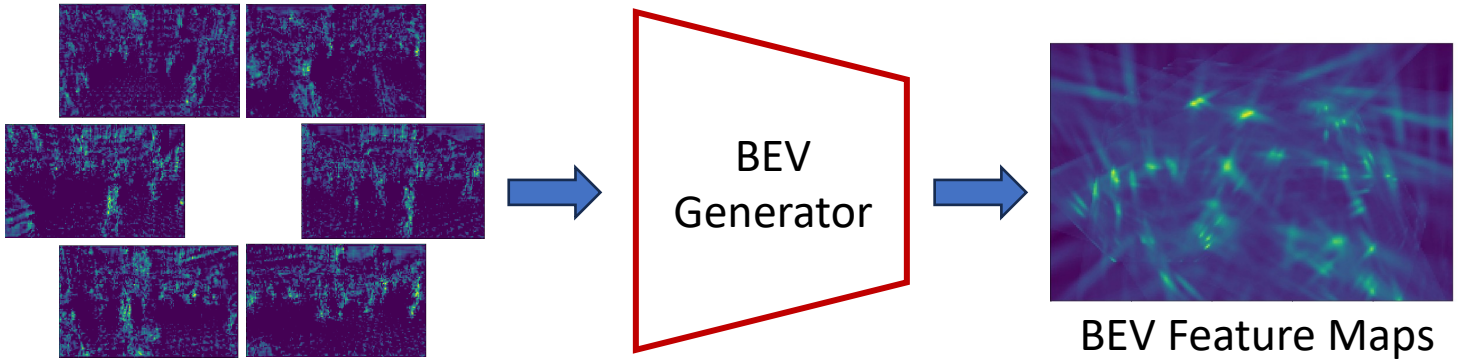
Voxelization



https://towardsdatascience.com/how-to-voxelize-meshes-and-point-clouds-in-python-ca94d403f81d

Related Work

2D to BEV Feature Transform

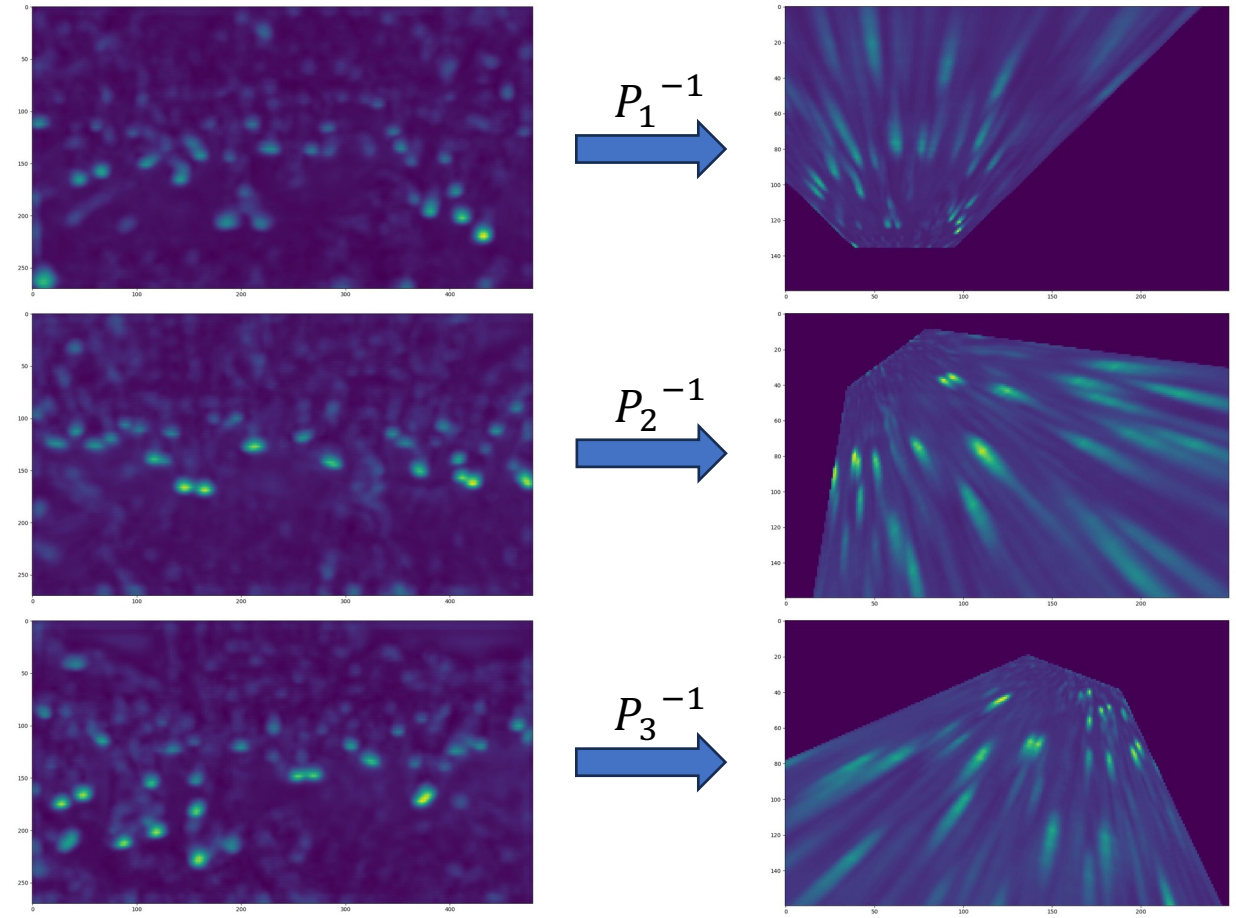


Multi-view Feature Maps
B x N x C x H_f x W_f

BEV Feature Maps
B x N x C x H_g x W_g

Related Work

Inverse Perspective Mapping (MVDet ECCV 2020)



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

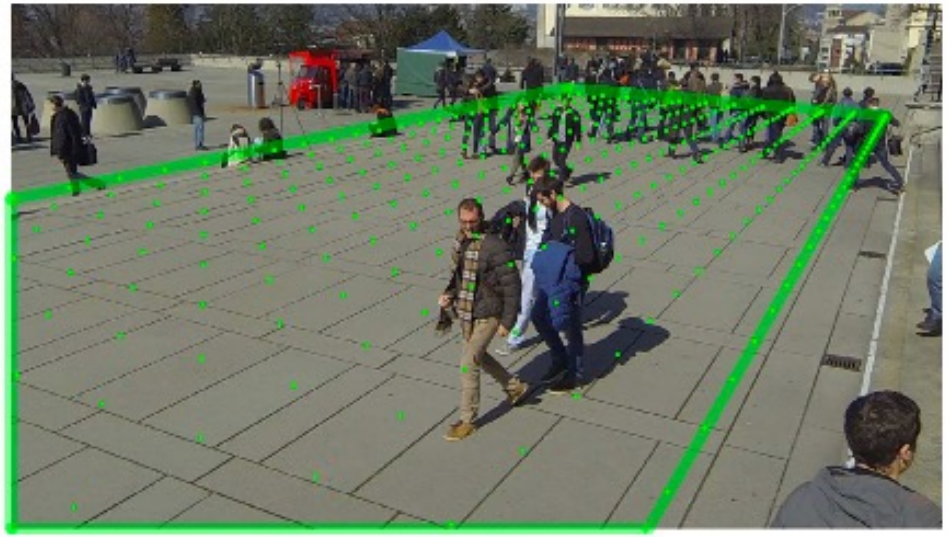
$$P = \mathrm{K}[R \quad T]$$

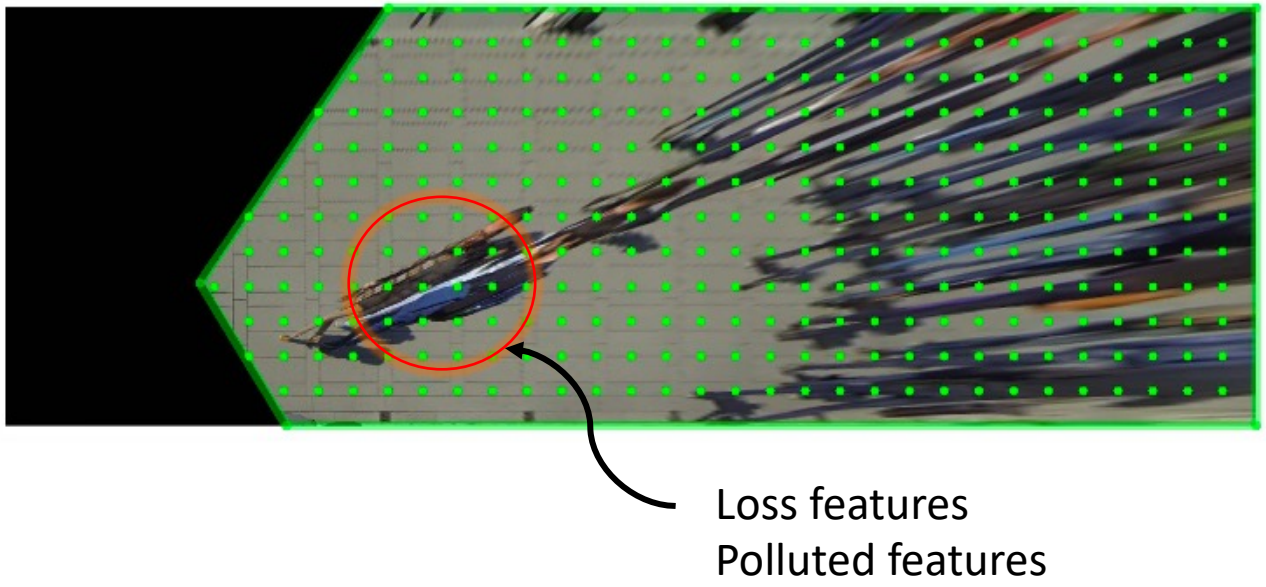
$$X_w = P^{-1}X_i$$

Hou, Yunzhong, Liang Zheng, and Stephen Gould. "Multiview detection with feature perspective transformation." *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*. Springer International Publishing, 2020.

Related Work

Inverse Perspective Mapping (MVDet ECCV 2020)

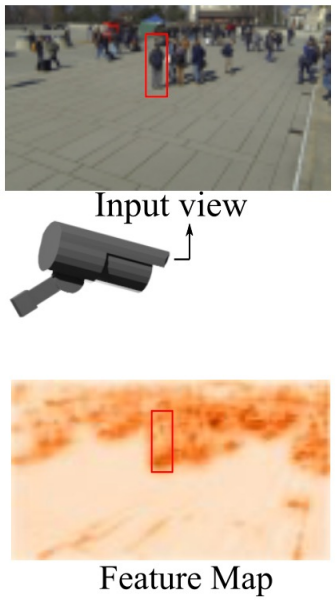


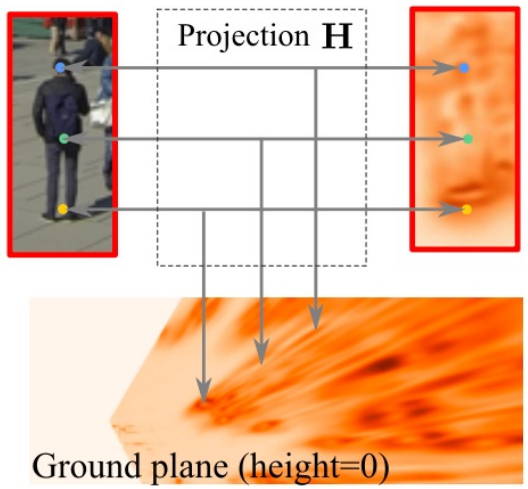


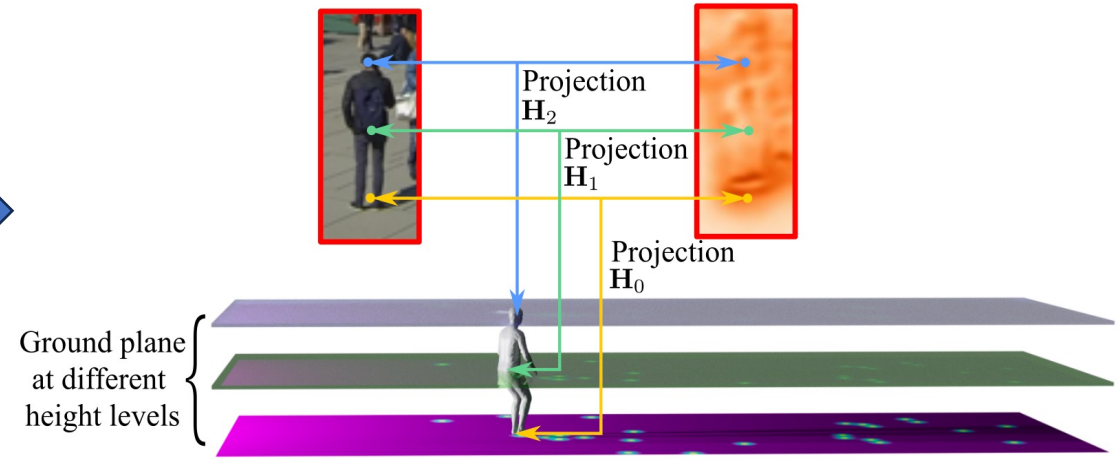
Hou, Yunzhong, Liang Zheng, and Stephen Gould. "Multiview detection with feature perspective transformation." *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*. Springer International Publishing, 2020.

Related Work

Inverse Perspective Mapping with Multiple Homographies (SHOT ICCV 2021)







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

0m
0.9m
1.8m

L. Song, J. Wu, M. Yang, Q. Zhang, Y. Li and J. Yuan, "Stacked Homography Transformations for Multi-View Pedestrian Detection," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 6029-6037, doi: 10.1109/ICCV48922.2021.00599.

Related Work

3D Feature Lifting

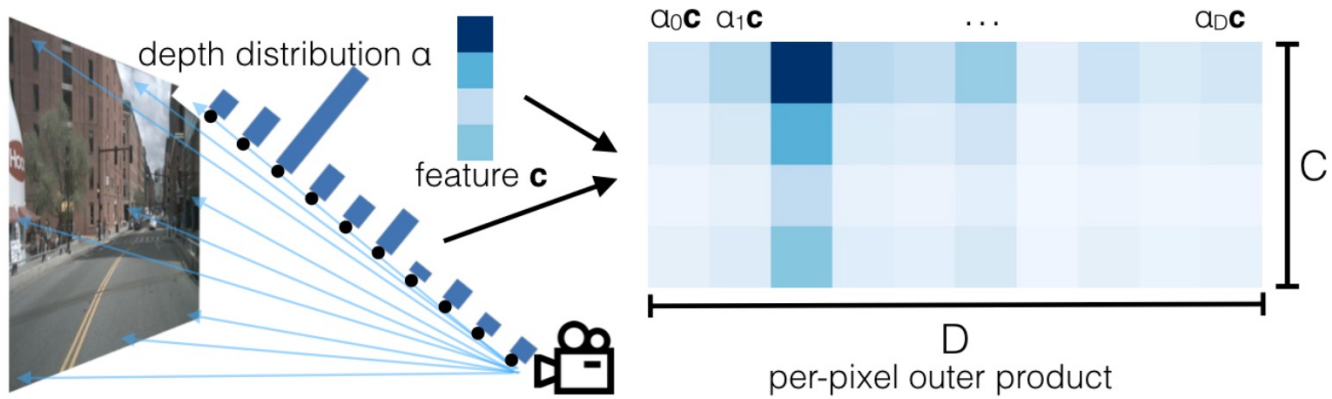
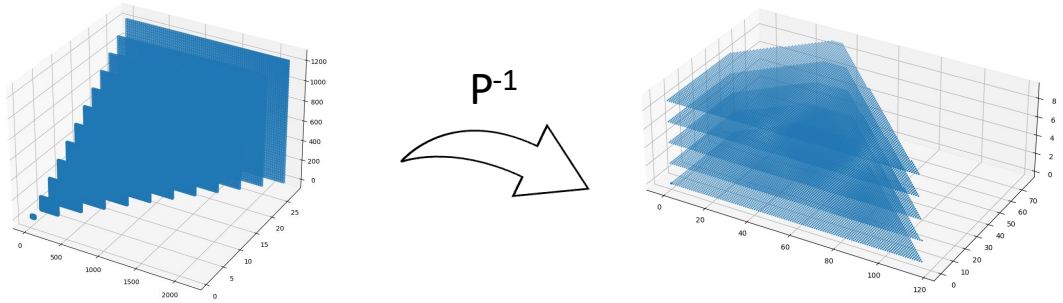
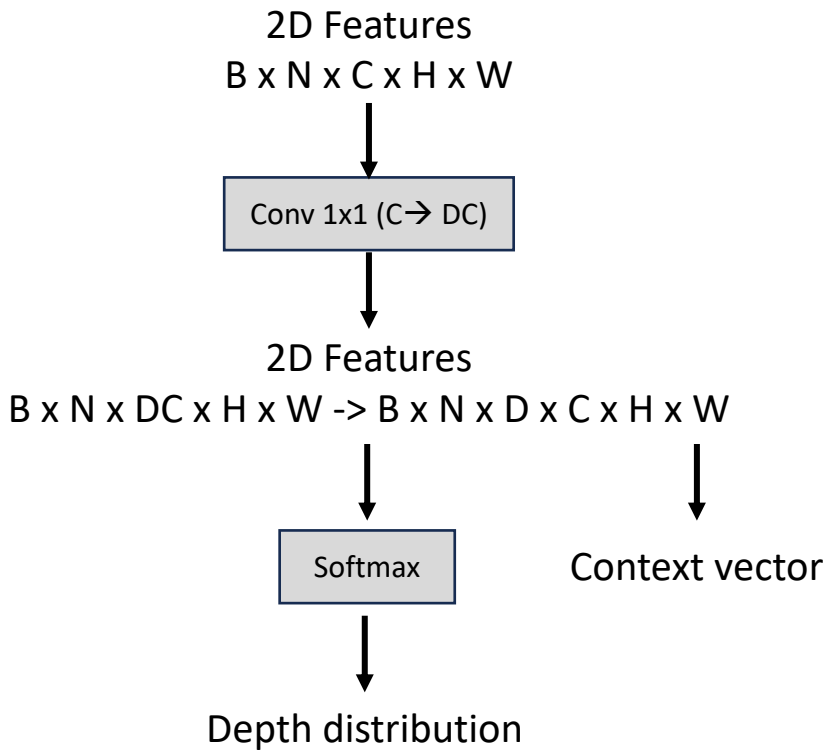


Fig. 3: We visualize the "lift" step of our model. For each pixel, we predict a categorical distribution over depth $\alpha \in \triangle^{D-1}$ (left) and a context vector $\mathbf{c} \in \mathbb{R}^C$ (top left). Features at each point along the ray are determined by the outer product of α and $\mathbf{c}$ (right).



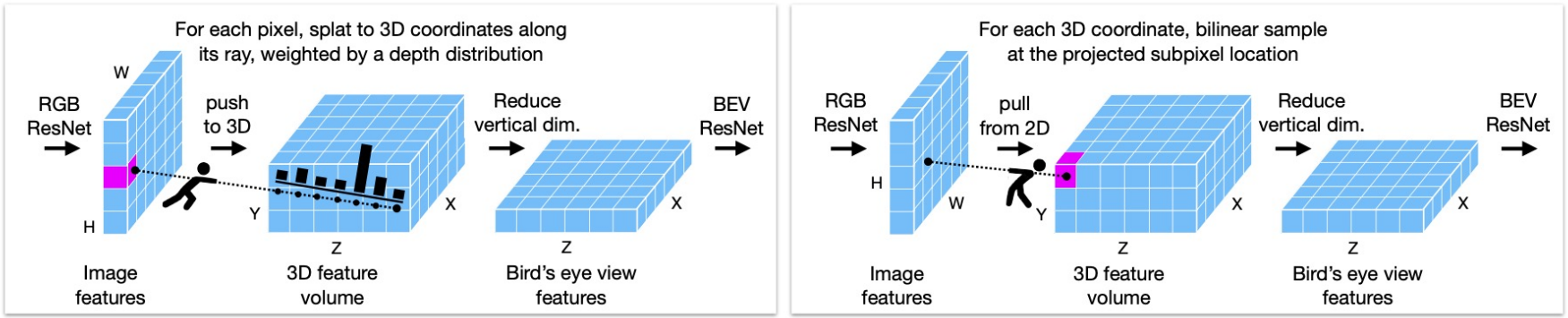


$$z\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathrm{K}[R \quad T]\begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix}$$

Philion, Jonah, and Sanja Fidler. "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d." *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer International Publishing, 2020.

Related Work

3D Feature Pulling



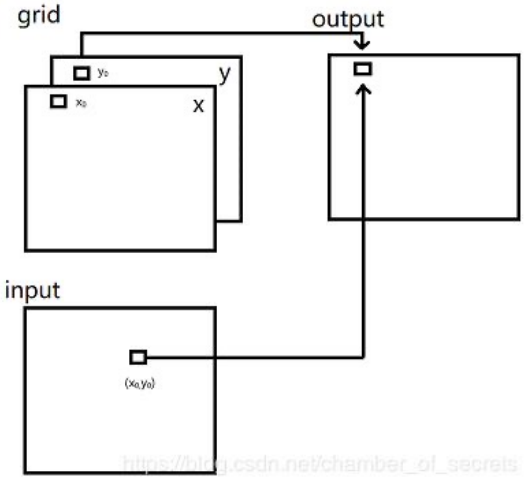


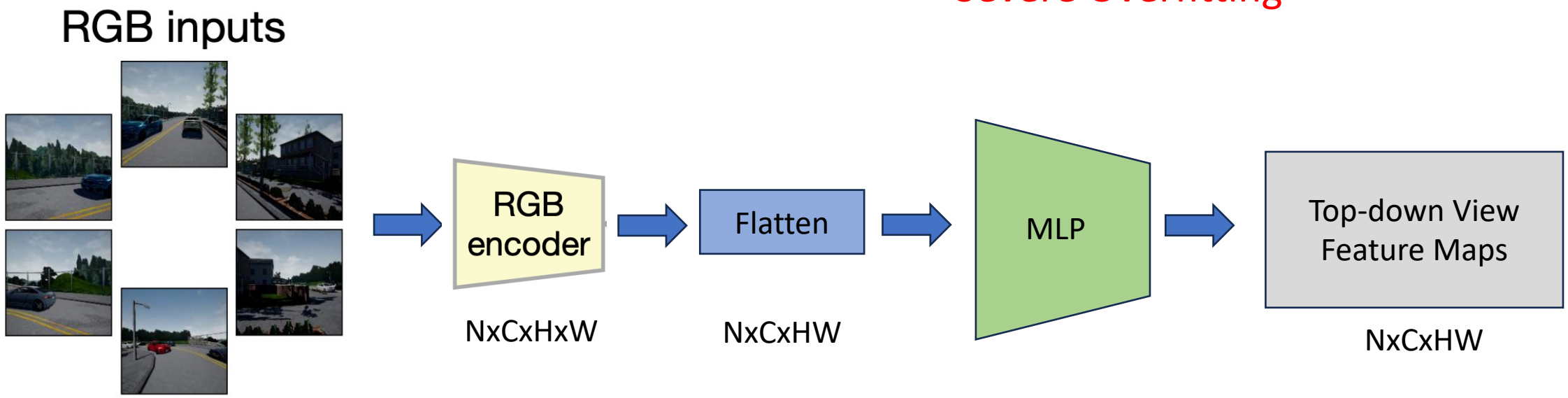
Fig. 1: **2D-to-BEV architecture, illustrated with two lifting strategies.** The left panel shows the Lift-Splat approach [9]: in this method, each 2D feature is "pushed" to 3D, filling voxels that intersect with its ray. The right panel shows our bilinear sampling approach: in this method, each 3D voxel "pulls" a feature from the 2D map, by projection and subpixel sampling.

Harley, Adam W., et al. "Simple-BEV: What really matters for multi-sensor BEV perception?." *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023.

Related Work

MLP-based Un-projection

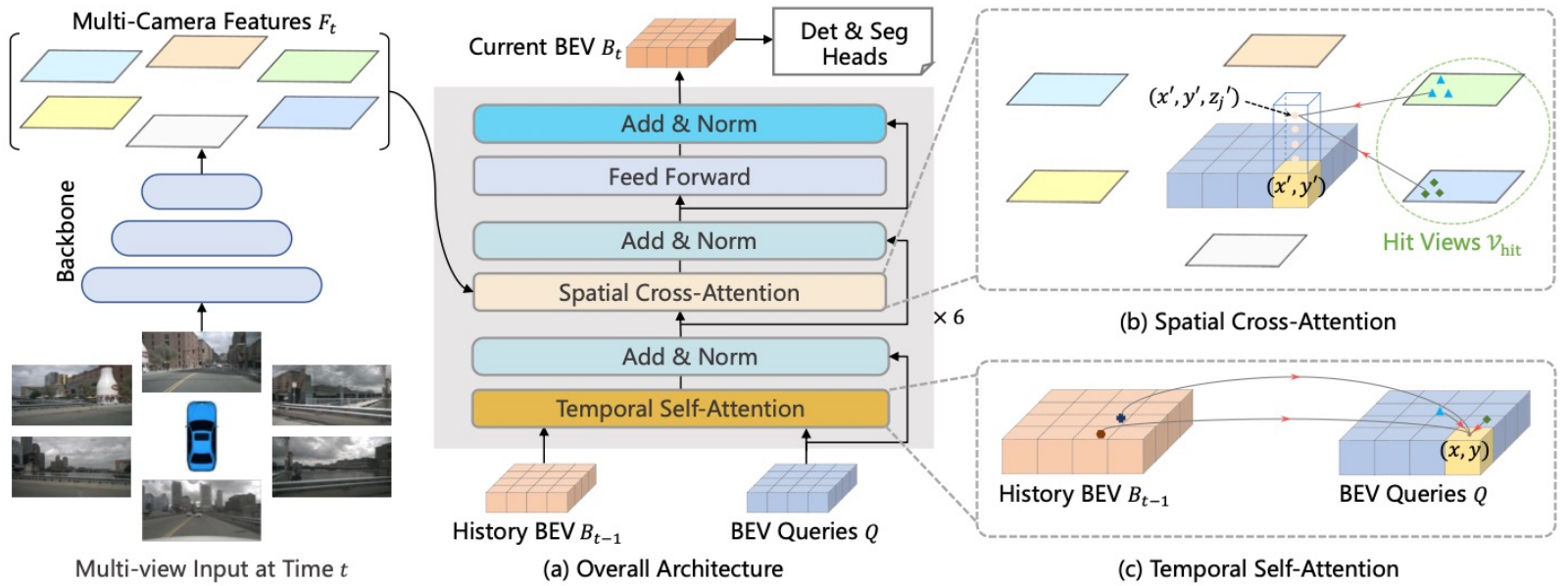
Severe Overfitting



Pan, Bowen, et al. "Cross-view semantic segmentation for sensing surroundings." *IEEE Robotics and Automation Letters* 5.3 (2020): 4867-4873.

Related Work

Geometry-aware Transformer-like Models (Explicit-based)



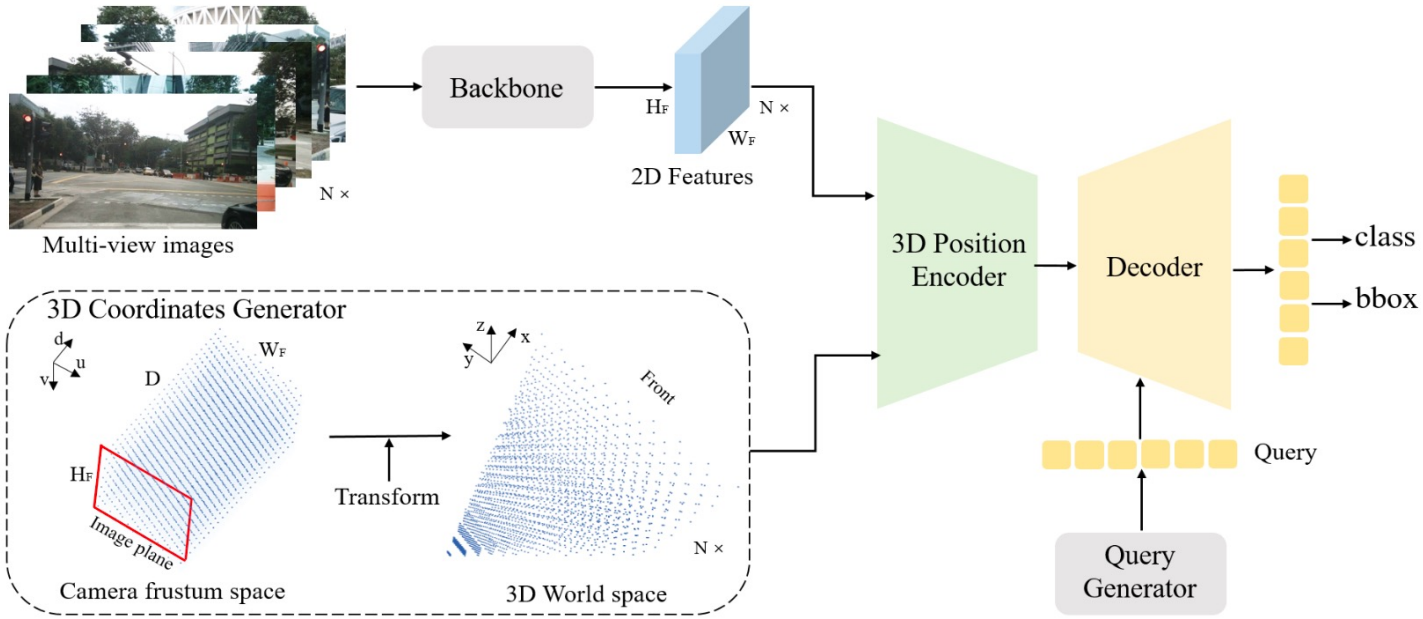
$$\text{DeformAttn}(q, p, x) = \sum_{i=1}^{N_{\text{head}}} \mathcal{W}_i \sum_{j=1}^{N_{\text{key}}} \mathcal{A}_{ij} \cdot \mathcal{W}_i' x(p + \Delta p_{ij}),$$

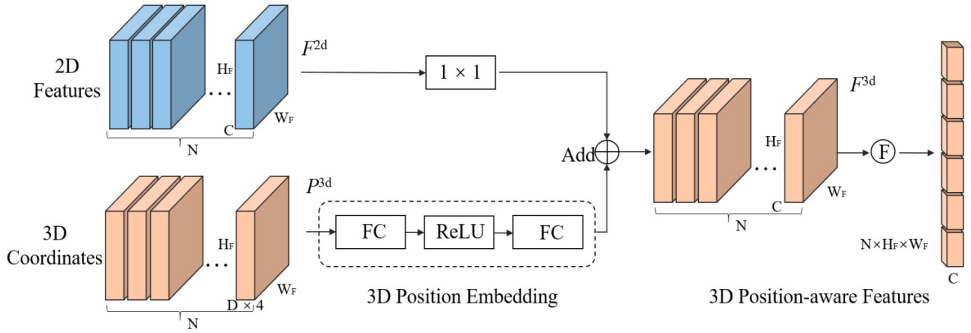
$$\text{SCA}(Q_p, F_t) = \frac{1}{|\mathcal{V}_{\text{hit}}|} \sum_{i \in \mathcal{V}_{\text{hit}}} \sum_{j=1}^{N_{\text{ref}}} \text{DeformAttn}(Q_p, \mathcal{P}(p, i, j), F_t^i),$$

Li, Zhiqi, et al. "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers." *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*. Cham: Springer Nature Switzerland, 2022.

Related Work

Geometry-aware Transformer-like Models (Implicit-based)

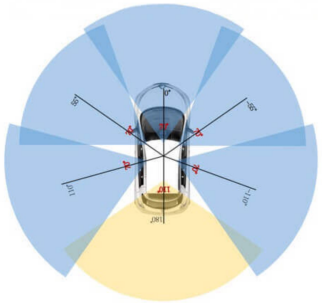




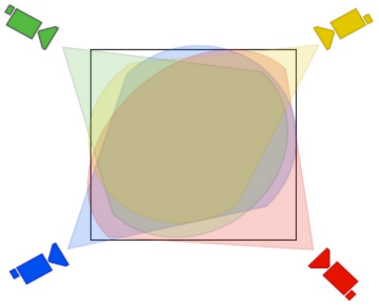
$$p_{i,j}^{3d} = K_i^{-1} p_j^m \qquad P^{3d} = \{P_i^{3d} \in R^{(D\times 4)\times H_F \times W_F}, i = 1, 2, \ldots, N\}$$

$$F_i^{3d} = \psi(F_i^{2d}, P_i^{3d})$$

Autonomous Driving

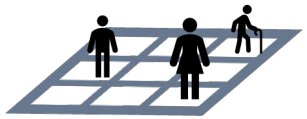


Our case



Liu, Yingfei, et al. "Petr: Position embedding transformation for multi-view 3d object detection." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022.

Related Work



2D to BEV Feature Transformation Summary

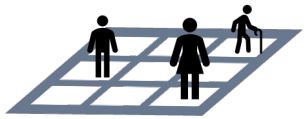
- Learnable
  - MLP
  - Deformable attention
  - 3D Feature Lifting
- Parameter-free
  - Inverse Perspective Mapping
  - 3D Feature Pulling

TABLE I: Effect of lifting strategy.

2D-to-BEV strategy	IOU
Unweighted splatting	43.1
Depth-based splatting [9]	44.4
Deformable attention [5]	46.5
Bilinear sampling (ours)	47.4
Multi-scale deform. attn. [5]	**48.9**

Harley, Adam W., et al. "Simple-BEV: What really matters for multi-sensor BEV perception?." *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023.

Previous Work

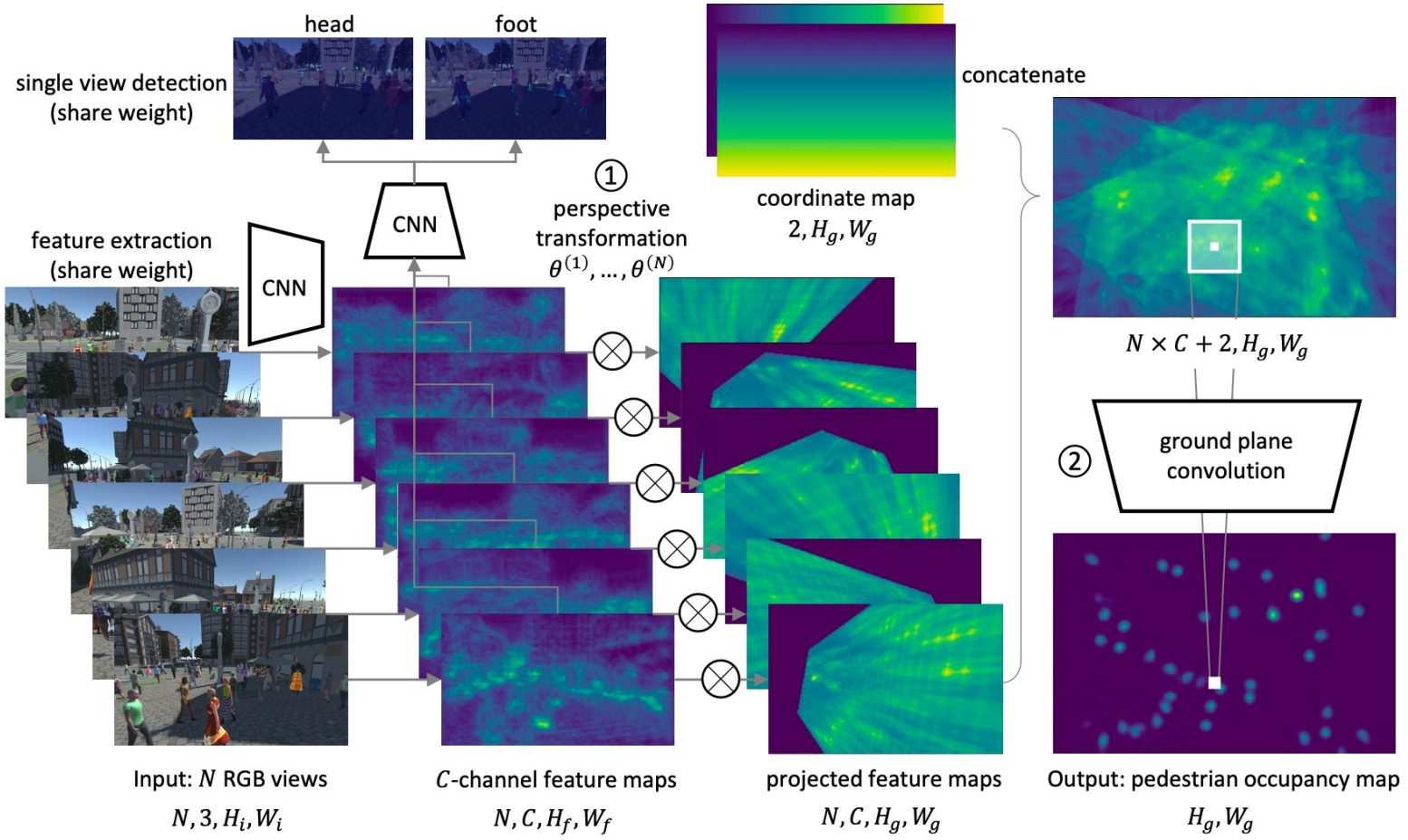


Requirements for Multi-view Detection System

- Dataset size is small (400 frames) -> Overfitting -> Model size needs to be small
- Crowded scene -> High resolution feature map
- Generalization performance

Previous Work

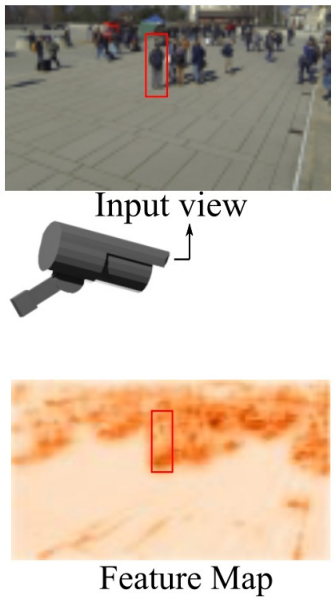
MVDet (ECCV 2020)

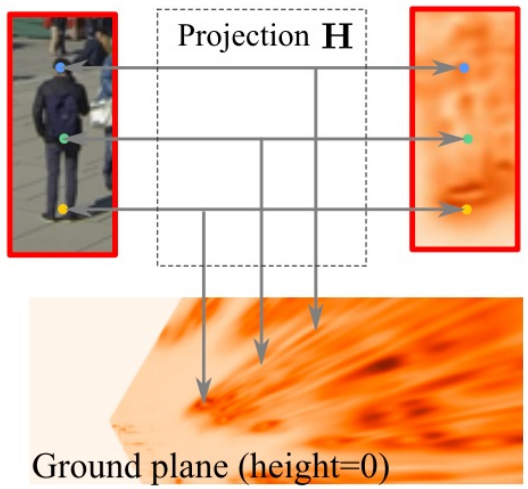


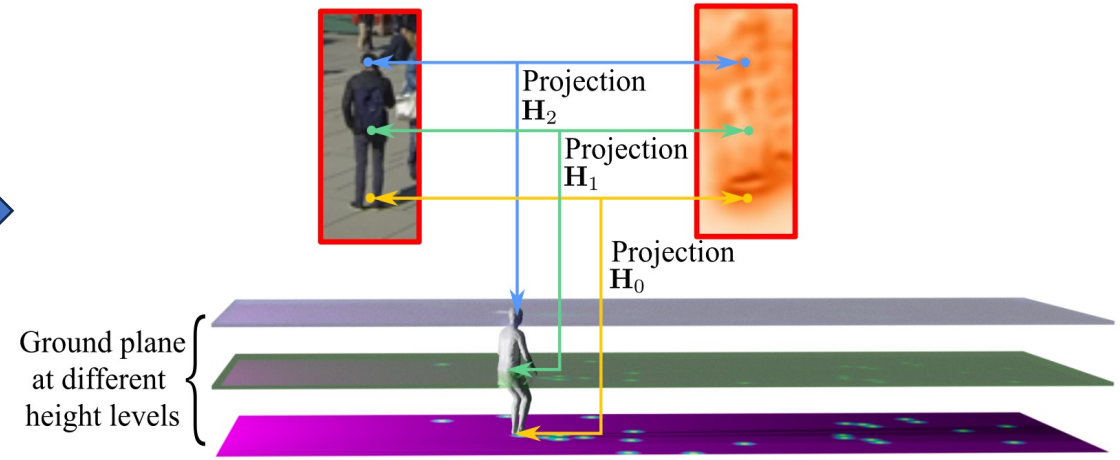
Hou, Yunzhong, Liang Zheng, and Stephen Gould. "Multiview detection with feature perspective transformation." *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*. Springer International Publishing, 2020.

Previous Work

SHOT (ICCV 2021)







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

0m
0.9m
1.8m

	MODA	MODP	P	R
MVDet	83.9	79.6	96.8	86.7
SHOT	88.3 (+4.4)	82.0	96.6	91.5

L. Song, J. Wu, M. Yang, Q. Zhang, Y. Li and J. Yuan, "Stacked Homography Transformations for Multi-View Pedestrian Detection," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 6029-6037, doi: 10.1109/ICCV48922.2021.00599.

Previous Work

3DROM (ECCV 2022)

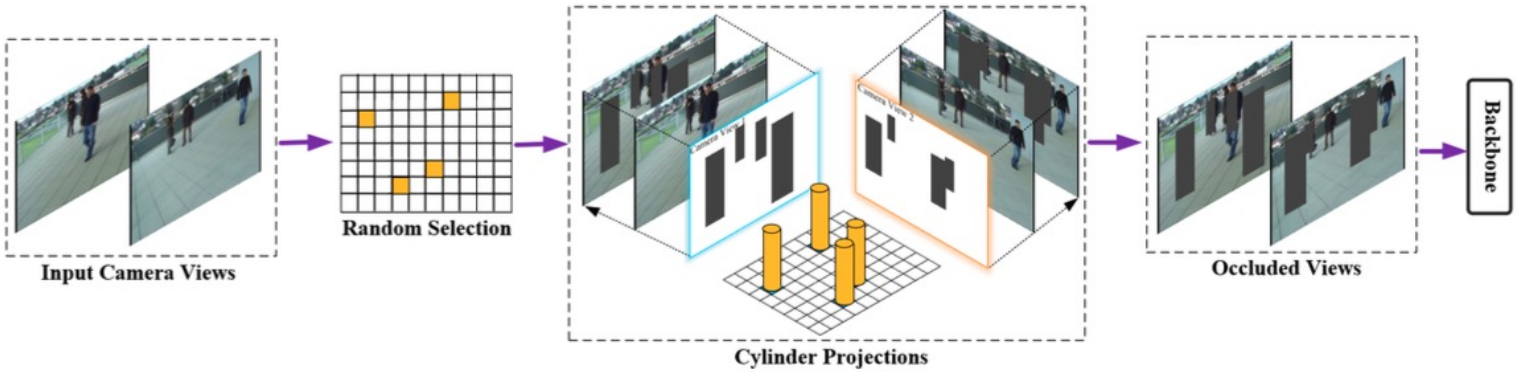


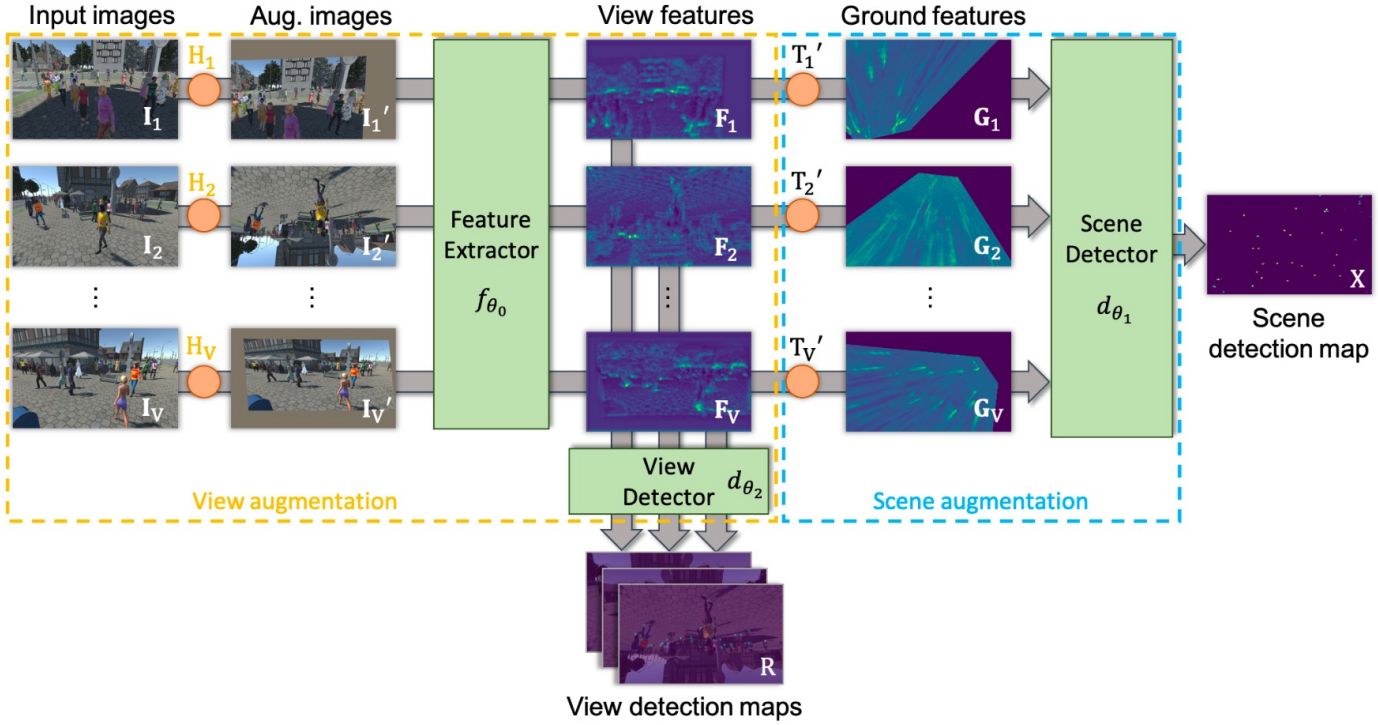
Fig. 2. A schematic diagram of the 3D Random Occlusion method.

	MODA	MODP	P	R
MVDet	83.9	79.6	96.8	86.7
SHOT (Multiple Homographies)	88.3 (+4.4)	82.0	96.6	91.5
3DROM (SHOT + 3D Random Erasing)	95.0 (+6.7)	84.9	99.0	96.1

Qiu, Rui, et al. "3d random occlusion and multi-layer projection for deep multi-camera pedestrian localization." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022.

Previous Work

MVAug (WACV 2023)



	MODA	MODP	P	R
MVDet	83.9	79.6	96.8	86.7
MVAug	95.3 (+11.4)	89.7	99.4	95.9

Engilberge, Martin, et al. "Two-level Data Augmentation for Calibrated Multi-view Detection." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023.