Enhancing Multi-view Pedestrian Detection Through Generalized 3D Feature Pulling

Sithu Aung[1,2], Haesol Park[1], Hyungjoo Jung[1], Junghyun Cho[1,2,3]
[1]KIST, Republic of Korea [2]UST, Republic of Korea [3]Yonsei-KIST, Republic of Korea
{sithu, haesol, jhj0220, jhcho}@kist.re.kr

Aung Sithu

MS-Student

9.11.23

will be presented in

UST 과학기술연합대학원대학교 UNIVERSITY OF SCIENCE & TECHNOLOGY

JAN 4-8 WACV 2024 WAIKOLOA HAWAII

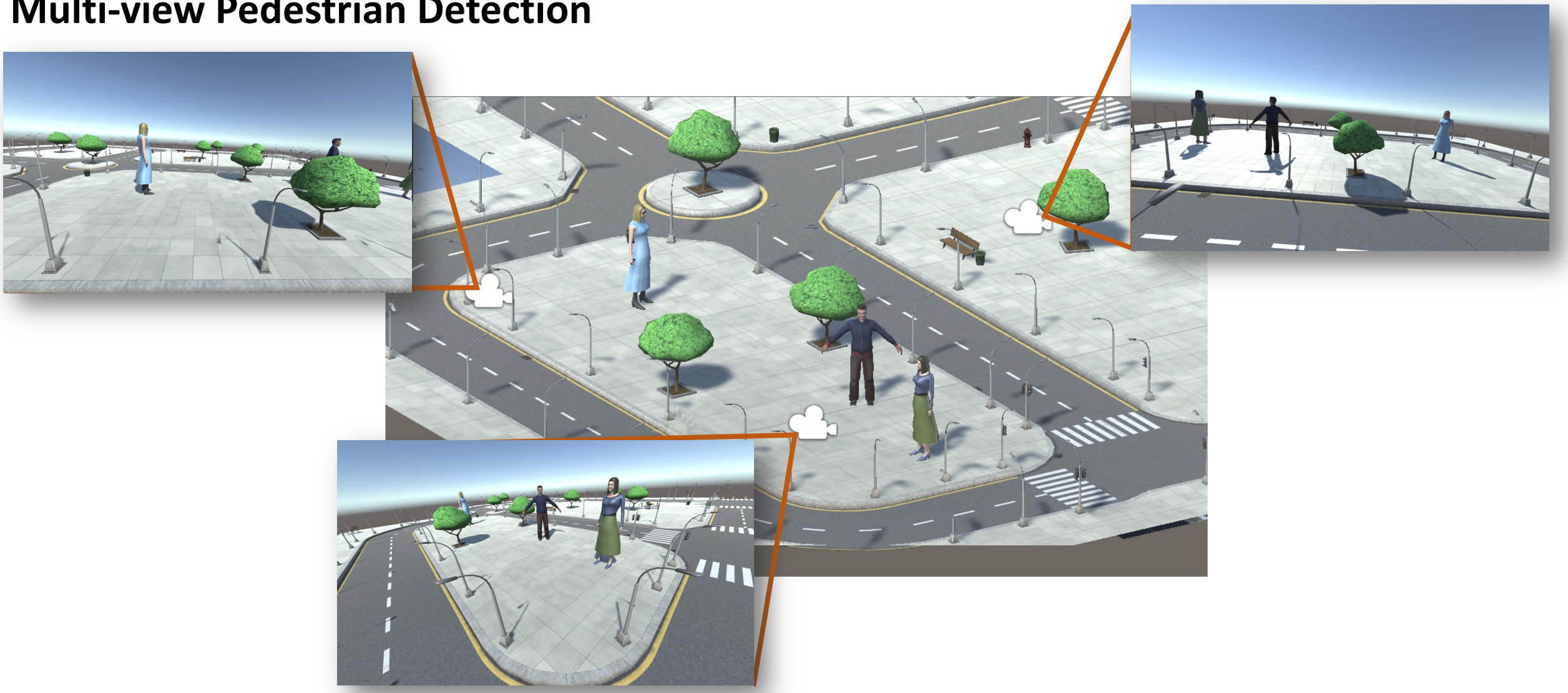
KIST 한국과학기술연구원 Korea Institute of Science and Technology

Contents

- Background
- Method
- Training Setups

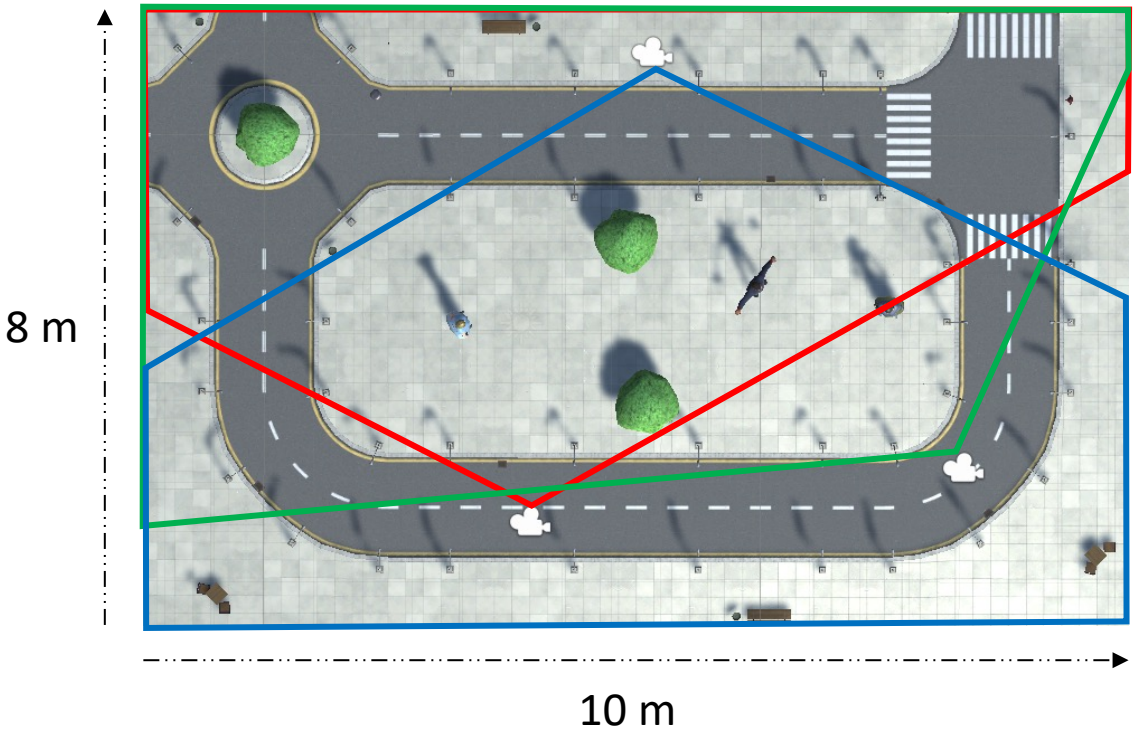
Background

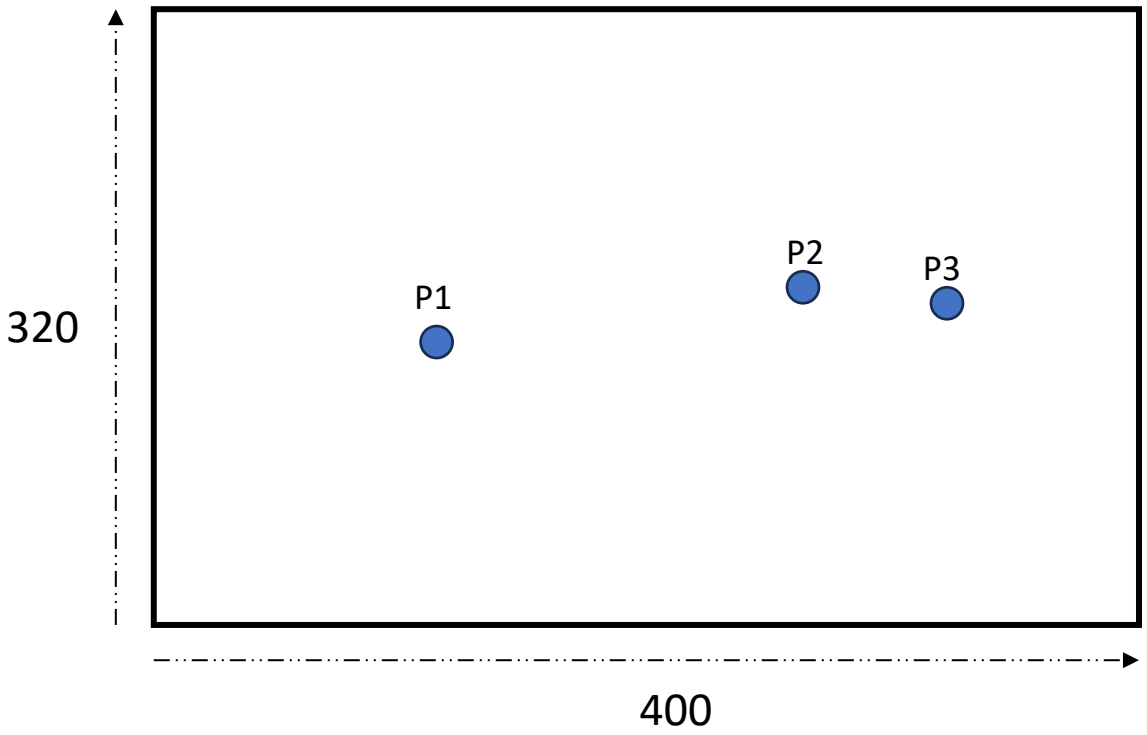
Multi-view Pedestrian Detection



Background

Localization within Ground Plane

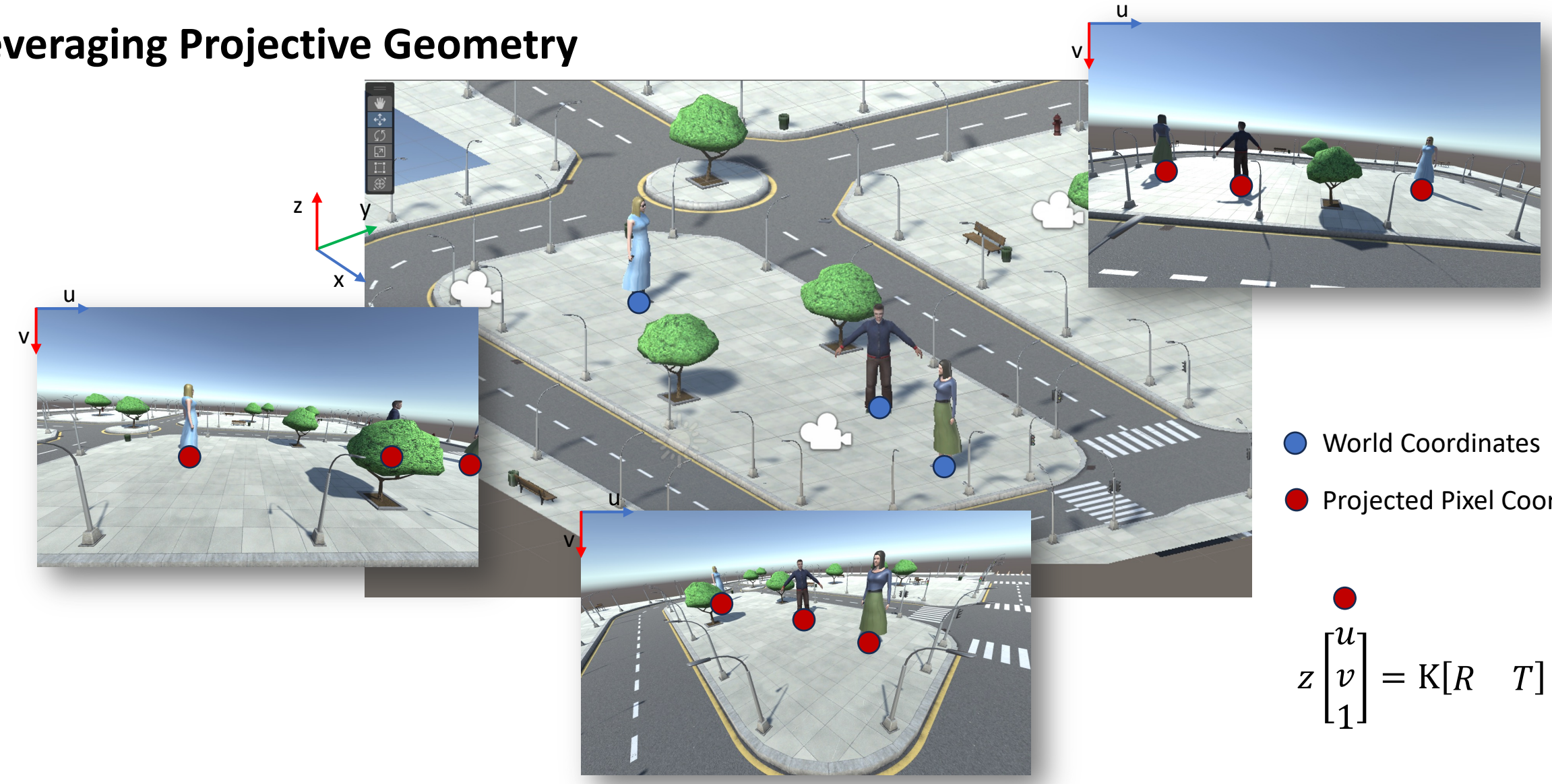




Discretized with 2.5 cm resolution

Background

Leveraging Projective Geometry

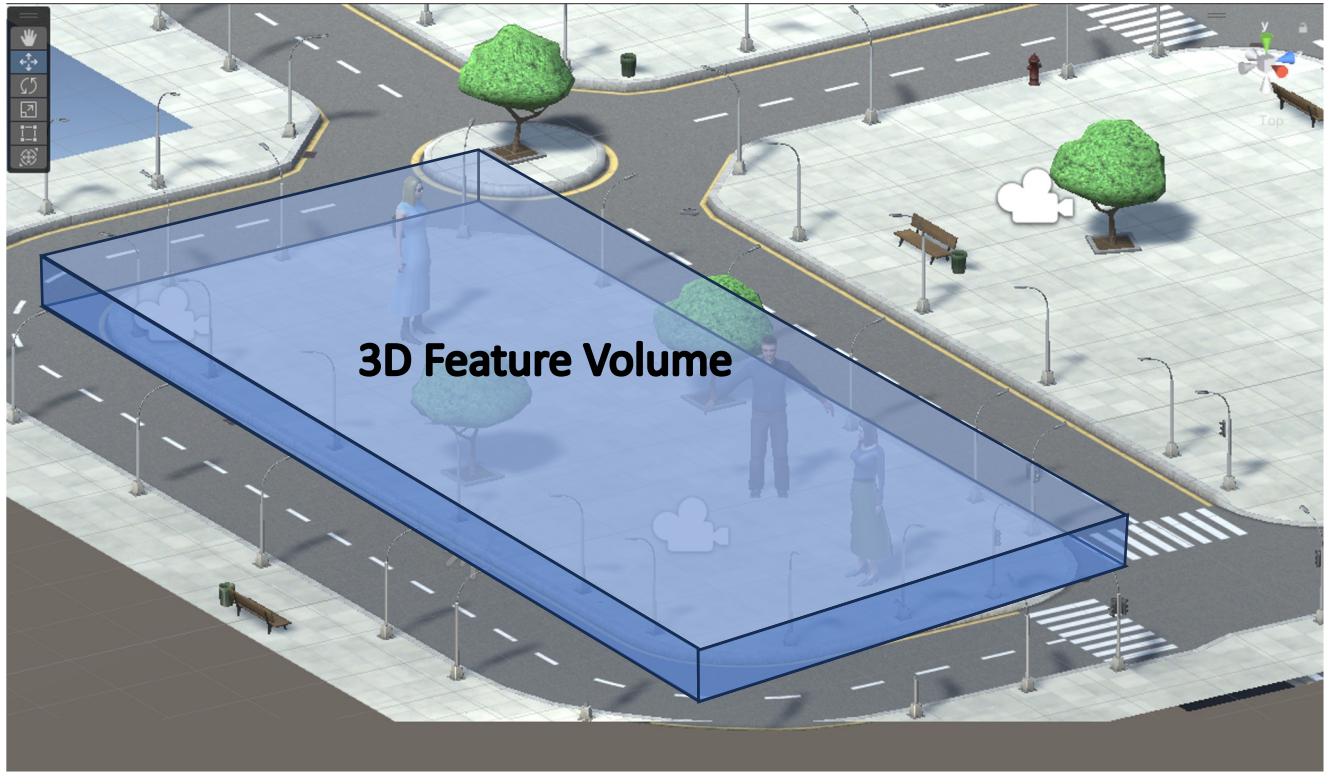


- World Coordinates
- Projected Pixel Coordinates

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

Method

Main Idea



Method

Overview

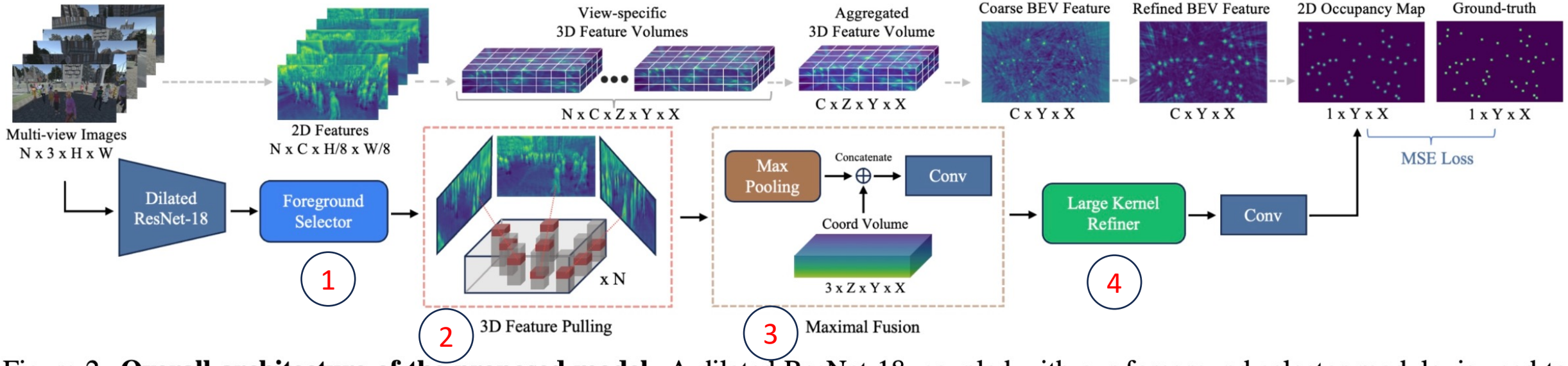


Figure 2. **Overall architecture of the proposed model**. A dilated ResNet-18, coupled with our foreground selector module, is used to extract multi-view features. In the 3D feature-pulling, a sub-pixel 2D feature is pulled for each 3D voxel using projection and bilinear sampling. A maximal fusion module 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.

Method

1. Foreground Selector Module

Why need FSM?

- To extract only essential semantic information from 2D features.
- Facilitate a more comprehensive feature selection.

Construction of FSM

- Inspired by channel-attention module in the squeeze-and-excitation network.

$$I_i \in \mathbb{R}^{3\times H\times W} \longrightarrow F_i^{2d} \in \mathbb{R}^{C\times\frac{H}{8}\times\frac{W}{8}}$$

$$i = 1, 2, \dots, N$$

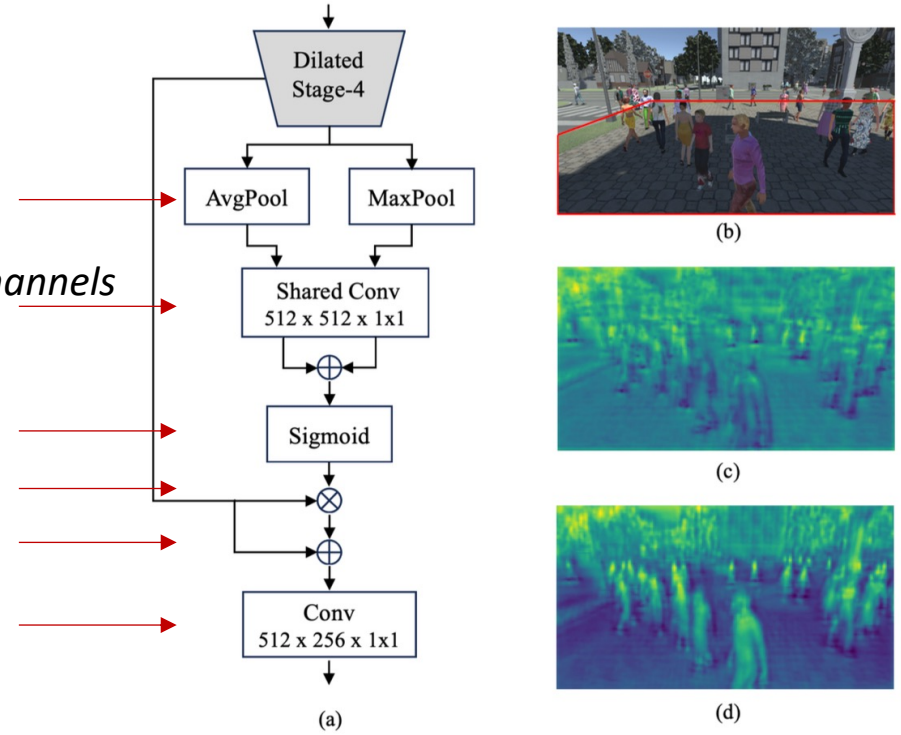


Figure 3. **Foreground selector module**. (a) Network structure of FSM. (b) Sample view. (c) Before FSM. (d) After FSM. The feature output is enhanced with the proposed module to focus on crucial foreground details, while filtering out unnecessary background information. The red area delineates the area of interest (AoI). Features outside of this region will not be utilized.

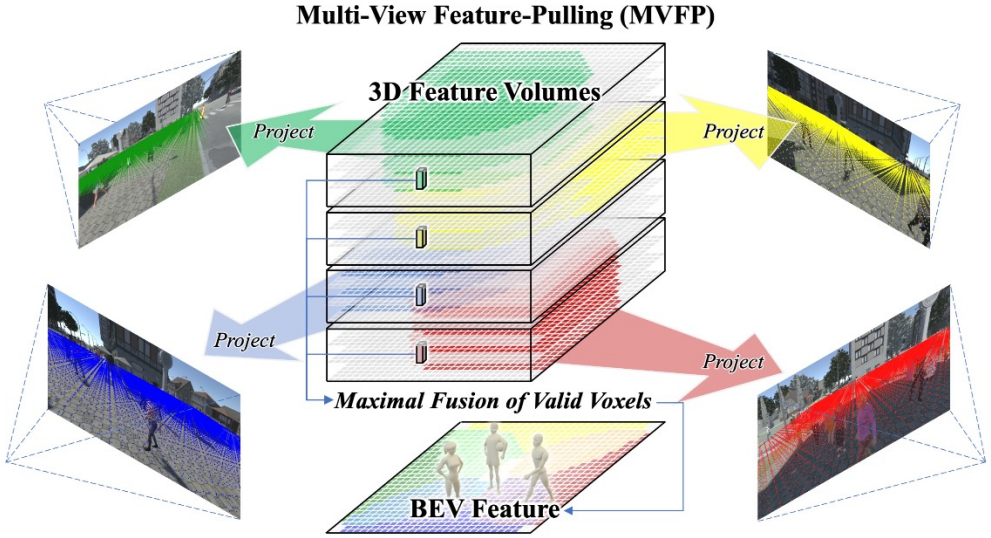
Method

2. 3D Feature Pulling

2D Feature Maps $F_i^{2d} \in \mathbb{R}^{C \times \frac{H}{8} \times \frac{W}{8}}$

3D Voxel Volumes $V_i \in \mathbb{R}^{Z \times Y \times X}$

$$\begin{bmatrix} u \\ v \end{bmatrix} = \Pi\ K'[R \quad T]\ D_g \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}, D_g = \begin{bmatrix} s_g & 0 & 0 & x_{min} \\ 0 & s_g & 0 & y_{min} \\ 0 & 0 & s_g & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$



$\begin{bmatrix} u \\ v \end{bmatrix}$ = 2D pixel coordinates in feature maps

$\begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}$ = 3D grid coordinates in voxel volume

D_g = Transformation matrix (World to Grid)

Π = perspective mapping

K' = scaled intrinsic matrix

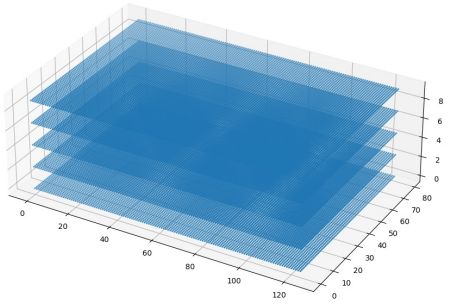
$[R \quad T]$ = extrinsic matrix

s_g = grid size

$x_{min}\ y_{min}$ = lower bounds of ground plane

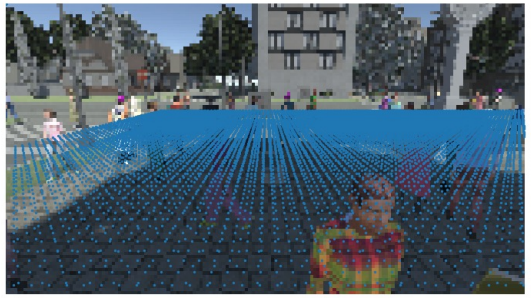
Method

2. 3D Feature Pulling (Continued…)



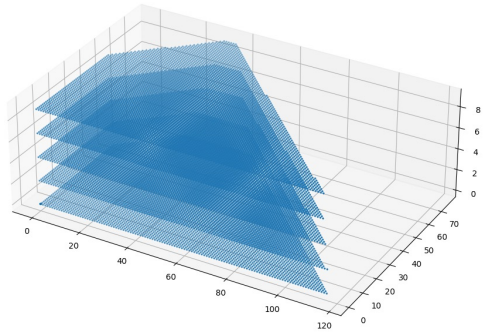
3D voxel grid covering whole AoI

project



Valid pixel coordinates for a view

Apply mask



Valid 3D voxel grid for a view

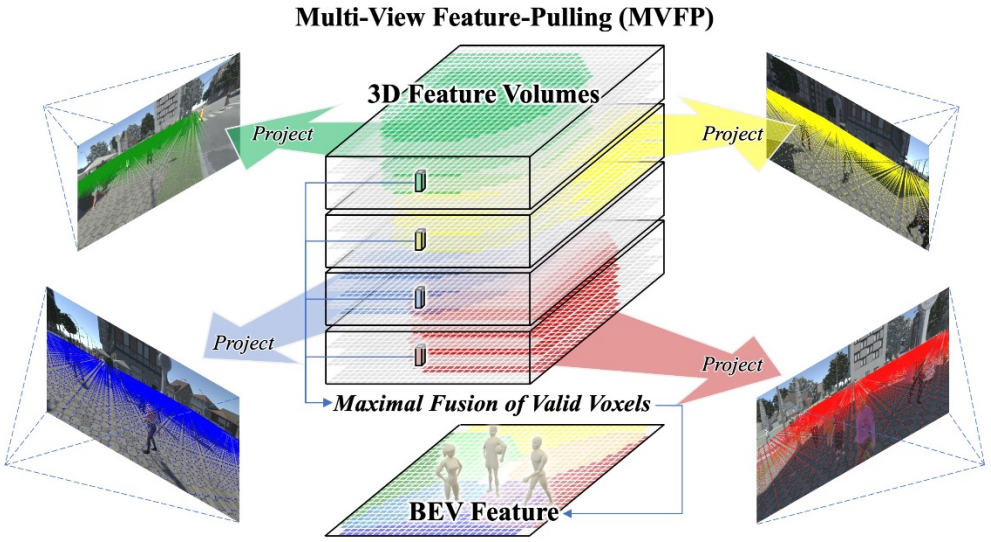
Binary Mask to indicate whether a voxel is inside the camera frustrum $M_i \in \mathbb{R}^{Z \times Y \times X}$

$$M_i(x,y,z) = \begin{cases} 1, & if\ 0 \le u \le \frac{W}{8}\ and\ 0 \le v \le \frac{H}{8} \\ 0, & otherwise \end{cases}$$

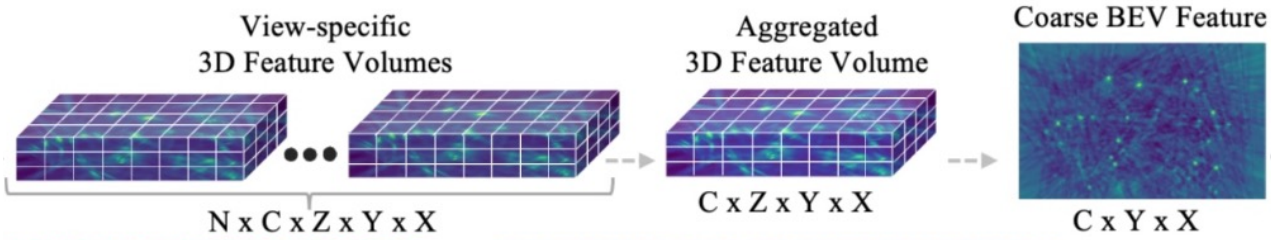
Pull feature for each valid voxel $V_i(x,y,z) = \begin{cases} F_i^{2d}(u,v), & if\ M_i(x,y,z) = 1 \\ 0, & otherwise \end{cases}$

Method

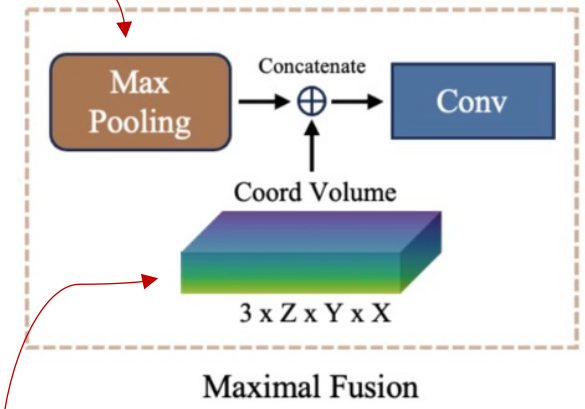
3. Maximal Fusion Module



$$V(x, y, z) = \max_{i}^{N} V_i$$



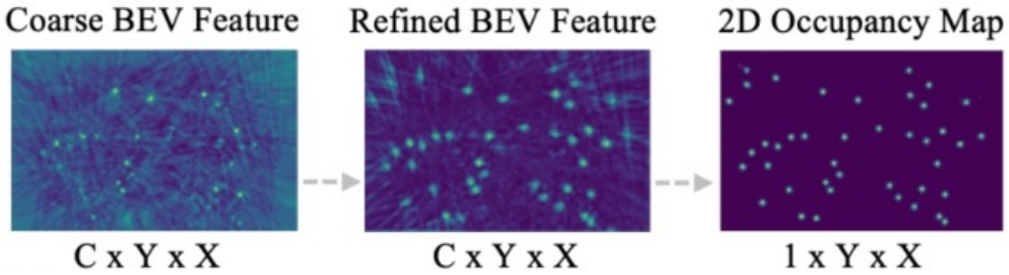
To work with arbitrary camera setups



To provide positional information of the 3D coordinates into the pulled 2D features

Method

4. Large Kernel Refiner Module



- To handle misalignments caused by imprecise calibration data.
- To consolidate comprehensive body information, harnessed from multiple perspective views into a more concise cluster of features within the BEV plane.

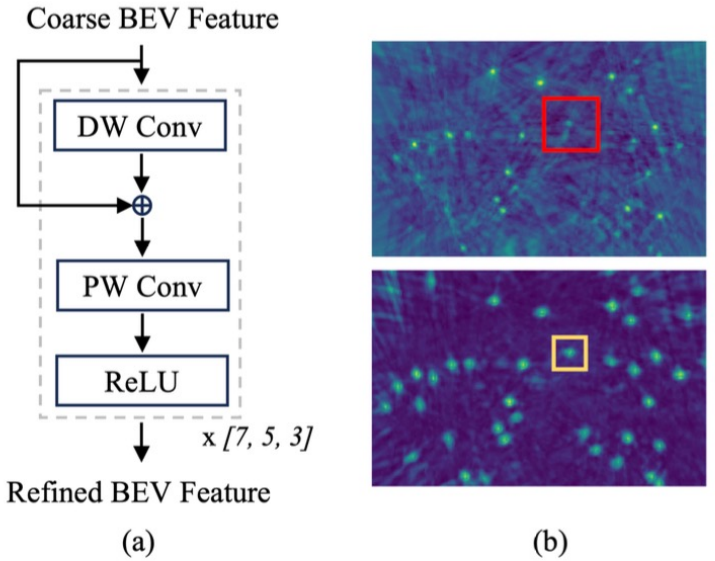
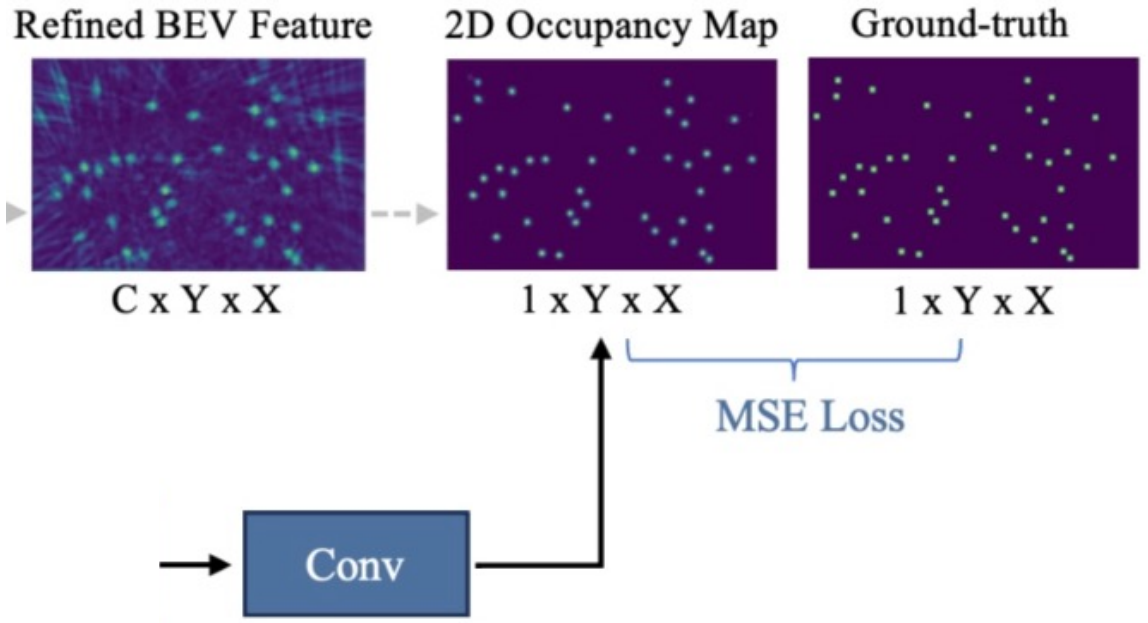


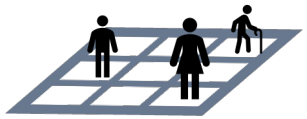
Figure 4. **Proposed large kernel refiner module**. (a) The figure illustrates a single block of the module. We apply three consecutive blocks, each with large kernel sizes of [7, 5, 3], which were chosen based on empirical results. (b) The results demonstrating the benefits of our proposed module. This module gradually refines and collects the scattered multi-view features, as showcased in the red box, leading to a more concise feature representation, as depicted in the yellow box.

Method

Head


Refined BEV Feature
2D Occupancy Map
Ground-truth
C x Y x X
1 x Y x X
1 x Y x X
MSE Loss
Conv

Training Setups



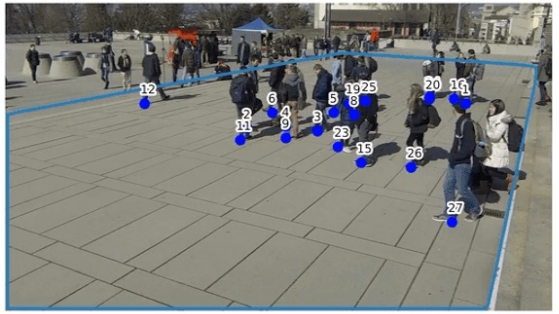
Backbone	**Dilated ResNet-18**
Channel Dimension	256
Voxel Size	10cm x 10cm x 20cm
GPU	4x Nvidia A100 GPUs
Optimizer	AdamW
Weight decay	1e-4
Batch size	4 (1 on each GPU)

Next Seminar

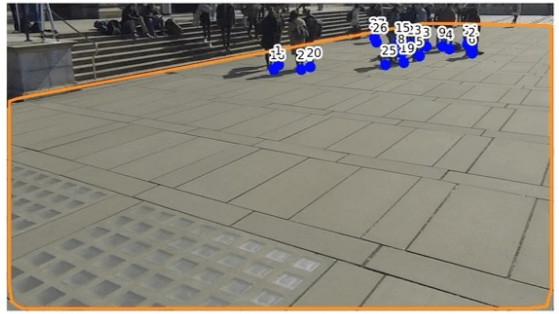
Experiment & Results

Ours

Camera 1

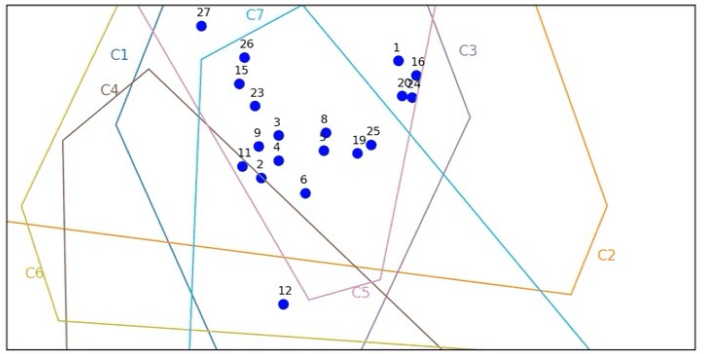


Camera 2

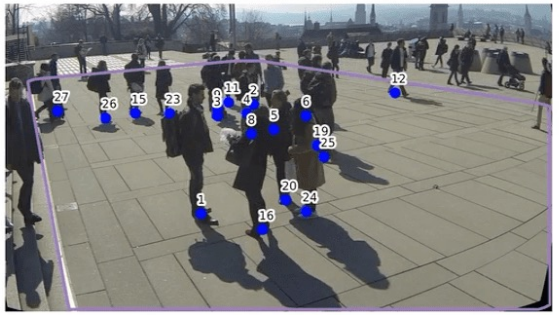


Frame: 1890

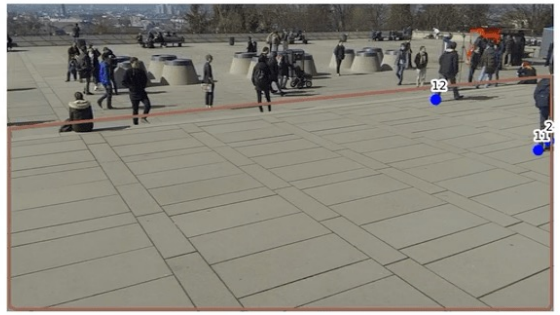
Ground Plane



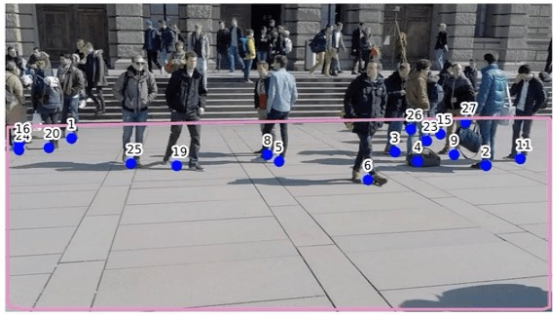
Camera 3



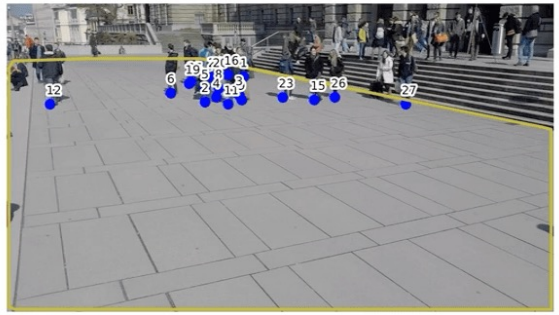
Camera 4



Camera 5



Camera 6



Camera 7