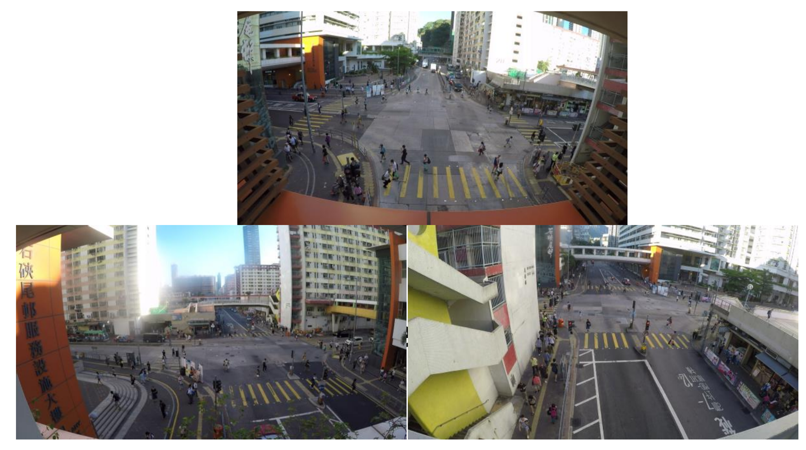
電腦科學系
Department of Computer Science
香港城市大學
City University of Hong Kong

3D Crowd Counting via Multi-View Fusion with 3D Gaussian Kernels

Qi ZHANG, Antoni. B. CHAN (City University of Hong Kong)

AAAI-20
Thirty-Fourth AAAI Conference on Artificial Intelligence
February 7-12, 2020, Hilton New York Midtown, New York, New York USA

Introduction

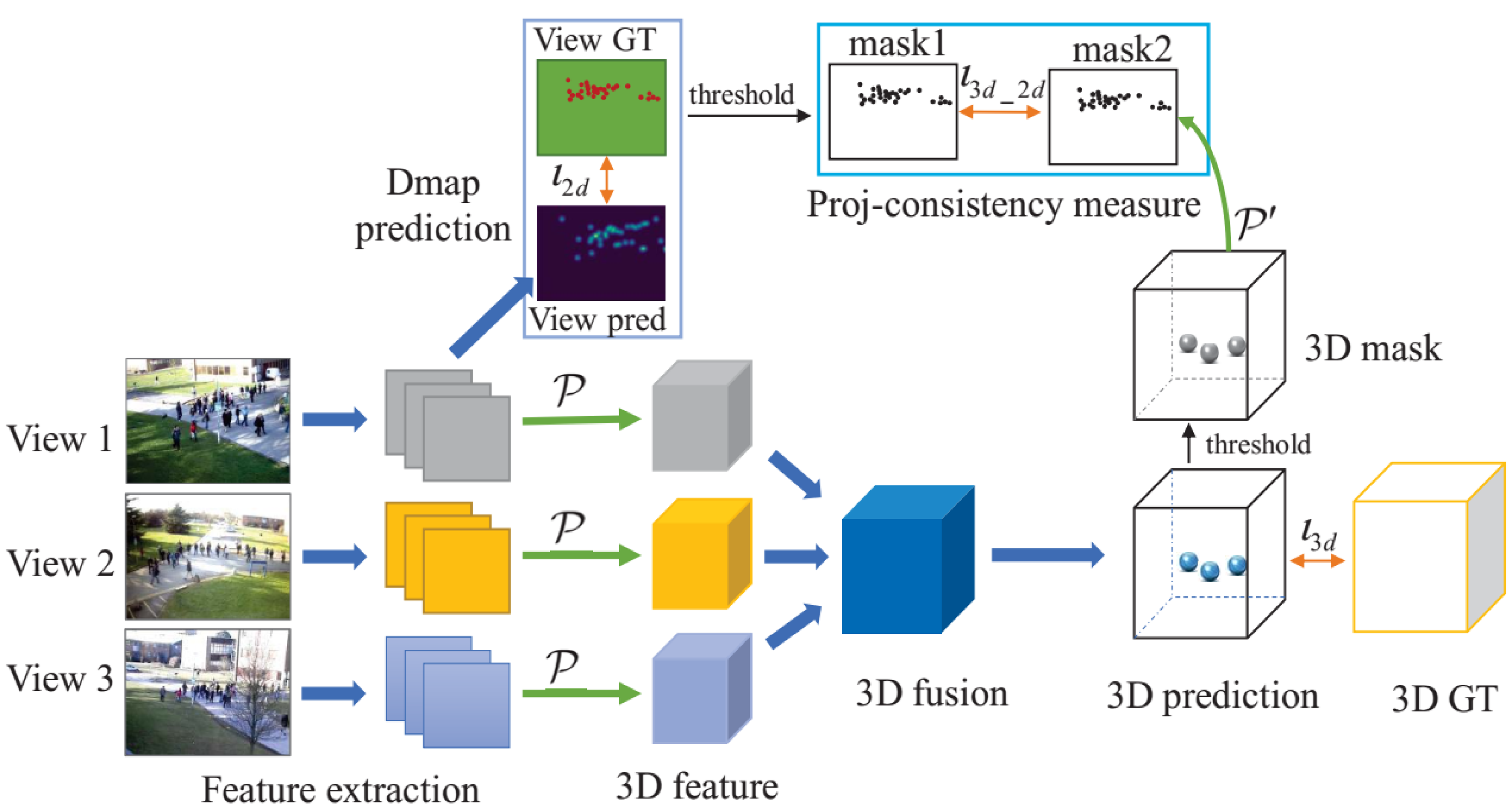


Problems with single-view counting:
- Limited field-of-view of single camera;
- Low resolution in the farther place;
- Severe occlusion, e.g., humans, buildings.

Solution: Multi-view counting
- Previous work fuse multiples views to estimate ground-plane 2D density maps.
- We propose counting with 3D density maps.

3D Crowd Counting

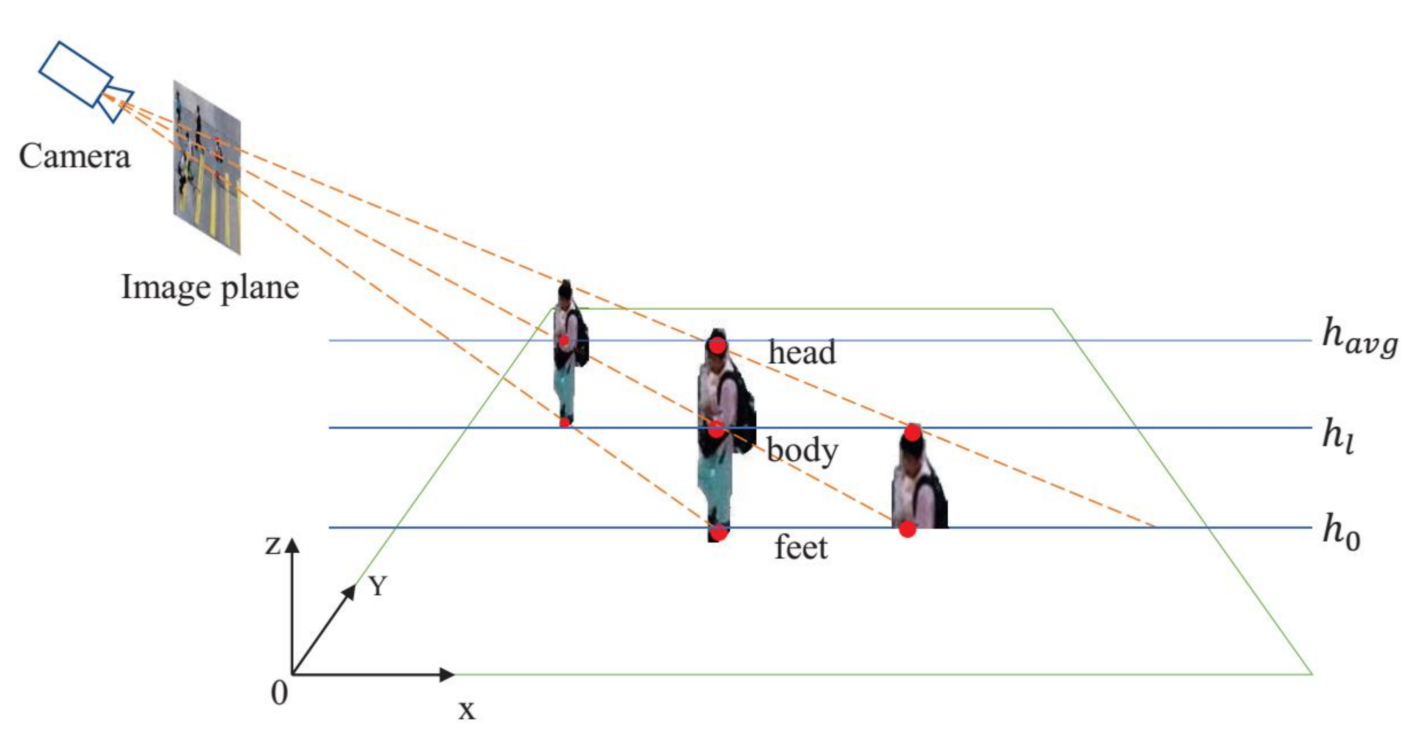
- Assume fixed camera, calibrated and synchronized;
- **Pipeline**: 3D projection, 3D fusion and 3D density map prediction;
- **Losses**: Camera prediction, 3D prediction, and 3D-2D projection consistency measure.



3D Crowd Counting Pipeline

1) 3D projection (multi-height projection)

- The multi-height projection puts feature of the person's body along z-dimension.
- Body features align to form a 3D representation for a person.



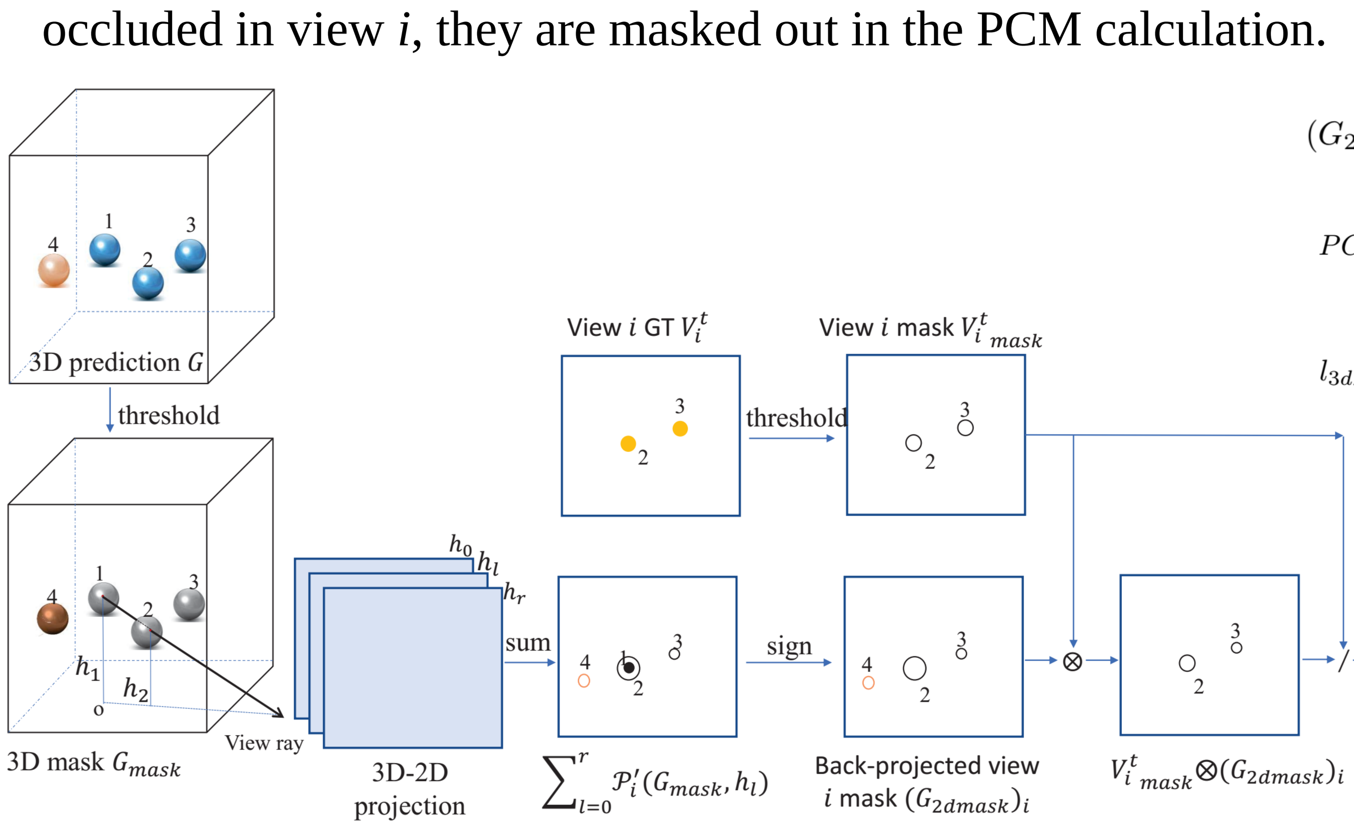
2) 3D-2D projection consistency loss *PCM*:

- Encourages consistency between 3d prediction and 2d ground-truth.
- Example: There are 4 people in the 3D prediction, while only Persons 2 and 3 are visible in the 2D view *i*. Since Person 1 is occluded by Person 2, and Person 4 is totally occluded in view *i*, they are masked out in the PCM calculation.

$$(G_{2dmask})_i = \text{sign}(\sum_{l=0}^{r} \mathcal{P}_i'(G_{mask}, h_l)),$$

$$PCM_i = \frac{\| V^t_{i\ mask} \otimes (G_{2dmask})_i \|_1}{\| V^t_{i\ mask} \|_1 + \alpha},$$

$$l_{3d\_2d} = \sum_{i=1}^{N} (1 - PCM_i).$$



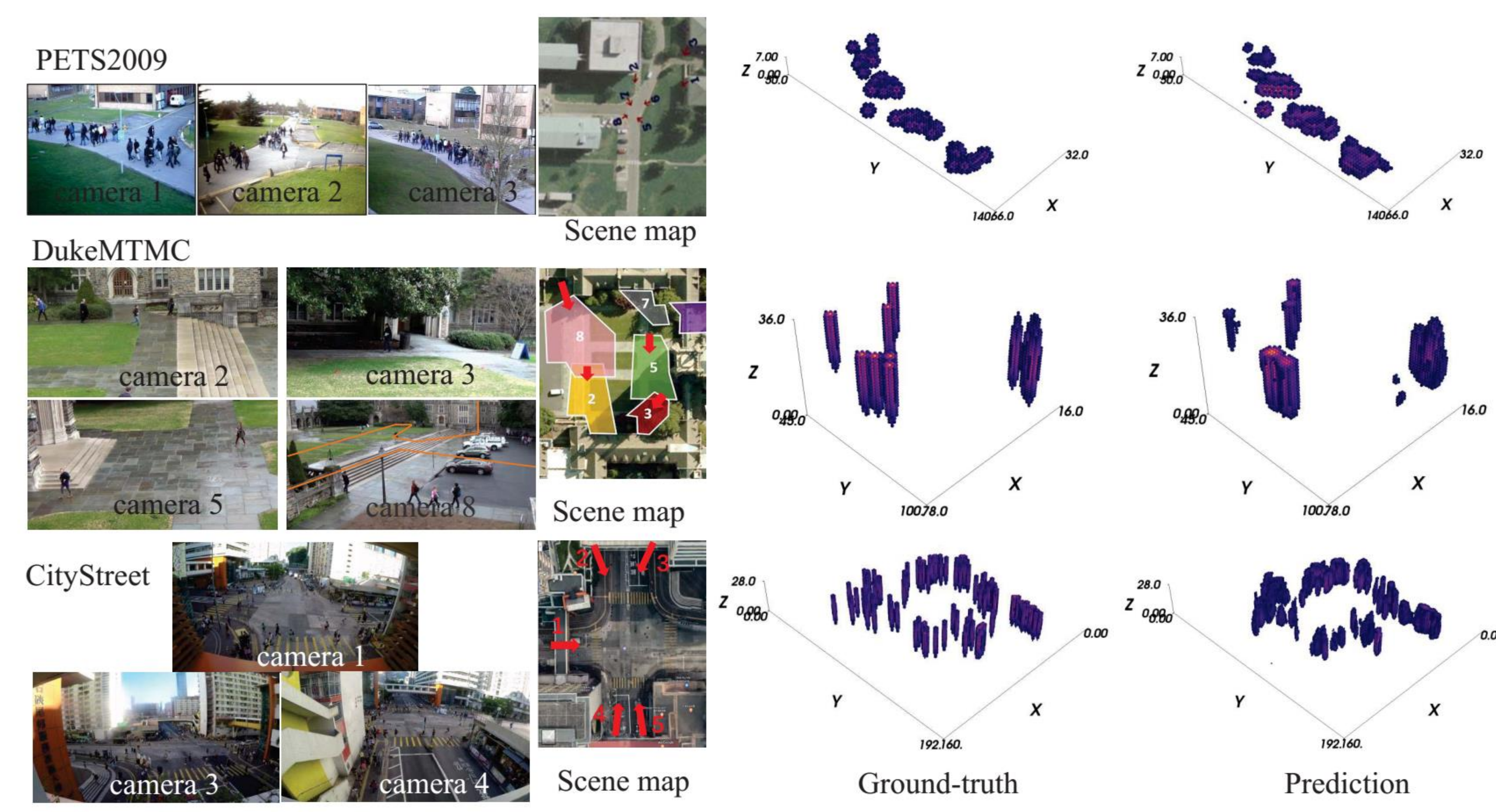
Loss Function

$$l_{all} = \underbrace{l_{3d}}_{\text{3D prediction loss}} + \beta \underbrace{l_{2d}}_{\text{2D prediction loss}} + \gamma \underbrace{l_{3d\_2d}}_{\text{3D-2D consistency loss (PCM)}},$$

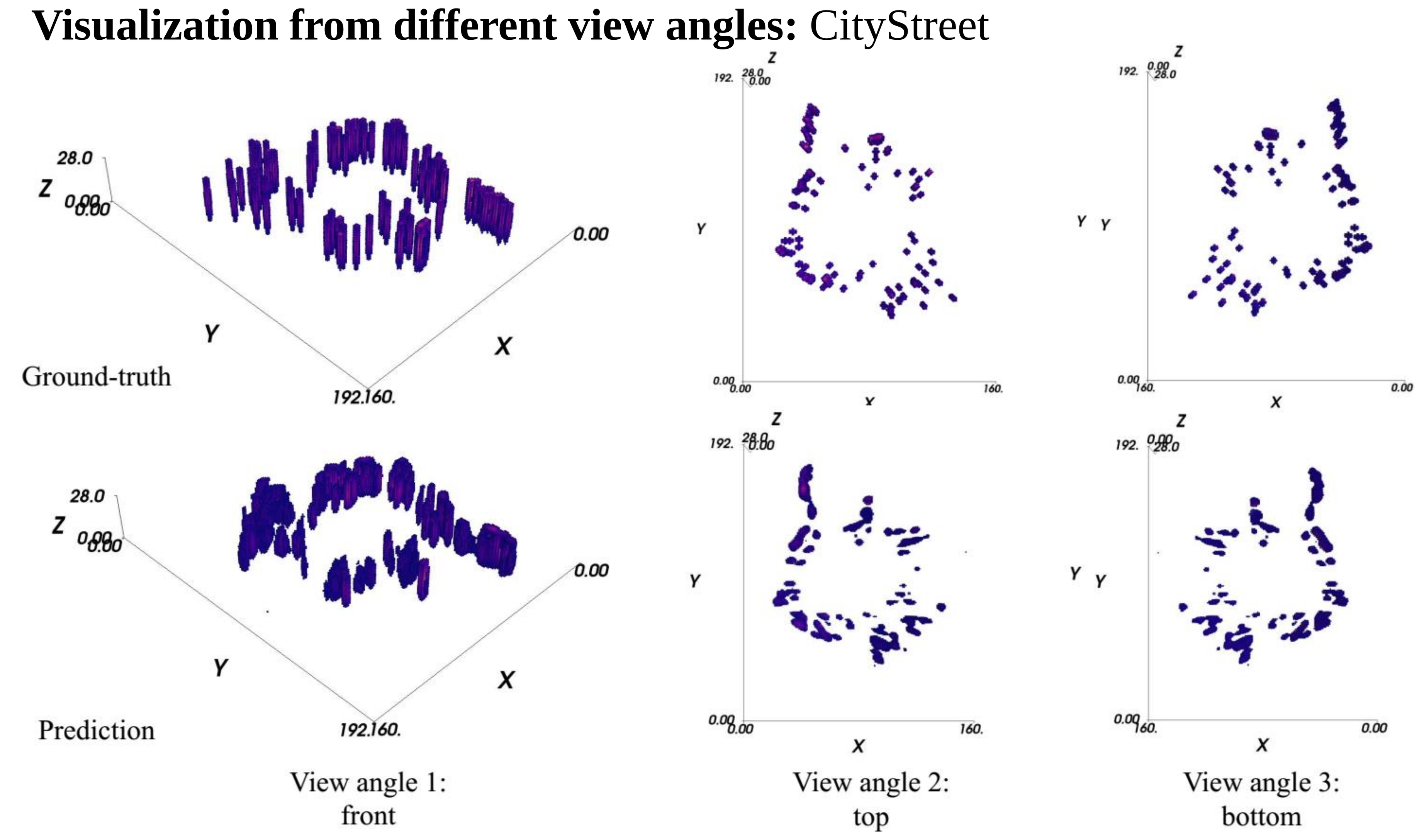
where β = 0.01 and γ are hyperparameters for weighting the contributions of each term.

Datasets and Results

Datasets: PETS2009, DukeMTMC and CityStreet



Visualization from different view angles: CityStreet



Experiments and Ablation Study

1) Counting performance (mean absolute error; lower is better):

Dataset	PETS2009	DukeMTMC	CityStreet
Dmap weighted	8.32	2.12	9.36
Detection+ReID	9.41	2.20	27.60
Late fusion (Zhang and Chan 2019)	3.92	1.27	8.12
Naïve early fusion (Zhang and Chan 2019)	5.43	1.25	8.10
MVMS (Zhang and Chan 2019)	3.49	**1.03**	8.01
3D counting (ours)	**3.15**	1.37	**7.54**

We perform best on two datasets and achieve comparable result on one dataset.

2) Ablation study on loss weights and ground-truth setting:

Dataset	PETS2009			DukeMTMC			CityStreet		
$n*h$	3D	3D+2D	3D+2D+PCM	3D	3D+2D	3D+2D+PCM	3D	3D+2D	3D+2D+PCM
7*40cm	4.12	3.20	**3.15** ($\gamma=100$)	1.82	1.71	1.65 ($\gamma=10$)	8.98	8.49	8.35 ($\gamma=100$)
14*20cm	4.88	4.57	4.24 ($\gamma=10$)	2.12	1.63	1.49 ($\gamma=3$)	8.72	7.89	7.71 ($\gamma=30$)
28*10cm	5.34	4.27	4.21 ($\gamma=1$)	2.15	1.41	**1.37** ($\gamma=0.5$)	7.87	7.58	**7.54** ($\gamma=10$)

γ is the hyperparameter for *PCM* loss. n is the number of voxels in the z-dimension (height), and h is the voxel height in the 3D world. The number of voxels in z-dimension is slightly larger (9, 18, 36) for DukeMTMC.

Future Work

- A more practical dataset for 3D counting;
- Cross-scene ability of the model.

(Our project) (Our lab)