

Wide-Area Crowd Counting via Ground-Plane Density Maps and Multi-View Fusion CNNs: Supplemental

Qi Zhang

Antoni B. Chan

Department of Computer Science, City University of Hong Kong

qzhang364-c@my.cityu.edu.hk, abchan@cityu.edu.hk

1. Rationale of the scale selection

The scale-selection Eq. (2) is derived from the projection equations. Let H be the 3D object height, z the zoom factor of the image pyramid, (s_r, d_r) and (s_0, d_0) the scale/depth of the object at a reference and observed scale, and Y the 2D object height at scale 0. Using a normalized camera ($f=1$), the projection equations give

$$z^{s_r} Y = \frac{H}{d_r}, z^{s_0} Y = \frac{H}{d_0} \Rightarrow d_r z^{s_r} = d_0 z^{s_0} \quad (1)$$

$$\Rightarrow s_r - s_0 = \log_z \frac{d_0}{d_r} \Rightarrow s_0 = s_r - \log_z \frac{d_0}{d_r}. \quad (2)$$

(s_r, d_r) are selected manually for fixed selection, and optimized as parameters (b, k) for learned selection (note that d_r can be absorbed into b in (3)).

2. Training details

We use 2 training stages. Stage 1 includes the main scene-level task and auxiliary tasks for each view (red boxes in Fig. 1). For late fusion, auxiliary losses are applied between the predicted and GT density maps for each view. For early fusion, an auxiliary branch of 3 layers of FCN predicts density maps for each view before the auxiliary loss. The learning rate is set to $1e-4$. In Stage 2, the auxiliary tasks are removed, leaving only the scene-level task. FCN-7 (either density map estimator or feature extractor) is fixed and the fusion and scale selection parts are trained. The loss function is the pixel-wise squared error between the ground-truth and predicted density maps. The learning rate is set to $1e-4$, and decreases to $5e-5$ during training. After training the two stages, the model is fine-tuned end-to-end. The training batch-size is set to 1 in all experiments.

3. Additional experiment results

We present additional experiment results, which were not included in the main paper due to space.

Fixed soft scale selection. The proposed two scale selection modules can be regarded as “fixed discrete” and “learnable soft” scale selection. We perform an extra ab-

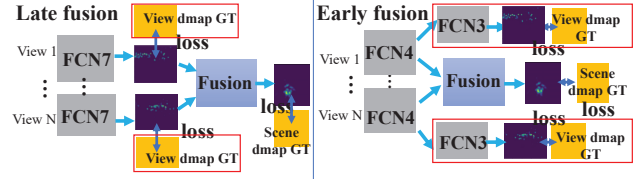


Figure 1: Architecture and loss functions for Stage 1 training. The modules and losses in red boxes are removed in Stage 2.

lation study on the third variant version “fixed soft” scale selection (see Table 1).

CSR-net backbone. We use the CSR-net [3] instead of the FCN-7 as the backbone of the proposed multi-view counting framework. The counting results of the 3 models on Street dataset can be found in Table 2. We get slightly better results and these results are consistent with the FCN-7 backbone: MVMS fusion improves over late/naive fusion.

Detection+ReID with Detect or ReID ground-truth.

In crowd scenes, detection methods are limited by severe occlusions among the crowd, while ReID methods are hindered by detection errors, partial occlusions, scale changes between cameras, and low image-patch resolution. To illustrate the difficulties, we use the ground-truth inter-camera associations (i.e., the best possible ReID) on the people detections and get counting MAE 30.3 on CityStreet, which is worse than our density-map fusion methods. Likewise, we apply ReID [4] on the ground-truth person boxes (i.e.,

Dataset	fixed discrete	fixed soft	learnable soft
PETS	3.82	3.59	3.49
Street	7.80	8.55	8.01

Table 1: MVMS model selection module settings comparison: fixed discrete, fixed soft and learnable soft.

Method	Dmap weighted	Late	Naive	MVMS
mae	9.36	8.36	8.19	7.89

Table 2: CSR-Net backbone counting performance on Street.

the best possible detector), and get counting MAE 13.7. Integrating multi-view detection and ReID for multi-view crowd counting would be interesting future work, and our dataset could serve as a test-bed.

Single camera for scene-level counting. The comparisons of scene counting using single camera views and multi-view methods are presented in Table 3, 4 and 5. In general, for each method, the lowest scene-level MAE (underlined in each row) is achieved by using multiple camera views, which shows that using multiple cameras improves the counting performance for wide-area scenes, since a single camera may not be able to reliably see the whole scene. For the “Dmap” method, the smallest MAE on PETS2009 and DukeMTMC is achieved by camera 3 and camera 8, respectively. The reason is that more people can be seen in the field-of-views of camera 3 or camera 8, and if the method cannot properly fuse the multi-view information from multiple cameras well, the error after fusion will increase compared to the single camera view prediction.

Performance on DukeMTMC Test Hard set. The scene-level counting results on the DukeMTMC Test Hard set are presented in Table 6. Using the same sampling method for the DukeMTMC training set, 200 frames of each view are extracted from the Test Hard set. Compared to the training set, the Test Hard set contains more crowds. Similarly, the region R2 (see paper Fig. 6) is excluded in the testing. Our fusion model can achieve better scene-level counting results than the baselines. Among our methods, late fusion has slightly lower error than MVMS.

Comparison with traditional multi-view counting method. We compare our method with a traditional multi-view counting method “Hybrid” [1] in Table 7. [1] proposed two approaches (head detector and count regression) by fusing hand-crafted features (corner points or Harr feature) from multiple cameras for multi-view counting. Similar to [1], we use PETS2009 S1L1 13_57 (view 1 and 2) for training and 13_59 (view 1 and 2) for testing. Our fusion models all achieve better performance than the multi-view counting method based on traditional hand-crafted low-level features, and MVMS with learnable scale selection achieves the best scene-level counting performance.

4. Example results

First, we show each view’s distance ratio map (the log transformed ratio of the pixel distance and the reference distance) in Fig. 2, and the scale selection masks of the fixed scale selection and the learnable scale selection in MVMS model in Fig. 3, 4 and 5.

Fig. 6 presents the predicted ground-plane density maps and the predicted scene-level counts. See the supplemental videos for more example results of the estimated ground-plane density maps.

References

- [1] Fabio Dittrich, Luiz ES de Oliveira, Alceu S Britto Jr, and Alessandro L Koerich. People counting in crowded and outdoor scenes using a hybrid multi-camera approach. *arXiv preprint arXiv:1704.00326*, 2017. 2, 4
- [2] James Ferryman and Ali Shahrokni. Pets2009: Dataset and challenge. In *2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, pages 1–6. IEEE, 2009. 3
- [3] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1091–1100, 2018. 1
- [4] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2197–2206, 2015. 1
- [5] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016. 3

Method (view)	PETS 2009 [2]			
	1	2	3	scene
Dmap (camera 1)	3.37	7.93	9.69	11.16
Dmap (camera 2)	2.59	5.59	7.62	9.09
Dmap (camera 3)	4.08	6.46	5.84	7.28
Dmap weighted (multiview)	2.63	5.56	6.18	7.51
Detection+ReID (camera 1)	8.60	12.90	15.51	16.27
Detection+ReID (camera 2)	7.67	11.19	13.81	14.57
Detection+ReID (camera 3)	8.35	12.37	14.61	15.40
Detection+ReID (multiview)	-	-	-	9.41
Late fusion (multiview)	-	-	-	3.92
Naïve (multiview)	-	-	-	5.43
MVMS (multiview)	-	-	-	3.49

Table 3: Comparison of single view and multi-view counting on PETS2009. Density map (Dmap) and Detection+ReID are applied to only single camera views. The columns show the mean absolute error (MAE) for each camera view and the whole scene.

Method (view)	DukeMTMC [5]				
	2	3	5	8	whole
Dmap (camera 2)	0.62	2.30	2.23	4.08	5.19
Dmap (camera 3)	2.39	0.91	1.52	6.88	8.03
Dmap (camera 5)	1.75	1.24	0.98	5.57	6.72
Dmap (camera 8)	4.48	6.09	5.38	1.41	1.93
Dmap weighted (multiview)	0.66	0.64	0.74	1.30	2.12
Detection+ReID (camera 2)	2.06	2.78	3.29	2.45	3.51
Detection+ReID (camera 3)	2.29	0.25	1.93	5.68	7.20
Detection+ReID (camera 5)	2.13	1.40	0.96	4.86	6.38
Detection+ReID (camera 8)	2.25	1.97	1.85	3.58	5.10
Detection+ReID (multiview)	-	-	-	-	2.20
Late fusion (multiview)	-	-	-	-	1.27
Naïve (multiview)	-	-	-	-	1.25
MVMS (multiview)	-	-	-	-	1.03

Table 4: Comparison of single view and multi-view counting on DukeMTMC. Density map (Dmap) and Detection+ReID are applied to only single camera views. The columns show the mean absolute error (MAE) for each camera view and the whole scene.

Dataset	City Street			
Scene	1	3	4	whole
Dmap (camera 1)	10.16	11.29	11.69	12.50
Dmap (camera 3)	14.03	12.55	16.48	17.94
Dmap (camera 4)	19.08	15.74	21.56	23.15
Dmap weighted (multiview)	9.58	12.43	15.62	11.10
Detection+ReID (camera 1)	41.38	37.83	44.16	45.80
Detection+ReID (camera 3)	36.48	32.94	39.23	40.87
Detection+ReID (camera 4)	25.97	22.90	28.57	30.03
Detection+ReID (multiview)	-	-	-	27.60
Late fusion (multiview)	-	-	-	8.12
Naïve (multiview)	-	-	-	8.10
MVMS (multiview)	-	-	-	8.01

Table 5: Comparison of single view and multi-view counting on City Street. Density map (Dmap) and Detection+ReID are applied to only single camera views. The columns show the mean absolute error (MAE) for each camera view and the whole scene.

Method	Dmap weighted	Detection+ReID	Late fusion		Naïve early fusion	MVMS	
setting	-	-	with	without	-	fixed	learnable
MAE	4.17	5.88	2.73	2.75	2.82	2.96	2.81

Table 6: Experiment results on DukeMTMC Test Hard set.

setting	Hybrid [1]	Late fusion (with)	Naive early fusion	MVMS (learnable)
MAE	2.03	1.53	1.57	1.44

Table 7: Experiment results on PETS S1L1 (view 1 and 2) comparing with the traditional multi-view method.

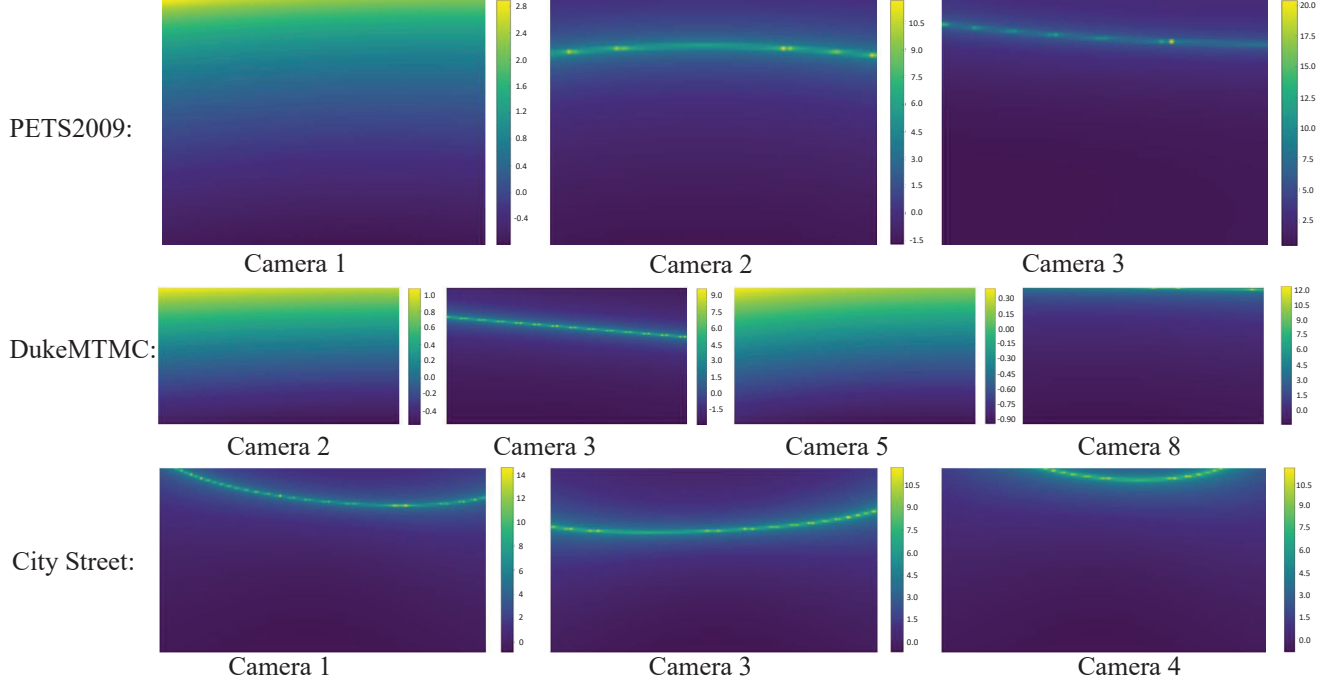


Figure 2: The distance ratio map of PETS 2009, DukeMTMC and City Street.

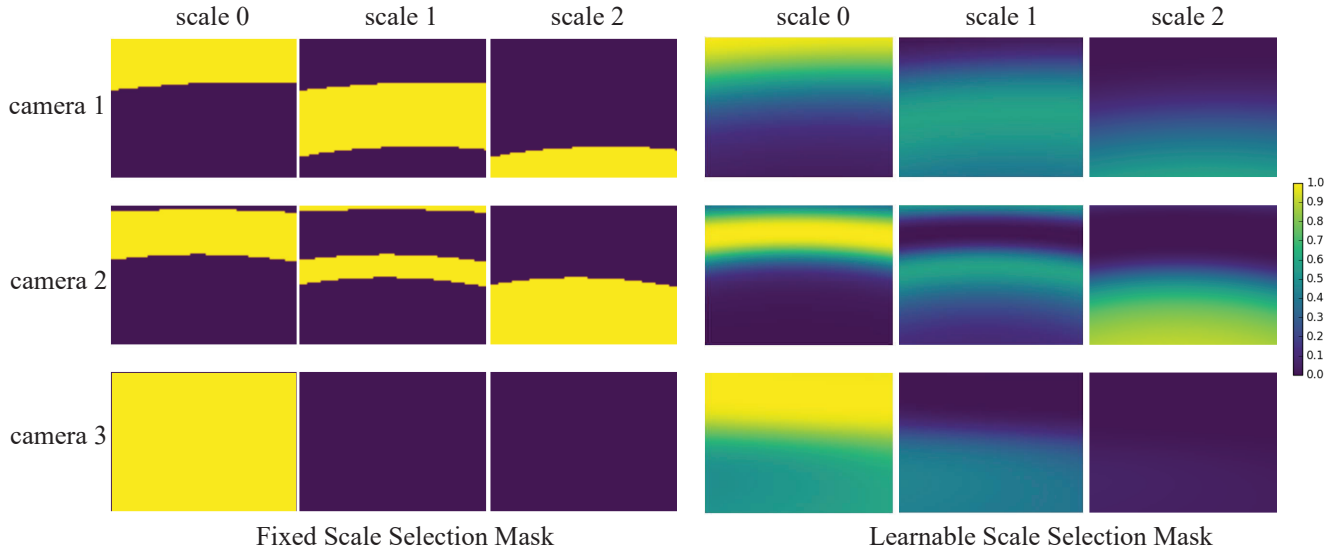


Figure 3: The selection mask of fixed scale selection (left) and learnable scale selection (right) for PETS2009.

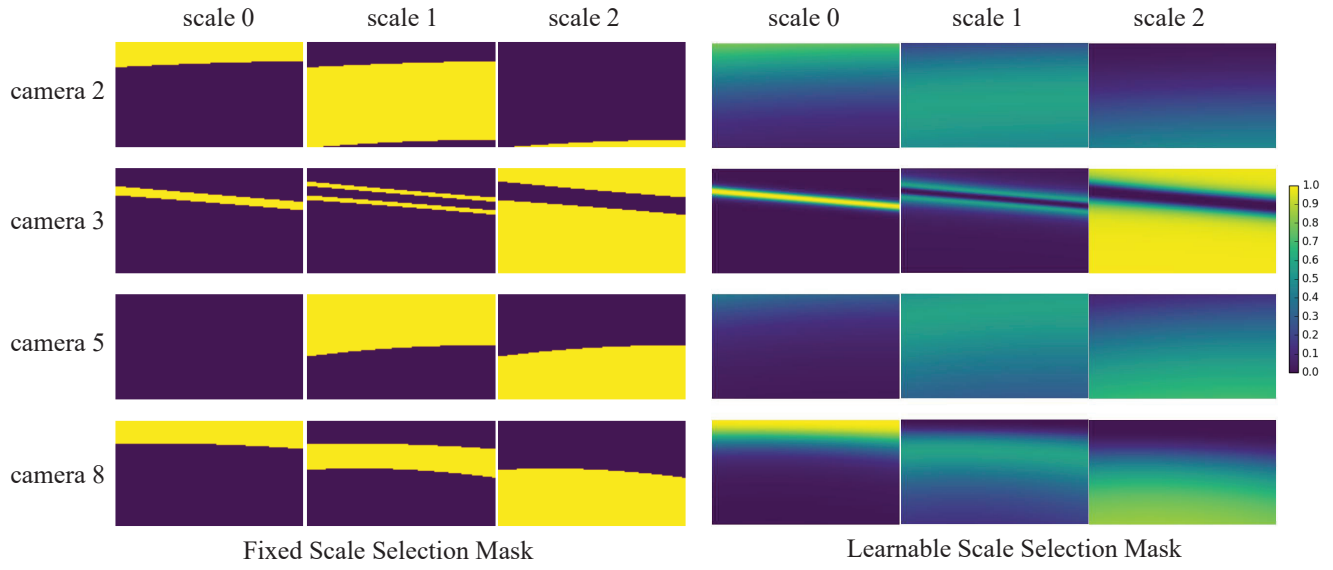


Figure 4: The selection mask of fixed scale selection (left) and learnable scale selection (right) for DukeMTMC.

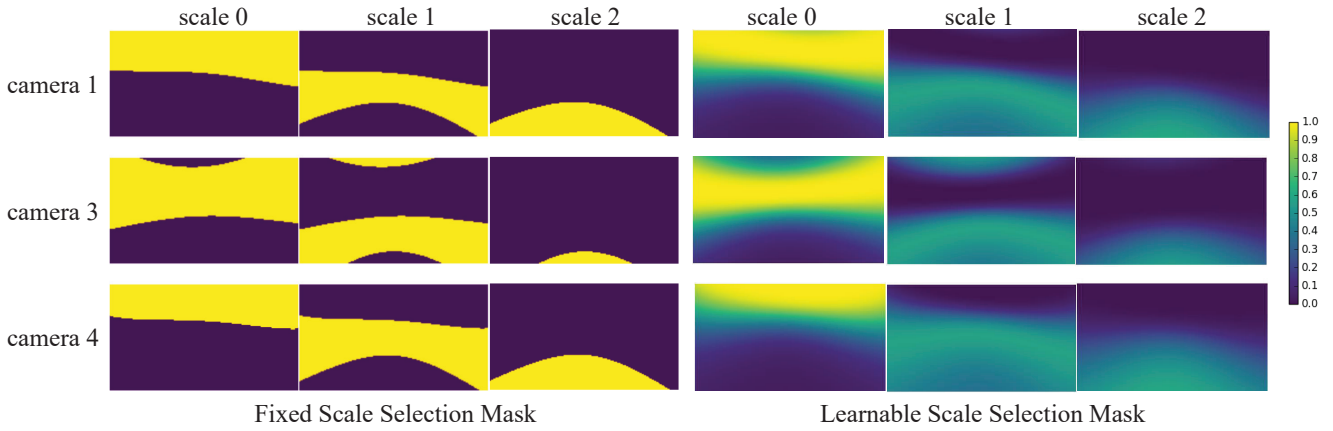


Figure 5: The selection mask of fixed scale selection (left) and learnable scale selection (right) for City Street.

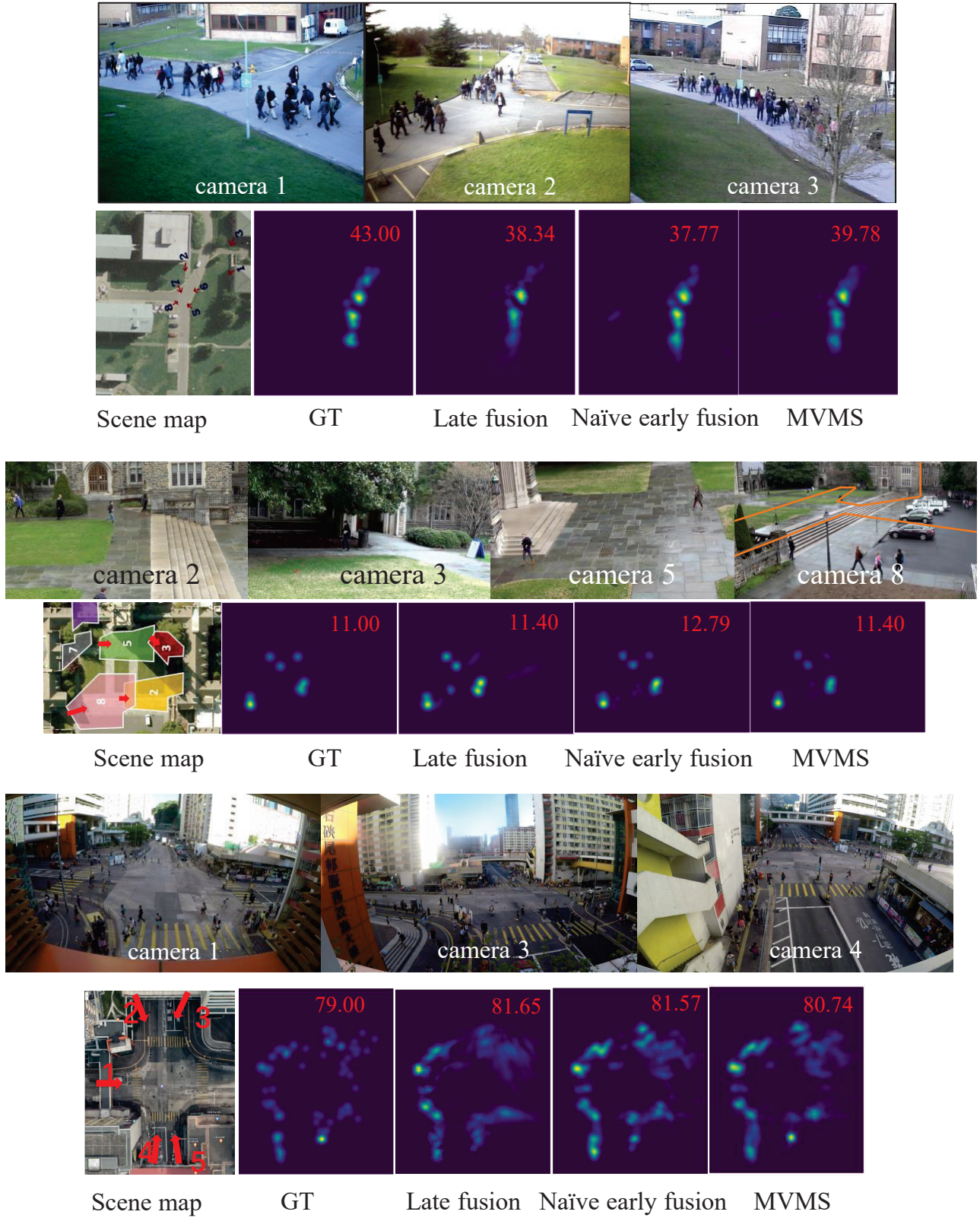


Figure 6: The example of the estimated ground-plane density maps of the proposed 3 multi-view fusion models on the 3 datasets.