Supplementary Material
Bringing Generalization to Deep Multi-View Pedestrian Detection

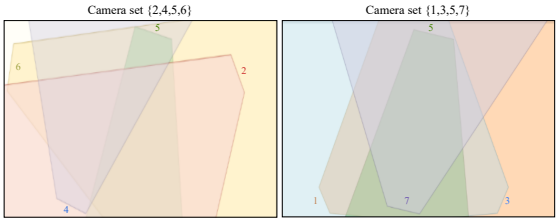

Figure 1. Camera splits of WildTrack dataset for changing camera configuration experiment.

1. Choice of Loss Function

Method	ImageNet (pre-train)	MODA	MODP	Prec	Recall
MSE	✓	57.3(±0.2)	72.6(±0.0)	75.6(±0.1)	84.5(±0.05)
CC	✓	55.5(±5.5)	**74.2**(±0.4)	72.1(±4.4)	**89.5**(±2.6)
KL	✓	62.5(±0.1)	73.4(±0.04)	**89.1**(±0.0)	71.3(±0.0)
KLCC	✓	**69.4**(±0.6)	72.96(±0.2)	83.74(±0.5)	86.14(±0.3)

Table 1. Choice of Loss Function: we present an ablation study for our proposed method on the scene generalization experiment. Overall, the model trained with both KL-Divergence and Cross-Correlation achieves the best performance.

We ablate the choice of the loss function in Table 1 for the scene generalization experiment. We consider the Mean Squared Error (MSE), KL-Divergence(KL), Pearson Cross-Correlation (CC), as well as our chosen loss function (KL+CC). We find that the combination of KL-Divergence and Pearson Cross-Correlation achieves significantly better results than any other loss function.

2. Qualitative results

First we show the predicted occupancy maps of MVDet, MVDeTr, SHOT and our method and compare them with the ground truth, in the traditional setting. Subsequently, qualitative results are shown w.r.t to three generalization abilities obtained from both the WildTrack and MultiViewX datasets.

2.1. WildTrack Dataset

The traditionally evaluated results which contains occupancy maps of ground truth, our method, MVDet, MVDeTr and SHOT are shown in Fig. 2. The occupancy map from our method which uses average pooling, KLCC loss function and ImageNet pretraining gives us more accurate localization as compared to the base MVDet architecture. The results (maps) are competitive when compared to SHOT and MVDeTr. The maps obtained using MVDeTr are sharper and focused, however, it also has more false positives.

Varying number of cameras: The output occupancy map for varying number of cameras are shown in Fig. 3. WildTrack consists of seven cameras, we show the results inferred with three cameras upto six cameras. As the number of views are increasing, we get an accurately localized occupancy map.

Changing camera configurations: The output occupancy map for cross subset evaluation are shown in Fig. 5. Here, we have the occupancy maps for a model trained on one set and tested on other set. For example, trained on camera views one, three, five and seven and tested on cameras two, four, five and six or vice-versa like the camera splits shown in Fig. 1. Clearly the pre-training is improving localization in both the methods. Furthermore, our method with average pooling is better at disambiguating the occlusions and also giving brighter outputs (resulting in sharp maxima's).

2.2. MultiViewX Dataset

In this subsection the qualitative results for MultiViewX dataset are been shown. We consider similar configurations as in the WildTrack dataset. The obtained results clearly indicates the improvements our method brings over the MVDet, MVDeTr and SHOT model and observations are similar to that of the WildTrack dataset. Fig. 2 shows the traditionally evaluated results.

Varying number of cameras: The output occupancy map for varying number of cameras are shown in Fig. 6. MultiViewX consists of six cameras, we show the results inferred with three cameras upto five cameras. As the number of views are increasing, we get an accurately localized occupancy map.

Changing camera configurations: The output occupancy map for cross subset evaluation are shown in Fig. 7. Here, we have the occupancy maps for a model trained on

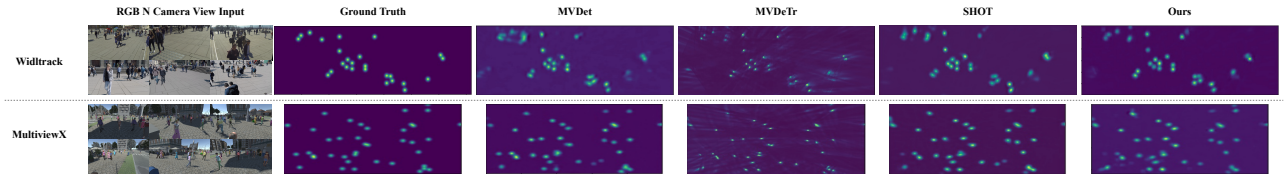

Figure 2. Sample frames from WildTrack and MultiViewX dataset with corresponding occupancy maps of ground truth, our result MVDet, MVDeTr and SHOT for comparison. We can see the clusters forming in the MVDet predictions, in contrast our method gives much sharper and distinct predictions.

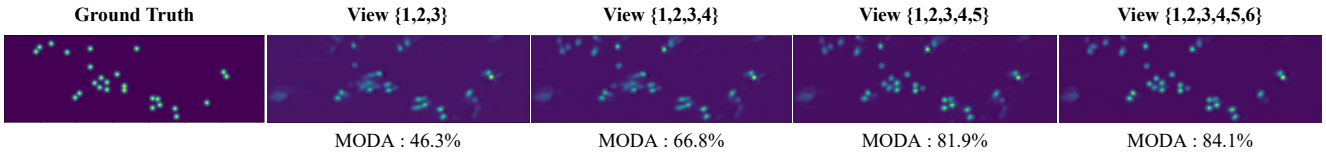

Figure 3. Occupancy maps for varying number of cameras on WildTrack dataset when trained on seven cameras and tested on varying subsets of the cameras.

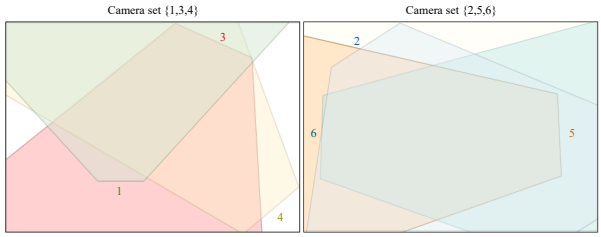

Figure 4. Camera splits of MultiViewX dataset for changing camera configuration experiment shown in Table 2.

one set and tested on other set. For example, trained on camera views one, three, and four and tested on cameras two, five and six or vice-versa, the camera splits are shown in Fig. 4 and their results are shown in Table 2.

			Inference on {1,3,4}				Inference on {2,5,6}			
		Method	MODA	MODP	Prec	Recall	MODA	MODP	Prec	Recall
Trained on camera set	{1,3,4}	MVDet	72	76.1	93.5	77.4	46.3	66.4	94.5	49.1
		MVDeTr	**77.4**	**85.1**	97.9	79	60.4	71.3	95.4	63.5
		SHOT	74.3	76.3	94.1	**79.3**	37.3	67	67.5	**72.1**
		Ours	67.7	76.4	96.2	70.5	59.6	73.4	94.7	63.2
		Ours (DropView)	67.3	75.3	**98.4**	68.5	**62.9**	**73.6**	**96.3**	65.4
	{2,5,6}	MVDet	34.3	66.2	93.8	36.7	77.6	77.4	93.8	83.1
		MVDeTr	51.1	72.1	**94.9**	54	**83.1**	**87.1**	**97.8**	**85**
		SHOT	47.3	**73**	94.2	50.3	80.7	78.7	96.1	84.1
		Ours	45.8	71.8	94.5	48.6	76.1	78.7	95.9	79.5
		Ours (DropView)	**53.4**	71.6	88.2	**61.6**	75.2	77.4	92.8	81.5

Table 2. Experiments on the MultiViewX dataset with changing camera configurations

2.3. Scene Generalization

The qualitative results of output occupancy map for cross-dataset evaluation are shown in Fig. 8, when we train on synthetic dataset (MultiViewX) and test on real dataset (WildTrack). First four occupancy maps are the outputs of MVDet, MVDeTr, SHOT and our method when tested on only 6 views of WildTrack dataset for having a fair comparison with other methods. We also show the output occupancy map when tested on all the views of WildTrack dataset. Our method provides accurately localized occupancy maps and disambiguate the occlusions as compared to other methods.

3. GMVD Dataset Charachteristics

The GMVD dataset includes seven distinct scenes, one indoor (subway) and six outdoors. We vary the number of total cameras in each scene and provide different camera configurations within a scene. Additional salient features of GMVD include daytime variations *(morning, afternoon, evening, night)* and weather variations *(sunny, cloudy, rainy, snowy)*. We generate multiple short sequences for each scene while randomly varying the daytime and the weather. The generation of multiple random sequences ensures diversity, as different pedestrians (with different clothing and appearance) are picked in each case, there are approximately 2800 person indentities as shown in Fig. 9. The dataset also includes significant variations in lighting conditions. Table 3 shows the comparison of our dataset with exisiting ones based on the ground plane grid area (region of interest ROI) in meters being used for multi-view detection and provides the dimensions to generate Top View (Bird's Eye View) representation of the ROI and the density of the pedestrians in the scene per frame basis (defined by crowdedness). Fig. 10 shows the generated annotations in terms of bounding boxes and ground truth occupancy map after synchronised camera calibrations.

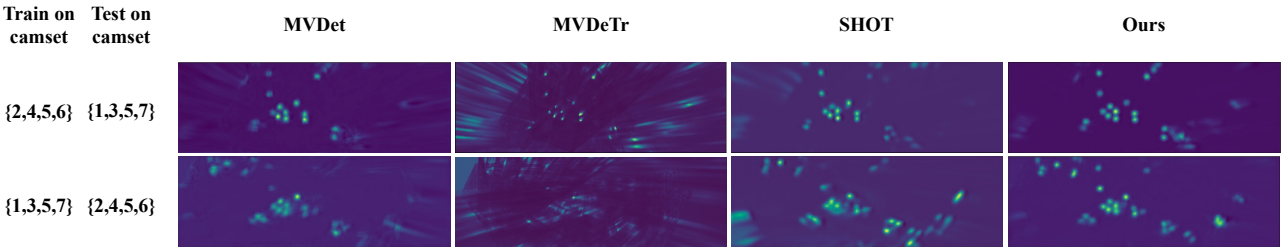

Figure 5. Result occupancy maps for cross subset evaluation from WildTrack dataset.

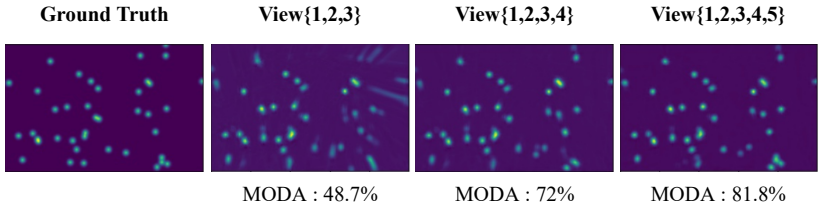

Figure 6. Occupancy maps for varying number of cameras on MultiViewX dataset when trained on seven cameras and tested on varying subsets of the cameras.

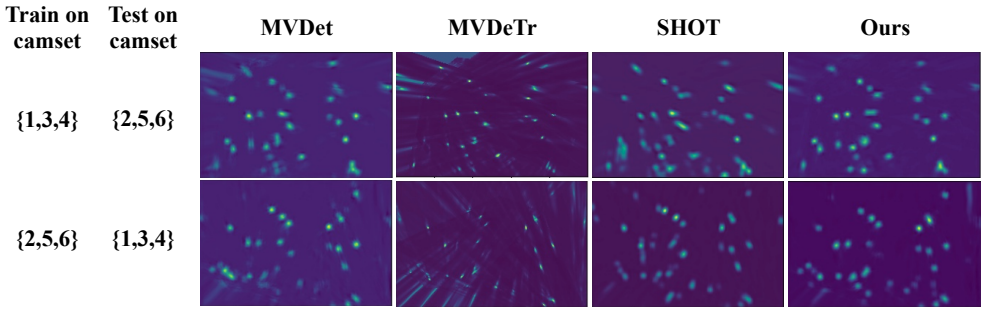

Figure 7. Result occupancy maps for cross subset evaluation from MultiViewX dataset.

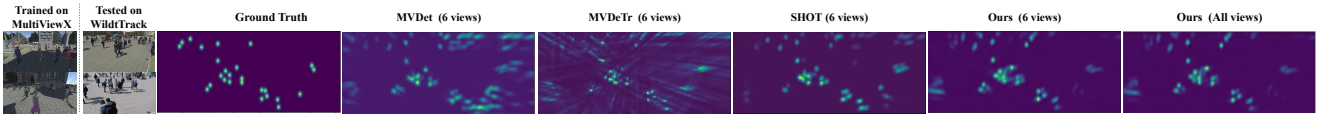

Figure 8. Occupancy maps obtained on inference from WildTrack dataset where the models where trained on the synthetic dataset (MultiViewX).

Table 3. Region of Interest (top view area) for various scenes of GMVD Dataset compared with WildTrack and MultiViewX

Dataset	Grid Area	Top View Dimensions	Crowdedness
WildTrack	12×36 m^2	480×1440	20 person/frame
MultiViewX	16×25 m^2	640×1000	40 person/frame
GMVD(ours)			
GTA scene 1	20×30 m^2	800×1200	20-50 person/frame
GTA scene 2	30×12 m^2	1200×480	20-50 person/frame
GTA scene 3	25×25 m^2	1000×1000	20-50 person/frame
GTA scene 4	29×19 m^2	1160×760	20-50 person/frame
GTA scene 5	28×27 m^2	1120×1080	20-50 person/frame
GTA scene 6	33×31 m^2	1320×1240	20-50 person/frame
Unity scene 1	16×25 m^2	640×1000	40 person/frame
Unity scene 2	16×25 m^2	640×1000	40 person/frame

Figure 9. Samples of various person identities are shown from both Unity and GTAV, there are total 2800 person identities which are included in GMVD Dataset.

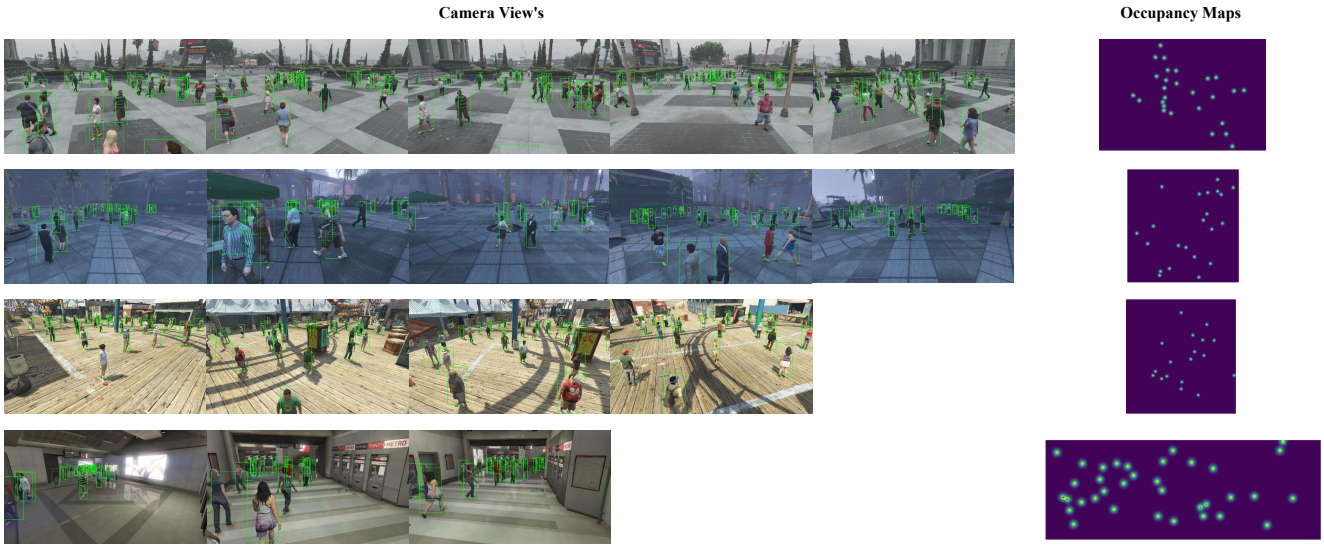

Figure 10. Synchronized camera calibration and sample ground truth annotations generated are shown in terms of bounding boxes in respective camera view's and the top view occupancy maps for GMVD Dataset.