Supplementary Material for 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

In this document, we further provide additional explanations and experimental results. Specifically, in Sec. 1, we report about the details of training configurations and datasets. Sec. 2 provides more ablation studies about the proposed model. In Sec. 3, we present the analysis of multi-frame results.

1. Experimental Details

1.1. Training hyperparameters

Along with the hyperparameters settings provided in Tab. 1, a weight decay of 1×10^{-4} is used to regularize the model for all experiments. Also, during the testing phase, a non-maximum suppression is applied to filter out predictions that fall below the predefined confidence threshold.

We train the model with a batch size of 1, containing three to eight views per batch depending on the dataset, as specified in Tab. 2. The memory requirement varies based on factors such as the chosen Z-dimension, the number of views, the ground plane size and the downsampled resolution. Detailed information about the memory requirements for various Z-dimensions is provided in Tab. 4.

1.2. Datasets comparison

Tab. 2 shows the comparison between three datasets. Apart from the differences shown in the table, certain parameters such as image size and grid cell size remain the same across the datasets, using 1920 x 1080 and 2.5 cm^2, respectively.

Evaluation	Train-set	Test-set	lr	epochs	threshold
Same-domain	WildTrack	WildTrack	5e-5	8	0.4
	MultiviewX	MultiviewX	1e-4	20	0.4
Scene generalization	MultiviewX	WildTrack	2e-5	15	0.4
	GMVD	WildTrack	2e-5	8	0.4
	GMVD	GMVD	1e-5	5	0.3

Table 1. **Hyperparameters used for training the proposed model on different evaluation protocols** *threshold* means the confidence threshold used in non-maximum suppression.

Figure 1. **Sample sequences from GMVD GTAV Scene 1**. The dataset provides variance for the same scene by recording time difference (Night, Shady, Sunny) and weather condition difference (Rain, Snow).

The GMVD dataset contains seven scenes with both indoor and outdoor environments, each with their own ground plane size and camera setup. Specifically, *GTAV Scene 5* is used as a testing split. Notably, this scene also introduces an additional layer of complexity by offering two different camera configurations: one with 6 cameras and the other with 8 cameras, testing the model's adaptability to varying camera setups. In addition to the challenges presented by varying scenes and camera configurations, GMVD dataset also incorporates differing weather conditions and recording times for each individual scene, as illustrated in Fig. 1.

1.3. Ground plane downsampling

To maintain consistency with previous methods and to reduce memory usage, we also increase the grid cell size from 2.5 cm^2 to 10 cm^2, leading to a 4 times downsampled ground plane resolution as described in in Tab. 2. The use of downsampled discrete coordinates introduces a truncation error due to the discretization of grid coordinates, as opposed to using continuous meter-based measurements and results in decrease on the MODP score which accesses the localization precision of each true positive detection.

Dataset	Scene	Frames	Cameras	Ground Plane Area	Original Grid Size	Downsampled Grid Size	Crowdedness
WildTrack	Real Scene	400	7	12 x 36 m^2	480 x 1440	120 x 360	20 person/frame
MultiviewX	Unity Scene	400	6	16 x 25 m^2	640 x 1000	160 x 250	40 person/frame
GMVD	GTAV Scene 1	1034	5	20 x 30 m^2	800 x 1200	200 x 300	20 person/frame
	GTAV Scene 2	1000	3	30 x 12 m^2	1200 x 480	300 x 120	30 person/frame
	GTAV Scene 3	1014	5	25 x 25 m^2	1000 x 1000	250 x 250	30 person/frame
	GTAV Scene 4	182	5	28 x 27 m^2	1120 x 1080	280 x 270	20 person/frame
	GTAV Scene 5	1012	6, 8	29 x 19 m^2	1160 x 760	290 x 190	30 person/frame
	GTAV Scene 6	1030	7	33 x 31 m^2	1320 x 1240	330 x 310	30 person/frame
	Unity Scene	723	6	16 x 25 m^2	640 x 1000	160 x 250	40 person/frame

Table 2. **Dataset statistics between three multi-view pedestrian detection datasets.** Crowdedness shows the average number of persons involved in the scene. For GMVD Scene 5, two configurations (6 or 8 cameras) are available for the same scene.

Refiner	MODA	MODP	Prec.	Recall
Dilated [1,2,4]	92.9	78.5	96.8	96.0
7x7x7	93.2	77.2	**97.3**	95.8
5x5x5	93.1	78.3	96.1	97.0
3x3x3	93.1	**79.0**	96.8	96.2
7x5x3	**94.1**	78.8	96.4	**97.7**

Table 3. **Performance of large kernel refiner module with different refinement mechanism on the WildTrack dataset.** The top row shows three stacks of dilated convolutions with different dilation rates of [1, 2, 4] with 3×3 kernel size. The following rows demonstrate the utilization of different large kernel sizes within the LKR module.

Z	Height	MODA	MODP	Prec.	Recall	Memory
1	1.6 m	92.3	77.5	96.7	95.6	9GB
4	1.6 m	93.0	77.9	96.2	96.7	10GB
8	1.6 m	**94.1**	78.8	96.4	**97.7**	15GB
16	1.6 m	**94.1**	78.6	**97.3**	96.8	27GB
7	1.4 m	93.8	**79.1**	97.0	96.7	14GB

Table 4. **Performance of 3D feature-pulling mechanism with different Z-dimensions on the WildTrack dataset.** Larger Z-dimension performs better until $Z = 16$.

2. More ablation studies

2.1. Effectiveness of gradual refinement in Large Kernel Refiner module

Our large kernel refiner module aims to gather an individual's dispersed features on the BEV plane and generate a concise and accurate representation. To achieve this, we have systematically explored the impact of employing large kernel convolutions, both in terms of dilation rates and kernel sizes in Tab. 3.

Dilated convolutions result in suboptimal performance compared to using larger kernel sizes. Smaller kernel sizes improve individual location accuracy on the ground plane, while larger kernel sizes exhibit the advantage of reducing false positive rates. Given our focus on accurate person identification across multiple views, we opt for a gradual refinement strategy. This approach yields superior results in terms of MODA and recall scores, aligning well with our goal of achieving robust multi-view detection performance.

2.2. Choice of Z-dimension in 3D feature-pulling

In Tab. 4, we have investigated the impact of varying the Z-dimension in the 3D feature-pulling mechanism. As the Z-dimension increases gradually, the overall performance also improves. However, there is no significant improvement byeond a Z-dimension of 16, even though a substantial increase in memory consumption. When reducing the height to 1.4 m with $Z = 7$, there is a slight drop in performance compared to the setting with 1.6 m with $Z = 8$. Therefore, we choose a Z-dimension of 8 with a height of 1.6 m which provides the optimal balance between accuracy and computational efficiency.

2.3. Analysis on Maximal Fusion Module

We have conducted experiments to test the impact of employing different aggregation mechanisms and the omission of the "Coord Volume" in the maximal fusion module, as presented in Tab. 5. By using an average pooling instead of max. pooling, we observe a minor decrease of 0.3% in MODA score, with the most notable decline in recall score, amounting to 1.7%.

Additionally, when the "Coord Volume" is excluded, the MODA score experiences a significant reduction of 0.9%, which underscores the importance of including positional information in our model's design.

2.4. Additional view-level augmentations

We have experimented adding view-level augmentations to the proposed framework, including random horizontal flipping, cropping and scaling, similar to MVDeTr. Unlike

	MODA	MODP	Prec.	Recall
Avg. pooling	93.8	78.6	**97.8**	96.0
Max. pooling	**94.1**	**78.8**	96.4	**97.7**
w/o. Coord. Volume	93.2	78.3	96.0	97.3
w/. Coord. Volume	**94.1**	**78.8**	**96.4**	**97.7**

Table 5. **Performance of maximal fusion module with different configurations on the WildTrack dataset.** The first two rows show the results with different pooling mechanisms, while the last two rows show the results with or without "Coord Volume" technique.

previous approaches, we applied the same transformation to all views. The results, as shown in Tab. 6, indicate only minimal improvements during same-domain testing on the WildTrack dataset. Our model already achieves a high level of performance, and augmentations provide only marginal additional benefit. Moreover, the WildTrack dataset includes missed annotations, which might have contributed to reaching an upper bound in the same-domain evaluation.

In Tab. 7, we have tested the performance of the proposed model on MultiviewX scene generalization experiment with or without augmentations. The results indicate that incorporating augmentations yields a 1.5% increase in the MODA score. This demonstrates that augmentations play a beneficial role in scene generalization. It can be noted that further optimization of hyperparameters and potentially employing more robust augmentations, such as in 3DROM and MVAug could lead to even better results.

2.5. Recovering truncation error with offset head

As reported in Sec. 1.3, there is a truncation error introduced due to the downsampling of the ground plane to a lower resolution. To address this issue, we adopt an additional offset head to regress the omitted decimal part,

	MODA	MODP	Prec.	Recall
w/o. augmentations	94.1	78.8	96.4	**97.7**
w/. augmentations	**94.2**	**79.5**	**97.5**	97.0

Table 6. **Adding augmentations to the same-domain testing on the WildTrack dataset.** Trained and tested on the same scene of WildTrack dataset.

	MODA	MODP	Prec.	Recall
w/o. augmentations	82.6	**76.2**	89.6	93.4
w/. augmentations	**84.1**	74.5	**90.4**	**94.1**

Table 7. **Adding augmentations to the scene generalization evaluation with the MultiviewX dataset.** Trained on a synthetic dataset (MultiviewX) and tested on a real dataset (WildTrack).

	MODA	MODP	Prec.	Recall
w/o. offset	**94.1**	78.8	96.4	**97.7**
w/. offset	94.0	**80.1**	**97.4**	96.9

Table 8. **Performance of the proposed model with or without an offset head.** Utilizing an offset head improves in MODP score but slightly impacts the localization accuracy.

	Params (M)	Time (ms)
Backbone	11.2	2.4
FSM	0.33	0.22
3DFP	-	0.36
MFM	0.13	0.18
LKR	0.06	0.29

Table 9. **Computational cost analysis of each module (tested on A100 GPU).** *Backbone* is the dilated ResNet-18 network. *3DFP* refers to the 3D feature pulling mechanism. *FSM*, *MFM* and *LKR* are the foreground selector, maximal fusion and large kernel refiner modules, respectively.

following a similar approach as in MVDeTr. The results presented in Tab. 8 indicate that this strategy improves the MODP score by 1.3%, leading to enhanced precision in individual location estimation, while it decreases both MODA and recall scores. Since we prioritize on achieving larger MODA and recall scores, we exclude the use of an offset head in our implementation.

2.6. Computational cost analysis

In Tab. 9, a detailed computational cost analysis of each individual module is presented. Notably, the inclusion of each proposed module does not significantly increase computational complexity. The most substantial GPU memory requirement and inference cost is attributed to the backbone network and the 3D feature-pulling process.

3. Multi-frame analysis

In Fig. 2, we conduct a multi-frame analysis that compares the localization accuracy on the WildTrack dataset using the model trained on GMVD dataset. Our approach demonstrates a lower count of false positives and missed detections, showcasing its robustness in managing identity disappearance across multiple frames.

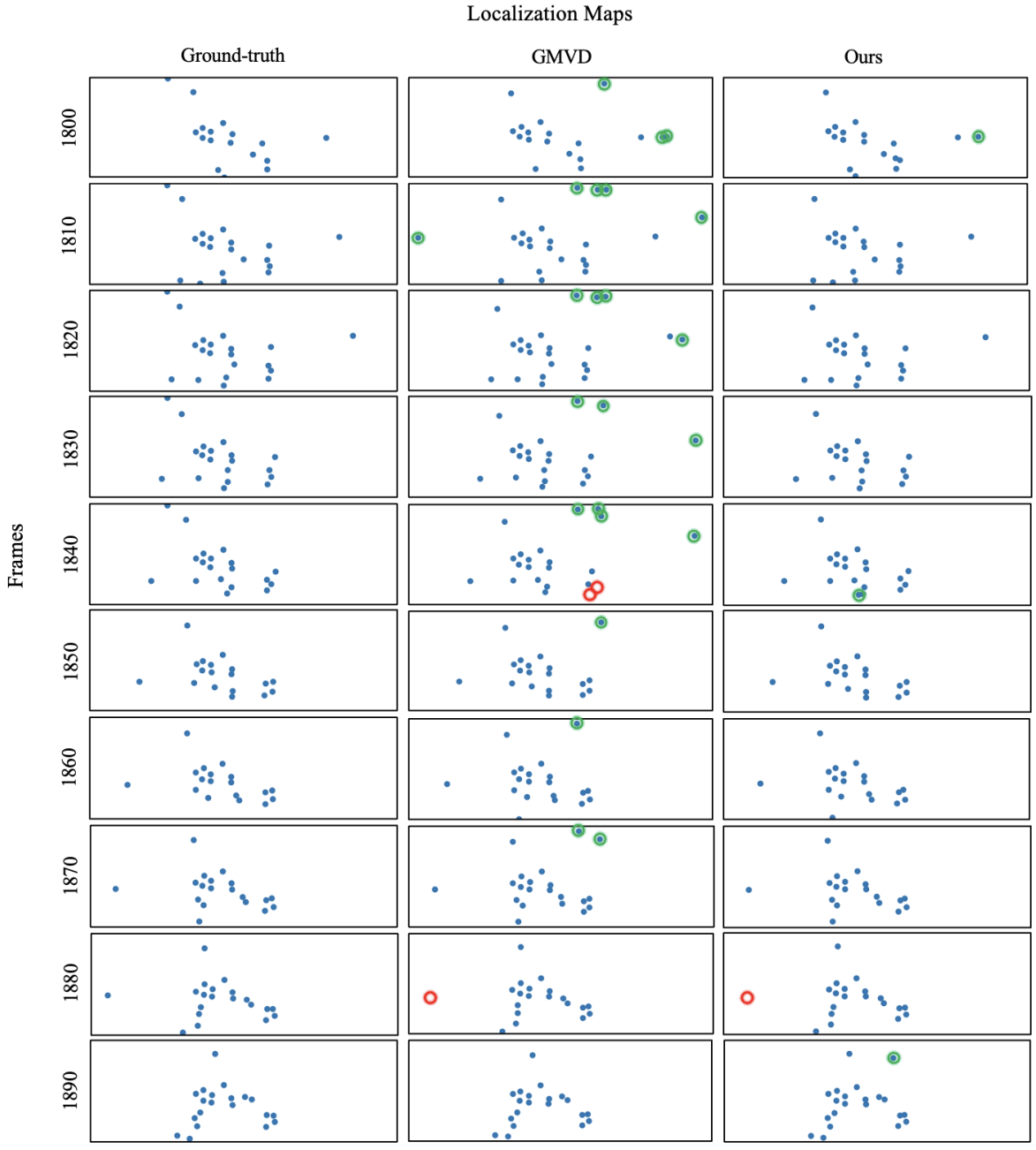

Figure 2. **Multi-frame analysis between GMVD model and our method**. The model is trained on GMVD dataset and tested on WildTrack dataset. Green circles denote false positives and red circles denote missed detections.