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

Abstract

The main challenge in multi-view pedestrian detection is integrating view-specific features into a unified space for comprehensive end-to-end perception. Prior multi-view detection methods have focused on projecting perspective-view features onto the ground plane, creating a "bird's eye view" (BEV) representation of the scene. This paper proposes a simple but effective architecture that utilizes a non-parametric 3D feature-pulling strategy. This strategy directly extracts the corresponding 2D features for each valid voxel within the 3D feature volume, addressing the feature loss that may arise in previous methods. The proposed framework introduces three novel modules, each crafted to bolster the generalization capabilities of multi-view detection systems. Through extensive experiments, the efficacy of the proposed model is demonstrated. The results show a new state-of-the-art accuracy, both in conventional scenarios and particularly in the context of scene generalization benchmarks.

1. Introduction

In recent years, there has been a growing interest in the exploration of 3D object detection methods that leverage multi-camera setups, particularly in the field of autonomous driving research [10, 12, 13, 22]. This work focuses on detecting and identifying pedestrians within a specific region both in indoor and outdoor environments equipped with multiple CCTV cameras.

Conventional methods [4, 9, 14, 15] address this problem by generating predictions from single cameras and associating distinct features of individuals across camera views. However, occlusion remains a prominent challenge, leading to identity conflations. Several approaches [3, 7, 17, 19] have been proposed to address the global association problem and distinguish different identities by localizing in ground plane and projecting back onto the camera views, leveraging the known calibration data. However, these methods rely on inverse perspective mapping (IPM), which can cause the loss of information along the human body and result in the mixture of features from different individuals, as depicted in Fig. 1a. Additionally, they encounter difficulties in accurately inferring the locations of distant pedestrians, as the inverse projection can cause elongated features, affecting the retrieval of features from individuals located farther away. To overcome these limitations, we introduce the use of an efficient 3D feature-pulling mechanism, as shown in Fig. 1b, that converts multi-view features into a unified 3D space, providing a more comprehensive representation of the scene.

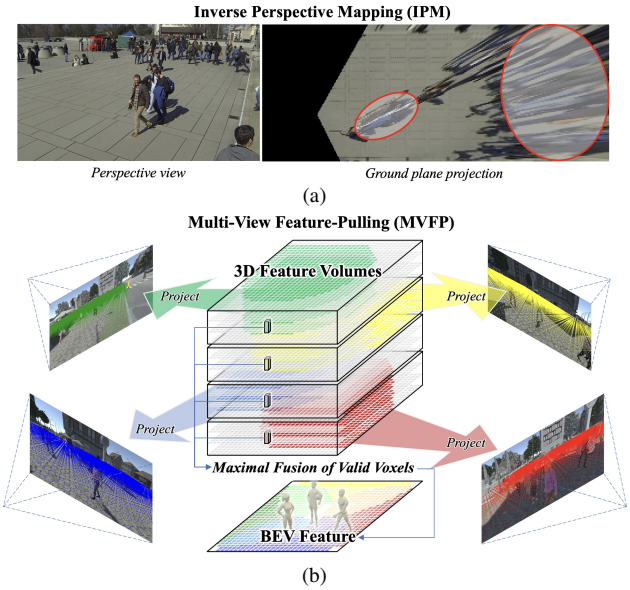

Figure 1. **Illustration of Multi-view Feature-Pulling (MVFP) method compared to Inverse Perspective Mapping (IPM)**. (a) IPM leads to the loss of features along the human body, resulting in contamination of features from other individuals. (b) MVFP uses 3D feature volumes that pull relevant 2D features from their corresponding views and subsequently aggregate them through maximal fusion, effectively addressing both occlusion and feature loss.

In addition to the challenges posed by the inverse projection mechanism, existing multi-view detection methods

have a significant limitation: overfitting to their training datasets. As highlighted in [20], the overfitting hampers the model's ability to generalize across diverse scenes and varying camera setups. A fundamental goal of this research is to create a model capable of generalizing from synthetic data to real-world scenarios. This is particularly crucial in situations where acquiring ground-truth data for real scenes is challenging, such as in densely populated indoor settings or sparsely distributed outdoor environments. To address this challenge, we meticulously analyze the components of the multi-view detection system and develop a model that demonstrates improved performance on unseen domains.

Our proposed method incorporates a "foreground selector" module to judiciously extract relevant semantic information directly from the perspective views in addition to the new feature transformation mechanism. This selective strategy fine-tunes attention to specific details of an individual's body, enhancing the performance of the 3D feature-pulling process. Additionally, we propose a "maximal fusion" module that accounts for the heterogeneous nature of CCTV cameras, preserving the most pertinent and dominant features for each individual across different camera views. To further enhance robustness against calibration errors, we invent a "large kernel refiner" module that captures intricate details and spatial relationships in the scene. With the aid of these modules, our method is able to achieve a significant improvement in performance. We summarize our key contributions as follows:

- We propose **Multi-view Feature-Pulling (MVFP)** method, which leverages efficient 3D feature-pulling mechanism, mitigating the loss of multi-view features that might occur in the process of global association.
- We introduce novel "foreground selector", "maximal fusion", and "large kernel refiner" modules specifically tailored for multi-view detection systems to complement the 3D feature-pulling mechanism, enhancing the overall system performance.
- We assess the performance of our model across various scenes, including WildTrack, MultiviewX, and GMVD datasets. We achieve state-of-the-art performance in both same-domain testing and scene generalization evaluation.

2. Related works

2.1. Multi-view pedestrian detection

MVDet [7] introduced pioneering research in multi-view detection by projecting perspective view features onto the ground plane and computing a pedestrian occupancy map through spatial aggregation. Building upon this, SHOT [19] incorporated multiple homographies to project features at various heights, thus improving performance and reducing the distortion caused by a single homography projection. MVDeTr [6] proposed a deformable attention mechanism that allows the aggregation of features from different positions and cameras to effectively handle the problem of shadow-like features. Subsequently, it incorporated view-level augmentations, comprising flipping, cropping, and scaling, to reduce overfitting and improve the diversity of the dataset. MVAug [3] further implemented scene-level augmentations, which apply geometric transformations to the projected ground plane features. Similar to random erasing [23] used in 2D object detection, 3DROM [17] introduced a cylinder-like random occlusion in the 3D space to increase the robustness of the model.

All of the above-mentioned methods are tailored to operate within the same scene with a fixed camera setup. GMVD [20] introduced a novel dataset that encompasses a wide range of scenes, each characterized by distinct camera configurations. Its model architecture is built upon MVDet [7], while introducing the use of average pooling for spatial aggregation instead of a learnable layer, enabling it to adapt effectively to different camera setups. Nevertheless, it still relies on the utilization of inverse projection, which leads to the loss of valuable information alongside human bodies and results in distorted patterns and shadow-like features, as illustrated in Fig. 1a.

2.2. 2D to 3D feature transformation

Geometry-aware Transformer-like Models. An emerging trend in multi-camera 3D object detection is the adoption of transformer models to explicitly construct a 3D feature volume using a deformable attention mechanism [10, 22]. However, our main focus is on crowded scenes where the accurate detection of pedestrians is highly important. Given the large dimensions of the ground plane and its associated resolution, the application to use attention mechanisms can lead to high computational complexities.

3D Feature Lifting. The concept of lifting 2D features into a 3D space was first introduced in autonomous driving domain [16]. Through a trainable layer, this method estimates depth distributions on a per-pixel basis along the camera rays and subsequently unprojected to their corresponding 3D locations. As reported in [5], this particular lifting strategy, despite incorporating a learnable layer to generate depth-dependent features, exhibits similar performance to the straightforward unprojection technique. Furthermore, it is worth noting that our goal lies on detecting pedestrians within a specific area of interest (AoI). Given this context, adopting a per-pixel depth-based feature estimation could potentially lead to extracting extraneous features outside the designated AoI.

3D Feature Pulling. Another promising approach for translating 2D features into a 3D context is through 3D feature pulling [5, 18, 21]. This technique involves generating 3D voxel coordinates that are projected onto the

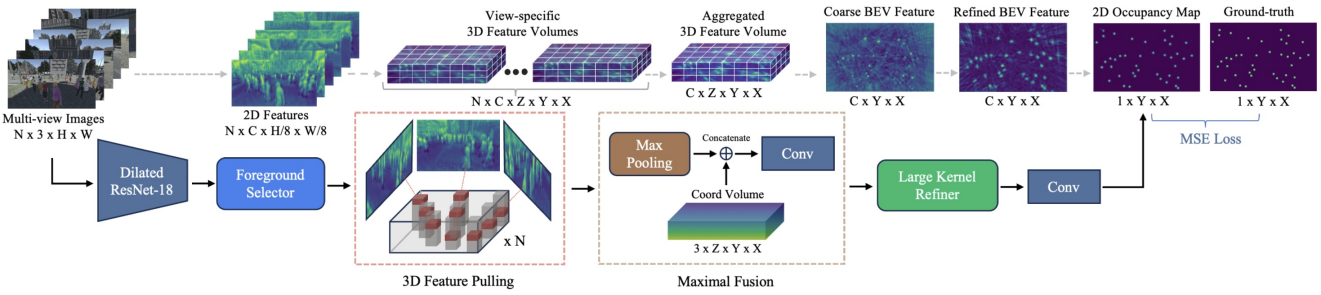

Figure 2. **Overall architecture of the proposed model**. A dilated ResNet-18, coupled with our foreground selector module, is used to extract multi-view features. In the 3D feature-pulling, a sub-pixel 2D feature is pulled for each 3D voxel using projection and bilinear sampling. A maximal fusion module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

2D space, creating a mapping between the 3D voxels and their corresponding pixel coordinates. From this, we can retrieve the sub-pixel 2D features for each 3D coordinate through bilinear sampling. The advantage of this approach is its non-parametric nature, which makes it particularly suited for multi-view detection systems due to its potential to counter overfitting issues. Another noteworthy aspect is that it selectively retrieves the relevant 2D multi-view features, specifically focusing on AoI. Despite its characteristics aligning well with the requirements of our task, this approach has remained unexplored in the field of multi-view pedestrian detection.

3. Methodology

3.1. Overall architecture

The overall architecture of the proposed model is illustrated in Fig. 2. Given a collection of images $I = \{I_i \in R^{3\times H\times W},\ i = 1, 2, \cdots, N\}$ from N views with H and W being the image height and width, the model employs a dilated ResNet-18 to extract the multi-view features $F^{2d} = \{F_i^{2d} \in R^{C\times H/8\times W/8},\ i = 1, 2, \cdots, N\}$, where C is the number of channels and 8 represents the downsampling factor.

A foreground selector module (Sec. 3.2) is utilized to extract meaningful semantic information and transfer them into a unified space with the 3D feature pulling mechanism (Sec. 3.3). This creates a 3D feature volume for each view, denoted as $F^{3d} = \{F_i^{3d} \in R^{C\times Z\times Y\times X},\ i = 1, 2, \cdots, N\}$, where X and Y represent the width and height of the ground plane and Z denotes the assumed vertical dimension.

A maximal fusion module (Sec. 3.4) is used to aggregate feature volumes from multiple views and reduce the vertical dimension to form a BEV feature map $F^{bev} \in R^{C\times Y\times X}$. The accumulated coarse BEV feature is further refined with novel large kernel refiner module (Sec. 3.5) and passed through a final layer to generate a 2D occupancy map.

A mean squared error (MSE) is used to calculate the loss between ground-truth $\mathbf{X}$ and predicted $\hat{\mathbf{X}}$ occupancy maps. To maintain the simplicity of our proposed model, we omit the use of any subsidiary heads such as a single-view detection head [7] for additional supervision in the perspective view or an offset head [6] to recover truncation error caused by downsampling of the ground plane to lower resolution.

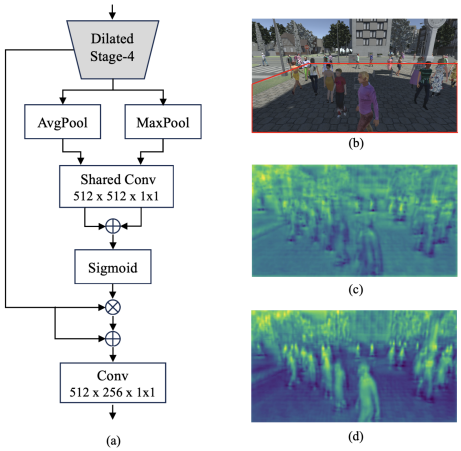

Figure 3. **Foreground selector module**. (a) Network structure of FSM. (b) Sample view. (c) Before FSM. (d) After FSM. The feature output is enhanced with the proposed module to focus on crucial foreground details, while filtering out unnecessary background information. The red area delineates the area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

3.2. Foreground selector module

Given that the 3D feature pulling mechanism is a parameter-free module, it is necessary to extract only the essential semantic information from 2D features and facilitate a more comprehensive feature selection. Hence, we propose a foreground selector module inspired by the channel-attention module in the squeeze-and-excitation network [8].

The structure of the foreground selector module is shown in Fig. 3. The module applies pooling mechanisms to create a global feature representation and extract a relation-

ship between different channels with a shared convolutional layer. A sigmoid function will compute channel-wise attention scores and element-wise multiply them with the original 2D features to create a weighted feature representation that captures foreground information. Before reducing the dimension of the feature maps, a skip connection is used to enhance the representation power of the network. By incorporating the foreground selector module into our model, we can ensure that only the relevant and important cues from 2D features are channeled to the 3D feature-pulling mechanism, leading to improved performance.

3.3. 3D feature pulling

In the 3D feature pulling mechanism, we first create a predefined 3D voxel volume $V_i(x, y, z) \in R^{Z \times Y \times X}$ using discrete grid coordinates and project it into each camera's view as follows:

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

where u and v represent the pixel coordinates and Π denotes the perspective mapping that transforms 3D points into 2D points on the image plane. K' is the scaled intrinsic matrix, considering the feature map downsampling and $[R \quad T]$ is the extrinsic matrix that describes the position and orientation of the camera in the world space. D_g represents a transformation matrix that converts the world coordinates to discrete grid coordinates, with s_g denoting the grid size and x_{min} and y_{min} denoting the lower bounds of the ground plane that define the minimum values for the x and y coordinates.

Using the mapping between 3D grid coordinates and projected 2D pixel coordinates, a binary mask M_i can be created, indicating the validity of each 3D coordinate with respect to the camera's frustum as

$$M_i(x, y, z) = \begin{cases} 1, & if \; 0 \leq u \leq \frac{W}{8} \; and \; 0 \leq v \leq \frac{H}{8} \\ 0, & otherwise \end{cases} \quad (2)$$

By applying a valid mask to the voxel volume, we can filter out the features corresponding to 3D coordinates lying beyond the camera's frustum. From which, we can pull the 2D features F_i^{2d} corresponding to each valid voxel through bilinear sampling, ultimately yielding a 3D feature volume $F_i^{3d} \in R^{C \times Z \times Y \times X}$ for each view:

$$V_i(x, y, z) = \begin{cases} sampling(F_i^{2d}(u, v)), & if \; M_i(x, y, z) = 1 \\ 0, & otherwise \end{cases} \quad (3)$$

3.4. Maximal fusion module

We propose a maximal fusion module to aggregate 3D feature volumes, addressing the challenges of achieving scene generalization with different camera setups and mitigating the computational complexity introduced by learnable layers. First, a max-pooling operation is used to extract the largest value from each voxel, thereby effectively selecting the most relevant and informative features from different camera perspectives. This emphasizes the most dominant features while omitting imperceptible or less informative ones, and allows us to handle an arbitrary number of views in random order. The aggregation process via max-pooling can be expressed as:

$$V(x, y, z) = \max_{i}^{N} V_i \quad (4)$$

Furthermore, we extend the CoordConv technique [11] from 2D pixel coordinates to 3D grid coordinates to provide 3D positional information, enhancing the model's understanding of the spatial relationships between the pulled 2D features and 3D coordinates. The introduced "Coordinate Volume" is seamlessly incorporated into the already aggregated 3D feature volume. A convolutional layer is applied to diminish the vertical dimension Z, resulting in a 2D BEV feature map, denoted as $F^{bev} \in R^{C \times Y \times X}$, tailored for localization within the ground plane.

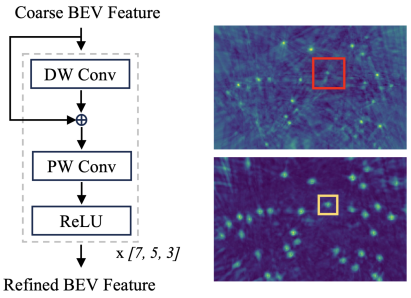

Figure 4. **Proposed large kernel refiner module**. (a) The figure illustrates a single block of the module. We apply three consecutive blocks, each with large kernel sizes of [7, 5, 3], which were chosen based on empirical results. (b) The results demonstrating the benefits of our proposed module. This 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.

3.5. Large kernel refiner module

We propose a novel "large kernel refiner" module to handle misalignments caused by imprecise calibration data and to consolidate comprehensive body information, harnessed from multiple perspective views, into a more concise cluster of features within the BEV plane. The architecture of the refiner module is depicted in Fig. 4. This module leverages large kernel-size convolutions to amass clusters of features, thereby mitigating undesired artifacts and enhancing the quality of the BEV feature representation.

However, the use of large kernel sizes can significantly increase the computational overhead. To minimize such

Figure 5. **Comparison between different datasets**. First column: samples from the training split. Second column: samples from the testing split. Top row (greyed): real view samples from WildTrack. Subsequent rows: synthetic view samples from MultiviewX and GMVD. GMVD contains a total of seven scenes with one scene being the testing scene. Only three scenes are visualized in the figure.

concerns, we adopt a depthwise separable convolution [2], greatly reducing the number of parameters required while preserving the representation capacity. In addition, we introduced a residual connection between the depthwise and pointwise convolutions, which facilitates the direct propagation of information and enables the model to learn effectively while minimizing the risk of information loss. This module serves as a key component in the context of 2D occupancy map prediction in a multi-view detection system.

4. Experiments

4.1. Experimental setup

WildTrack [1] is a real-world multi-camera dataset of images captured with seven cameras at a resolution of 1920×1080 and comprising 400 frames annotated at every two frames per second. The dataset covers a region of 12 m×36 m quantized into a 480×1440 grid using a resolution of 2.5 cm^2 with an average coverage of 3.74 cameras for each person and 20 individuals per frame.

MultiviewX [7] is a synthetic dataset that closely follows the style of WildTrack, maintaining the same resolution and frame-count but using one less camera. It accommodates 40 pedestrians per frame while having a slightly lower coverage of 16 m×25 m, which is equivalent to 640×1000 grids using the same grid cell size of 2.5 cm^2.

GMVD [20] is a large-scale synthetic dataset encompassing a distinct array of scenes captured with varying camera setups. In addition, the temporal and weather conditions are also diverse, which makes the dataset more challenging compared with a constrained environment like WildTrack or MultiviewX. Apart from the size of the ground plane and the number of cameras used, which vary for each scene, other parameters follow those of the MultiviewX dataset.

For the WildTrack and MultiviewX datasets, the first 360 frames are used for training and the remaining ones for evaluation. This means that a single scene is used for both training and evaluation, which may lead to overfitting. In contrast, the training and evaluation splits for the GMVD dataset present a different environment for each phase. For a visual representation, refer to Fig. 5, which illustrates a comparative view of the three datasets.

Evaluation metrics. Following prior methods, we use four metrics to evaluate the ground plane occupancy map: MODA (Multiple Object Detection Accuracy), MODP (Multiple Object Detection Precision), precision, and recall. MODA is computed by $1 - \frac{FP+FN}{N}$, where N being the number of ground-truth pedestrians, FP and FN being the false positives and false negatives. MODP is calculated by $\frac{\sum 1-d[d<t]/t}{TP}$, where d is the distance from a detection to its ground-truth and t is a threshold set to 0.5 m to ascertain true positives. Precision and recall can be computed by $\frac{TP}{FP+TP}$ and $\frac{TP}{N}$. MODA and recall mainly gauge correct localization accuracy while MODP assesses the localization precision of each detection. Precision evaluates the true positive rate relative to all predictions. We emphasize MODA and recall as the key indicators to judge the performance of a multi-view system since accurate identification of total pedestrians count within the scene is more important than precisely pinpointing their individual locations.

4.2. Implementation details

We employ a dilated ResNet-18 to extract multi-view features, resulting in a feature map that is downscaled by a factor of 8 compared to original image size. The channel dimension is reduced to 256 within the foreground selector module to reduce memory cost in the subsequent layers. A voxel size of 10 cm×10 cm×20 cm is used in the X, Y, and Z dimensions of the voxel volume. This translates to $Z = 8$ grids along a vertical height of 1.6 m, while also resulting in a grid size that is 4 times downsampled compared to the original ground plane resolution. The network is optimized using an Adam optimizer and a cosine annealing scheduler.

Method	WildTrack				MultiviewX			
	MODA	MODP	Prec.	Recall	MODA	MODP	Prec.	Recall
MVDet [7]	88.2	75.7	94.7	93.6	83.9	79.6	96.8	86.7
GMVD* [20]	86.7	76.2	95.1	91.4	88.2	79.9	96.8	91.2
SHOT [19]	90.2	76.5	96.1	94.0	88.3	82.0	96.6	91.5
MVDeTr† [6]	91.5	**82.1**	**97.4**	94.0	93.7	**91.3**	**99.5**	94.2
MVAug† [3]	93.2	79.8	96.3	97.0	95.3	89.7	99.4	95.9
3DROM† [17]	93.5	75.9	97.2	96.2	95.0	84.9	99.0	96.1
Ours*	**94.1**	78.8	96.4	**97.7**	**95.7**	85.1	98.4	**97.2**

Table 1. **Comparison against state-of-the-art methods**. Same-domain testing on WildTrack and MultiviewX datasets. * shows the method that works with different camera setups while other methods are configured to work on a fixed camera setup as in training. † shows the method that uses additional augmentations. Our method outperforms previous methods with larger MODA and recall scores, accurately identifying the number of individuals involved in the scene.

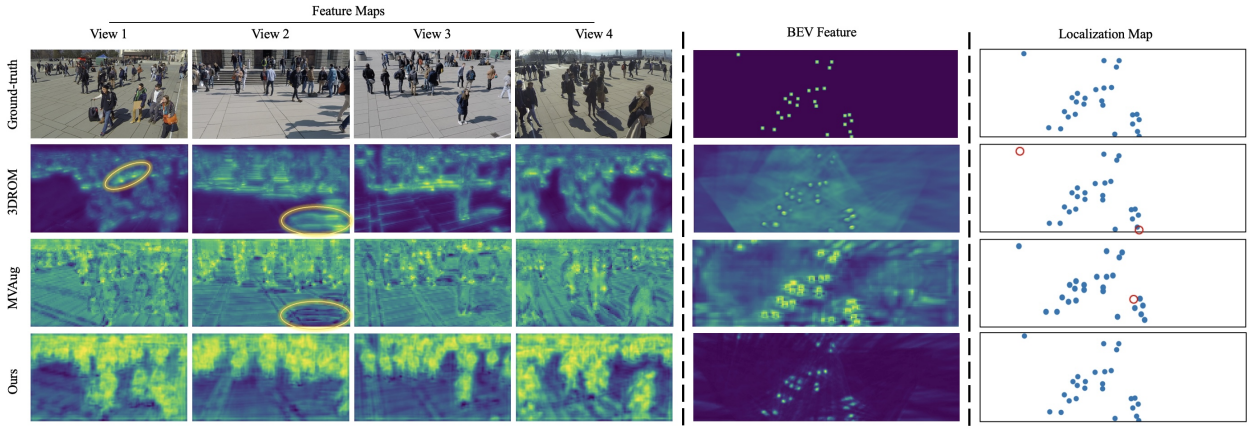

Figure 6. **Qualitative comparison between 3DROM [17], MVAug [3], and our method on the WildTrack dataset**. The initial four columns provide visual representations of example views and the corresponding extracted feature maps. The central column displays the aggregated BEV feature, while the final column illustrates ground plane localizations. Yellow ovals draw attention to the shadow of the person, while red circles emphasize instances of missed detections.

We utilized four Nvidia A100 GPUs for expedited experimentation while our model can be trained on a GPU with a memory capacity exceeding 15GB. Further specifics regarding the training hyperparameters are provided in Supp. Material. To highlight the effectiveness of our proposed model, we purposely abstained from utilizing augmentation methods such as in [3, 6, 17], though we directly compared with these methods in our experiments. The impact of the augmentations to our model can be seen in Supp. Material.

4.3. Comparison with state-of-the-arts

Quantitative evaluation. In Tab. 1, we compare our proposed model against the previous state-of-the-art methods on the WildTrack and MultiviewX datasets. Our method outperforms the published methods on both datasets, demonstrating its superiority in the same domain testing. Remarkably, despite not using any augmentation techniques, our method surpasses heavily augmented methods such as 3DROM [17] and MVAug [3], achieving higher MODA and recall scores on both datasets. On the other hand, our method underperforms on the MODP score compared to MVDeTr [6] since we did not recover the truncation error introduced by the downsampling of the ground plane with an additional offset head. Furthermore, compared to GMVD [20], which works with different camera setups, our method exhibits a significant performance advantage with increased MODA scores of 7.4% and 7.5% on WildTrack and MultiviewX, respectively. These results underline the effectiveness and robustness of our proposed approach, even when compared to methods specifically designed for fixed camera setups.

Qualitative evaluation. In Fig. 6, we perform a visual comparison between our method and the state-of-the-art 3DROM [17] and MVAug [3] models on the WildTrack dataset. Our approach stands out in capturing comprehensive and distinctive whole-body information, while 3DROM tends to extract ground plane features and shadows and MVAug faces difficulties in distinguishing between foreground and background, resulting in less precise feature extraction. As mentioned in Sec. 3.5, the refinement mechanism within our method aids in eliminating excessive artifacts and creating a more concise cluster of features for each

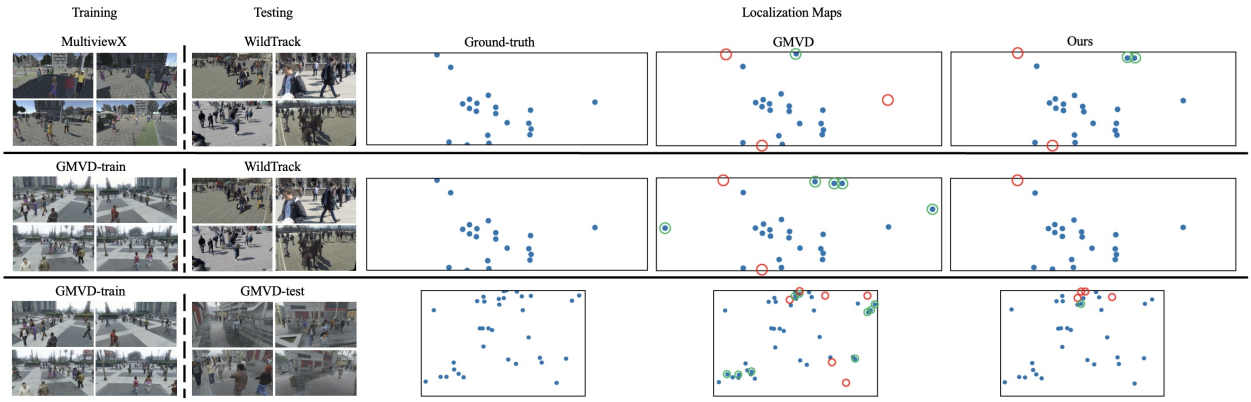

Figure 7. **Qualitative comparison between GMVD [20] and our method on scene generalization**. The first column illustrates samples from the training set, while the second column visualizes samples from the testing set. Subsequent columns depict the ground truth and predicted localization maps. Red circles denote missed detections (false negatives) while green circles denote false positives.

Method	NV_t	NV_i	MODA	MODP	Prec	Recall
MVDet [7]	6	6	17.0	65.8	60.5	48.8
MVAug [3]	6	6	26.3	58.0	71.9	50.8
MVDeTr [6]	6	6	50.2	69.1	74.0	77.3
SHOT [19]	6	6	53.6	72.0	75.2	79.8
GMVD [20]	6	6	66.1	72.2	82.0	84.7
3DROM [17]	6	6	67.5	65.6	**94.5**	71.7
Ours	6	6	**76.7**	**74.9**	85.2	**92.8**
GMVD [20]	6	7	70.7	73.8	89.1	80.6
Ours	6	7	**82.6**	**76.2**	**89.6**	**93.4**

Table 2. **Scene generalization evaluation with the MultiviewX dataset**. Trained on a synthetic dataset (MultiviewX) and tested on a real dataset (WildTrack). Camera 7 of the WildTrack dataset was discarded in the first group. NV_t and NV_i represent the number of views in training and inference, respectively.

individual, leading to more accurate ground plane localization. In contrast, 3DROM struggles to detect pedestrians near the boundaries of the ground plane, and MVAug faces challenges in distinguishing closely positioned individuals.

4.4. Scene generalization performance

Our primary goal is to apply the trained model, which has been developed using synthetic data, in real-world scenarios. Hence, we experiment a scene generalization benchmark which focuses on training with a synthetic dataset and testing on a real dataset. This evaluation scenario is crucial for determining the model's robustness and ability to handle the variations and complexities present in real environments that differ from the synthetic training data.

MultiviewX Benchmark: In this evaluation scenario, the model is trained on MultiviewX and evaluated on WildTrack. The results are presented in Tab. 2. For methods designed primarily for a fixed camera setup, we remove one extra view from WildTrack during evaluation while retaining the total number of individuals present in the scene, which is the same experimental setting as in [20]. When assessed with six camera views, our method outperforms the preceding methods, even surpassing GMVD [20] with seven cameras. Upon evaluation with seven cameras, our approach achieves a significant boost in MODA and recall scores with 82.6 and 93.4, respectively. It is noteworthy to mention that MVDet [7] and MVAug [3] suffer substantial accuracy reduction during scene generalization due to the limitations of a single-layer projection, which overly relies on the memorization of the ground plane structure from the training data, facing difficulties in new scenarios. Conversely, methods such as SHOT [19] and 3DROM [17], which incorporate multi-layer projections showcase improved generalization performance. In the supplementary material, we also demonstrate that the generalization performance can be further elevated with the adoption of augmentation methods.

GMVD Benchmark: In Tab. 3, we perform scene generalization evaluation using a large-scale GMVD dataset, which comprises a wide range of camera setups varying from three to eight views per scene. Consequently, we opt to only compare our results against the GMVD [20] model in this context since we believe that manipulating (adding or dropping) frames may not yield meaningful insights for methods designed for fixed camera setups. When trained on the GMVD train-set and subsequently evaluated on the GMVD test-set, our method exhibits remarkable performance, achieving an improvement of 5.1% in MODA score compared to the GMVD baseline. Moreover, when evaluated on the WildTrack test set, our model achieves a MODA score of 85.6 (only 2.6% lower than the MVDet [7] result from the same-domain testing in Tab. 1 and 3% higher than our result trained on the MultiviewX dataset in Tab. 2). While both GMVD and our method highlight the substantial impact of leveraging a large and diverse synthetic dataset, our method manages to surpass GMVD performance significantly, even when trained on a comparatively smaller dataset like MultiviewX.

Method	NV_t	GMVD					WildTrack				
		NV_i	MODA	MODP	Prec.	Recall	NV_i	MODA	MODP	Prec.	Recall
GMVD [20]	3,5,6,7	6,8	68.2	76.3	91.5	75.5	7	80.1	75.6	90.9	89.1
Ours	3,5,6,7	6,8	**73.3**	**76.5**	**93.0**	**79.2**	7	**85.6**	**78.0**	**91.8**	**94.0**

Table 3. **Scene generalization evaluation with the GMVD dataset**. Trained on GMVD train-set and tested on GMVD test-set and real dataset (WildTrack). NV_t and NV_i represent the number of views in training and inference, respectively.

Qualitative Comparison: We further validate the generalization performance by comparing the localization maps shown in Fig. 7. The first two rows provide visualizations of the results obtained on the WildTrack dataset, where the model was trained using the MultiviewX and GMVD datasets, respectively. The last row illustrates the outcomes when the model was trained on the GMVD train-set and evaluated on its test-set. Comparing the first and second rows, both GMVD [20] and our method show improvements when trained on a larger dataset, evident by the reduction in missed detections. In terms of false positives, GMVD appears to predict a higher number of false positive detections compared to our method; on the other hand, our model demonstrates fewer false positives and offers more accurate detection results with fewer missed detections that closely align with the ground-truth.

4.5. Ablation Study

We provide an ablation study in Tab. 4 to systematically evaluate the individual contributions of each component within our proposed model. The maximal fusion module is an inseparable part of the overall system and therefore is not dissected in this analysis. The first row represents the baseline result, achieved by setting $Z = 1$ in the 3D feature-pulling mechanism (3DFP), replacing the foreground selector module (FSM) with a 1×1 conv. layer for dimensional reduction, and swapping the large kernel refiner (LKR) module with a 3×3 conv. layer for spatial aggregation. The baseline performance slightly lags behind MVDet [7] since naive a 3×3 conv. layer for spatial aggregation, instead of 3-layer dilated convolutions in MVDet may lack the requisite learning capacity to capture the complex spatial relationships present within the aggregated BEV feature map.

Out of the three components, it is evident that 3DFP and LKR both make substantial and comparable contributions to performance enhancement, while FSM appears to have a lesser impact. When combined, the deployment of 3DFP and LKR achieves a MODA score of 93.2, outperforming the results attained by using either component in isolation. The best result is achieved by integrating FSM, which fosters a synergy amplifying the efficacy of the 3D feature-pulling mechanism. More ablation analyses are provided in Supp. Material. In summation, the incorporation of the 3D feature pulling mechanism, along with the introduction of novel modules, each plays a substantial role in elevating the overall performance of a multi-view detection system.

3D FP	FSM	LKR	MODA	MODP	Prec	Recall
			86.7	75.3	95.9	90.5
✓			92.3	78.7	**96.9**	95.3
	✓		89.7	76.1	95.5	94.1
		✓	92.2	77.5	96.8	95.4
✓		✓	93.2	78.1	**96.9**	96.2
✓	✓	✓	**94.1**	**78.8**	96.4	**97.7**

Table 4. **Ablation results on the WildTrack dataset**. *3D FP* refers to the 3D feature pulling mechanism, where Z is set to 8. *FSM* and *LKR* are the foreground selector module and large kernel refiner module, respectively.

5. Conclusion

This research focuses on a strategic approach to effectively transfer 2D features into 3D space by leveraging a parameter-free 3D feature-pulling mechanism. A detailed examination of each constituent element within the multi-view detection system led to the development of three novel modules, which have demonstrated their pivotal role in significantly enhancing the overall performance. Notably, the proposed model showcases adaptability across diverse scenes with varying camera setups.

However, there are still avenues for improvements to reduce the domain gap between the source domain (synthetic data) and the target domain (real data), stemming from factors like lighting conditions, varying appearances of pedestrians, and camera placements. Furthermore, it is important to note that while our model was initially designed to excel on small datasets, its performance on larger datasets, such as GMVD, might not be as optimal. The model's capacity to effectively capture the heightened complexity present in the larger datasets could be a potential area for enhancement.

Acknowledgements This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT)(RS-2023-00227592, Development of 3D Object Identification Technology Robust to Viewpoint Changes, 50%) and the Korea Institute of Science and Technology (KIST) Institutional Programs (Project No. 2E32301, 50%).

References

[1] Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5030–5039, 2018. 5

[2] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 5

[3] Martin Engilberge, Haixin Shi, Zhiye Wang, and Pascal Fua. Two-level data augmentation for calibrated multi-view detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 128–136, 2023. 1, 2, 6, 7

[4] Yiyang Gan, Ruize Han, Liqiang Yin, Wei Feng, and Song Wang. Self-supervised multi-view multi-human association and tracking. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 282–290, New York, NY, USA, 2021. Association for Computing Machinery. 1

[5] Adam W Harley, Zhaoyuan Fang, Jie Li, Rares Ambrus, and Katerina Fragkiadaki. Simple-bev: What really matters for multi-sensor bev perception? In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2759–2765. IEEE, 2023. 2

[6] Yunzhong Hou and Liang Zheng. Multiview detection with shadow transformer (and view-coherent data augmentation). In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 1673–1682, New York, NY, USA, 2021. Association for Computing Machinery. 2, 3, 6, 7

[7] Yunzhong Hou, Liang Zheng, and Stephen Gould. Multiview detection with feature perspective transformation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 1–18. Springer, 2020. 1, 2, 3, 5, 6, 7, 8

[8] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 3

[9] Yuntae Jeon, Dai Quoc Tran, Minsoo Park, and Seunghee Park. Leveraging future trajectory prediction for multi-camera people tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5398–5407, June 2023. 1

[10] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022. 1, 2

[11] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. *Advances in neural information processing systems*, 31, 2018. 4

[12] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, pages 531–548. Springer, 2022. 1

[13] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petrv2: A unified framework for 3d perception from multi-camera images. *arXiv preprint arXiv:2206.01256*, 2022. 1

[14] Elena Luna, Juan C. SanMiguel, José M. Martínez, and Pablo Carballeira. Graph neural networks for cross-camera data association. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(2):589–601, 2023. 1

[15] Quang Qui-Vinh Nguyen, Huy Dinh-Anh Le, Truc Thi-Thanh Chau, Duc Trung Luu, Nhat Minh Chung, and Synh Viet-Uyen Ha. Multi-camera people tracking with mixture of realistic and synthetic knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5495–5505, June 2023. 1

[16] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020. 2

[17] Rui Qiu, Ming Xu, Yuyao Yan, Jeremy S Smith, and Xi Yang. 3d random occlusion and multi-layer projection for deep multi-camera pedestrian localization. In *European Conference on Computer Vision*, pages 695–710. Springer, 2022. 1, 2, 6, 7

[18] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2397–2406, 2022. 2

[19] Liangchen Song, Jialian Wu, Ming Yang, Qian Zhang, Yuan Li, and Junsong Yuan. Stacked homography transformations for multi-view pedestrian detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6029–6037, 2021. 1, 2, 6, 7

[20] Jeet Vora, Swetanjal Dutta, Kanishk Jain, Shyamgopal Karthik, and Vineet Gandhi. Bringing generalization to deep multi-view pedestrian detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 110–119, 2023. 2, 5, 6, 7, 8

[21] Enze Xie, Zhiding Yu, Daquan Zhou, Jonah Philion, Anima Anandkumar, Sanja Fidler, Ping Luo, and Jose M Alvarez. M ^{2} bev: Multi-camera joint 3d detection and segmentation with unified birds-eye view representation. *arXiv preprint arXiv:2204.05088*, 2022. 2

[22] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, Jie Zhou, and Jifeng Dai. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17830–17839, June 2023. 1, 2

[23] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020. 2