CaMuViD: Calibration-Free Multi-View Detection

Amir Etefaghi Daryani* M. Usman Maqbool Bhutta Byron Hernandez Henry Medeiros
University of Florida
{amir.etefaghidar, m.maqboolbhutta, bhernandezosorio, hmedeiros}@ufl.edu

Abstract

Multi-view object detection in crowded environments presents significant challenges, particularly for occlusion management across multiple camera views. This paper introduces a novel approach that extends conventional multi-view detection to operate directly within each camera's image space. Our method finds objects bounding boxes for images from various perspectives without resorting to a bird's eye view (BEV) representation. Thus, our approach removes the need for camera calibration by leveraging a learnable architecture that facilitates flexible transformations and improves feature fusion across perspectives to increase detection accuracy. Our model achieves Multi-Object Detection Accuracy (MODA) scores of 95.0% and 96.5% on the Wildtrack and MultiviewX datasets, respectively, significantly advancing the state of the art in multi-view detection. Furthermore, it demonstrates robust performance even without ground truth annotations, highlighting its resilience and practicality in real-world applications. These results emphasize the effectiveness of our calibration-free, multi-view object detector.

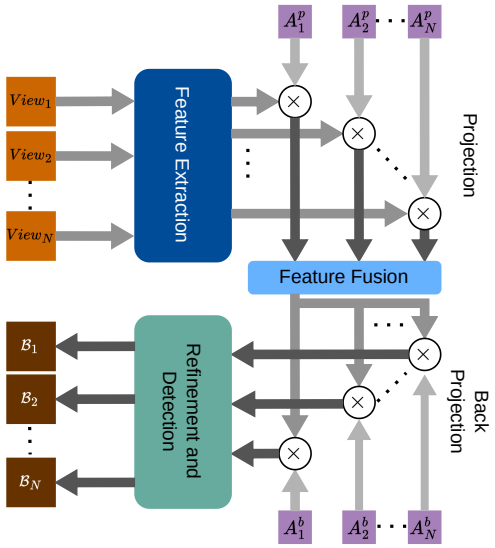


Figure 1. Overview of the Calibration-free Multi-View Detection (CaMuViD) architecture. Our method uses N camera views to extract feature maps and estimate the projection (A_i^p) / back projection matrices (A_i^b) that predict bounding boxes $\mathcal{B}_i$ for each view.

1. Introduction

Occlusions continue to represent a significant barrier to the accurate detection and tracking of objects using computer vision techniques. Various strategies have been employed to address this problem for single-view scenarios, such as part-based detection [38, 47] and detectors trained with specialized losses [42, 46]. Alternatively, numerous approaches make use of additional information from RGB-D sensors [14, 15], LIDAR point clouds [7], or multiple RGB camera views [6, 11]. In this study, we consider the problem of *multi-view pedestrian detection*, which is the recognition of pedestrians from several RGB camera perspectives.

Multi-view pedestrian detection uses synchronized frames from multiple calibrated cameras that observe a common region of interest with partially overlapping fields of view. The calibration parameters of the cameras provide the connection between 2D image coordinates and 3D world locations. The world plane, typically a BEV, encompasses points in the 3D world at a fixed height, generally $z = 0$, i.e., the ground plane. Assuming average human dimensions, bounding boxes from various image views can be projected to the real world and aggregated at each location on the ground plane. This enables multi-view pedestrian detection methods to assess ground-plane pedestrian occupancy. However, this approach faces significant challenges, primarily due to view obstructions, particularly in densely populated environments. The dynamic nature of occlusions and the lack of adequate criteria to determine optimal camera perspectives for multi-object detection further complicate matters.

Traditional methods [12, 18, 27, 28] address the challenges of multi-view pedestrian detection by making predictions from individual cameras and matching distinct features of pedestrians across different views. However, occlusions still pose significant challenges, often leading to association errors. To solve the global association problem, several

strategies [10, 17, 30, 34] have been proposed to identify individuals on the ground plane and project these identifications back to the camera views using camera calibration data. However, these methods often face difficulties in accurately determining the positions of distant pedestrians, as inverse projections can distort features, complicating the extraction of reliable features from targets farther from the camera.

Additionally, a significant limitation of most existing multi-view detection methods is their tendency to overfit, as highlighted in [39]. This overfitting severely restricts the models' ability to generalize to new environments and diverse camera configurations. A key objective of this research is to devise a model that can effectively generalize from synthetic datasets to real-world scenarios. This is especially important in environments where acquiring accurate ground-truth and camera calibration data is difficult, such as densely occupied indoor spaces or sparsely populated outdoor areas.

In response to these challenges, we introduce a novel and effective method for Calibration-Free Multi-View Detection (CaMuViD). Our approach brings a fresh perspective to the field by deviating from conventional techniques in two significant aspects. First, instead of relying on camera calibration parameters for multi-view feature projection, we use a camera calibration-free feature projection. Second, and most importantly, we present a comprehensive framework that generates bounding boxes directly in the camera views without requiring any identification on the ground plane, facilitating the detection of people far from the cameras. As Fig. 1 illustrates, at inference time, CaMuViD not only generates the detected bounding boxes of pedestrians across multiple camera views but also estimates the corresponding feature projection and inverse projection matrices across these views.

We empirically demonstrate the efficacy of CaMuViD through an extensive evaluation on two benchmark datasets specifically designed for multi-view pedestrian detection and association. On the Wildtrack dataset [6], CaMuViD substantially outperforms state-of-the-art methods, achieving a Multi-Object Detection Accuracy (MODA) of 95.0%. On the MultiviewX dataset[17], it achieves similar results, with a MODA score of 96.5%.

In this paper, we provide the following contributions:

1. A novel framework for multi-view pedestrian detection based on calibration-free feature projection across cameras observing a common region of interest.
2. A multi-view pedestrian detection model based on a fine-tuned Internimage (deformable convolution model) as a backbone [41] using a new learning technique to enable the estimation of feature projection and back projection matrices.
3. An extensive evaluation on two publicly available datasets demonstrating state-of-the-art performance according to various object detection evaluation metrics.
4. The source code is publicly available on https://github.com/amiretefaghi/CaMuViD.

2. Related Work

This section discusses the main recent contributions to monocular object detection, i.e., object detection using a single camera perspective. It then examines key techniques and methodologies employed in multi-view pedestrian detection.

2.1. Monocular detection

Monocular object detection is a fundamental task in computer vision due to its relevance in real-world applications such as autonomous driving [43], robotics [29], surveillance [33], and augmented reality [22]. This task entails identifying and localizing objects within a single image or video frame. The primary challenge in monocular object detection revolves around accurately estimating the 2D bounding boxes of multiple objects in the image while simultaneously recognizing their categories in the presence of frequent occlusions and cluttered background scenes. Over the years, numerous techniques to address these challenges have led to significant advancements in the field. These approaches can be broadly categorized into two main groups: a) two-stage detectors, and b) single-stage detectors.

Many state-of-the-art monocular object detection methods adopt a two-stage approach [48]. In the first stage, they generate a set of object proposals or candidate regions within the image. In the second stage, they refine these proposals, classify objects, and predict bounding box coordinates. Notable methods in this category include Faster R-CNN [31] and FPN [23]. In contrast, single-stage detectors simplify the detection pipeline by directly predicting object categories and bounding boxes from the image in a single pass. SSD (Single Shot MultiBox Detector) [26] and YOLO (You Only Look Once) [40] are widely used single-stage detectors for their real-time performance. Among single-stage methods, anchor-based approaches use predefined anchor boxes of various scales and aspect ratios to improve object localization [24, 26, 36, 45]. Anchor-free methods, on the other hand, do not rely on anchors, thus offering more flexibility [9, 19, 20]. DETR [4] represents a novel approach that redefines object detection as a set prediction task, introducing a complete end-to-end detection network based on a transformer architecture.

However, in real-world scenarios involving multiple camera perspectives, single-view models often struggle with occlusions and variations in object scale, limiting their scene understanding capability. Consequently, multi-view object detection methods have become increasingly important for handling these challenges and improving detection performance in complex multi-view scenarios.

2.2. Multi-view Pedestrian Detection

To address pedestrian detection under severe occlusion, many detection methods use multiple synchronized and calibrated camera views, which provide complementary perspectives of the scene. Camera calibration defines a mapping between each ground-plane location and the corresponding bounding boxes across various camera angles. This mapping enables the calculation of 2D bounding boxes using approximate human width and height in the 3D world. However, this 2D bounding box calculation assumes fixed human dimensions, which may not accurately reflect the actual height and width of pedestrians. Consequently, multi-view detection systems often assess their effectiveness using pedestrian occupancy maps on the ground plane. Integrating information from different views remains a primary challenge in multi-view detection, and a variety of methods have been proposed.

One approach to modeling the multi-view environment involves aggregating local appearance information based on spatial relationships among pixels, using methods such as mean-field inference [2, 11] and conditional random fields (CRFs) [2, 32]. Baque et al. [2] achieved state-of-the-art performance on the Wildtrack dataset [6] by constructing higher-order potentials that ensure consistency between CNN estimations and generated reference images. They further enhance the performance of their approach by training the CRF alongside the CNN in an integrated manner. Roig et al. [32] frame the problem of multi-class object detection in a multi-camera environment as an energy minimization task using CRFs. Rather than making independent predictions about object presence at specific image locations, they concurrently predict the labeling for the entire scene.

The MVDet [17] framework introduced a novel approach to multi-view detection by transforming perspective-view features onto the ground plane, where pedestrian occupancy maps were computed through spatial aggregation. SHOT [34] later improved this approach by using multiple homographies to project features at various heights, reducing distortions caused by single homography projections and enhancing detection accuracy. MVDeTr [16] further advanced these techniques by introducing a deformable attention mechanism, allowing for more effective feature aggregation across different positions and camera views, and mitigating issues like shadow artifacts. Additionally, it incorporated view-level augmentations, such as flipping, cropping, and scaling, to improve dataset variety and reduce overfitting. MVAug [10] added scene-level augmentations by applying geometric transformations to the features projected on the ground plane. Similarly, 3DROM [30], drawing from random erasing in 2D object detection, introduced random occlusion techniques in 3D space to make the model more robust. However, these methods often show poor generalization, tending to overfit to specific scenes and camera configurations.

GMVD [39] is the first attempt to improve the flexibility of multi-view detection models to different camera arrangements. It builds upon the MVDet [17] architecture but replaces the learnable layer for spatial aggregation with average pooling to more effectively adapt to different scenarios. In [39], the authors also present a new dataset that covers a wide range of scenes with varying camera configurations to evaluate their model. However, GMVD still relies on inverse projections, which can result in the loss of critical information introduce distortions and shadow-like artifacts. MVFP [1] introduces a nonparametric 3D feature pulling strategy, directly extracting the corresponding 2D features for each valid voxel within the 3D feature volume. This approach effectively addresses the feature distortion from previous methods, offering a simple but effective solution to enhance performance. To aggregate information from different views, all the previous methods used calibration matrices to project extracted features to a real-world plane and detected people from a BEV perspective. Unlike previous approaches, our method directly learns the relationships among features from different views, eliminating the dependence on camera calibration information.

3. Proposed Method

This work focuses on detecting occluded pedestrians across multiple camera views. Fig. 2 provides an overview of our proposed method. As described in detail below, our model uses fully connected networks (FCNs) to generate projection matrices based on the image feature maps obtained from the detection backbone. We employ a unique learning loss to enables our model to estimate both the projection and back projection matrices that relate the camera views.

3.1. Feature Extraction and Detection Network

As shown in Fig. 2, the input for our model is a set of images $\mathcal{I} = \{I_i\}_{i=1}^{N}$, where $I_i \in \mathbb{R}^{3\times H\times W}$ represents the i-th camera view, which has 3 channels, width W, and height H. A backbone network D then extracts the set of feature maps, $\mathcal{F} = \{F_i\}_{i=1}^{N}$, where $F_i \in \mathbb{R}^{C_f\times H_f\times W_f}$ is the feature map corresponding to the i-th image.

In our model, the backbone D is based on the InternImage [41] model pre-trained on the MS COCO [25] dataset. The most distinctive feature of InternImage is the use of deformable convolutions, which offer two key benefits. First, they provide the large effective receptive field required for tasks such as detection and segmentation. Second, they enable adaptive contextual aggregation, tailored to input and task requirements. By incorporating deformable convolutions, InternImage relaxes the strict inductive bias inherent in traditional CNNs, allowing it to learn more robust patterns from extensive data, similar to vision transformers [8]. This highlights the potential of deformable convolutions as a compelling backbone for large-scale vision models.

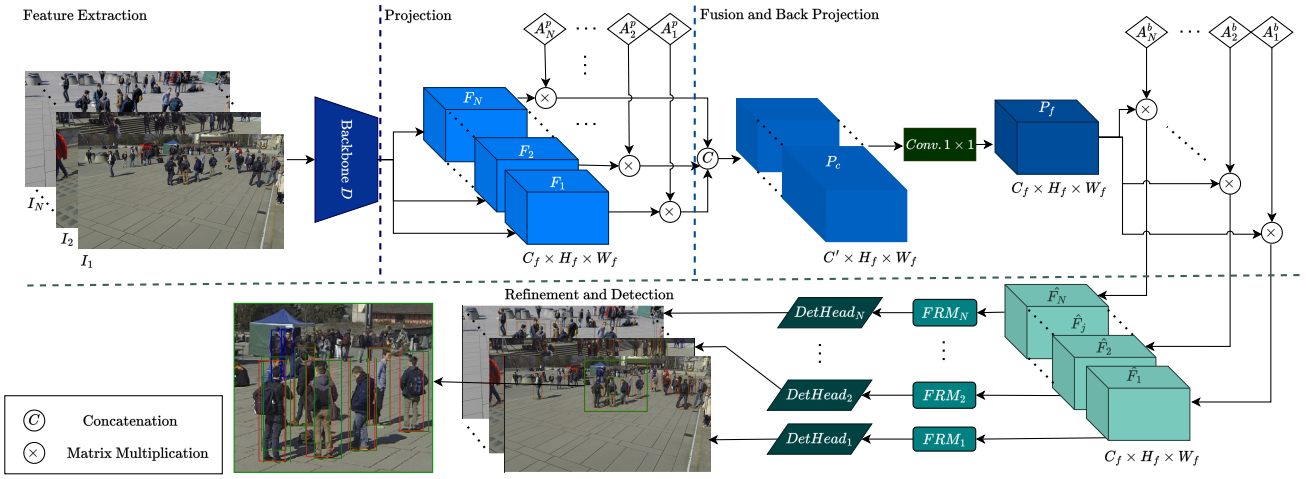


Figure 2. Approach Overview. For given input $\mathcal{I} = \{I_i\}_{i=1}^N$, our method extracts set of feature maps, $\mathcal{F} = \{F_i\}_{i=1}^N$, and uses these features to calculate projection $\mathcal{A}^p = \{A_i^p\}_{i=1}^N$ and back projection $\mathcal{A}^b = \{A_i^b\}_{i=1}^N$ matrices. With calculated projection matrices, the model is able to project the feature maps to common space for multi-view fusion and accumulate information from different views, and the back projection matrices give the model this ability to back project information $\hat{\mathcal{F}} = \{\hat{F}_i\}_{i=1}^N$ to each view and detect pedestrians with bounding boxes.

3.2. Feature Projection Matrix Estimation

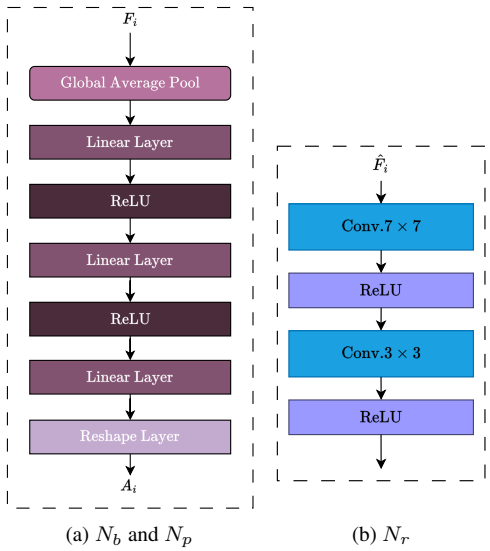
Instead of projecting the camera views to a world plane using camera matrices, we implicitly aggregate information from different views to detect people directly in the image space. As shown in Fig. 3, our method uses FCNs to estimate the set of projection matrices $\mathcal{A}^p = \{A_i^p \in \mathbb{R}^{C_f \times C_f}\}_{i=1}^N$, define the mapping from the i-th camera view to a common representation space for all the cameras. We estimate these projection matrices by feeding extracted feature maps in $\mathcal{F}$ to the fully connected network, N_p, as shown in Eq. (1).

$$A_i^p = N_p\left(F_i\right), \tag{1}$$

where $i \in 1, 2, ..., N$. The model uses A_i^p to project extracted features to the common space. This projection is performed by multiplying A_i^p and F_i, i.e.,

$$P_i = A_i^p \times F_i, \tag{2}$$

where P_i is the projected feature map corresponding to camera view i through channel-wise matrix multiplication A_i^p.



(a) N_b and N_p (b) N_r

Figure 3. FCN Architecture of a) projection and back projection matrices block, and b) the feature refinement block.

3.3. Feature Fusion and Back Projection

After projecting the features onto a common space, we fuse them to integrate information from multiple views. This fusion combines complementary data from each view, resulting in a more robust and comprehensive feature representation. Our fusion strategy concatenates the features from all views along the channel dimension, i.e.,

$$P_c = [P_1, P_2, ..., P_N], \tag{3}$$

where $P_c \in \mathbb{R}^{C' \times H_f \times W_f}$ and C' is $N \times C_f$. The main benefit of concatenation over summation is that it does not dilute information from distinct views. That is, a high-activation area summed to a low-activation area would force the two activations to the middle, which is undesirable. On the other hand, the concatenated feature map may contain redundant information. Hence, we apply a convolution layer with kernel size 1 to reduce the number of channels of the concatenated feature vector to the desired size and also act as learnable weighted summation, according to

$$P_f = N_c(P_c), \tag{4}$$

where N_c is the convolution layer. Once we have the fused feature map $P_f \in \mathbb{R}^{C_f \times H_f \times W_f}$, which contains information from all the camera views, we project this information back to each camera's image plane. To achieve this goal, we employ another FCN, N_b, to learn the back projection by generating the set of matrices $\mathcal{A}^b = \{A_i^b \in \mathbb{R}^{C_f \times C_f}\}_{i=1}^N$. We impose the condition that $A_i^b = (A_i^p)^{-1}$, which ensures that the model learns the back projection from the common representation space to each camera's image plane. That is,

$$A_i^b = N_b\left(F_i\right). \tag{5}$$

Fig. 3(a) shows the architecture of the N_c and N_b networks, which generate the feature projection and back-projection matrices, respectively.

After generating the back projection matrices, we can apply the back projection from the common space to each camera's view space by multiplying the back projection matrices and the fused feature maps,

$$\hat{F}_i = A_i^b \times P_f, \tag{6}$$

where $\hat{F}_i$ are the back-projected feature maps from the common space to camera view i, obtained through channel-wise matrix multiplication A_i^b. To satisfy the condition that $A_i^b = (A_i^p)^{-1}$, we employ the loss term presented in Eq. (7)

$$\mathcal{L}_{vp} = \sum_{i=1}^{N} |A_i^b \times P_i - F_i|. \tag{7}$$

While previous methods [21] proposed mechanisms to calculate projection matrices for image reconstruction, our method estimates both projection and back projection matrices for features (i.e., not images) from different views. Therefore, we use the multi-view loss in Eq. (7) to accurately estimate both the projection and back projection matrices.

3.4. Feature Refinement and Target Detection

With the features successfully fused and re-projected to their respective camera spaces, the model is now equipped with enriched, multi-view representations. However, to fully benefit from this process, an additional step is required to ensure the features are well-suited for accurate target detection. This is where the feature refinement module (FRM) plays a crucial role, fine-tuning the features before they proceed to the detection head. As shown in Fig. 3(b), this block contains two convolution layers followed by a ReLU activation function. We use a large kernel size for the first convolution layer to refine the re-projected feature maps based on a large area to improve feature summarization performance. After applying the FRM to the re-projected feature maps, they are used as inputs to the detection head, Eq. (8).

$$\mathcal{B} = DetHead(N_r(\hat{\mathcal{F}})), \tag{8}$$

where $\hat{\mathcal{F}}$ is set of $\{\hat{F}_i\}_{i=1}^N$, $\mathcal{B}$ is the set of predicted bounding boxes and N_r is the FRM network. Our detection head is based on the cascaded RCNN model [13], which an effective person detector in crowded scenes because of its multi-level detection capability.

3.5. Network Implementation and Training Details

We use InternImage-t [41] as the backbone of our detector network. We use the COCO pre-trained weight for the backbone and train our model independently on each dataset, namely Wildtrack [6] and MultiviewX [17] for 20 epochs using a learning rate of $1\mathrm{e}-4$. During training, we use the view projection loss from Eq. (7) in addition to standard object detection loss terms (bounding boxes and classification losses) with a weight of $1\mathrm{e}-4$. We downsize the input images to 640×1338 to limit memory consumption. As in previous works [1, 10, 16, 17, 30, 34], we used 90% of the first frames of each dataset for the training set and the remaining frames for the test set.

4. Experiments and Results

We compare the detection performance of our model with several state-of-the-art multi-view detection methods, namely, MVDet [17], Deep-Occlusion [2], DeepMCD [5], POM-CNN [11], RCNN & clustering [44], SHOT [34], MVDetr [16], MVAug [10], 3DROM [30], MVFP [1], and TrackTacular [37]. We also present an ablation analysis to elucidate the contributions of individual components of our approach. Our evaluation is based on five key object detection metrics obtained using two publicly available datasets, as explained in detail below.

4.1. Datasets

We evaluate our method on two of the most widely used publicly available multi-camera datasets for pedestrian detection: Wiltdtrack [6] and MultiviewX [17]. Both datasets provide high-resolution images (1080×1920 pixels) from multiple calibrated cameras with partially overlapping fields of view, as well as ground truth bounding boxes and corresponding global identifiers for the pedestrians visible to all the cameras. Both datasets employ a grid-based annotation system to precisely locate and track objects on the ground plane. Additionally, both datasets provide annotated pedestrian bounding boxes at two frames per second, aligning their temporal granularity.

4.1.1. Wildtrack

Wildtrack covers a scene area spanning 12×36 meters observed by seven cameras. The ground plane is quantized into a 480×1440 grid, with each grid cell measuring 2.5 centimeters square. Approximately 20 people are visible in each frame on average, and the region of interest is covered by an average of 3.74 cameras per grid cell.

Table 1. Detection performance evaluation results. Results marked with a * were reported in [17].

	Wildtrack					MultiviewX				
Method	MODA ↑	MODP ↑	Prec. ↑	Rec. ↑	F_1 ↑	MODA ↑	MODP ↑	Prec. ↑	Rec. ↑	F_1 ↑
RCNN & clustering [44]	11.3	18.4	68.0	43.0	52.7	18.7*	46.4*	63.5*	43.9*	51.9
POM-CNN [11]	23.2	30.5	75.0	55.0	63.5	-	-	-	-	
DeepMCD [5]	67.8	64.2	85.0	82.0	83.5	70.0*	73.0*	85.7*	83.3*	84.5
Deep-Occlusion [2]	74.1	53.8	95.0	80.0	86.8	75.2*	54.7*	97.8*	80.2*	88.1
MVDet [17]	88.2	75.7	94.7	93.6	94.1	83.9	79.6	96.8	86.7	91.5
SHOT [34]	90.2	76.5	96.1	94.0	95.0	88.3	82.0	96.6	91.5	94.0
MVDetr [16]	91.5	**82.1**	**97.4**	94.0	95.7	93.7	**91.3**	**99.5**	94.5	96.9
MVAug [10]	93.2	79.8	96.3	97.0	96.6	95.3	89.7	99.4	95.9	97.6
3DROM [30]	93.5	75.9	97.2	96.2	96.7	95.0	84.9	99.0	96.1	97.5
MVFP [1]	94.1	78.8	96.4	97.7	97.0	95.7	82.1	98.4	97.2	97.8
TrackTacular [37]	93.2	77.5	97.3	95.8	96.5	**96.5**	75.0	99.4	97.1	98.2
Ours	**95.0**	80.9	96.3	**98.6**	**97.4**	**96.5**	89.3	97.9	**98.6**	**98.3**

4.1.2. MultiviewX

MultiviewX is a synthetic dataset generated using the Unity engine and human models from PersonX [35]. It covers a smaller area than Wildtrack: 16×25 meters. Like Wildtrack, it also employs a 2.5 centimeter square grid resolution on the ground plane. MultiviewX features six cameras with partially overlapping fields of view, and an average of 4.41 cameras covers each grid cell in the region of interest.

4.2. Evaluation Metrics

We evaluate our method in terms of its Multi-Object Detection Accuracy (MODA) and Multi-Object Detection Precision (MODP) [3], as well as its precision, recall, and F_1 score. It is important to note that our method performs pedestrian detection solely on the image plane. Hence, unlike previous methods that detect pedestrians from a BEV perspective, our approach only produces bounding boxes. Consequently, there is a subtle but important distinction in the determination of True Positives (TP), False Positives (FP), and False Negatives (FN). In BEV, detections are mapped to a global coordinate frame, clustered, and compared with the ground truth (GT). A detection from any single camera that matches the GT is considered a TP, while redundant detections from other views are not treated as FPs. Our approach replicates this procedure without projecting bounding boxes onto the ground plane. During the evaluation, detections that meet an IoU threshold, it is 0.45 in our case, with GT bounding boxes are assigned to the GT using the Hungarian algorithm. We aggregate all detections assigned to GT boxes and unassigned detections from all views for each frame and then check whether they correspond to TPs, FPs, or FNs. If an ID is assigned to at least one detection, it is counted as a TP; otherwise, it is considered an FN. Detections without corresponding GT are projected onto the world plane

(a) Camera view 1 (b) Camera view 2 (c) Camera view 5 (d) Camera view 6

Figure 4. Qualitative detection results (green represents ground truth and red detections). The first row corresponds to the Wildtrack dataset and the second row shows frames from the MultiviewX dataset.

and clustered to combine identical detections from different views into a single point, counted as an FP.

4.3. Multi-View Object Detection Performance

As shown in Tab. 1, our method outperforms several state-of-the-art methods by a significant margin on both datasets. On Wildtrack, we obtain a 0.9% MODA improvement over the cutting-edge MVFP [1]. On MultiviewX, the improvement is 0.8%. Our method also achieves the best results in both datasets in terms of recall and F_1 score. As already mentioned, our MODP is computed based on IoU. In contrast, BEV-based approaches compute Euclidean distances in the world plane, which makes our MODP more strict. Despite this, the MODP achieved by our model is competitive under those considerations.

Fig. 4 illustrates the performance of our method on various camera views from both datasets. In addition to detecting virtually all of the pedestrians within the region of interest, our method also detects individuals who are mistakenly not represented in the ground truth annotations. People with red bounding boxes but no corresponding green bounding boxes in Fig. 4 lack ground truth annotations in several camera views, yet our method successfully detects them, as well as people with blue bounding boxes are detected, but they are out of field of view. This indicates that our method does not rely solely on ground truth annotations to learn coherent multi-view appearance features.

4.4. Cross-dataset Performance

In this evaluation, the model is trained on the MultiviewX dataset and tested on WildTrack. This procedure assesses the model's robustness and ability to handle the diverse and complex conditions found in real environments, which show noticeable differences from the synthetic training data. The results are shown in Tab. 2. For methods designed for fixed camera setups, we exclude one view from WildTrack in the evaluation while maintaining the total number of individuals in the scene, following the experimental setup from [39]. With six camera views, our method outperforms prior approaches, even exceeding MVFP [1], which uses seven cameras. In this experiment, we can observe a more notorious effect of the IoU-based MODP. The actual person width and height deviations from the fixed values used in WILDTRACK and MultiviewX for ground truth annotations also contribute to the MODP reduction. As mentioned in [1], it is important to note that methods such as MVDet [17] and MVAug [10] experience significant accuracy drops during scene generalization due to their reliance on a single-layer projection, which depends heavily on memorizing the ground plane structure from the training data, leading to challenges in new scenarios. In contrast, our method, which uses extracted feature maps for projection estimation,demonstrates better generalization.

Table 2. Cross-Dataset performance comparison of different methods.

Method	MODA	MODP	Prec	Recall	F1
MVDet [17]	17.0	65.8	60.5	48.8	54.0
MVAug [10]	26.3	58.0	71.9	50.8	59.5
MVDeTr [16]	50.2	69.1	74.0	77.3	75.6
SHOT [34]	53.6	72.0	75.2	79.8	77.4
GMVD [39]	66.1	72.2	82.0	84.7	83.3
3DROM [30]	67.5	65.6	**94.5**	71.7	81.5
MVFP [1]	76.7	**74.9**	85.2	92.8	88.8
Ours	**86.4**	60.7	89.3	**98.1**	**93.5**

4.5. Camera Elimination Performance Analysis

This experiment evaluates the impact of removing individual camera views on various performance metrics in a multi-view pedestrian detection model based on Wildtrack. By systematically blanking out different cameras, we assessed how reducing the number of active cameras influences key metrics, including MODA, MODP, Precision, Recall, and F1 score, as shown in Tab. 3. The results show that reducing the number of camera views negatively impacts all metrics, with the most significant declines observed in MODA and MODP when four or more views are removed. MODA, which reflects detection accuracy by penalizing missed detections and false positives, is notably lower when fewer than four cameras are active. Configurations with only one or two cameras achieve MODA scores 35.5% and 18.8% lower than the optimal configuration, respectively. In contrast, configurations with five or more cameras reach up to 95.6% MODA, indicating the importance of multiple perspectives for accurate detection. Similarly, MODP, which measures the precision of bounding box localization, improves with additional camera views, reaching a maximum of 80.2% when five views are active, compared to only 65.0% with a single view. This suggests that more views help the model accurately localize pedestrians, reducing the margin of error in bounding box placements. Precision and Recall also show improvements with more camera views. Precision remains high across configurations, but it is maximized at 96.6% when multiple views are used, showing that false positives are minimized with comprehensive scene coverage. Recall, however, is more sensitive to the number of cameras, rising from 60.2% with a single view to 99.3% with five active views. This underscores that additional views reduce the likelihood of missing detections, especially in occluded or crowded areas. Finally, the F1 score is highest (98.0%) when six cameras are active, reflecting the combined benefits of high accuracy and broad coverage. These findings highlight that using multiple camera views is essential in multi-view pedestrian detection to achieve robust performance across all metrics. Multi-view camera configuration mitigate occlu-

Table 3. Camera elimination model performance.

C1	C2	C3	C4	C5	C6	C7	MODA	MODP	Precision	Recall	F1
✓	✗	✗	✗	✗	✗	✗	60.1	65.0	99.8	60.2	75.1
✓	✓	✗	✗	✗	✗	✗	77.8	71.5	99.8	77.9	87.5
✓	✓	✓	✗	✗	✗	✗	90.6	76.4	98.4	92.1	95.1
✓	✓	✓	✓	✗	✗	✗	93.8	79.4	98.3	95.5	96.9
✓	✓	✓	✓	✓	✗	✗	93.8	80.0	96.8	96.9	97.0
✓	✓	✓	✓	✓	✓	✗	**95.6**	80.2	**96.6**	**99.3**	**98.0**
✓	✓	✓	✓	✓	✓	✓	95.0	**80.9**	96.3	98.6	97.4

sions, provide diverse perspectives, and enhance the model's ability to detect and localize pedestrians accurately, ensuring reliable performance in complex environments.

4.6. Ablation Study

Our ablation study evaluates the effectiveness of different fusion strategies and the Feature Refinement Module (FRM). Three sets of experiments were performed: (1) using summation instead of concatenation for feature fusion without FRM, (2) using concatenation for feature fusion without FRM, and (3) using concatenation for fusion with FRM.

As shown in Tab. 4, the summation-based fusion method without FRM achieved a MODA of 93.8% and an F1 score of 96.9%. While this approach provides reasonable results, we observed a slight improvement when switching to concatenation-based fusion, which increased the MODA to 94.6% and the F1 score to 97.3%. This improvement suggests that concatenating feature maps before detection allows for a richer feature representation compared to summation, leading to enhanced performance in detection metrics. Finally, in our last experiment, we introduced the FRM following the concatenation operation. This module further improves performance, achieving a MODA of 95.0% and an F1 score of 98.3%. The increase in both precision and recall highlights the FRM's ability to refine fused features effectively, resulting in more accurate detections. Additionally, the MODP score saw an increase from 80.2% to 80.9%, indicating better alignment and quality of the detected objects. Overall, these results demonstrate that concatenation-based fusion, combined with feature refinement through the FRM, significantly improves the detection performance compared to using summation or concatenation alone.

Fig. 5 compares projected, fused, and reprojected feature maps using two strategies. In geometry-based methods, feature maps are projected onto the world ground plane using camera parameters, fused with other views, and then reprojected back onto image space for detection. However, this process loses spatial details, particularly pedestrians' appearance information, and projective transformations distort the activation maps, causing misalignment across views and reducing their effectiveness for pedestrian detection. In contrast, our method maintains consistency across camera views

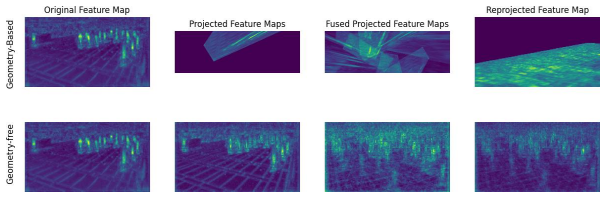


Figure 5. Comparison between geometry-based and geometry-free. Geometry-free: matrix-multiplying estimated projection ($\mathcal{A}^p$) and back-projection ($\mathcal{A}^b$) by the feature maps. Geometry-based: we project the feature maps to the world plane and backproject it by using the camera matrices.

by operating directly in image space. To further clarify, we conducted an ablation study by replacing the A^p and A^b matrices with camera parameter-based projections, leading to a significant performance drop: MODA 9.1% on Wildtrack.

5. Conclusion

We present a novel framework for multi-view detection. Our model learns feature projection and back-projection matrices for all camera perspectives, advancing our understanding of occlusion modeling and enhancing individual identification across camera views. This method establishes a new benchmark in multi-view pedestrian detection and achieves precise bounding box predictions within the image space for each distinct camera view.

Our model achieves an impressive MODA of 95.0% on the Wildtrack dataset and 96.5% on MultiviewX, marking significant progress over existing models and underscoring the robustness and efficacy of our approach. Qualitatively, our model consistently identifies individuals, even in cases where ground truth annotations are absent, illustrating its adaptability and resilience in complex multi-view scenarios.

Looking forward, we aim to extend our model to incorporate multi-view association, which would enable the tracking of individuals across multiple cameras and potentially across sequential frames. This development will enrich the model's contextual understanding from various perspectives, enhancing its capability to detect, associate, and track individuals. By integrating multi-view associations, the model can offer a more comprehensive understanding of pedestrian dynamics in dense scenes, ultimately broadening its utility in real-world applications.

Table 4. Ablation study results.

Fusion	FRM	MODA	MODP	Precision	Recall	F1
Sum	✗	93.8	80.3	95.5	98.4	96.9
Concat	✗	94.6	80.2	95.4	**99.3**	97.3
Concat	✓	**95.0**	**80.9**	**96.3**	98.6	**97.4**

Acknowledgement

This research was partially funded by the U.S. Department of Homeland Security under Grant Award 22STESE00001-04-00 and and the National Science Foundation, United States, Grant #2224591. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

References

[1] Sithu Aung, Haesol Park, Hyungjoo Jung, and Junghyun Cho. Enhancing multi-view pedestrian detection through generalized 3d feature pulling. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*, 2024. 3, 5, 6, 7

[2] Pierre Baqué, François Fleuret, and Pascal Fua. Deep occlusion reasoning for multi-camera multi-target detection. In *IEEE International Conference on Computer Vision, ICCV*, pages 271–279, 2017. 3, 5, 6

[3] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP J. on Image and Video Processing*, 2008. 6

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference Computer Vision, (ECCV)*, pages 213–229, 2020. 2

[5] Tatjana Chavdarova and François Fleuret. Deep multi-camera people detection. In *International Conference on Machine Learning and Applications (ICMLA)*, pages 848–853, 2017. 5, 6

[6] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5030–5039, 2018. 1, 2, 3, 5

[7] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3D object detection network for autonomous driving. In *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, pages 6526–6534, 2017. 1

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations, (ICLR)*, 2021. 3

[9] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *IEEE/CVF International Conference on Computer Vision, (ICCV)*, pages 6568–6577, 2019. 2

[10] Martin Engilberge, Haixin Shi, Zhiye Wang, and Pascal Fua. Two-level data augmentation for calibrated multi-view detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, 2023. 2, 3, 5, 6, 7

[11] François Fleuret, Jérôme Berclaz, Richard Lengagne, and Pascal Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30 (2):267–282, 2008. 1, 3, 5, 6

[12] Yiyang Gan, Ruize Han, Liqiang Yin, Wei Feng, and Song Wang. Self-supervised multi-view multi-human association and tracking. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, 2021. 1

[13] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, 2014. 5

[14] Saurabh Gupta, Ross B. Girshick, Pablo Andrés Arbeláez, and Jitendra Malik. Learning rich features from RGB-D images for object detection and segmentation. In *European Conference Computer Vision, (ECCV)*, pages 345–360, 2014. 1

[15] Judy Hoffman, Saurabh Gupta, Jian Leong, Sergio Guadarrama, and Trevor Darrell. Cross-modal adaptation for RGB-D detection. In *International Conference on Robotics and Automation, (ICRA)*, pages 5032–5039, 2016. 1

[16] Yunzhong Hou and Liang Zheng. Multiview detection with shadow transformer (and view-coherent data augmentation). In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, 2021. 3, 5, 6, 7

[17] Yunzhong Hou, Liang Zheng, and Stephen Gould. Multiview detection with feature perspective transformation. In *European Conference on Computer Vision (ECCV)*, pages 1–18, 2020. 2, 3, 5, 6, 7

[18] Yuntae Jeon, Dai Quoc Tran, Minsoo Park, and Seunghee Park. Leveraging future trajectory prediction for multi-camera people tracking. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023. 1

[19] Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, and Jianbo Shi. Foveabox: Beyond anchor-based object detector. *CoRR*, abs/1904.03797, 2019. 2

[20] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. *Int. J. Comput. Vis.*, 128(3):642–656, 2020. 2

[21] Bingzhen Lei, Xiuqing Li, Jun Zhang, and Junhai Wen. Calculation of projection matrix in image reconstruction based on neural network. In *2018 3rd International Conference on Computational Intelligence and Applications (ICCIA)*, pages 166–170, 2018. 5

[22] Xiang Li, Yuan Tian, Fuyao Zhang, Shuxue Quan, and Yi Xu. Object detection in the context of mobile augmented reality. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 156–163, 2020. 2

[23] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, pages 936–944, 2017. 2

[24] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):318–327, 2020. 2

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 3

[26] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. In *European Conference Computer Vision (ECCV)*, pages 21–37, 2016. 2

[27] Elena Luna, Juan C. SanMiguel, José María Martínez Sanchez, and Pablo Carballeira. Graph neural networks for cross-camera data association. *IEEE Trans. Circuits Syst. Video Technol.*, 2023. 1

[28] 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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023. 1

[29] Pekka Pääkkönen and Daniel Pakkala. Evaluation of human pose recognition and object detection technologies and architecture for situation-aware robotics applications in edge computing environment. *IEEE Access*, 2023. 2

[30] 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 *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part X*, 2022. 2, 3, 5, 6, 7

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 2

[32] Gemma Roig, Xavier Boix, Horesh Ben Shitrit, and Pascal Fua. Conditional random fields for multi-camera object detection. In *International Conference on Computer Vision, (ICCV)*, pages 563–570, 2011. 3

[33] Abubakar Siddique and Henry Medeiros. Tracking passengers and baggage items using multiple overhead cameras at security checkpoints. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2022. 2

[34] 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)*, 2021. 2, 3, 5, 6, 7

[35] Xiaoxiao Sun and Liang Zheng. Dissecting person re-identification from the viewpoint of viewpoint. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[36] Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, (CVPR)*, pages 10778–10787, 2020. 2

[37] Torben Teepe, Philipp Wolters, Johannes Gilg, Fabian Herzog, and Gerhard Rigoll. Lifting Multi-View Detection and Tracking to the Bird's Eye View. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 5, 6

[38] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning strong parts for pedestrian detection. In *International Conference on Computer Vision, (ICCV)*, pages 1904–1912, 2015. 1

[39] Jeet Vora, Swetanjal Dutta, Kanishk Jain, Shyamgopal Karthik, and Vineet Gandhi. Bringing generalization to deep multi-view pedestrian detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACV 2023 - Workshops, Waikoloa, HI, USA, January 3-7, 2023*, 2023. 2, 3, 7

[40] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7464–7475, 2023. 2

[41] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. InternImage: Exploring large-scale vision foundation models with deformable convolutions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14408–14419, 2023. 2, 3, 5

[42] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. Repulsion loss: Detecting pedestrians in a crowd. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, (CVPR)*, pages 7774–7783, 2018. 1

[43] Lingzhi Xu, Wei Yan, and Jiashu Ji. The research of a novel WOG-YOLO algorithm for autonomous driving object detection. *Scientific reports*, 13(1):3699, 2023. 2

[44] Yuanlu Xu, Xiaobai Liu, Yang Liu, and Song-Chun Zhu. Multi-view people tracking via hierarchical trajectory composition. In *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, pages 4256–4265, 2016. 5, 6

[45] Tong Yang, Xiangyu Zhang, Zeming Li, Wenqiang Zhang, and Jian Sun. Metaanchor: Learning to detect objects with customized anchors. In *Advances in Neural Information Processing Systems, (NeurIPS)*, pages 318–328, 2018. 2

[46] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. Occlusion-aware R-CNN: detecting pedestrians in a crowd. In *European Conference Computer Vision, (ECCV)*, pages 657–674, 2018. 1

[47] Chunluan Zhou and Junsong Yuan. Multi-label learning of part detectors for heavily occluded pedestrian detection. In *IEEE International Conference on Computer Vision, (ICCV)*, pages 3506–3515, 2017. 1

[48] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 2023. 2