This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Enhanced Multi-View Pedestrian Detection Using Probabilistic Occupancy Volume

Reef Alturki Adrian Hilton Jean-Yves Guillemaut
University of Surrey, UK.
{r.alturki, a.hilton, j.guillemaut}@surrey.ac.uk

Abstract

Occlusion poses a significant challenge in pedestrian detection from a single view. To address this, multi-view detection systems have been utilized to aggregate information from multiple perspectives. Recent advances in multi-view detection utilized an early-fusion strategy that strategically projects the features onto the ground plane, where detection analysis is performed. A promising approach in this context is the use of 3D feature-pulling technique, which constructs a 3D feature volume of the scene by sampling the corresponding 2D features for each voxel. However, it creates a 3D feature volume of the whole scene without considering the potential locations of pedestrians. In this paper, we introduce a novel model that efficiently leverages traditional 3D reconstruction techniques to enhance deep multi-view pedestrian detection. This is accomplished by complementing the 3D feature volume with probabilistic occupancy volume, which is constructed using the visual hull technique. The probabilistic occupancy volume focuses the model's attention on regions occupied by pedestrians and improves detection accuracy. Our model outperforms state-of-the-art models on the MultiviewX dataset, with an MODA of 97.3%, while achieving competitive performance on the Wildtrack dataset.

1. Introduction

Over the past few years, object detection has gained significant interest in the research community, as it plays an essential role in many high-level computer vision tasks. Extensive research has been conducted on object detection using single cameras. Nevertheless, object detection remains challenging in crowded, highly cluttered, and large-scale environments due to frequent occlusions that prevent objects from being fully captured. As a result, detection performance may suffer from increased missed detections.

To overcome occlusion, multi-camera setups with overlapping fields of view are often used, providing comprehensive coverage from multiple perspectives that helps mitigate occlusions, as objects occluded in one camera's view can be captured by another [31]. Camera calibrations are typically provided to enable the aggregation of data from multiple perspectives.

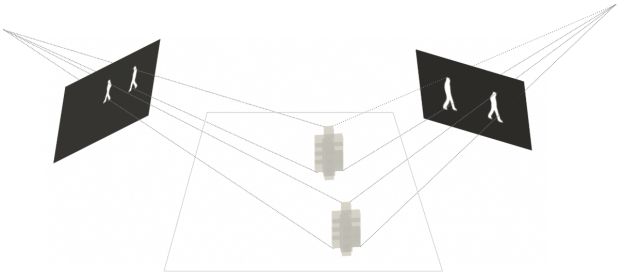


Figure 1. An example illustrating the occupancy volume of the scene, reconstructed using the visual hull technique, which highlights the voxels corresponding to the regions with high probability of being occupied by pedestrians.

Recent approaches in multi-view pedestrian detection project the feature maps extracted from all views into the Bird's Eye View (BEV), and then conduct the detection [15, 16]. However, these approaches are based on perspective projection, which results in severe distortion on the ground plane. This distortion causes the pedestrian's features to spread out from their actual locations resembling shadows, thereby impacting the detection of pedestrians located farther away. Additionally, it can lead to the loss of information along the pedestrian's body [1].

To overcome these limitations, [1, 32] used the 3D feature-pulling method [12], initially proposed as a lifting technique for autonomous vehicle perception, to project the features to a unified 3D space. This method creates a unified 3D feature volume representing the scene by sampling the features from the images for each voxel, effectively addressing the issues arising from perspective projection. When this method is applied to images captured from overlapping camera views, it performs triangulation of image features at the voxel granularity [32]. However, while it produces

a 3D feature volume of the entire scene, it does not specifically account for the potential locations of pedestrians. This results in insufficient focus on the potential locations occupied by pedestrian, which can hinder its effectiveness for pedestrian detection.

Shape-from-silhouettes technique is a specific category of object reconstruction that estimates the shape of an object by intersecting the volume regions defined by silhouettes from multiple views. Each silhouette confines the object to a specific region in 3D space forming cones. The intersection of cones estimates the volumetric description of the object. The Visual Hull (VH) of an object is defined as the largest possible volume that represents the given silhouettes of an object by intersecting these cone-shaped regions. It provides the maximum boundary of the object's shape, which becomes increasingly accurate as more silhouettes are used [18]. Probabilistic approaches [11] have also been proposed to account for uncertainty in the silhouettes.

In this paper, we propose a new model that efficiently incorporates traditional 3D reconstruction techniques with deep learning to enhance multi-view pedestrian detection. Specifically, leveraging the 3D feature-pulling method, the model creates a unified 3D feature volume representing the scene using feature maps from individual views, while incorporating a probabilistic occupancy volume constructed using the visual hull technique [18], which provides complementary information about the locations of pedestrians in the scene. An example of the visual hull of pedestrians is presented in Figure 1. To reconstruct the visual hull of pedestrians, we extract high-quality masks from individual views and employ the 3D feature-pulling method to these mask images to construct the visual hull. The visual hull provides the model with a more precise estimate of voxel occupancy by pedestrians, refining its focus on regions occupied by pedestrians, thereby limiting the area of interest to relevant spatial locations. The visual hull is multiplied by the 3D feature volume to give more weight to voxels occupied by pedestrians, the result is concatenated back with the 3D feature volume. Furthermore, it is important to note that the visual hull technique is independent of learnable parameters. Consequently, it does not introduce significant computational overhead, making it an attractive solution for real-time applications.

The main contributions of this paper are as follows:

- We propose a model that leverages a traditional 3D reconstruction technique to enhance deep multi-view pedestrian detection, a novel technique that has not been explored in existing literature.
- We propose an efficient method for aggregating the reconstructed pedestrians with the unified scene representation, focusing the model on the potential locations of pedestrians while maintaining stable feature representation.
- We evaluate the effectiveness of our model both quantitatively and qualitatively by comparing it to state-of-the-art methods. On MultiviewX dataset, our model outperforms existing methods across all metrics, achieving an MODA of 97.3%. On Wildtrack dataset, our model stands among the top performers and demonstrates competitive performance.

2. Related Work

Single-view detection The field of object detection has undergone considerable evolution over the years. With the rise of deep learning, object detection methods are typically categorized into two primary types: two-stage and one-stage detectors. Two-stage detectors, such as Fast R-CNN[10] and Faster R-CNN [26], first extract regions of interest (RoIs), which are then passed into a network for classification and regression. Mask-RCNN [14] builds upon Faster R-CNN by incorporating a mask prediction head, enabling it to simultaneously detect objects and predict their masks. On the other hand, one-stage detectors [5, 8, 19, 25] skip the region of interest extraction step, directly classifying and regressing the candidate anchor boxes. While general object detection focuses on detecting various objects, several methods were specifically developed for pedestrian detection [22, 30]. Moreover, occlusion remains a critical challenge in pedestrian detection. To tackle this challenge, part-based models have been developed to address the partial occlusion problem [23, 33]. However, the challenge of occlusion remains a significant obstacle in pedestrian detection in a single-view setup, particularly in crowded scenes where pedestrians frequently overlap.

Multi-view detection Multi-view pedestrian detection systems typically use synchronized frames from multiple calibrated cameras that cover overlapping regions, which are then aggregated to perform pedestrian detection. Early approaches [4, 28] were focused on the probabilistic modeling of objects before the advancements brough by deep learning. Several methods [2, 27] used conditional random field (CRF) for aggregating information from multiple views. In [35], the visual hull technique was employed for counting people in crowds. With the advent of deep learning, a new generation of methods emerged. MVDet [16] employs a fully convolutional approach, extracting feature maps from the input images and projecting them onto a common ground plane using perspective transformation, and then applying convolutions with a large receptive field to effectively capture neighboring locations from multiple camera views. This method led to significant improvements and established itself as the foundation for all subsequent approaches. However, its main limitation is the distortions introduced by the perspective transformation, causing pedestrian features to spread out from their locations on the ground plane, resembling shadows of the actual objects [15]. Various approaches have been proposed to ad-

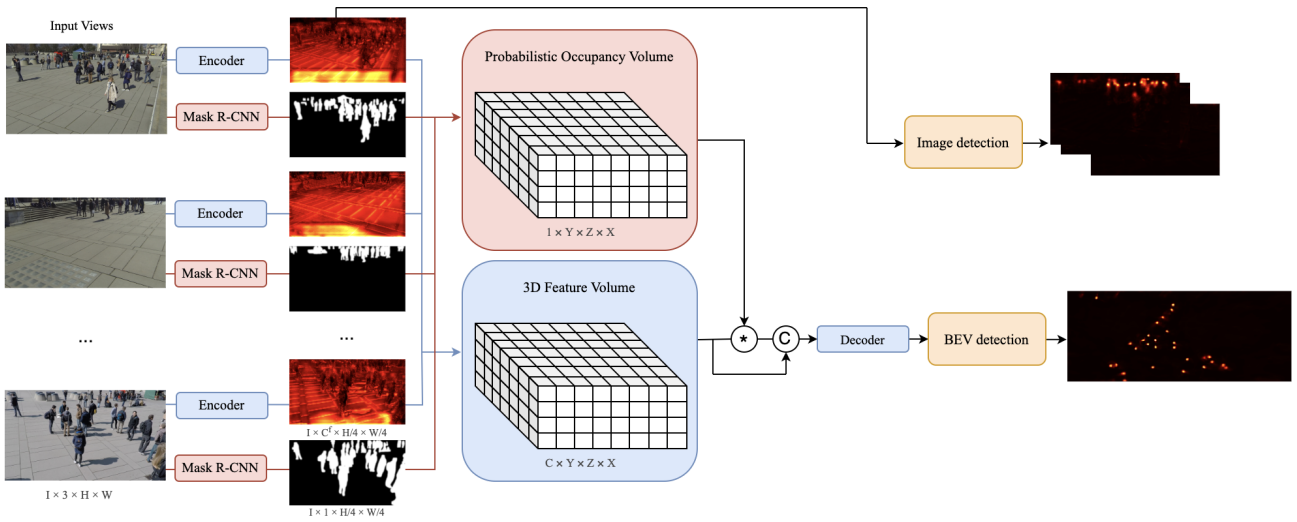


Figure 2. Overview of our model pipeline. The input views are fed into an encoder to extract feature maps, and Mask R-CNN is applied to the input views to yield silhouettes for pedestrians. The 3D feature-pulling is applied to the feature maps to create the 3D feature volume, it is also applied to Mask R-CNN to compute the probabilistic occupancy volume, which is multiplied with the 3D feature volume and the result is concatenated to the 3D feature volume. The resulting feature is compressed in the vertical dimension and fed into the decoder. C denotes concatenation while * represents element-wise multiplication.

dress this limitation [15, 20, 29]. MVDeTr [15] proposed a deformable attention mechanism to enable adaptive aggregation of features from different cameras and positions, allowing shadow-like features to be relocated to their original locations. Moreover, they introduced view-consistent data augmentation techniques, including cropping, flipping, and scaling to prevent overfitting while maintaining multi-view consistency. SHOT [29] proposed projecting the feature maps onto a stack of homographies at different heights to approximate the 3D projection, with a soft selection module predicting the likelihood score of each homography transformation for every pixel according to its semantics. MVTT [20] utilizes single-view detection results to encode the full features of each pedestrian by applying Region of Interest (RoI) pooling to the detected bounding boxes, followed by projecting the encoded features onto the predicted foot positions on the ground plane. On the other hand, 3DROM [24] proposed a data augmentation technique that involves adding additional occlusions by randomly placing 3D cylindrical objects on the ground plane, thereby reducing overfitting and making the model more robust to occlusion. MVFP [1] employs the 3D feature-pulling strategy [12] which constructs a 3D feature volume of the scene by sampling the respective 2D features for each valid voxel. This approach effectively mitigates feature loss along the pedestrian's body that occurs in methods relying on perspective projection.

Although MVFP achieves state-of-the-art results, it creates a unified 3D feature volume of the whole scene without considering the potential locations where pedestrians may be located. Moreover, none of the existing methods have integrated the visual hull into a deep learning framework for multi-view detection. In contrast to prior methods, we introduce a novel approach that leverages visual hull constraints within a deep learning framework to focus the model on potential pedestrian locations. By using the 3D feature-pulling technique for multi-view aggregation and visual hull reconstruction, our model boosts pedestrian detection accuracy.

3. Methodology

The overall structure of the proposed model is shown in Figure 2. Given a set of synchronized images from I calibrated cameras, these images are augmented and fed into the encoder to extract the feature maps. Next, the feature maps are projected to a common BEV space using the 3D feature-pulling method [12]. This creates a 3D feature volume for each view by sampling the 2D image features for each voxel, then the volumes are aggregated across the views using valid-weighted average, ultimately resulting in a 3D feature volume representing the scene. Moreover, Mask R-CNN [14] is applied to the input images to extract high quality masks of the pedestrians, which are used to construct a probabilistic occupancy volume of the pedestrians by employing the 3D feature-pulling method. The probabilistic occupancy volume is computed using a Probabilistic Visual Hull (PVH) method, which provides the probability of occupancy for each voxel and is multiplied with the 3D feature volume to emphasize the voxels occupied by pedestrians. The resulting feature is then concatenated with the original 3D feature volume to account for inaccuracies in camera

calibration and ensure a more stable feature representation. The BEV feature is then compressed in vertical dimension and fed through the decoder network.

3.1. Encoder

The encoder, based on [31], takes input images from I cameras, each having an input size of $3 \times H \times W$, and extracts the feature maps using ResNet. The ResNet consists of three blocks that progressively downsample the images by a factor of 2. The output features of each layer are upsampled and concatenated with the output features from the previous layer. These combined features are then passed through a convolutional layer, producing the final feature map with dimensions of $C^f \times H^f \times W^f$, where $H^f = H/4$, $W^f = W/4$, and C^f represents the output feature channels.

3.2. 3D Feature-Pulling

The features maps from all I cameras are projected to a common BEV space using the 3D feature-pulling method [12], which produces a 3D feature volume representing the scene. The process involves creating a 3D voxel coordinate volume $V_i(x, y, z) \in \mathbb{R}^{Y \times Z \times X}$ for each camera i. The dimensions Y and X are determined by the size of the ground plane $[H^g, W^g]$, while Z represents the presumed vertical dimension. The coordinate volume is projected onto each camera's view using the pinhole camera model [13]. The translation between 2D image pixel coordinates (u, v) and 3D locations (x, y, z) is computed with:

$$d \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = K[R|t] \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} = P^{(i)} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} \tag{1}$$

where d denotes a scaling factor accounting for the downsampling of the feature map, and $P^{(i)} = K[R|t]$ is a 3×4 transformation matrix for camera i. In particular, K is the intrinsic parameter matrix, while $[R|t]$ denotes the extrinsic parameter matrix, where R specifies the rotation and t specifies the translation.

To address the issue of 3D voxels lying outside the camera's frustum, a validity mask B_i is calculated for each voxel with respect to camera i. The mask determines if the voxel's projected 2D coordinates fall within the image bounds:

$$B_i(x, y, z) = \begin{cases} 1, & \text{if } u \in [0, W^f] \text{ and } v \in [0, H^f] \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

The validity mask is used to eliminate the features of voxels falling outside the camera's frustum. For each valid voxel, bilinear sampling is used to extract the 2D features C^f sub-pixel accurate to the voxel, resulting in a 3D feature volume $\mathcal{F}_i^f \in \mathbb{R}^{C \times Y \times Z \times X}$ for each view i:

$$\mathcal{F}_i^f = \begin{cases} \text{sampling}(C_i^f(u, v)) & \text{if } B_i(x, y, z) = 1 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

To aggregate the 3D feature volumes across the views, we apply average pooling on valid voxels across the different views. This pooling operation results in a single 3D feature volume $\mathcal{F}^f \in \mathbb{R}^{C \times Y \times Z \times X}$, which serves as a unified representation of the scene.

3.3. Probabilistic Occupancy Volume

We leverage the visual hull technique [18] to assist the model in identifying the locations of pedestrians in the scene, which is represented in the form of a probabilistic occupancy volume. To achieve this, we use Mask R-CNN [14] to obtain high-quality masks of pedestrians, which serve as silhouettes for visual hull reconstruction. To ensure proper alignment with the feature maps, the silhouette images are downsampled to match their size, with blurring applied first to reduce aliasing artifacts in downsampling. The 3D feature-pulling method is then applied to the silhouette images to compute the visual hull. Specifically, a 3D voxel coordinate volume is created for each view, and each voxel samples the corresponding 2D point from the silhouette images using bilinear sampling, resulting in a 3D occupancy volume $\mathcal{F}_i^o \in \mathbb{R}^{1 \times Y \times Z \times X}$ for each view i. Next, the visual hull is created by evaluating the occupancy of each voxel. A voxel (x, y, z) is considered part of the visual hull if the total number of non-zero values sampled across all valid views is equal to the number of views where the voxel's projection lies within the camera's frustum. This condition is illustrated as:

$$V_{\text{h}}(x, y, z) = \begin{cases} 1 & \text{if } \sum_{i=1}^{N} \mathbb{1}\left(\mathcal{F}_i^o(x, y, z) > 0\right) = N_{\text{v}}(x, y, z), \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $\mathbb{1}\left(\mathcal{F}_i^o(x, y, z) > 0\right)$ evaluates to 1 if the sampled value is non-zero. The total number of views where a voxel's projection lies within the camera's frustum is calculated using validity mask values across all views. For each voxel (x, y, z), this is given by:

$$N_{\text{v}}(x, y, z) = \textstyle\sum_{i=1}^{N} B_i(x, y, z) \tag{5}$$

Once the voxels that are part of the visual hull are identified, we use the Probabilistic Visual Hull (PVH), a variant of the traditional visual hull proposed in [9]. Instead of relying on binary values $\{0, 1\}$, PVH calculates the occupancy probability for a given voxel by multiplying the sampled values across all valid views, as follows:

$$\mathcal{F}^o = \begin{cases} \prod_{i=1}^{N_{\text{v}}(x,y,z)} \mathcal{F}_i^p(x, y, z) & \text{if } V_h(x, y, z) = 1 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

This results in a 3D volume, $\mathcal{F}^o \in \mathbb{R}^{1 \times Y \times Z \times X}$, that provides a refined estimate of voxel occupancy. We multiply the occupancy volume $\mathcal{F}^o$ by the 3D feature volume $\mathcal{F}^f$, ensuring that the features from the voxels occupied by people are preserved and given more weight. This selective

weighting enhances the model's capability to detect and precisely localize people on the ground plane. The result of the multiplication is concatenated to the original 3D feature volume. The concatenation of the occupancy volume with the original 3D feature volume aids in addressing misalignments from inaccurate calibration data, ensuring more stable feature representation. Finally, the vertical dimension Z is compressed using a convolutional layer, producing a 2D BEV feature map, $C^g \times H^g \times W^g$, optimized for localization on the ground plane.

3.4. Decoder

Prior approaches [6, 31, 32] have employed Feature Pyramid Networks on the BEV feature to introduce a large receptive field for each location, enabling effective aggregation of location and appearance features of pedestrians. Our work builds on the approaches [6, 32], which utilize a ResNet-18 decoder. In this framework, the BEV feature is downsampled by a factor of 2 in each layer of the ResNet. Each layer's output is upsampled to match the size of the previous larger output. Afterward, the two features are concatenated along the channel dimension and passed through a deformable convolution layer [37], which alleviates spatial misalignment across scales in the fusion stage, leading to more accurate detection. The decoded output from the feature pyramid maintains the same dimensions as the input, while having a larger receptive field for each grid location.

3.5. Heads and losses

To compute the Probabilistic Occupancy Map (POM) on the ground plane, we adopt the CenterNet [36] architecture. In particular, we apply a center prediction head that reduces the BEV to a size of $1 \times H^g \times W^g$, producing a heatmap representing the POM on the ground plane. We also apply an offset prediction head to produce an offset map, $2 \times H^g \times W^g$, to recover the omitted factional part during the downsampling process. Focal Loss [21] is used to optimize the center head, while L1 Loss is used to optimize the offset head. Moreover, we incorporate detection heads for the image features that estimate the 2D bounding box centers and the approximated foot position at the bottom-center of the bounding box. This enables the model to allocate more attention to areas where pedestrians are located.

4. Experiments

4.1. Datasets

Wildtrack Dataset Wildtrack [3] is a dataset captured from a real-world scene using seven static cameras, each recording video at 1080×1920 resolution and a frame rate of 2 fps. These cameras are synchronized and calibrated with a highly overlapping field-of-view, covering a region of $12\,\text{m} \times 36\,\text{m}$. This area is discretized into a 480×1440 grid, yielding grid cells of $2.5\,\text{cm} \times 2.5\,\text{cm}$. Each frame contains an average of 20 pedestrians, and each location is monitored by around 3.74 cameras. The dataset consists of 400 frames, with the first 360 frames used to train the model and the last 40 frames used for testing.

MultiviewX Dataset MultiviewX [16] is a synthetic dataset generated using a game engine, designed to have similar properties to Wildtrack. It maintains the same resolution, frame rate, and frame count as Wildtrack, but uses one fewer camera. It covers an area of $16\,\text{m} \times 25\,\text{m}$, discretized into a 640×1000 grid with the same cell size. On average, it captures 40 pedestrians per frame, with around 4.41 cameras covering each location. Similarly, it comprises 400 frames, where the first 360 frames used to train the model, and the subsequent 40 frames used to test the model.

4.2. Evaluation Metrics

Multi-view detection systems typically evaluate the POM on the ground plane using Euclidean distance to compare the predicted and ground truth detections [3]. We use four evaluation metrics to assess the performance: Multiple Object Detection Accuracy (MODA), Multiple Object Detection Precision (MODP), Precision, and Recall. MODA is computed as:

$$\text{MODA} = 1 - \frac{FP + FN}{N}, \tag{7}$$

where FP is the number of false positives, FN is the number of false negatives, and N is the number of ground-truth pedestrians. MODP is calculated as:

$$\text{MODP} = \frac{\sum(1 - d[d < t]/t)}{TP}, \tag{8}$$

where d represents the distance between a detection and its corresponding ground truth, with t set to $0.5\,\text{m}$ as a threshold to determine true positives, and TP is the number of true positives. Moreover, Precision is defined as:

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{9}$$

while the Recall is defined as:

$$\text{Recall} = \frac{TP}{N}. \tag{10}$$

As done in prior works [15, 16], we use Multiple Object Detection Accuracy (MODA) as the key performance metric, as it considers both false positives and missed detections.

4.3. Implementation Details

The input images have dimensions of 720×1280 pixels. To prevent overfitting, a series of augmentations are applied

	WildTrack				MultiviewX			
	MODA	MODP	Precision	Recall	MODA	MODP	Precision	Recall
MVDet [16]	88.2	75.7	94.7	93.6	83.9	79.6	96.8	86.7
SHOT [29]	90.2	76.5	96.1	94.0	88.3	82.0	96.6	91.5
MVDeTr [15]	91.5	82.1	97.4	94.0	93.7	91.3	**99.5**	94.2
MVAug [7]	93.2	79.8	96.3	97.0	95.3	89.7	99.4	95.9
3DROM [24]	93.5	75.9	97.2	96.2	95.0	84.9	99.0	96.1
EarlyBird [31]	91.2	81.8	94.9	96.3	94.2	90.1	98.6	95.7
TrackTacular [32]	92.1	76.2	97.0	95.1	96.5	75.0	99.4	97.1
MVTT [20]	**94.1**	81.3	**97.6**	96.5	95.0	92.8	99.4	95.6
MVFP [1]	**94.1**	78.8	96.4	**97.7**	95.7	85.1	98.4	97.2
Ours	93.6	**82.4**	96.6	97.0	**97.3**	**95.0**	**99.5**	**97.9**

Table 1. Detection performance evaluation on Wildtrack and MultiviewX datasets agaisnt state-of-the-art models. The highest and second-highest performers are denoted by bold and underlined, respectively.

Configuration	MODA	MODP	Prec.	Recall
+VH	97.1	94.8	99.4	97.7
+ PVH	**97.3**	**95.0**	**99.5**	**97.9**

Table 2. Comparison of detection performance between the visual hull (VH) and probabilistic visual hull (PVH).

during training. Specifically, each image is randomly resized within a scale range of $[0.8, 1.2]$, shifted by a random offset from the center, and cropped to preserve its original size [12, 15]. The camera intrinsics matrices are adjusted accordingly to maintain multi-view consistency. In addition, some noise is introduced to the translation vector of the extrinsic matrix to avoid overfitting the detector, following the approach in [31, 32]. The voxel volume is defined with dimensions of $10\,\text{cm} \times 10\,\text{cm} \times 20\,\text{cm}$ along the X, Y, and Z axes. This coarse voxel grid resolution enhances resilience to minor calibration errors when estimating occupancy probability. For training, we use Adam optimizer with a one-cycle learning rate scheduler, setting the peak learning rate to 1×10^{-3}. The encoder and decoder are initialized using weights pre-trained on *ImageNet-1K*.

4.4. Quantitative Results

In Table 1, the performance of our model is compared to state-of-the-art methods on Wildtrack and MultiviewX datasets. Our model consistently outperforms existing methods by a large margin across all metrics on MultiviewX dataset, achieving an increase of $+0.8$ in MODA compared to TrackTacular [32]. On Wildtrack dataset, our model ranks among the top, securing a strong second place with an MODA of 93.6% and recall of 97.0%, in addition to the highest MODP of 82.4%. Moreover, the results on MultiviewX dataset show larger improvements across all methods compared to those on Wildtrack dataset. According to [32], the results might imply that the labeling accuracy on Wildtrack could be a restrictive factor, as they annotated using perspective transform [3].

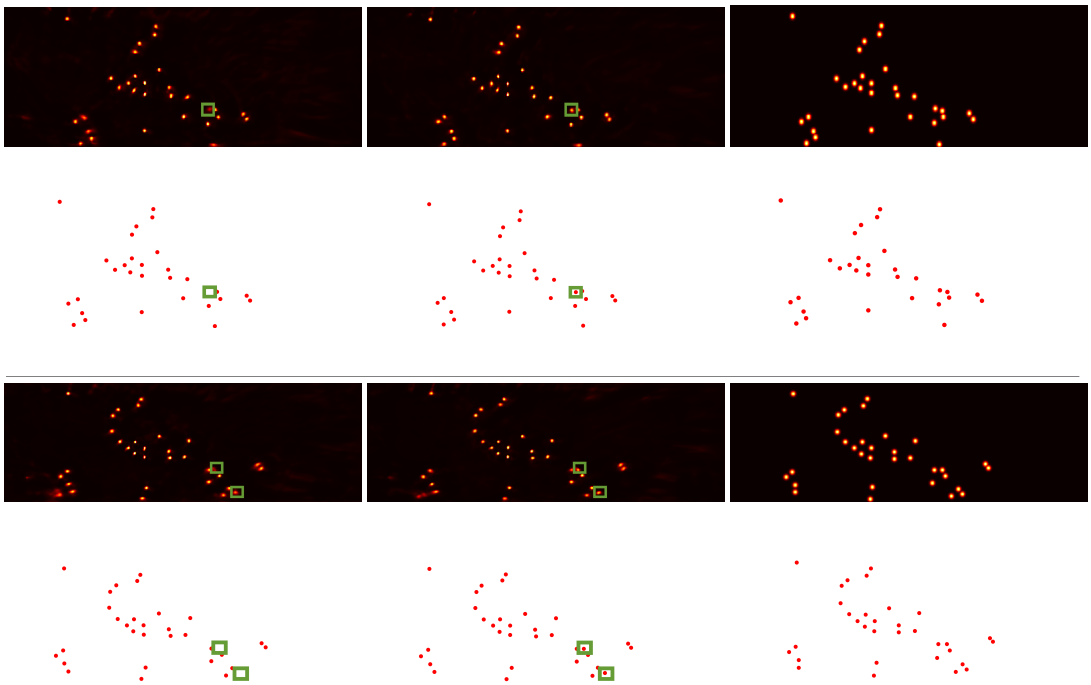
Configuration	MODA	MODP	Prec.	Recall
Basline	96.8	94.7	**99.5**	97.3
+PVH *Concat PVH*	97.1	94.5	99.4	97.7
+PVH *Mult PVH*	93.8	94.3	99.4	94.4
+PVH *Mult & Add PVH*	97.2	94.4	**99.5**	97.7
+PVH *Mult & Concat PVH*	**97.3**	**95.0**	**99.5**	**97.9**

Table 3. Comparisons of various PVH integration methods (concatenation, multiplication, and addition) with the original features, evaluating their impact on detection performance.

4.5. Ablation Studies

Influence of different visual hull techniques. We compare the performance of the standard visual hull (VH) and the Probabilistic Visual Hull (PVH) in Table 2 to compute the occupancy volume. Specifically, using the PVH increased MODA from 97.2% to 97.3%, highlighting a slight improvement in pedestrian detection accuracy. Moreover, the PVH also led to a slight enhancement in both MODP, Precision, and Recall, with overall performance being close, but slightly better than, that of the traditional VH.

Influence of different PVH integration techniques. We evaluate the impact of various PVH integration methods in Table 3. The first row presents the performance of the baseline model, which creates the 3D feature volume but does



(a) Baseline. (b) Our model. (c) Ground truth.

Figure 3. Comparison of detection performance between our model and the baseline which does not incorporate the PVH on Wildtrack dataset, showing the predicted heatmap and the corresponding localization map after applying NMS and thresholding to the predicted heatmap.

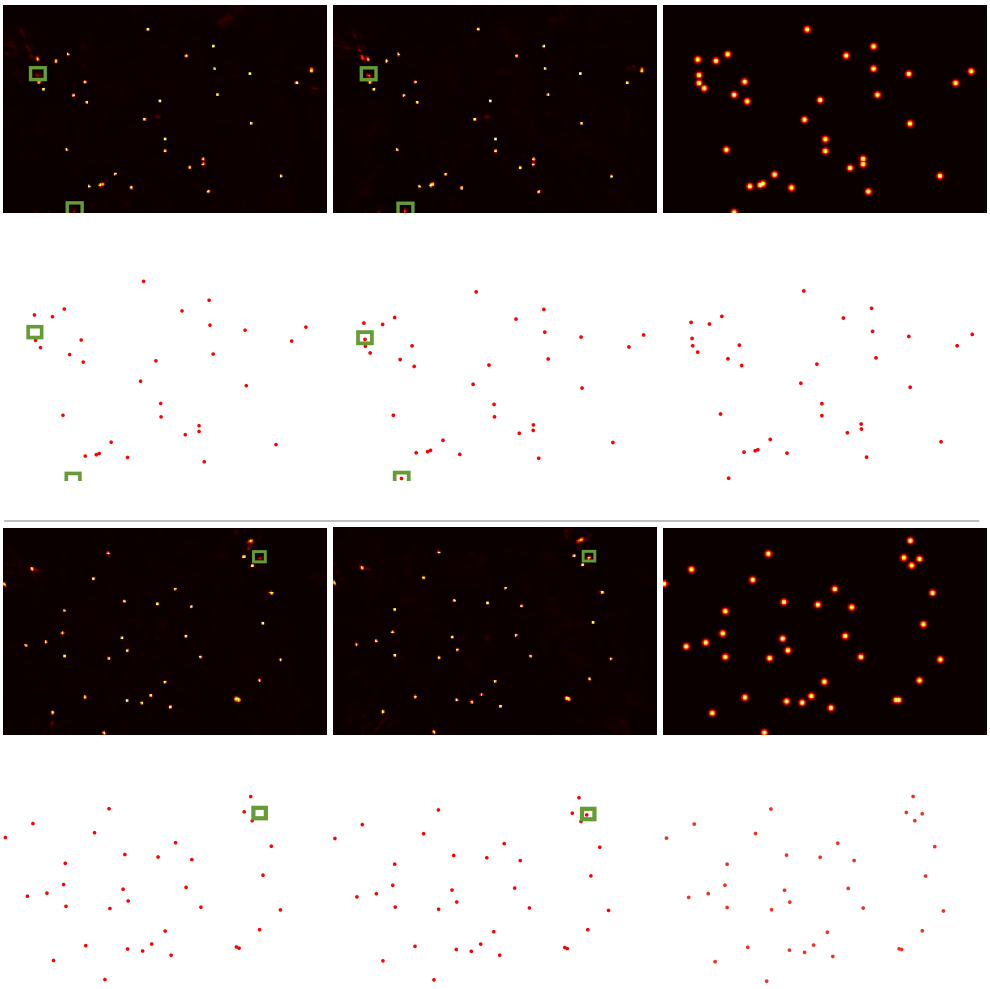
not incorporate the occupancy information. The second row shows the performance of the method where the PVH is concatenated with the 3D feature volume, leading to improved performance. The third row shows the performance of using the multiplication alone to integrate the PVH, without retaining the 3D feature volume. This results in the worst performance, with a significant drop in MODA. The fourth row illustrates the performance of adding the product of the PVH and the 3D feature volume back to the 3D feature volume, boosting the pedestrian features and yielding enhanced performance compared to the PVH concatenation alone. Notably, the best performance is achieved when the PVH is integrated through multiplication and concatenation, shown in the last row, where the PVH is first multiplied with the 3D feature volume to emphasize areas likely to be occupied by pedestrians. The resulting features are then concatenated with the 3D feature volume, ensuring reliable feature representation and providing the most effective enhancement.

4.6. Qualitative Analysis

To highlight the contribution of our proposed method, a qualitative comparison between the proposed model and the baseline is shown in Figures 3 and 4. Figure 3 presents the results using two frames from Wildtrack dataset, while Figure 4 represnt the results using two frames from MultiviewX dataset. The predicted heatmap and corresponding localization map are visualized on the ground plane. We compare the results of our model, which incorporates the PVH, with the baseline method that uses the 3D feature volume and does not incorporate the PVH. The regions highlighted in green boxes illustrate the impact of incorporating the PVH, clearly showing how it enhances both the accuracy of the heatmap and the localization map, thus leading to better detection of pedestrians. Furthermore, it is worth noting that our model produces a more concentrated and less distorted detection heatmap on Wildtrack dataset, further emphasizing its superior performance.

5. Conclusion

We introduced a novel framework for enhancing multi-view pedestrian detection using simple 3D reconstruction techniques. We leverage the visual hull technique to provide a precise estimate of voxel occupancy by pedestrians, thereby enhancing detection accuracy while maintaining computational efficiency. Comprehensive experiments on the two benchmark datasets demonstrate the effectiveness of our proposed model. Additionally, through detailed analysis, we have identified key areas for future work that could further improve detection accuracy.



(a) Baseline. (b) Our model. (c) Ground truth.

Figure 4. Comparison of detection performance between our model and the baseline which does not incorporate the PVH on MultiviewX dataset, showing the predicted heatmap and the corresponding localization map after applying NMS and thresholding to the predicted heatmap.

6. Future Work

While our model shows promising results, several areas remain for further improvement. One challenge is addressing the issue of phantom volumes during visual hull reconstruction, which represent the reconstructed 3D regions that do not correspond to pedestrians. To address this issue, a potential direction is to explore alternative reconstruction methods that are not limited to silhouette information. Meanwhile, although Mask R-CNN performs well in detecting pedestrians, its performance is limited with very small or large objects. It may also fail when object boundaries are unclear, leading to misclassification and affecting overall accuracy. One way to address this challenge is by investigating more recent segmentation models, such as SAM [17]. Moreover, as existing benchmark datasets remain relatively small, a potential future work direction is to evaluate the model on the large-scale Generalized MVD (GMVD) dataset [34]. Since the model is based on the visual hull technique, which does not depend on learnable parameters, it has strong potential for generalization across diverse scenes. It also maintains low computational complexity, though further investigation into its real-time performance, as well as the influence of the number of cameras, is essential to assess its practicality in dynamic environments.

7. Acknowledgements

This research was supported by the Saudi Arabia Cultural Bureau (SACB) in the UK and the UKRI EPSRC and BBC Prosperity Partnership AI4ME: Future Personalised Object-Based Media Experiences Delivered at Scale Anywhere EP/V038087.

References

[1] Sithu Aung, Haesol Park, Hyungjoo Jung, and Junghyun Cho. Enhancing multi-view pedestrian detection through generalized 3D feature pulling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1196–1205, 2024. 1, 3, 6

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

[3] 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, 6

[4] Adam Coates and Andrew Y Ng. Multi-camera object detection for robotics. In *2010 IEEE International conference on robotics and automation*, pages 412–419. IEEE, 2010. 2

[5] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6569–6578, 2019. 2

[6] Martin Engilberge, Weizhe Liu, and Pascal Fua. Multi-view tracking using weakly supervised human motion prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1582–1592, 2023. 5

[7] 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. 6

[8] Ali Farhadi and Joseph Redmon. Yolov3: An incremental improvement. In *Computer Vision and Pattern Recognition*, pages 1–6. Springer Berlin/Heidelberg, Germany, 2018. 2

[9] Andrew Gilbert, Matthew Trumble, Charles Malleson, Adrian Hilton, and John Collomosse. Fusing visual and inertial sensors with semantics for 3D human pose estimation. *International Journal of Computer Vision*, 127:381–397, 2019. 4

[10] Ross Girshick. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015. 2

[11] Kristen Grauman, Gregory Shakhnarovich, and Trevor Darrell. A bayesian approach to image-based visual hull reconstruction. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, pages I–I. IEEE, 2003. 2

[12] Adam W Harley, Zhaoyuan Fang, Jie Li, Rares Ambrus, and Katerina Fragkiadaki. Simple-bev: What really matters for multi-sensor bev perception? In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2759–2765. IEEE, 2023. 1, 3, 4, 6

[13] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 4

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017. 2, 3, 4

[15] 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*, pages 1673–1682, 2021. 1, 2, 3, 5, 6

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

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 8

[18] Aldo Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(2):150–162, 1994. 2, 4

[19] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018. 2

[20] Wei-Yu Lee, Ljubomir Jovanov, and Wilfried Philips. Multiview target transformation for pedestrian detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 90–99, 2023. 3, 6

[21] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2980–2988, 2017. 5

[22] Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, and Yinan Yu. High-level semantic feature detection: A new perspective for pedestrian detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5187–5196, 2019. 2

[23] Wanli Ouyang, Xingyu Zeng, and Xiaogang Wang. Partial occlusion handling in pedestrian detection with a deep model. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(11):2123–2137, 2015. 2

[24] 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. 3, 6

[25] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2016. 2

[27] Gemma Roig, Xavier Boix, Horesh Ben Shitrit, and Pascal Fua. Conditional random fields for multi-camera object de-

tection. In *2011 International Conference on Computer Vision*, pages 563–570. IEEE, 2011. 2

[28] Aswin C Sankaranarayanan, Ashok Veeraraghavan, and Rama Chellappa. Object detection, tracking and recognition for multiple smart cameras. *Proceedings of the IEEE*, 96(10):1606–1624, 2008. 2

[29] Liangchen Song, Jialian Wu, Ming Yang, Qian Zhang, Yuan Li, and Junsong Yuan. Stacked homography transformations for multi-view pedestrian detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6049–6057, 2021. 3, 6

[30] Tao Song, Leiyu Sun, Di Xie, Haiming Sun, and Shiliang Pu. Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 536–551, 2018. 2

[31] Torben Teepe, Philipp Wolters, Johannes Gilg, Fabian Herzog, and Gerhard Rigoll. Earlybird: Early-fusion for multi-view tracking in the bird's eye view. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 102–111, 2024. 1, 4, 5, 6

[32] 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*, pages 667–676, 2024. 1, 5, 6

[33] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning strong parts for pedestrian detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1904–1912, 2015. 2

[34] 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. 8

[35] Yang and Guibas. Counting people in crowds with a real-time network of simple image sensors. In *Proceedings ninth IEEE international conference on computer vision*, pages 122–129. IEEE, 2003. 2

[36] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 5

[37] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019. 5