3D Random Occlusion and Multi-Layer Projection for Deep Multi-Camera Pedestrian Localization

Rui Qiu[1,2], Ming Xu[1,2], Yuyao Yan[1], Jeremy S. Smith[2], and Xi Yang[1]

[1] School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China
{ming.xu, xi.yang01}@xjtlu.edu.cn
[2] Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, L69 3BX, UK
{rui.qiu, j.s.smith}@liverpool.ac.uk

Abstract. Although deep-learning based methods for monocular pedestrian detection have made great progress, they are still vulnerable to heavy occlusions. Using multi-view information fusion is a potential solution but has limited applications, due to the lack of annotated training samples in existing multi-view datasets, which increases the risk of overfitting. To address this problem, a data augmentation method is proposed to randomly generate 3D cylinder occlusions, on the ground plane, which are of the average size of pedestrians and projected to multiple views, to relieve the impact of overfitting in the training. Moreover, the feature map of each view is projected to multiple parallel planes at different heights, by using homographies, which allows the CNNs to fully utilize the features across the height of each pedestrian to infer the locations of pedestrians on the ground plane. The proposed 3DROM method has a greatly improved performance in comparison with the state-of-the-art deep-learning based methods for multi-view pedestrian detection. Code is available at `https://github.com/xjtlu-cvlab/3DROM`.

Keywords: Multi-view detection, Deep learning, Data augmentation, Perspective transformations

1 Introduction

Pedestrian detection plays an important role in the fields of tracking, person re-identification and crowd counting. In recent years, deep-learning based object detection methods have made significant progress in pedestrian detection. However, these deep monocular methods are not robust enough to detect heavily occluded pedestrians or localise partially occluded pedestrians on the ground. The solution to this problem lies in multi-view pedestrian detection. Compared with single-view pedestrian detection, multi-view methods can detect heavily occluded pedestrians more effectively and accurately [9].

Deep-learning based multi-camera detection methods need to be trained on sufficient annotated samples to achieve the desired performance. However, the

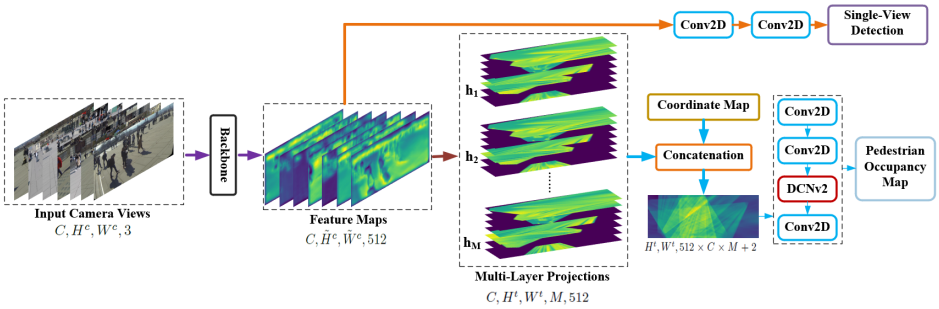

Fig. 1. The structure of 3DROM. h_1, h_2, ..., h_M represent the multi-layer projections at different heights.

limited ground truth data available in existing multi-view video datasets makes it difficult for the network to achieve the best performance in training, which limits deep learning methods from being widely used in multi-view pedestrian detection. The reason behind this is that the annotation of a multi-view pedestrian dataset is a tedious and time-consuming process. For example, with the help of an annotation tool specifically designed for multi-view datasets, it took a trained annotator an average of 10 minutes to annotate one frame with 7 views for the WILDTRACK dataset [3][6]. On the other hand, although monocular data augmentation methods, such as flipping, random cropping and Random Erasing [28], can relieve overfitting and improve the robustness of the networks to occlusion, these methods violate the homographic constraint among multiple views and cannot be used for multi-view pedestrian detection methods.

In this paper, on the basis of the MVDet framework [12], a data augmentation method is proposed to address this problem, in which occlusion boxes are randomly but consistently added to multiple camera views in the training. In this method, the ground-plane area of interest (AOI) is discretized into a grid of locations; 3D cylinders, of the average size of pedestrians, are placed at randomly selected locations on the ground plane and projected into each of the multiple camera views as filled rectangles. It reduces the risk of overfitting in the training and improves the robustness of pedestrian detection with heavy occlusions. In addition, the feature maps are projected to multiple planes parallel to the ground plane and at different heights. The multi-layer projection allows the different features (feet, torso and head) of each pedestrian to be projected to the same location in the top view but at different heights. This allows the features across the height of that pedestrian to be fully utilised in comparison with the ground-plane feature projection in MVDet. This proposed algorithm is referred to as 3DROM. A schematic diagram of the system architecture is shown in Fig. 1.

The contributions of this paper are twofold: (1) A data augmentation method is proposed for deep multi-view pedestrian detection, in which 3D random occlusions are generated and back-projected to multiple camera views. It can be

used to prevent overfitting and improve the detection performance with a limited number of multi-view training samples. To the best of our knowledge, this method is used for deep multi-view pedestrian detection for the first time. (2) A multi-layer projection method for the single-view feature maps is used to fully utilize the pedestrians' features across a range of heights. The locations of pedestrians can be inferred from the multi-height features, rather than only the ground-plane features, of the pedestrians.

2 Related Work

2.1 Multi-view Pedestrian Detection

A recent survey on multi-view pedestrian detection can be found in [18]. The state-of-the-art methods in this field can be categorised into top-down approach and bottom-up approach. The top-down approach divides the ground plane into a grid. Each location in this grid is thought of as the location of a potential pedestrian and is back-projected to individual views for finding the optimal match between foregrounds and a generative model. The bottom-up approach projects the foregrounds from the individual views to a reference view and analyses the overlaid foreground projections to determine the locations of pedestrians.

Top-Down Approach Fleuret et al. [9] estimated a probabilistic occupancy map through a generative model that represents each pedestrian as a filled rectangle of the average size of pedestrians. The occupancy probability was updated iteratively for finding the locations of the rectangles which cover more foreground pixels in all the views. On the basis of this point of view, Alahi et al. [4] formulated multi-view pedestrian detection as a linear inverse problem; Peng et al. [17] modelled pedestrians and their occlusion relationships by using a multi-view Bayesian network; Yan et al. [25] used a non-iterative logic minimization method to reduce false-positive detections. Chavdarova and Fleuret [7] proposed an end-to-end multi-view pedestrian detection network. They back-projected each ground-plane location to individual views and created a rectangle box at the corresponding positions. A CNN was used to extract features within these rectangles and infer the locations of pedestrians by using Multi-Layer Perception. Baqué et al. [5] proposed a method which combines CNNs and a Conditional Random Field. The CNN in the discriminative model extracts pedestrian features from individual views and uses Gaussian Mixture networks to classify the body parts as pedestrian features. Meanwhile, a generative model is used to model the occlusion relationships among pedestrians. The locations where the discriminative model fits the generative model well are thought of as the locations of pedestrians.

Bottom-Up Approach Khan and Shah [15][14] projected the foreground likelihood maps of individual views to a reference view using multi-plane homographies. Areas with heavily overlaid foregrounds are thought of as the locations of pedestrians. However, the foreground projections of different pedestrians may overlap, which leads to false positive detections. Eshel and Moses [8] projected the individual views to the head plane and detected pedestrians at the locations

where the intensities projected from different views are pixelwise correlated. Ge and Collins [10] modelled each pedestrian as a cylinder and used Gibbs sampling to find the locations of pedestrians. Utasi and Benedek [22] also used cylinders in the 3D space to model the foreground silhouettes, which was enhanced by pixel-level leg and head features, and determined the pedestrians' locations by using a 3D Bayesian Marked Point Process model. Xu et al. [24] detected pedestrians in individual views using Faster RCNN [19] and projected the foot points of the bounding boxes of the pedestrians to the ground plane. They clustered the projected foot points to determine the locations of pedestrians in the top view. Hou et al. proposed MVDet [12], an anchor-free end-to-end pedestrian detection network. This system uses ResNet18 [11] as the backbone to extract feature maps from individual views. The feature maps from multiple views are projected to the ground plane and concatenated there. Then a ground-plane classifier predicts the locations of pedestrians. This feature projection method is similar to that proposed by Zhang and Chan [26][27] for multi-camera crowd counting. On the basis of the MVDet framework, Song et al. [21] proposed the SHOT algorithm which projects the feature map of each individual view to multiple parallel planes. The multi-plane feature maps projected from the same view were weighted and summed into one feature map. Such feature maps from the multiple views are concatenated to predict a pedestrian occupancy map on the ground. When the multi-height feature maps were summed into a single feature map, it causes an information loss; whilst such multi-height feature maps are concatenated with no information loss in the 3DROM algorithm, which leads to an improved performance.

2.2 Data Augmentation

In deep-learning based methods, data augmentation methods are widely used to increase the number of training samples and improve the robustness by applying various transformations to existing samples [11][20][16]. One of these methods is to directly apply an image processing operation, such as flipping, folding, rotating, adding noise and Random Erasing, to existing samples. Random Erasing [28] overwrites each pixel in a randomly selected region of an image with a random colour. This method can be applied to the training of deep-learning based algorithms for image classification, person re-identification and object detection tasks. It improves the robustness of an algorithm to occlusion and reduces the risk of overfitting the samples in the training. In addition, Wang et al. [23] proposed a method for generating samples with occlusion and deformation using adversarial networks. These generated samples can improve the accuracy and robustness of Faster R-CNN in the detection of deformed or occluded objects. However, both methods are currently used in monocular detection only and cannot work well for deep end-to-end multi-view pedestrian detection without considering the geometrical relationship among multiple views.

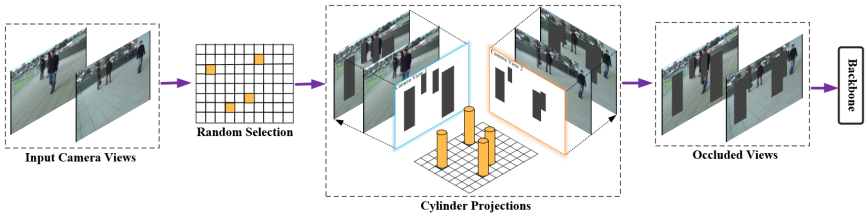

Fig. 2. A schematic diagram of the 3D Random Occlusion method.

3 Methodology

The motivation of our work is to address the performance improvement on deep multi-view pedestrian detection networks with a limited number of training samples. Robust pedestrian detection requires efficient network training with limited samples and an effective fusion method for multi-view features. We focus on reducing the risk of overfitting during the training and improving the utilization of the feature maps across multiple views to improve the detection performance in the MVDet framework.

3.1 Notations and Homography Estimation

Let C be the number of the cameras in a multi-view pedestrian dataset. The size of input image I_c, from camera view c ($c \in [1, C]$), is $H^c \times W^c$. (u^c, v^c) represents an image coordinate in view c. The size of the feature map F_c, extracted from camera view c, is $\widetilde{H}^c \times \widetilde{W}^c$. $H^t \times W^t$ is the size of the top view image. Assume the area of interest (AOI) on the ground plane is discretized into a grid of G locations. Let $\mathbf{X}_i$ be the coordinate of the i-th location ($i \in [1, G]$). Let S denote the set of the index numbers for the grid locations that have been selected to place the 3D occlusions.

Planar homography is the relationship between a pair of captured images of the same plane. Let $\mathbf{u}$ and $\mathbf{X}$ be the homogeneous image coordinates of the same point on a plane in camera view c and the top view. They are associated by the homography matrix $\mathbf{H}^{c,t}$ for that plane as follow:

$$\mathbf{X} \cong \mathbf{H}^{c,t}\mathbf{u}. \tag{1}$$

A 3×4 projection matrix can be calculated by using the intrinsic and extrinsic parameters of camera c: $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_3, \mathbf{m}_4]$. The homography matrix, from the top view t to camera view c, for the ground plane is:

$$\mathbf{H}_0^{t,c} = (\mathbf{H}_0^{c,t})^{-1} = [\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_4]. \tag{2}$$

The homography matrix, from the top view t to camera view c, for the plane parallel to the ground plane and at a height of h can be written as:

$$\mathbf{H}_h^{t,c} = [\mathbf{m}_1, \mathbf{m}_2, h\mathbf{m}_3 + \mathbf{m}_4] = \mathbf{H}_0^{t,c} + [\mathbf{0} \mid h\mathbf{m}_3], \tag{3}$$

where $[\mathbf{0}]$ is a 3×2 zero matrix.

3.2 3D Random Occlusion

Compared with single-view detection, multi-view pedestrian detection requires the use of geometric constraints to establish the correspondence among multiple views. Monocular data augmentation methods, such as flipping, cropping, rotation and Random Erasing, may affect the performance of multi-view detection algorithms, since they violate the homography constraint. Therefore, Algorithm 1 was developed as a 3D data augmentation method for the training of multi-view pedestrian detection algorithms.

The 3D Random Occlusion algorithm is based on the input camera views in the training, as shown in Algorithm 1. The process of 3D Random Occlusion is illustrated in Fig. 2. The ground plane is discretized into a grid of locations. The i-th location ($i \in [1, G]$) in the top view is associated with its corresponding location (u_i^c, v_i^c) in camera view c ($c \in [1, C]$) through the ground-plane homography $\mathbf{H}_0^{t,c}$. A 3D cylinder placed at the ith location on the ground plane is back-projected to a filled rectangle r_i^c sitting at location (u_i^c, v_i^c) in camera view c. The rectangle is designed to have the average height H_i^c and width W_i^c of the pedestrians standing at the ith location. H_i^c is calculated as follows: the i-th location in the top view is projected back to camera view c using the homographies, $\mathbf{H}_0^{t,c}$ and $\mathbf{H}_{h_a}^{t,c}$, for the planes at the heights of 0 cm and 180 cm; The vertical distance between the two projected points in view c is H_i^c; the average width $W_i^c = \alpha H_i^c$, where α is a constant ratio.

The inputs of Algorithm 1 are the images $I = \{I_1, I_2, \cdots, I_c\}$ from multiple camera views, the number of occlusions n per frame and the occlusion probability p of each frame to be selected to add 3D random occlusions. The n locations in the top view are selected to generate filled rectangles at the corresponding locations in all the views. To ensure that the occlusions are not too close to each other, the ground distance between each selected location and other cylinder occlusions must be greater than a threshold $d = 1$ meter. The selected locations are projected to all the views, by using homographies $\mathbf{H}_0^{t,c}$ and $\mathbf{H}_{h_a}^{t,c}$, to generate the filled rectangles with a constant pixel value Ω.

3.3 The Multi-Layer Projection of Feature Maps

Within the MVDet framework, the feature map F_c of view c is projected to the ground plane by using a homography transformation. The feature map in the output of the backbone network does not have the same size as the input image I_c of view c. However, it is resized to the same size afterwards. Therefore, the projected feature map $F_h^{c,t}$ from view c to the top view can be written as:

$$F_h^{c,t} = \mathbf{H}_h^{c,t}(F_c), \tag{4}$$

for a plane parallel to the ground and at a height of h.

The feature map on the ground plane is compromised when pedestrians' feet are occluded or their feet are off the ground. This may affect the model to infer the locations of the pedestrians. In [14], foreground likelihood maps are projected

to multiple planes parallel to the ground plane and at different heights, which can significantly reduce detection errors. The foreground likelihood map, which indicates how likely a pixel in an image belongs to foregrounds, is similar to the feature map in MVDet. We assume that each pedestrian occupies a specific location in the top view. The multi-layer projection of the feature maps of multiple views can provide the comprehensive feature information for any pedestrian standing at that location. Compared with the ground-plane projection used in MVDet, the top view CNN is able to infer the locations of pedestrians from a wider range of features.

Algorithm 1: 3D Random Occlusion at one frame

Input : Input image $I = \{I_1, I_2, \cdots, I_c\}$;
The number of occlusions n per frame;
Occlusion probability p;
Output: Occluded image $I^* = \{I_1^*, I_2^*, \cdots, I_c^*\}$.

$S = \phi$;
$I^* = I$;
$p_1 = \text{Rand}(0, 1)$;
if $p_1 > p$ **then**
  return I^*.
else
  $i = 0$;
  while $i < n$ **do**
    $k = \text{Rand}(1, G)$;
    if $\forall l \in S, \|\boldsymbol{X}_k - \boldsymbol{X}_l\|_2 < d$ **then**
      goto 9;
    else
      for *camera view* $c = 1$ *to* C **do**
        $(\mathbf{H}_0^{t,c}\mathbf{X}_k, \mathbf{H}_{h_a}^{t,c}\mathbf{X}_k) \Rightarrow (u_k^c, v_k^c, H_k^c, W_k^c)$;
        for $u = u_k^c - W_k^c/2$ *to* $u_k^c + W_k^c/2$ **do**
          for $v = v_k^c$ *to* $v_k^c + H_k^c$ **do**
            $I_c^*(u, v) = \Omega$;
      $S = S \cup \{k\}$;
      $i = i + 1$
  return I^*;

The multi-layer feature projection is illustrated in Fig. 3(a) and (b). The projected features (or silhouette) of a pedestrian is like the shadow of that pedestrian. When the features of a pedestrian are projected from multiple views to a specific plane, they intersect at the body parts of that pedestrian at the height of that plane. By using multi-plane feature projection, the features across the

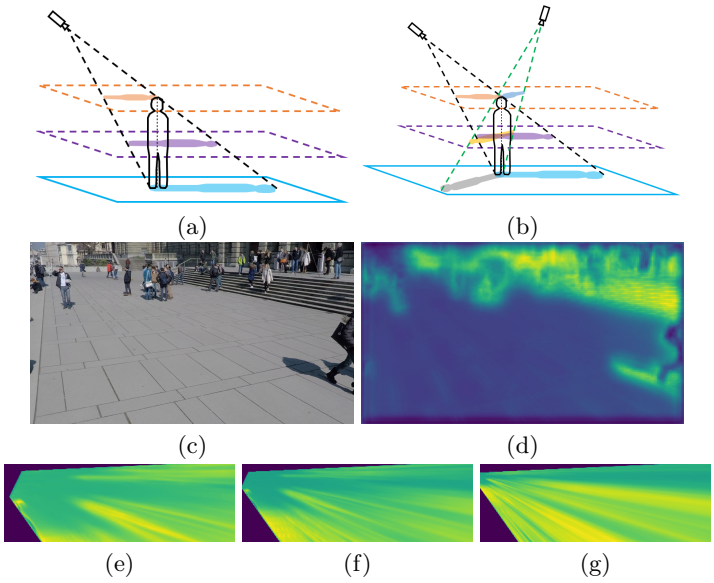

Fig. 3. The multi-layer projections: (a) from a single camera view, (b) from two camera views, and an example of the WILDTRACK dataset, where (c) is the original image of a camera view, (d) is its feature map, and (e)-(g) are the projected feature maps at the heights of 0cm, 90cm and 180cm, respectively.

height of each pedestrian can be utilized. An example of the multi-layer feature projection is shown in Fig. 3(c)-(g).

The projected feature maps and the ground-plane coordinate map are concatenated for the inference of pedestrian locations. The concatenated feature maps are denoted as:

$$F = \{F_h^{c,t}, c \in [1, C], h \in \{h_1, h_2, \ldots, h_M\}\}. \tag{5}$$

where M is the number of the parallel planes used for feature map projection.

Since the feature maps are projected to the top view with geometric deformation, a layer of DCNv2 [29] is added to the top view CNN to handle the geometric deformation in the projected feature maps. The DCNv2 layer is a complementary component used with the multi-layer projection in 3DROM.

3.4 Loss Function

The loss function is the same with that of the MVDet [12]. The network output is an occupancy probability map $\tilde{g}$. A Gaussian kernel $f(\cdot)$ is used to blur the ground-truth pedestrian occupancy map g. The loss of the top view L_t is the Euclidean distance between them:

$$L_t = \|\tilde{g} - f(g)\|_2. \tag{6}$$

The loss function of the single view detection in camera view c is:

$$L^c_{single} = \|\tilde{s}^c_{head} - f(s^c_{head})\|_2 + \|\tilde{s}^c_{foot} - f(s^c_{foot})\|_2, \tag{7}$$

where $\tilde{s}^c_{head}$ and $\tilde{s}^c_{foot}$ are the single-view likelihood maps for heads and feet, respectively; s^c_{head} and s^c_{foot} are the ground-truth location maps for heads and feet, respectively.

The overall loss for training 3DROM combines the single view loss L_{single} and the top view loss L_t. It can be written as:

$$L_{overall} = L_t + \frac{1}{C}\sum_{c=1}^{C} L^c_{single}. \tag{8}$$

4 Experimental Results

4.1 Experiment Setup

The proposed method has been evaluated on the EPFL WILDTRACK [3][6], MultiviewX [1][12] and EPFL Terrace datasets [2]. These three public video datasets have been widely used to evaluate multi-view pedestrian detection algorithms. Tab. 1 shows the detailed information of these datasets.

Table 1. Datasets used for performance evaluation.

Dataset	Input Resolusion	Feature Resolusion	Training Frames	Testing Frames	AOI ($m \times m$)	Top View Grid Size	Number of 3D Occlusions
WILDTRACK	1920×1080	270×480	360	40	12×36	120×360	25
MultiviewX	1920×1080	270×480	360	40	16×25	160×250	25
Terrace	360×288	360×288	300	200	5.3×5	220×150	20

The proposed 3DROM method is based on the MVDet framework. Therefore, most of the network parameters were set to the same values as those in MVDet. ResNet-18 was used as the backbone network without using a pre-trained model. The kernel of DCNv2 used in location regression in the top view was set to a size of 2×2. The setup of the input image size, the feature map size, the top view grid size and the number of 3D random occlusions for each dataset are shown in Tab. 1. The 3D random occlusions were added to each frame before the images were input to the backbone in the training.

For the training and testing on all the three datasets, the number of projection layers was set to $M = 5$. The feature maps were projected to 5 parallel planes at the heights of 0 cm, 15 cm, 30 cm, 60 cm and 90 cm, respectively. The batch size was set to 1. The occlusion probability p was set to 100%. All experiments were carried out using one RTX-3090 GPU.

4.2 Qualitative Performance Evaluation

The performance of 3DROM on three datasets is demonstrated in the qualitative evaluation. Fig. 4 shows the detection results at frame 3225 of the EPFL Terrace dataset with four camera views. The red rectangle on the ground shows the AOI region. The pedestrians outside the AOI were ignored in the detection and evaluation. The camera positions labelled in the top view are approximate ones. The colour points in the top view represent the detected pedestrians. Meanwhile, the colour of each point in the top view is consistent with the colour of the bounding boxes of the same pedestrian in all the camera views. As can be seen in Fig. 4, the pedestrian in the pink bounding box is completely occluded in C0, partially occluded in C1 and C2, and out of the field of view in C3. The 3DROM method can still infer the location of this pedestrian using limited pedestrian features, which demonstrates its strong detection capability in heavy occlusion.

Fig. 4. The detection results at frame 3225 of the EPFL Terrace dataset: from left to right, camera views C0, C1, C2, C3 and the top view. Each detected pedestrian is represented by a distinguished colour consistent across different views. The red rectangle on the ground is the AOI. The field of view of each camera is shown in the top view.

Fig. 5 shows the detection results at frame 1960 of the EPFL WILDTRACK dataset with seven camera views. The pedestrians stand in a group at the centre of the square and are occluded by each other. The 3DROM algorithm combines the feature information in the multi-view and multi-layer feature projections. These pedestrians are detected correctly by 3DROM.

Fig. 6 shows the detection results at frame 399 of the MultiviewX dataset with six camera views. A large number of pedestrians are standing very close to the border of the AOI in the top view with limited feature information. The use of multi-layer feature projection and 3D Random Occlusion in the training allows the 3DROM algorithm to detect such pedestrians accurately.

4.3 Quantitative Evaluation

The proposed method was evaluated using performance metrics Multiple Object Detection Accuracy (MODA) [13], Multiple Object Detection Precision (MODP) [13], Precision (Prec.) and Recall, which are widely used for multi-view pedestrian detection. The Hungarian algorithm was used to match the detected pedestrians and ground-truth pedestrians. A distance threshold $r = 0.5$m on the ground plane was used in the matching.

Fig. 5. The detection results at frame 1960 of the EPFL WILDTRACK dataset: (top row) from left to right, camera views C1, C2, C3 and C4; (bottom row) camera views C5, C6, C7 and the top view.

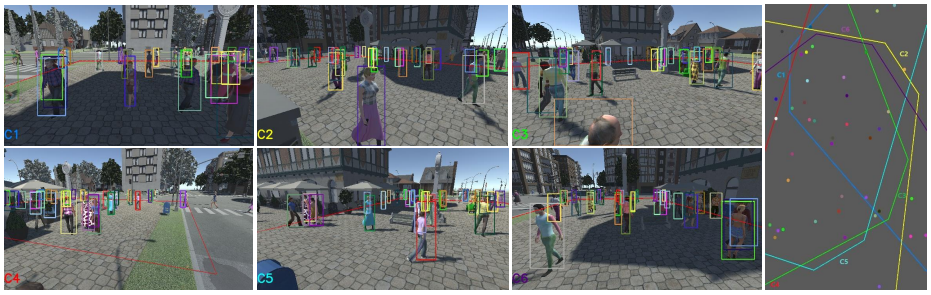

Fig. 6. The detection results at frame 399 on the MultiviewX dataset: (top row) from left to right, camera views C1, C2 and C3; (bottom row) camera views C4, C5, C6 and the top view.

The proposed method was compared with several state-of-the-art deep-learning based methods such as RCNN-2D/3D [24], POM-CNN [9], DeepMCD [7], Deep Occlusion [5], MVDet [12] and SHOT [21], as shown in Tab. 2 in which "Eval.' indicates who made the evaluation. The MODA of the 3DROM method is increased to 93.5%, 95.0%, and 94.8% in the evaluation on the WILDTRACK, MultiviewX, and Terrace datasets, respectively. Compared with the baseline algorithm MVDet that uses single-layer projections, the 3DROM increases the MODA by 5.3%, 11.1%, and 7.6%, respectively. Compared with the algorithm SHOT that partly uses multi-layer projections, the 3DROM increases the MODA by 3.3%, 6.7% and 7.7%, respectively. Meanwhile, the 3DROM achieves the best performance in almost all the four performance metrics.

Ablation Study. In order to evaluate the contributions of each component in our model, an ablation study was carried out. The results are shown in Tab. 3, in which M denotes the multi-layer projection and R represents 3D Random Occlusion. As seen from the result, whichever component is added on the baseline MVDet, the performance can have a significant boost in all three datasets. When both components are used in 3DROM, the models are driven to find more robust

Table 2. Performance comparisons of deep multiview pedestrian detection.

Methods	Eval.	MODA	MODP	Prec.	Recall
MultiviewX Dataset					
RCNN-2D/3D [24]	[12]	0.187	0.464	0.635	0.439
DeepMCD [7]	[12]	0.700	0.730	0.857	0.833
Deep Occlusion [5]	[12]	0.752	0.547	0.978	0.802
MVDet [12]	[12]	0.839	0.796	0.968	0.867
SHOT [21]	[21]	0.883	0.820	0.966	0.915
3DROM	ours	**0.950**	**0.849**	**0.990**	**0.961**
EPFL WILDTRACK Dataset					
RCNN-2D/3D [24]	[5]	0.113	0.184	0.68	0.43
POM-CNN [5]	[5]	0.232	0.305	0.75	0.55
DeepMCD [7]	[5]	0.678	0.642	0.85	0.82
Deep Occlusion [5]	[5]	0.741	0.538	0.95	0.80
MVDet [12]	[12]	0.882	0.757	0.947	0.936
SHOT [21]	[21]	0.902	**0.765**	0.961	0.940
3DROM	ours	**0.935**	0.759	**0.972**	**0.962**
EPFL Terrace Dataset					
RCNN-2D/3D [24]	[5]	-0.11	0.28	0.39	0.50
POM-CNN [5]	[5]	0.58	0.46	0.80	0.78
Deep Occlusion [5]	[5]	0.71	0.48	0.88	0.82
MVDet [12]	ours	0.872	0.700	0.982	0.888
SHOT [21]	ours	0.871	0.703	0.989	0.881
3DROM	ours	**0.948**	**0.705**	**0.997**	**0.951**

Table 3. Ablation study of 3DROM.

	MultiviewX Dataset				WILDTRACK Dataset				Terrace Dataset			
Methods	MODA	MODP	Prec.	Recall	MODA	MODP	Prec.	Recall	MODA	MODP	Prec.	Recall
MVDet	0.839	0.796	0.968	0.867	0.882	0.757	0.947	0.936	0.872	0.700	0.982	0.888
MVDet+M	0.900	0.837	0.975	0.924	0.912	**0.769**	0.959	0.953	0.894	0.689	0.983	0.911
MVDet+R	0.898	0.830	0.986	0.912	0.923	0.768	0.964	0.959	0.915	**0.709**	0.994	0.920
3DROM	**0.950**	**0.849**	**0.990**	**0.961**	**0.935**	0.759	**0.972**	**0.962**	**0.948**	0.705	**0.997**	**0.951**

features across multiple views, and multi-layer projection can provide sufficient features. These two components do not conflict but work better together.

Choice of Projection Layers. To illustrate the benefits of using five-layer feature projections, a validation study was carried out on the Terrace dataset. As reported in Tab. 4, when more than one layer is used in the feature map projection, MODA increases with the number of the projection layers. The experiments show that the feature projection, by using the planes below the waist height (100 cm), leads to better results than that using the planes equidistantly selected between 0 cm and the average pedestrian height 180 cm. This can be interpreted as follows: as can be seen in Fig. 3(a) and (b), in comparison with the ground-plane projection, the feature projection of a pedestrian on a higher plane tends to move towards the underlying camera in the top view, which projects the features, for the pedestrians who are outside of the AOI, into the AOI of the

top view. Therefore, by using a projection plane at the pedestrians' heights, the features, extracted from the distant pedestrians, disturb the pedestrian detection within the AOI. The use of the projection planes below the waist is a good trade-off between the benefits of using multiple planes and the side effects.

Table 4. Validation of the number of projection layers (with 3D Random Occlusion applied).

Layers	Heights (cm)	MODA	MODP	Prec.	Recall
1	0	0.915	0.709	0.994	0.920
2	0, 180	0.867	0.688	0.972	0.893
	0, 60	0.934	0.708	0.996	0.937
3	0, 90, 180	0.892	0.700	0.973	0.918
	0, 60, 90	0.936	0.710	**0.997**	0.938
4	0, 60, 120, 180	0.902	0.697	0.988	0.912
	0, 30, 60, 90	0.943	**0.712**	0.996	0.946
5	0, 45, 90, 135, 180	0.901	0.682	0.966	0.934
	0, 15, 30, 60, 90	**0.948**	0.705	**0.997**	**0.951**

Validation of 3D Random Occlusion. To investigate the role of the frequency to use 3D Random Occlusion, we tried different values of the occlusion probability p for using 3D Random Occlusion. As reported in Tab. 5, MODA increases with p and reaches the maximum value when $p = 100\%$ in all three datasets. We further compared the 3D Random Occlusion with the related Random Erasing method which was applied to each camera view independently. In this experiment, the optimal settings of Random Erasing [28] proposed by the authors were used. In Tab. 6, The MODA decreases after 3D Random Occlusion is replaced by Random Erasing in all three datasets. This experiment shows the 3D Random Occlusion method can simulate the effect of Random Erasing in 3D space and is specifically designed for multi-view detection.

Fig. 7 shows the validation of the number of 3D random occlusions. When occlusions are too few, the risk of overfitting increases in the training. On the other hand, too many occlusions will cover most pedestrians so that the network cannot learn effective features well. The most appropriate number of occlusions

Table 5. Validation of the occlusion probability (with 5-layer projection applied).

	MultiviewX Dataset				WILDTRACK Dataset				Terrace Dataset			
p	MODA	MODP	Prec.	Recall	MODA	MODP	Prec.	Recall	MODA	MODP	Prec.	Recall
0%	0.900	0.837	0.975	0.924	0.882	0.757	0.947	0.936	0.894	0.689	0.983	0.911
30%	0.927	**0.852**	**0.991**	0.936	0.920	0.757	**0.975**	0.944	0.924	0.697	0.982	0.941
50%	0.934	0.851	0.978	0.956	0.923	0.748	0.961	0.962	0.941	0.694	0.983	**0.957**
70%	0.941	0.846	0.984	0.956	0.928	0.742	0.967	0.960	0.944	**0.706**	0.994	0.949
100%	**0.950**	0.849	0.990	**0.961**	**0.935**	**0.759**	0.972	**0.962**	**0.948**	0.705	**0.997**	0.951

Table 6. A comparison of data augmentation schemes (with 5-layer projection applied).

	MultiviewX Dataset				WILDTRACK Dataset				Terrace Dataset			
Methods	MODA	MODP	Prec.	Recall	MODA	MODP	Prec.	Recall	MODA	MODP	Prec.	Recall
w/o Augmentation	0.900	0.837	0.975	0.924	0.882	0.757	0.947	0.936	0.894	0.689	0.983	0.911
Random Erasing	0.927	0.847	0.983	0.943	0.920	**0.766**	0.953	**0.967**	0.923	0.692	0.980	0.943
3DROM	**0.950**	**0.849**	**0.990**	**0.961**	**0.935**	0.759	**0.972**	0.962	**0.948**	**0.705**	**0.997**	**0.951**

used in training correlates with the average number of pedestrians per frame and the density of pedestrians. Since the WILDTRACK and MultiviewX datasets contain more pedestrians than the Terrace, this number is greater.

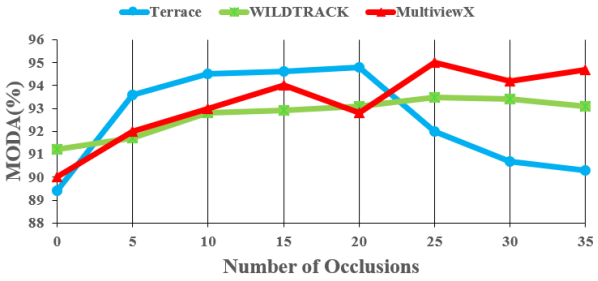

Fig. 7. Parameter validation on the number of occlusions.

5 Conclusions and Future Work

In this paper, we have proposed 3DROM for deep multiview pedestrian detection, which is based on the MVDet framework. 3D Random Occlusion provides extra training samples to the multi-view pedestrian detection network to improve the robustness in occlusion and prevent overfitting. In addition, by learning the multi-layer feature information, 3DROM can fully utilize the limited feature information from each camera view and improve pedestrian detection performance. The greatly improved performance of the 3DROM has been demonstrated in comparison with state-of-the-art methods. Future work is to find a more efficient way to fuse large-scale features and improve the across-dataset generalizability in deep-learning based multi-view pedestrian detection.

Acknowledgments This work was supported by National Natural Science Foundation of China (NSFC) under Grant 60975082 and Xi'an Jiaotong-Liverpool University under Grant RDF-17-01-33, RDF-19-01-21 and FOSA2106045.

References

1. MultiviewX. https://github.com/hou-yz/MVDet
2. Terrace. https://www.epfl.ch/labs/cvlab/data-pom-index-php/
3. WILDTRACK. https://www.epfl.ch/labs/cvlab/data/data-wildtrack/
4. Alahi, A., Jacques, L., Boursier, Y., Vandergheynst, P.: Sparsity driven people localization with a heterogeneous network of cameras. Journal of Mathematical Imaging and Vision **41**(1), 39–58 (2011)
5. Baqué, P., Fleuret, F., Fua, P.: Deep occlusion reasoning for multi-camera multi-target detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 271–279 (2017)
6. Chavdarova, T., Baqué, P., Bouquet, S., Maksai, A., Jose, C., Bagautdinov, T., Lettry, L., Fua, P., Van Gool, L., Fleuret, F.: WILDTRACK: A multi-camera hd dataset for dense unscripted pedestrian detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5030–5039 (2018)
7. Chavdarova, T., Fleuret, F.: Deep multi-camera people detection. In: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 848–853. IEEE (2017)
8. Eshel, R., Moses, Y.: Tracking in a dense crowd using multiple cameras. International Journal of Computer Vision **88**(1), 129–143 (2010)
9. Fleuret, F., Berclaz, J., Lengagne, R., Fua, P.: Multicamera people tracking with a probabilistic occupancy map. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(2), 267–282 (2007)
10. Ge, W., Collins, R.T.: Crowd detection with a multiview sampler. In: European Conference on Computer Vision. pp. 324–337. Springer (2010)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
12. Hou, Y., Zheng, L., Gould, S.: Multiview detection with feature perspective transformation. In: European Conference on Computer Vision. pp. 1–18. Springer (2020)
13. Kasturi, R., Goldgof, D., Soundararajan, P., Manohar, V., Garofolo, J., Bowers, R., Boonstra, M., Korzhova, V., Zhang, J.: Framework for performance evaluation of face, text, and vehicle detection and tracking in video: data, metrics, and protocol. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(2), 319–336 (2008)
14. Khan, S.M., Shah, M.: Tracking multiple occluding people by localizing on multiple scene planes. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(3), 505–519 (2009)
15. Khan, S.M., Shah, M.: A multiview approach to tracking people in crowded scenes using a planar homography constraint. In: European Conf. Computer Vision. pp. 133–146 (2006)
16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems **25**, 1097–1105 (2012)
17. Peng, P., Tian, Y., Wang, Y., Li, J., Huang, T.: Robust multiple cameras pedestrian detection with multi-view Bayesian network. Pattern Recognition **48**(5), 1760–1772 (2015)
18. Qiu, R., Xu, M., Yan, Y., Smith, J.S.: A methodology review on multi-view pedestrian detection. In: Pedrycz, W., Chen, S.M. (eds.) Recent Advancements in Multi-View Data Analytics, pp. 317–339. Springer (2022), https://doi.org/10.1007/978-3-030-95239-6_12

19. Ren, S., He, K., Girshick, R., Sun, J.: 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)
20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
21. Song, L., Wu, J., Yang, M., Zhang, Q., Li, Y., Yuan, J.: Stacked homography transformations for multi-view pedestrian detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6049–6057 (2021)
22. Utasi, Á., Benedek, C.: A bayesian approach on people localization in multicamera systems. IEEE Transactions on Circuits and Systems for Video Technology **23**(1), 105–115 (2012)
23. Wang, X., Shrivastava, A., Gupta, A.: A-Fast-RCNN: Hard positive generation via adversary for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2606–2615 (2017)
24. Xu, Y., Liu, X., Liu, Y., Zhu, S.: Multi-view people tracking via hierarchical trajectory composition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4256–4265 (2016)
25. Yan, Y., Xu, M., Smith, J.S., Shen, M., Xi, J.: Multicamera pedestrian detection using logic minimization. Pattern Recognition **112**, 107703 (2021)
26. Zhang, Q., Chan, A.B.: Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8297–8306 (2019)
27. Zhang, Q., Lin, W., Chan, A.B.: Cross-view cross-scene multi-view crowd counting. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 557–567 (2021)
28. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 13001–13008 (2020)
29. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets v2: More deformable, better results. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9308–9316 (2019)