Multiview Detection with Cardboard Human Modeling

Jiahao Ma[†,*,1,2], Zicheng Duan[*,3], Liang Zheng[1], and Chuong Nguyen[2]

[1] Australian National University
[2] CSIRO DATA61
[3] Australian Institute for Machine Learning, University of Adelaide
jiahao.ma@anu.edu.au
zicheng.duan@adelaide.edu.au

Abstract. Multiview detection uses multiple calibrated cameras with overlapping fields of view to locate occluded pedestrians. In this field, existing methods typically adopt a "human modeling - aggregation" strategy. To find robust pedestrian representations, some intuitively incorporate 2D perception results from each frame, while others use entire frame features projected to the ground plane. However, the former does not consider the human appearance and leads to many ambiguities, and the latter suffers from projection errors due to the lack of accurate height of the human torso and head. In this paper, we propose a new pedestrian representation scheme based on human point cloud modeling. Specifically, using ray tracing for holistic human depth estimation, we model pedestrians as upright, thin cardboard point clouds on the ground. Then, we aggregate the point clouds of the pedestrian cardboard across multiple views for a final decision. Compared with existing representations, the proposed method explicitly leverages human appearance and reduces projection errors significantly by relatively accurate height estimation. On four standard evaluation benchmarks, our method achieves very competitive results. The code and data are available at https://github.com/Jiahao-Ma/MvCHM.

Keywords: Multi-view detection, Pedestrian detection

1 Introduction

Multiview detection, a.k.a. multi-camera detection, usually refers to detecting objects using images from multiple viewpoints. This setup is especially advantageous when the scene is under heavy occlusion, which causes difficulties for monocular detection systems.

Existing methods in this field adopt two general steps: human feature modeling and aggregation. The former aims to leverage the scene geometry to extract

[*]Equal contribution. [†]Corresponding author.

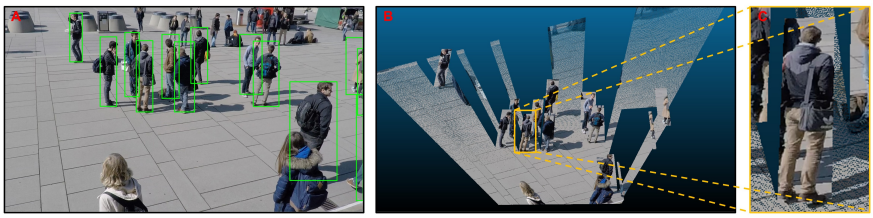

Fig. 1: Illustration of modeling humans as cardboard point clouds. **A**: Our system detects pedestrians in the 2D bounding boxes. **B**: After computing the depth of pedestrians and the ground plane by ray tracing, we project them to the 3D space. **C**: We zoom in on the pedestrian in the yellow box in Figure B. In this procedure, cardboard modeling refers to the upright, thin cardboard-like human point clouds of size $w \times h \times 1$, where w and h are the width and height, respectively, and 1 means 1 channel (a single plane of thickness 1). The cardboard human point clouds reflect the 3D position, height and appearance of each pedestrian and are later aggregated to find human locations.

discriminative descriptors for pedestrians (and the scene). The latter fuses what is extracted from all the viewpoints and locates pedestrians on a common ground plane. This paper focuses on improving the first step, especially on how to leverage the provided calibration parameters to model pedestrians.

Literature broadly has two modeling strategies. Some intuitively and simply use 2D perception results such as 2D bounding boxes [2, 16, 37], segmentation/foreground pixels [7, 38] to represent individual pedestrians, which are later clustered on the ground plane. While these methods have strong generalization ability and are interpretable, they merely use mathematical geometry relations to cluster pedestrians' positions without considering their *appearance* feature, typically leading to inaccurate aggregation outcomes. Others leverage camera calibration to project features of *entire image frames* onto the ground plane, which are used to collectively represent pedestrians [10, 11, 21, 26, 32]. Compared with the first strategy, these approaches use both pedestrian location and appearance features, aggregating frame features to obtain improved performance. However, they exhibit inaccuracies in feature projections due to a missing estimation of pedestrian height, resulting in pixels along identical vertical lines in the 3D coordinate system not being accurately projected onto the same location on the ground plane.

Considering the above discussions, we introduce a new multiview detection method with cardboard human modeling referred to as MvCHM. In a nutshell, we first detect pedestrians in each camera view in a plain way, estimate their standing points and then build human point clouds, which are defined as cardboard human modeling, after estimating the depth of the standing point (the location where a pedestrian stands) and head. The point clouds from all the views are fed into a neural network for aggregation and location regression. This pipeline is illustrated in Fig. 2, where an interesting component is the cardboard-like human point clouds made up of only one channel of pixels, as shown in Fig.1. Compared with existing works, they contain more accurate human appearance and location to be further vectorized by the neural network.

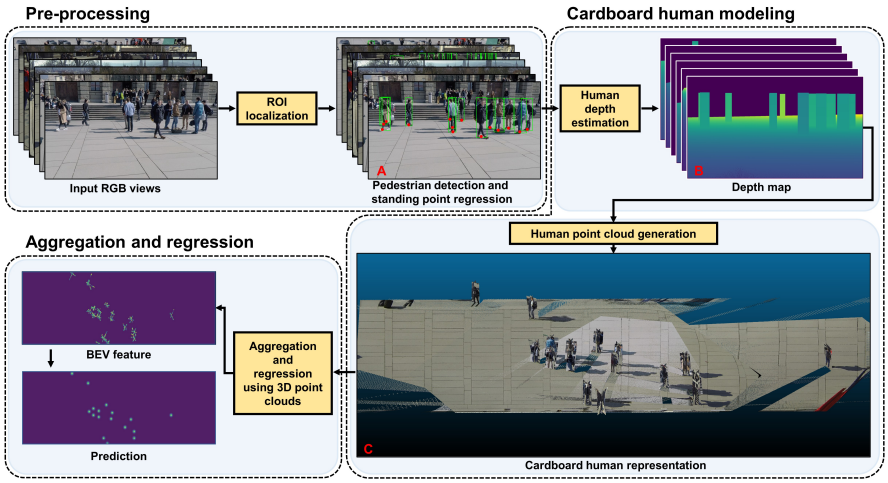

Fig. 2: The proposed system pipeline. First, given input RGB images from each view, we apply 2D object detection to obtain per-view pedestrian detection results. Then, through keypoint detection, we find the *standing point* of each detected bounding box. Next, we estimate the depth of the standing point and head of each detected person and fill the whole body region with the interpolated depth. With the estimated depth we project the detection results into the 3D space and generate human cardboard point clouds (with only one channel). Finally, we use an aggregation and regression network to find the occupancy heat map on the bird's-eye-view (BEV) plane.

Our method has a few advantages. First, compared with the "2D perception + clustering" methods [2, 7, 16, 37, 38], we incorporate sufficient human appearance features into this pipeline. Second, compared with the "projection + aggregation" methods [10, 11, 21, 26, 32], our method significantly reduces projection errors caused by inaccurate height through the cardboard modeling process and thus provides more accurate human appearance features.

We evaluate our system on four multiview pedestrian detection benchmarks WildTrack [4], MultiviewX [11] and their extension. We show that the proposed point clouds processing procedures enabled by the aggregation network give very competitive results.

2 Related Work

2.1 Feature-projection based multiview detection

Generally, feature-projection-based methods [10, 11, 21, 26, 32] project multiview high-resolution feature maps to the ground plane, concatenate these features and regress object positions from the features. Hou *et al.* [11] project convolution feature maps to the ground plane via a perspective transformation and adopt a full convolution network to aggregate the concatenated feature maps. Motivated by [11], Song *et al.* [32] introduces stacked homography transformation to project

frame features to the ground plane at different height levels. Hou *et al.* [10] deal with shadow-like distortions in different cameras and positions via a transformer structure. To align features along the vertical direction of objects, Ma *et al.* [21] voxelize 3D features with fixed-size before aggregation. As mentioned above, projection-based methods are not accurately 3D aware, so would likely encode noisy image content (*e.g.*, background and misaligned human) in the project features.

2.2 2D-Perception based multiview detection

The other line of methods [2,7,16,37,38] intuitively utilize 2D perception results to model each pedestrian, which are clustered on the ground plane. We call them 2D-perception-based methods. For example, Lima *et al.* [16] forfeit training and instead estimate the standing point within each 2D detection bounding box and predict the 3D coordinate of pedestrians by solving the clique cover problem. Yan *et al.* [38] calculates the likelihood of pedestrian presence in each foreground region and clusters pedestrian positions via minimizing a logic function. Fleuret *et al.* [7] estimate the probabilities of pedestrian occupancy via a probabilistic occupancy map. To aggregate multiview detection results, mean-field inference [2, 7] and conditional random field (CRF) [2, 29] can be exploited. Our work also starts from using 2D perception results, *i.e.,* 2D bounding boxes, but differs from existing works in that we explicitly consider the human appearance and use regression to find human locations (similar to the projection-based methods).

2.3 Estimating point clouds in 3D object detection

In 3D object detection, some existing methods generate scene point clouds using depth estimation. Since the point clouds are not provided by LiDAR, they are often called pseudo LiDAR point clouds. Wang *et al.* [34] show that a key to closing the gap between image- and LiDAR-based 3D object detection may simply be 3D representations. MF3D [36] estimates disparity maps to obtain pseudo LiDAR and fuses input RGB images with front-view features obtained by the disparity map. Mono3D-PLiDAR [35], a two-stage 3D object detection pipeline, converts input images into point clouds via DORN [8] and applies Frustum PointNets [24] to localize 3D objects. While these works use end-to-end pixel-wise depth estimation methods, we calculate the depth value of detected pedestrian regions via the *ray tracing* technique given camera poses, a ground plane, and pedestrians standing points on the ground.

2.4 Existing human modeling methods

Various 3D human body modeling techniques that incorporate appearance features and 3D geometry have been widely utilized in other areas. Carving-based 3D reconstruction methods [13,23,31] utilize the visual silhouettes obtained from multiple cameras to generate 3D carvings of the human body, whereas multiview

pose estimation methods [33, 39] leverage 3D voxels for precise keypoint detection. Dense pose estimation methods [6, 9, 12, 30] rely on dense surface meshes to model targets. However, these modeling techniques have certain limitations. They necessitate either perfectly overlapped view fields or dense annotations, rendering them unsuitable for real-world multiview detection scenarios where camera views are not perfectly aligned and only sparse labels, such as bounding boxes and standing points, are available.

3 Preliminaries: Ray tracing to compute 3D coordinates of a 2D point

Ray tracing technique models light transport: a light ray emerges at the light source, reflects on objects, and goes into the camera. Formally, ray tracing is formulated as:

$$P = O + tD. \tag{1}$$

This formula computes the *3D coordinates of a reflection point* on an object, denoted as $P = \left[P_x, P_y, P_z\right]^T$. $O = [O_x, O_y, O_z]^T$ is the 3D position of the camera, or origin; $D = [D_x, D_y, D_z]^T$ is the direction of the ray; $t \in \mathbb{R}$ is the distance between the camera and the reflection point on the object. where O and D are accessible with camera pose. Using Eq. 1, we compute the 3D coordinates and thus the depth of the standing point and the head of each pedestrian.

4 Proposed system

As shown in Fig. 2, our system consists of pre-processing (Section 4.1), human modeling (Section 4.2) and aggregation (Section 4.3), where *human modeling is our main contribution*. Below we will detail these steps with a focus on the human modeling process, including human depth estimation and human point clouds generation.

4.1 Pre-processing

Pre-processing, also denoted as ROI localization in Fig. 2 A, aims to find 1) pedestrian regions in the shape of bounding boxes and 2) the standing point of each person. Results from both steps will be used in Section 4.2 for depth estimation and pedestrian height calculation.

Standing point estimation. This could be considered naively as the middle bottom of the detected bounding box of a pedestrian. However as shown in Figure 4, this approach is inaccurate, leading to errors in 3D space. Therefore, we define the standing point as the midpoint between the person's feet. We first adopt a 2D detector CrowdDet [5] to detect bounding boxes. Then following [15, 22], we use a regression neural network to obtain the positions on the ground where pedestrians stand. Essentially, the detected bounding boxes are used as input, and the output is a single standing point. In the implementation, we use the

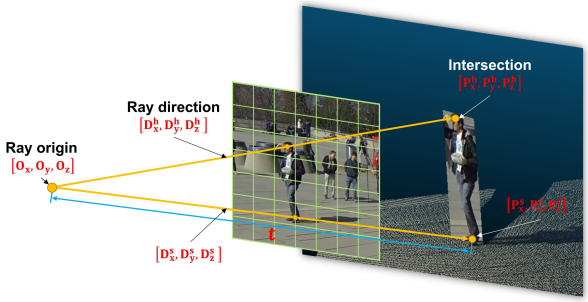

Fig. 3: Illustration of ray tracing. According to the reversibility of the ray, we define the camera center (ray origin) as O, the ray direction as D, the object 3D location as P, and the distance between the camera and reflection points on the object as t. First, we assume the pedestrian standing point is on the ground surface where $P_z = 0$, then given the camera matrix and the 2D coordinates of the standing point in the image, the depth of the standing point can be accurately calculated, the depth of the head is further calculated by substitution, finally, the depth of the rest of body region is linearly interpolated. Detailed derivations are presented in the supplementary material.

global standing point annotations provided by the benchmarks and mainly use [15] for regression, with a comparison with [22].

4.2 From 2D to 3D: Cardboard human modeling

In this section, we describe the proposed cardboard modeling that transforms 2D bounding boxes into 3D point clouds shaped as standing cardboard on the ground plane. Specifically, based on the located ROI, a human in a 2D image is modeled as a cardboard-like point clouds of size $w \times h \times 1$ (refer Fig. 1) in the 3D space. These point clouds reflect a pedestrian's appearance, height and 3D spatial position, and will be used for human feature extraction and localization (Section 4.3). Generating the cardboard human is simple: we calculate the depth of each pedestrian using ray tracing, and then project the pedestrian into the 3D space.

Human depth estimation. Due to the lack of pixel-wise human depth annotations, it is infeasible to estimate accurate depth for each human pixel. To get around this problem, we compute the depth of the standing point and the head using the ray tracing technique [1] (Section 3). The two depth values are subsequently used to interpolate the depth of other pixels in the bounding box in a linear way.

We now leverage Eq. 1 to find the 3D coordinates of the standing point and the head, denoted as $P_{\text{standpoint}} = [P^{\text{s}}_{\text{x}}, P^{\text{s}}_{\text{y}}, P^{\text{s}}_{\text{z}}]$ and $P_{\text{head}} = [P^{\text{h}}_{\text{x}}, P^{\text{h}}_{\text{y}}, P^{\text{h}}_{\text{z}}]$, respectively, as shown in Fig 3. On the one hand, to compute $P_{\text{standpoint}}$, we assume that all pedestrians are standing on the ground plane with $P^{\text{s}}_{\text{z}} = 0$. This assumption intuitively holds in normal scenarios. Based on the ground plane assumption above, we just need to calculate O and D to get $P_{\text{standpoint}}$. We first use the provided camera pose to obtain the camera 3D position O, then we calculate the ray direction D through the standing point with the image

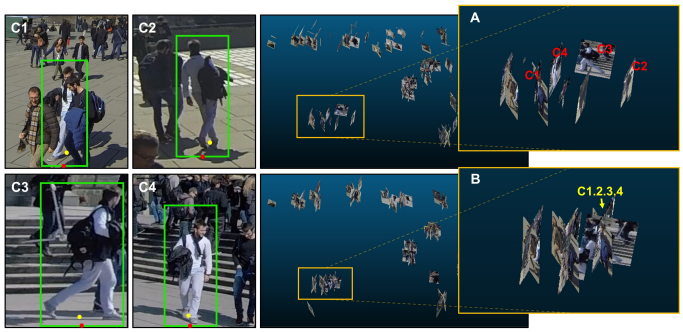

Fig. 4: An illustration of the benefit of standing point estimation. Figures C1 - C4 record four different views capturing the same person. In each view, the green box is the 2D detection result of the person, the red dot is the bottom center of the detected box, and the yellow dot represents the estimated standing point using [15]. Figures A and B show the projection results when regarding the bottom center or the estimated point as the standing point. In figure A and B, each human cardboard is marked with the corresponding view number. We observe that the projected cardboards form denser clusters in figure B, which validates the effectiveness of standing point estimation.

coordinates of the standing point and camera intrinsic, and then obtain $P_{\mathrm{x}}^{\mathrm{s}}$ and $P_{\mathrm{y}}^{\mathrm{s}}$ by substituting O and D into Eq. 1. Next, to find the pedestrian head in a 3D coordinates system, we assume the standing point and head of each pedestrian have the same x and y coordinates, *i.e.,* $P_{\mathrm{x}}^{\mathrm{s}} = P_{\mathrm{x}}^{\mathrm{h}}$, $P_{\mathrm{y}}^{\mathrm{s}} = P_{\mathrm{y}}^{\mathrm{h}}$. Similar to calculating the z coordinate of the standing point $P_{\mathrm{standpoint}}$, we solve the z coordinate of the head $P_{\mathrm{z}}^{\mathrm{h}}$ by substituting $P_{\mathrm{x}}^{\mathrm{h}}$ or $P_{\mathrm{y}}^{\mathrm{h}}$ into Eq. 1. After computing the 3D coordinates of the standing point and head in the world coordinate system and then converting them into the camera coordinate system, we default the value of the Z of the camera coordinate to the depth value.

Finally, we use linear interpolation to fill the rest of the detection region with a rough depth value, the generated depth map is shown in Fig. 2B. More derivation details are provided in Section 4 of the supplementary material.

Point clouds generation. After assigning each pixel in the pedestrian region with a depth value, we project the pedestrian region from the 2D to 3D space as point clouds according to the intrinsic and extrinsic parameters and estimated depth. Projection details are shown in Section 5 of the supplementary materials. Our experiments in Section 4.6 show that the ground plane point clouds introduce noisy features and additional computational cost, leading to poorer model performance. Therefore, we only project the pedestrian region to the 3D space.

4.3 Aggregation and regression using 3D point clouds

Aiming to aggregate features from multiple views, we process point clouds into feature vectors, using the network proposed in [14]; features are then concatenated to regress pedestrian position on the ground plane. Specifically, we discretize point clouds into an evenly spaced grid in the BEV plane, creating a set

of pillars (voxels with ultimate spatial extend in the Z direction [14]). Then, we randomly sample the point clouds in each pillar and adopt PointNet [25] to extract high-dimensional features (pillar feature) in each pillar. Based on the pillar representations, we follow [41] to flatten pillar features to the BEV plane and regress the final pedestrian position. Similar to [3, 11], we represent binary ground truth pedestrian occupancy as Gaussian distribution maps. We use the focal loss [18] as position regression loss:

$$\mathcal{L}_{reg} = -\alpha(1-p)^{\gamma} log_p \ , \tag{2}$$

where α and γ are two hyper-parameters. We use the same values of α and γ as [18].

4.4 Experimental settings

Dataset. We compare our method on two standard multiview pedestrian benchmarks [4, 11], and two newly created datasets Wildtrack+ and MultiviewX+.

Wildtrack [4] is a real-world multiview pedestrian detection benchmark capturing people on a square of 12 $\times$ 36 meters with 7 calibrated cameras. The image resolution is 1080 $\times$ 1920, making it the largest in the community, and the square is discretized to a 480 $\times$ 1440 grid. The dataset contains 400 images, the first 360 frames for training and the last 40 for testing.

Wildtrack+ is an extension of the Wildtrack [4] dataset, in which we additionally annotate the unlabelled pedestrians outside the detection area. Note that labels inside the detection area remain unchanged. The new annotations allow us to train a 2D detector on Wildtrack instead of borrowing an off-the-shelf detector trained on other datasets.

MultiviewX [11] is a synthetic dataset created by Unity for pedestrian detection in crowded scenes. This dataset covers an area of 16 $\times$ 23 meters with 6 synchronized cameras. The ground plane is quantized into a 640 $\times$ 1000 grid, and the resolution is 1080 $\times$ 1920. It also has 400 frames with the last 40 frames for testing.

MultiviewX+ is newly generated using the same Unity engine following the same labeling mechanism as MultiviewX [11]. Compared with the MultiviewX dataset, our MultiviewX+ dataset *1)* additionally annotates the pedestrians outside the detection area to train 2D detectors locally *2)* introduces new characters different from that in MultiviewX *3)* provides more accurate camera calibrations.

Evaluation metrics. Four metrics are used: Multiple Object Detection Accuracy (MODA), Multiple Object Detection Precision (MODP), Precision, and Recall. Specifically, MODA accounts for the normalized missed detections and false positives and MODP assesses the localization precision. We estimate the empirical precision and recall, calculated by $P = TP/(TP + FP)$ and $R = TP/(TP + FN)$ respectively. We view MODA as the primary indicator. A threshold of 0.5 meters is used to decide true positives.

For evaluating detection models on the Wildtrack and Wildtrack+ datasets, we observe severe annotation missing near the border of the detection area,

Method	Wildtrack*				MultiviewX			
	MODA	MODP	Precision	Recall	MODA	MODP	Precision	Recall
RCNN & clustering [37]	11.9§	18.1§	66.1§	44.9§	18.7	46.4	63.5	43.9
Deep-Occlusion [2]	-	-	-	-	75.2	54.7	97.8	80.2
MVDet [11]	88.7	73.6	93.2	95.4	83.9	79.6	96.8	86.7
SHOT [32]	90.8	77.7	96.0	94.3	88.3	82.0	96.6	91.5
MVDeTr [10]	92.1	84.1	96.1	94.5	93.7	91.3	99.5	94.2
3DROM [26]	93.9	76.0	97.7	96.2	95.0	84.9	99.0	96.1
MvCHM (ours)	95.3	84.5	98.2	97.1	93.9	88.3	98.5	94.8
	Wildtrack+*				**MultiviewX+**			
RCNN & clustering [37]	10.1§	17.2§	65.1§	42.3§	19.9§	48.9§	64.1§	44.0§
Deep-Occlusion [2]	-	-	-	-	-	-	-	-
MVDet [11]	87.8	74.9	95.1	90.7	84.5	80.9	96.4	85.2
SHOT [32]	90.2	77.5	95.7	94.1	88.5	82.7	97.1	90.2
MVDeTr [10]	92.2	84.2	96.3	94.1	93.8	91.5	99.6	93.9
3DROM [26]	93.8	77.1	96.9	96.1	95.2	85.1	99.2	96.7
MvCHM (ours)	94.6	84.7	98.3	96.6	93.8	87.9	98.6	95.3

Table 1: Comparison with the state-of-the-art methods on the standard evaluation benchmarks. For each metric, the best, second best, and third best numbers (in percentage) are highlighted in red, blue and green, respectively. Our method yields state-of-the-art performances on the Wildtrack/Wildtrack+ datasets and very competitive results on the MultiviewX/MultiviewX+ datasets. On the Wildtrack and MultiviewX datasets, due to the lack of pedestrian training labels outside the detection area, we adapt a pre-trained 2D detector in the ROI localization procedure mentioned in Section 4.1, while on the Wildtrack+ and MultiviewX+ datasets, we train a 2D detector using the proposed complete annotations, all other methods follow the same training scheme for fair comparisons. ***** denotes that we use a mask to reduce the effect of the inaccurate labeling, details are discussed in the evaluation Section 4.6. § indicates the results are from our implementation.

leading to an accuracy drop for existing methods. To reduce the impact of missing labels, we mask the border area on both regressed and ground truth heatmap during evaluation, and as a result, all the compared methods now have higher accuracy. The Section 1 in the supplementary materials provides more details of the mask.

4.5 Implementation details

We train the pedestrian detector on the Widltrack+ dataset and MultiviewX+ dataset while borrowing the best-trained model provided by CrowdDet [5] on the Wildtrack dataset and the MultiviewX dataset. In training the pedestrian detector, we use the Earth Mover's Distance loss (EMD Loss) and Set NMS [5] which are shown to improve robustness against occlusions. For standing point estimation, we apply the MSPN [15] network and train it with the provided standing point ground truths provided in all four datasets. When constructing human point clouds, to avoid projection noise, we directly remove the background and merely project pixels in each bounding box to the 3D space. To

train the aggregation and regression network, we use an Adam optimizer with L2 regularization of 5×10^{-3}. α and γ in Eq. 2 are set to be 2 and 4, respectively. The learning rate is set to 2×10^{-4}. During the evaluation, the heatmap thresholds are set to be 0.8, 0.86, 0.8, 0.8 on the Wildtrack, Wildtrack+, MultiviewX, and MultiviewX+ datasets respectively. We conduct all experiments on a single RTX-3090 Ti GPU.

4.6 Evaluation

Comparison with the state-of-the-art methods. Table 1 summarizes this comparison. On the Wildtrack dataset, our pipeline achieves state-of-the-art performance: MODA=95.3%, MODP = 84.5%, Precision = 98.2%, and Recall = 97.1%. Regarding MODA, our method is 1.4% higher than the second best method 3DROM [26] based on feature projection. On the Wildtrack+ dataset, our approach outperforms other methods with similar margins.

Regarding the MultiviewX and MultiviewX+ datasets, our method is slightly outperformed by the previous state-of-the-art method [26] but remains very competitive. The main reason is that the camera positions in the MultiviewX are lower than that in the Wildtrack dataset, which causes the cameras to look in a relatively horizontal direction, making it difficult to capture the pedestrians' feet. The detection results on the evaluation benchmarks are visualized in Fig. 7 in the supplementary materials.

Necessity of estimating the standing point. We perform an ablation study on this module in Fig. 6B. For convenient, the standing point is denoted as *SP* in Fig. 6B. During the Pre-processing process introduced in Section 4.1, the standing point estimator (MSPN) regresses the standing point of each person (yellow dot in Fig. 4 C1∼4), where "*W/o SP* estimation" indicates directly regarding the bottom center of the pedestrian bounding box as the standing point (red dot in Fig. 4 C1∼4). From Fig. 6B and Tab. 2, we observe that without the standing point estimation step, system accuracy drops significantly from 95.3% to 42.1%. A probable reason for this drop is that the bottom centers of detection bounding boxes usually do not stably indicate the human position (refer to the comparison in A and B of Fig. 4 for the scattered centers).

Importance of having human appearance features. As mentioned before, 2D detection-based methods undesirably discard human appearance features, which is unavoidable due to their method designs [2, 7, 16, 37, 38]. In Fig. 5, we conduct ablation studies to verify the importance of integrating human appearance features. In this figure, *w/o feature* means we directly remove RGB from normal point clouds, *i.e.*, changing each point from $[x, y, z, r, g, b]$ into $[x, y, z]$. *"full black"*, *"full white"* and *"mean value"* replace the RGB pixels on the human with black pixels, white pixels and the mean RGB value, respectively. Therefore, these four variants of our method merely encode human location. Ablation results in Fig. 5A and Tab. 2 indicate the importance of having human appearance features and these results further validate our design since our method exceeds other variants with a clear margin.

Method	Detector	SP	AF	MODA
MvCHM	✓			20.4
	✓		✓	42.1
	✓	✓		78.1
	✓	✓	✓	95.3

Table 2: Modular ablation study reported on the Wildtrack dataset. **SP**: Standing Point estimation, **AF**: Appearance Feature.

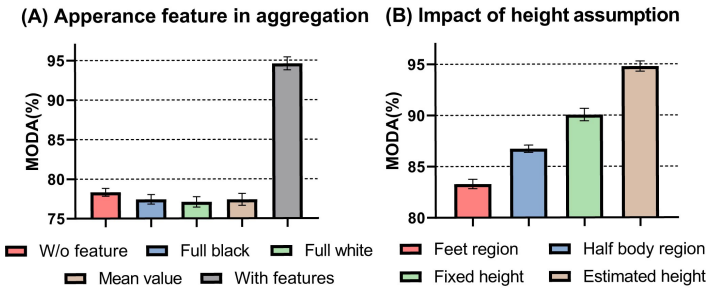

Fig. 5: (A) Ablation study on integrating human appearance feature. (B) Comparing with variants in human height estimation.

Comparison of various pedestrian detectors and keypoint detectors. 2D human detection and standing point detection are two important components of our system. In Fig. 6A, we compare CrowdDet [5] used in our system with SSD [20], YOLO-v3 [27], Faster RCNN [28], and RetinaNet [19] on the Wildtrack dataset. We find that the multiview detection performance has the same trend as 2D detection accuracy. For example, the best 2D detection method CrowdDet also gives the highest MODA in multiview detection. These results suggest that 2D detection has a profound influence on our method. On the other hand, we compare MSPN [15] used in our system with Hourglass [22] as Fig. 6B shown. We find that MSPN with a higher standing point estimation accuracy contributes to better system performance. This is because correct standing point estimation plays an important role in constructing cardboard humans as the actual position on the ground plane.

Comparing different human height estimates. In Fig. 5B, we compare a few variants in human height estimation. "Fixed height" of 1.8m is used in some existing feature projection-based methods [10, 11, 21, 32], which inevitably introduces noise given its inaccuracy. Moreover, we expect insufficient human description if we consider half of the body or only the foot region. These considerations are verified in this experiment, where using the whole body region found by 2D detection yields the highest MODA accuracy. Using the feet region only is the worst variant because too little appearance is integrated. This experiment confirms that relatively accurate height estimates are beneficial for appearance feature extraction and avoiding background noise.

Impact of point clouds sampling rate and ground plane modeling. By default, for each cardboard we use 50% of its points; we also discard all the ground plane points. Here we evaluate how these two aspects (both related

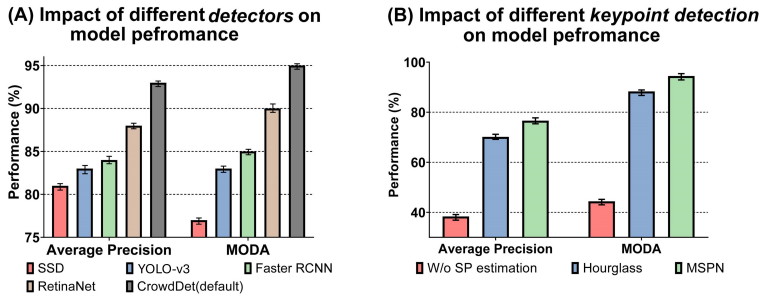

Fig. 6: (A) Comparing various pedestrian detectors on their 2D detection accuracy and overall system performance. (B) Ablation study on having keypoint estimation modules. Verify the impact of standing point detection accuracy on the overall system performance. The results are reported on the Wildtrack dataset.

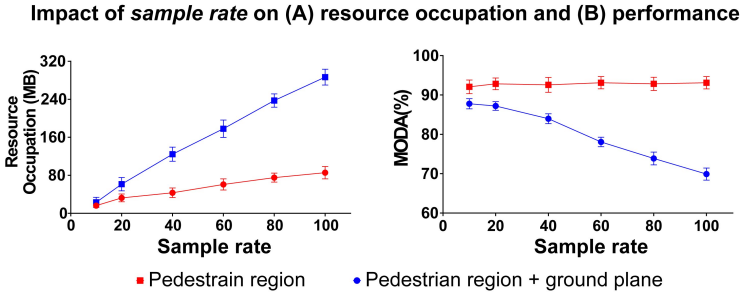

Fig. 7: Comparing the performance of additional 3D modeling ground plane in our system. Two aspects are considered: (A) GPU memory consumption and (B) detection accuracy MODA (%).

to point clouds) impact our system, in Fig. 7, where "Sample rate" means the preservation rate of the point clouds. In Fig. 7A, when we gradually increase the point clouds sampling rate, the GPU memory consumption increases linearly, and modeling the ground plane would incur additional memory costs because the ground plane itself takes up a considerable amount of memory. On the other hand, Fig. 7B shows that detection accuracy remains stable with increased sample rates when the ground plane is not included in the modeling; conversely, the performance declines when it is included.

This is probably because the modeling of the ground plane introduces noise which compromises our system. Hence, considering both memory consumption and accuracy, we choose to remove the ground plane and to use 50% of the points for each cardboard.

Different human appearance descriptors in cardboard modeling. By default, our method simply encodes human appearance features with thin cardboard-like point clouds. In this section, we explore the impact of applying different human appearance representation strategies on model performance. We compare our point clouds descriptor with two other types of human feature descriptors, namely, the re-ID features from an off-the-shelf person re-identification (re-ID) model [40] and the feature from the Feature Pyramid Network (FPN) [17]

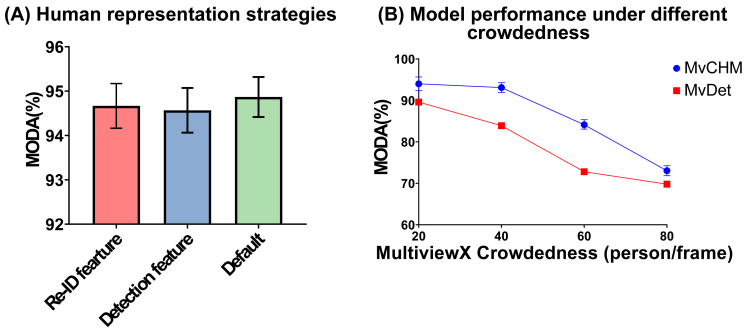

Fig. 8: (A): Comparing multiple human feature representation strategies with the default point clouds representation. (B): Comparison with the feature-projection-based method MVDet [11] under various pedestrian density levels on the MultiviewX dataset.

used in our detection model [5]. Both variants follow the same training and test protocol as our system. Results are summarized in Fig. 8A. We observe that models with different feature representations perform similarly because all of them include similar human feature appearances to some extent.

Impact of pedestrian density. In Fig. 8B, we evaluate our system under various levels of crowdedness on the MultiviewX dataset and compare it with MVDet [11]. We find decreased detection accuracy with increasing crowdedness, which is consistent with the findings in [11]. In fact, a crowded scene deteriorates the 2D detector and adds noise to the subsequent feature learning process. Furthermore, our method consistently outperforms MVDet, indicating the robustness of the proposed system.

5 Discussion

Ours *vs.* projection-based methods: less susceptible to projection noise. As mentioned in previous sections, feature projection-based methods [10, 11, 21, 32] suffer from inaccurate projections due to wrong human height estimation. In the latter case, part of the pedestrian torso is wrongly projected on the planes with an inaccurate height, and the projected pedestrian features are intermingled with background features or noise.

Our method estimates an accurate height using bounding boxes before projecting the pedestrian region to the corresponding 3D space. Thereby, the proposed method effectively recovers the pixel's 3D position along the Z axis and separates human features from the background, which alleviates projection noise.

Ours *vs.* clustering-based methods: can take advantage of human appearance. By extracting features from the human point clouds, we seamlessly integrate human appearance into the system. In comparison, existing 2D-perception and clustering-based methods [2, 7, 16, 37, 38] merely use pedestrian 2D position features to predict target position. The drawback of not using human appearance features is experimentally analyzed in Section 4.6.

Ours *vs.* 3D-carving-based methods: robust 3D in single-camera, occluded scenes. Unlike traditional 3D carving methods [13,23,31] that require

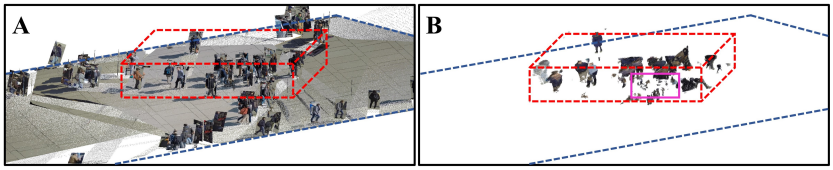

Fig. 9: 3D carving (B) can only reconstruct and detect people in the red region shared by at least two cameras, while our method (A) can succeed with only one camera. Also, the 3D carving [13,23,31](B) fails in the occluded scenes, like the artifacts in the purple box.

objects to be captured by at least two cameras and also fail in the presence of occlusion, our proposed cardboard human modeling can generate accurate 3D representations from a single camera in occluded scenes. Refer to Fig.9 for comparison. This expands its potential applications to large ground planes where camera fields of view may not overlap and allows for reliable 3D reconstruction in challenging scenarios.

Multiview detection *vs.* multiview pose estimation: sparse labels *vs.* rich annotations. Compared to pose estimation methods [6, 9, 12, 30, 33, 39]that necessitate rich annotations such as dense surface labels or human joint labels, the multiview detection task provides considerably sparser labels such as bounding boxes, which pose estimation techniques are unsuitable for, while our proposed approach is specifically customized to fit this situation.

Two major performance influencers. Our system effectiveness relies on 2D detection and standing point (keypoint) detection performance, demonstrated by experiments using various detectors and keypoint estimators in Fig. 6. Stronger detectors lead to higher accuracy. Instead, we don't claim these models as our contribution. The novelty lies in the 3D cardboard representation of humans. If we use the same detectors and pose estimators but a different human representation, the multiview detection accuracy will drop significantly (see Fig. 5). As such, the use of these existing models does not affect our novelty.

Acknowledgment. We would like to express our gratitude to Yunzhong Hou for his valuable guidance and suggestions, which greatly contributed to the success of this work.

6 Conclusion

We propose a new pedestrian representation for multiview detection, modeling humans as cardboard-like point clouds. This approach leverages scene geometry to fuse height and appearance features, and less noise is included compared with previous feature projection based methods. Our system is evaluated on two existing multiview detection datasets and their extension datasets where we report very competitive results compared with previous state-of-the-art methods.

References

1. Appel, A.: Some techniques for shading machine renderings of solids. In: Proceedings of the April 30–May 2, 1968, spring joint computer conference. pp. 37–45 (1968)
2. 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)
3. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7291–7299 (2017)
4. 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)
5. Chu, X., Zheng, A., Zhang, X., Sun, J.: Detection in crowded scenes: One proposal, multiple predictions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
6. Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: Retinaface: Single-shot multi-level face localisation in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5203–5212 (2020)
7. 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)
8. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2002–2011 (2018)
9. Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7297–7306 (2018)
10. Hou, Y., Zheng, L.: Multiview detection with shadow transformer (and view-coherent data augmentation). In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 1673–1682 (2021)
11. Hou, Y., Zheng, L., Gould, S.: Multiview detection with feature perspective transformation. In: European Conference on Computer Vision. pp. 1–18. Springer (2020)
12. Jiang, B., Hong, Y., Bao, H., Zhang, J.: Selfrecon: Self reconstruction your digital avatar from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5605–5615 (2022)
13. Khan, S.M., Shah, M.: A multiview approach to tracking people in crowded scenes using a planar homography constraint. In: Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part IV 9. pp. 133–146. Springer (2006)
14. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12697–12705 (2019)
15. Li, W., Wang, Z., Yin, B., Peng, Q., Du, Y., Xiao, T., Yu, G., Lu, H., Wei, Y., Sun, J.: Rethinking on multi-stage networks for human pose estimation. arXiv preprint arXiv:1901.00148 (2019)

16. Lima, J.P., Roberto, R., Figueiredo, L., Simoes, F., Teichrieb, V.: Generalizable multi-camera 3d pedestrian detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 1232–1240 (June 2021)
17. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
18. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
19. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
20. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016)
21. Ma, J., Tong, J., Wang, S., Zhao, W., Zheng, L., Nguyen, C.: Voxelized 3d feature aggregation for multiview detection. arXiv preprint arXiv:2112.03471 (2021)
22. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European conference on computer vision. pp. 483–499. Springer (2016)
23. Possegger, H., Sternig, S., Mauthner, T., Roth, P.M., Bischof, H.: Robust real-time tracking of multiple objects by volumetric mass densities. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2395–2402 (2013)
24. Qi, C.R., Liu, W., Wu, C., Su, H., Guibas, L.J.: Frustum pointnets for 3d object detection from rgb-d data. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 918–927 (2018)
25. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660 (2017)
26. Qiu, R., Xu, M., Yan, Y., Smith, J.S., Yang, X.: 3d random occlusion and multi-layer projection for deep multi-camera pedestrian localization. arXiv preprint arXiv:2207.10895 (2022)
27. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
28. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
29. Roig, G., Boix, X., Shitrit, H.B., Fua, P.: Conditional random fields for multi-camera object detection. In: 2011 International Conference on Computer Vision. pp. 563–570. IEEE (2011)
30. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2304–2314 (2019)
31. Sekii, T.: Robust, real-time 3d tracking of multiple objects with similar appearances. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4275–4283 (2016)
32. 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)

33. Tu, H., Wang, C., Zeng, W.: Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 197–212. Springer (2020)
34. Wang, Y., Chao, W.L., Garg, D., Hariharan, B., Campbell, M., Weinberger, K.Q.: Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8445–8453 (2019)
35. Weng, X., Kitani, K.: Monocular 3d object detection with pseudo-lidar point cloud. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019)
36. Xu, B., Chen, Z.: Multi-level fusion based 3d object detection from monocular images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2345–2353 (2018)
37. Xu, Y., Liu, X., Liu, Y., Zhu, S.C.: Multi-view people tracking via hierarchical trajectory composition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4256–4265 (2016)
38. Yan, Y., Xu, M., Smith, J.S., Shen, M., Xi, J.: Multicamera pedestrian detection using logic minimization. Pattern Recognition **112**, 107703 (2021)
39. Ye, H., Zhu, W., Wang, C., Wu, R., Wang, Y.: Faster voxelpose: Real-time 3d human pose estimation by orthographic projection. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI. pp. 142–159. Springer (2022)
40. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3702–3712 (2019)
41. Zhou, Y., Tuzel, O.: Voxelnet: End-to-end learning for point cloud based 3d object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4490–4499 (2018)