Pattern Analysis and Applications (2023) 26:1059–1070
https://doi.org/10.1007/s10044-023-01168-6

INDUSTRIAL AND COMMERCIAL APPLICATION

MVDet: multi-view multi-class object detection without ground plane assumption

Sola Park[1] · Seungjin Yang[1] · Hyuk-Jae Lee[1]

Received: 26 September 2022 / Accepted: 19 April 2023 / Published online: 13 June 2023

Abstract

Although many state-of-the-art methods of object detection in a single image have achieved great success in the last few years, they still suffer from the false positives in crowd scenes of the real-world applications like automatic checkout. In order to address the limitations of single-view object detection in complex scenes, we propose MVDet, an end-to-end learnable approach that can detect and re-identify multi-class objects in multiple images captured by multiple cameras (multi-view). Our approach is based on the premise that incorrect detection results in a specific view can be eliminated using precise cues from other views, given the availability of multi-view images. Unlike most existing multi-view detection algorithms, which assume that objects belong to a single class on the ground plane, our approach can classify multi-class objects without such assumptions and is thus more practical. To classify multi-class objects, we propose an integrated architecture for region proposal, re-identification, and classification. Additionally, we utilize the epipolar geometry constraint to devise a novel re-identification algorithm that does not require assumptions about ground plane assumption. Our model demonstrates competitive performance compared to several baselines on the challenging MessyTable dataset.

Keywords Multi-view object detection · Automatic checkout · Epipolar geometry · Re-identification · Multi-view classification

1 Introduction

With the significant progress in deep learning, several methods based on deep learning have been proposed in various applications of computer vision, such as automatic checkout [1, 2], autonomous driving [3], and robotics [4]. Among these applications, automatic checkout employs different types of sensors including cameras, LIDARs, microphones, and scales to automatically recognize the items bought by a customer in a store, without the need for scanning barcodes. In comparison with other types of sensors, camera-based automatic checkout is not only less expensive but also advantageous in utilizing the recent promising outcomes of computer vision research.

However, Rigner et al. [2] have shown that the performances of some representative vision-based detection algorithms, such as Mask R-CNN [5], YOLO [6], and RetinaNet [7], were degraded when the scenes become crowded in automatic checkout. While several methods have attempted to solve this issue in a single-view setting [8–10], a single-view clue is insufficient to address the occlusion problem in complex scenes. To improve detection accuracy under occlusion, recent approaches have utilized depth information [11–13], LIDAR point cloud [14, 15], or multiple camera views (multi-view) [16–19]. In this paper, we focus on multi-class object detection from multiple RGB camera views.

In this study, the multi-view images capture the overlapping fields of view and are characterized by the intrinsic and extrinsic camera parameters. Each scene contains multiple classes of objects, and the positions of the cameras are randomized per every scene for a general setting. The objective is to locate objects in each view (region proposal), compare the objects across views to determine if

✉ Hyuk-Jae Lee
hjlee@capp.snu.ac.kr

Sola Park
sapark@capp.snu.ac.kr

Seungjin Yang
sjyang@capp.snu.ac.kr

[1] Department of ECE, Seoul National University, 1 Gwanak-ro, Seoul 08826, Republic of Korea

they have the same identity (re-identification), and classify the re-identified objects (classification).

Existing multi-view detection studies have two limitations: (i) they cannot process multi-class objects, and (ii) they assume ground plane for re-identification. For instance, Cai et al. [20] have proposed to separately use the state-of-the-art methods in the fields of single-view detection and multi-view re-identification for multi-view detection. However, this approach cannot determine the final class if the object's classes are different in each view. Similarly, other studies [17, 19, 21] have integrated the detection and re-identification procedures in an end-to-end manner, but they also assume only one class, and all objects in the scenes are considered to be of the same class without classification. Furthermore, multi-view pedestrian detection methods [17, 21], which are the main focus of multi-view detection research, assume a reference ground plane, which is not applicable to scenarios where objects do not stand on the ground plane, such as automatic checkout and unmanned stores. Therefore, a general model that performs multi-class object detection and re-identification without ground plane assumption is required.

Therefore, this paper proposes a method to address two issues in object detection: multi-class classification and re-identification without ground plane assumption. For multi-class classification, the proposed method simultaneously performs region proposal, re-identification, and classification in an end-to-end manner. Specifically, we use faster R-CNN's region proposal network to locate objects in each view, and then, a view embedding network (VEN) trained with triplet loss [22] to re-identify the region proposal boxes. Finally, a classification network determines their class after pooling the re-identified regions. We save time and memory use by sharing the initial features in three stages. In addition, to improve detection accuracy at inference time, the re-identification and classification networks are trained on the incomplete detection results generated by the region proposal network instead of the error-free ground truth.

For re-identification without ground plane assumption, the proposed method adopts epipolar geometry to deal with occlusions and view variations. We calculate the embedding distance of a pair of region proposal boxes using the features extracted by VEN, and re-identify them as the same instance if they have the smallest distance across views and satisfy the epipolar constraint. Without ground plane assumption, our method achieves accurate re-identification.

Our model has been extensively tested on the challenging MessyTable dataset [20], which contains complex scenes with multi-view multi-class objects. The studies demonstrate that our model improves the detection performance (MODA) of faster R-CNN by +16% point. Moreover, our jointly optimized model outperforms the simple combination of detection and re-identification by +21% MODA and +25.9% AP, respectively. In summary, our contributions are as follows:

- We propose MVDet, an end-to-end learnable object detector that is capable of handling multi-class objects in multi-view scenarios. To the best of our knowledge, this is the first attempt in this domain.
- We have developed a novel algorithm that can re-identify objects across multi-view images under the epipolar geometry constraints, without relying on the ground plane assumption. Our method can be applied to various scenarios where objects are not on the ground plane, such as automatic checkout and unmanned stores.
- Our proposed MVdet outperforms the single-view detection model and separately optimized multi-view detection models by jointly optimizing the region proposal, re-identification, and classification networks, without the ground plane assumption.

2 Related works

2.1 Single-view object detection

Object detection on a single image has made significant progress with deep learning, with methods such as Faster R-CNN [23] proposing regions where objects are expected to be and performing classification on those regions. Other detectors like YOLO [24], SSD [25], and EfficientDet [26] combine both steps by simultaneously localizing and classifying objects. However, accurate object detection a single image is limited when objects are partially or completely occluded.

2.2 Multi-view re-identification

Multi-view re-identification research has primarily focused on person retrieval, with many studies exploiting the parts of the objects [27–30]. In contrast, FaceNet [22] has introduced a triplet loss to minimize the distance between an anchor and a positive input, while maximizing the distance between the anchor and a negative input. ASNet [20], which performs well on the MessyTable dataset, leverages context of instances in complex scenarios. However, these approaches significantly suffer when inaccurate single-view detection results are used as input since they are trained and optimized on accurate ground truth boxes.

2.3 Multi-view classification

Research on 3D object detection has focused on classifying a group of 2D images that represent a 3D object [31–35]. One relevant study to our paper is MVCNN [36], which combines the multi-view images and performs pooling for classification. However, this method only classifies bundles of images with the same class and does not provide complete multi-view detection since it does not locate individual objects.

2.4 Multi-view object detection

Roig et al. [16] have proposed multi-class object detection under multi-camera settings by applying conditional random fields to object detection results. In comparison with this study, which does not contain re-identification, several studies [17, 18, 21] have suggested the simultaneous localization and re-identification for the single class object in multi-view images. Baque et al. [17] have demonstrated the effectiveness of integrating CNN and conditional random field to improve the robustness to occlusion in multi-view multi-target detection. Chavdarova et al. [18] have used a predefined occlusion mask to partially mask input images during training and fuse multi-view features. Hou et al. [21] have proposed an anchor-free multi-view pedestrian detection using perspective transformation of the feature map. However, these studies assumed a ground plane and could not handle multi-class objects. There have also been approaches to recognize instances without ground plane assumption using image appearance and geometric information of cameras in the multi-view setting [19, 37], but they could not handle multi-class objects. To address these limitations, we propose a multi-view multi-class object detector without a ground plane assumption.

3 Method

In this section, we provide a detailed introduction to our method, which has two distinct characteristics: a network capable of learning multi-class multi-view object detection in an end-to-end manner, and an epipolar geometry-based re-identification algorithm without the assumption of a ground plane. As shown in Fig. 1, our model takes a multi-view

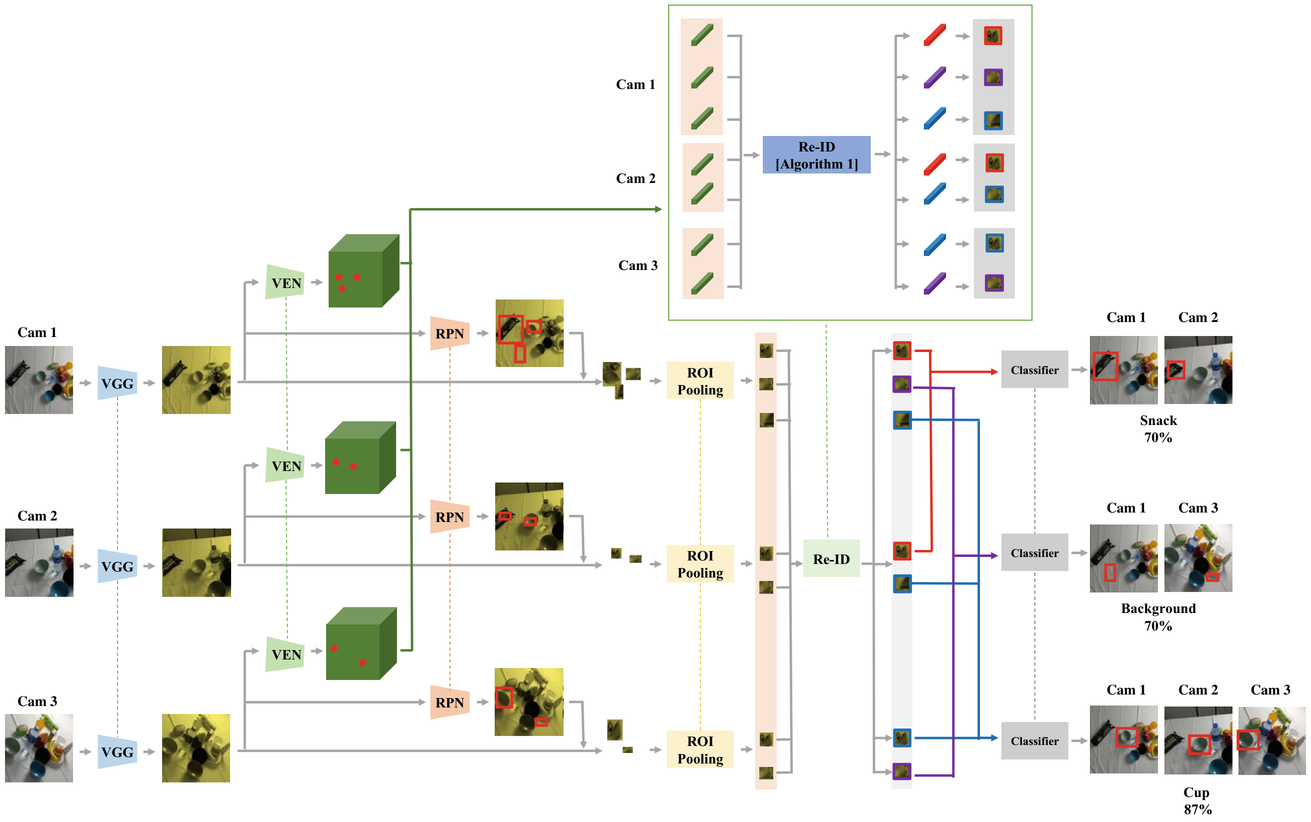

Fig. 1 MVDet is a multi-view multi-class object detector that comprises three networks: RPN, VEN, and classifier. All three networks share the initial features that are extracted from VGG. The RPN network generates region proposal boxes using the shared features, which are then passed through the ROI pooling layer and resized to the same size. Meanwhile, the VEN network extracts view embedding features from the shared features and then, selects the view embedding vectors corresponding to the region proposal boxes. Re-identification across the views is performed using these vectors and epipolar geometry, as specified in Algorithm 1. Finally, the classifier network determines the classes for the sets of re-identified boxes. The dotted lines indicate that the weights of the networks are shared for each camera

RGB image as input and outputs the position and class of the objects in each view. We explain the architecture of the multi-class multi-view object detector in Sect. 3.1 and the re-identification algorithm based on epipolar geometry in Sect. 3.2.

3.1 Architecture of MVDet

MVDet consists of a region proposal (RPN), view embedding (VEN), and classification networks, as shown in Fig. 1. These networks work together to detect, re-identify, and classify objects in multiple views simultaneously. In the following sections, we explain each network in detail.

3.1.1 RPN

In the first step of MVDet, the RPN generates box-shaped region proposals where objects are likely to be located, similar to the Faster R-CNN approach [23]. To extract the feature maps from the multi-view images that will be shared in the subsequent steps, the VGG network [38] is used. We choose VGG because it is a lightweight network, but other networks could be used as well. The weights of the VGG network are shared across all cameras to ensure memory efficiency and to extract features that are robust to changes in viewpoint. Following this, the RPN network locates regions in the feature map that corresponds to the predicted object locations, and the features within these regions are cropped and resized to a uniform size using pooling layer.

3.1.2 VEN

VEN is a network that extracts view embedding features from a shared feature map. When the region proposal boxes extracted by the RPN are provided across all views, the view embedding vector that corresponds to each region proposal box is indexed from the view embedding feature. The distances between these vectors across all views are used as the similarity between objects, and these distances are utilized for re-identification with epipolar geometry. This section provides a detailed explanation of VEN.

View embedding. VEN extracts the embedding feature map $\mathrm{Y} \in \mathbb{R}^{\mathrm{W\times H\times A\times F_2}}$ from the shared feature map X $\in \mathbb{R}^{\mathrm{W\times H\times F_1}}$, where W and H represent the number of rows and columns of grid cells in an image, and A denotes the number of anchors in a grid cell. F_1 represents the channel size of the shared feature, and F_2 is the dimension of the view embedding vector. The view embedding vector $y \in \mathbb{R}^{F_2}$ corresponding to a specific region proposal box is chosen from Y based on the coordinates of the grid cell and anchor index of that region proposal box. Indexing the view embedding feature map to find the embedding vectors that correspond to the region proposal boxes is a faster process than embedding all of the cropped region proposal images from the beginning. This speeds up the running time of the algorithm.

VEN training. Inspired by the approach used in FaceNet [22], our goal is to ensure that the embedding distance between region proposal boxes belonging to the same object is small, while the distance between boxes of different instances is large. In other words, a region proposal box B^a (anchor) of a specific object in a particular view should be closer to the region proposal boxes B^p (positive) of the same object in other views than to the region proposal boxes B^n (negative) of any other object. To achieve this, we aim to minimize the VEN loss, which is formulated as follows:

$$\sum_{i}^{N}[\|f(B_i^a)-f(B_i^p)\|_2^2-\|f(B_i^a)-f(B_i^n)\|_2^2+\alpha], \tag{1}$$

where f is a mapping function from view embedding feature map Y to view embedding vector using the index of the given region proposal box B. α is a margin between positive and negative pairs, and N is the number of samples.

To train the embedding vectors of the region proposal boxes using Eq. (1), we first sample an anchor in a specific view from the region proposal boxes that has the largest Intersection Over Union (IOU) with a ground truth object in that view. Then, we search for the positive and negative samples corresponding to this anchor in the other views. The positive sample is the region proposal box with the largest IOU with the anchor. The negative sample, on the other hand, is randomly selected from the region proposal boxes whose embedding distances to the anchor are longer than the embedding distance between the anchor and the positive sample.

VEN architecture. VEN architecture consists of two ReLU-activated convolutional layers, each with a 3×3 kernel and 512 output channels, and one sigmoid-activated convolutional layer with a 1×1 kernel and $A * F_2$ output channels. The resulting feature map $Z \in \mathbb{R}^{W\times H\times A*F_2}$ is then reshaped to the embedding feature map $Y \in \mathbb{R}^{W\times H\times A\times F_2}$, followed by L_2 normalization.

The technical specifications for VEN are as follows:

- The size of the shared feature map is W × H × 512 ($F_1 = 512$).

- The size of the embedding feature map is W × H × 9 × 128 ($A = 9, F_2 = 128$).
- The margin between positive and negative pairs (α) is set to 0.3.
- We exclude region proposal boxes with an IOU smaller than 0.3 during VEN training.
- We limit the number of VEN training samples to a maximum of 16 per iteration.

3.1.3 Classifier

After re-identifying the region proposal boxes using VEN, the next step is to classify them. For classification, we use a modified network based on faster R-CNN. Firstly, a 1×1 convolution operation is applied to the features of each region proposal box to reduce their channel size by 1/N, where N is the number of views. Then, the features of the same instances are concatenated. However, if a view is missing in a re-identified instance, the view is replaced with a black image.

Next, two fully connected layers with two dropout layers are used for pooling. The resulting feature map is passed through two parallel paths. The first path is a fully connected layer followed by softmax for object classification. The second path is a linearly activated fully connected layer for localization. The localization step refines the localization result obtained in the region proposal step.

Finally, after classification, a novel multi-view non-maximum suppression (NMS) is used. If the overlapping area between two objects of the same class is more than 0.3 in at least one view, the object with lower confidence is removed.

Our proposed classifier is designed to have shallow layers, which provides several advantages compared to existing multi-view classifiers with deep structures such as MVCNN [36]. One advantage is the efficient use of memory during training and inference, which is beneficial for practical applications. Moreover, our simple classifier can still achieve high accuracy because it utilizes the shared feature map that has already extracted the key features of objects, as opposed to starting from scratch for each view.

Additionally, our classifier includes a localization layer that refines the inaccurate region proposal results from RPN. This is in contrast to MVCNN, which only performs classification. The ability to refine the localization results can further improve the overall accuracy of our method.

3.2 Re-identification based on VEN and epipolar geometry

In this section, we explain our approach for re-identification without assuming a ground plane. We use the embedding vector from VEN and epipolar geometry for this task. We note that the term re-identification in this paper is slightly different from the term used in person re-identification. The latter refers to the task of retrieving person images in one view, given a query target person in another view. In contrast, re-identification in our paper refers to the task of associating the same instances when objects are detected across multiple views. Our approach to re-identification differs from person re-identification in that it deals with inaccurate detection results and associates objects in all views. The objective of this section is to describe our re-identification algorithm that leverages VEN and epipolar geometry.

To apply our re-identification algorithm, we first gather region proposal boxes from all views and select the top M boxes with the highest confidence scores. These M boxes are then defined as reference boxes, denoted as R_{top}. Given a reference box B_1 in the first view, we use its center coordinates to compute an epipolar line in the second view, using the intrinsic and extrinsic parameters of the camera. The pixel distances between all region proposal boxes and the epipolar line in the second view are then calculated to exclude unlikely region proposal boxes that are far from the epipolar line more than a certain threshold θ_1. This threshold is a hyperparameter, and we empirically determine its value. After excluding unlikely candidates, we identify the box B_2 with the smallest embedding distance from B_1 as the same instance.

Once a pair of matched region proposal boxes B_1 and B_2 are identified, two epipolar lines in the third view could be calculated from B_1 and B_2, respectively, using the camera's intrinsic and extrinsic parameters. A feasible match in the third view is a region proposal box that is close to the intersection of these two epipolar lines. We set a threshold θ_2 to exclude candidates that are too far from the intersection. Among the remaining region proposal boxes, the one with the shortest embedding distance from B_1 is considered as the same instance. This process is repeated for the rest of the views.

Algorithm 1 outlines the re-identification process. The EMD(a, b) function calculates the embedding distance between the region proposal boxes a and b. EPD_1(a, b) function calculates the pixel distance between the center coordinate of region proposal box a and the epipolar line derived from the region proposal box b. The EPD_2(a, b, c) function calculates the pixel distance between the center coordinate of region proposal box c and the intersection of two epipolar lines, which are derived from region proposal boxes a and b, respectively.

Algorithm 1 Re-identification

```
Require: $R_1$, $R_2$, ..., $R_N$ where each $R_i$ is a set of region proposal boxes
    for the $i_{th}$ view and N is the number of the views, $\theta_1$ and $\theta_2$ which are
    the thresholds to remove the candidates that do not meet the epipolar
    constraint.
Ensure: Re-identified region proposal boxes $I$
$R_{all} = \bigcup_{i=1}^{N} R_i$
$R_{top} = M$ region proposal boxes with the highest confidence in $R_{all}$
$I = []$
for all $a \in R_{top}$ do
    $I_{cur} = [a]$
    $i$ = view index of $a$
    $R_{rest} = R_{all}$ - $R_i$
    $\hat{b} = \underset{b \in R_{rest}}{\arg\min}\ \mathrm{EMD}(a, b)$ s.t. $\mathrm{EPD}_1(a, b) < \theta_1$
    if $\hat{b} \neq$ NULL then
        $I_{cur}$.append($\hat{b}$)
        $j$ = view index of $\hat{b}$
        for $k = 1, \ldots, N$ except for i and j do
            $\hat{c} = \underset{c \in R_k}{\arg\min}\ \mathrm{EMD}(a, c)$ s.t. $\mathrm{EPD}_2(a, \hat{b}, c) < \theta_2$
            if $\hat{c} \neq$ Null then
                $I_{cur}$.append($\hat{c}$)
            end if
        end for
    end if
    $I$.append($I_{cur}$)
end for
```

4 Experiments

4.1 Dataset

MessyTable [20] is a multi-camera object dataset designed for the instance re-identification task. It consists of 120 object classes with varying sizes, colors, and materials. The dataset comprises 5,579 scenes captured by nine synchronized cameras, with 6 to 67 instances randomly placed on a table under different lighting conditions and backgrounds. The camera poses are set randomly in 567 configurations. The scenes are categorized into three difficulty levels, with harder scenes featuring more occluded objects, similar-looking instances, or fewer instances in the overlapping field of cameras. A total of 50,211 images are labeled with 1,219,240 bounding boxes, each annotated by class and instance IDs. The dataset also provides the calibrated intrinsic and extrinsic camera parameters.

The annotations in MessyTable, including camera parameters, bounding boxes with class labels for objects, and instance IDs for each bounding box, are utilized to evaluate the effectiveness of the proposed multi-view detection model. However, due to high memory usage, only 16,737 images from three cameras are utilized in our experiments. The training, validation, and test sets are randomly divided in a 1:1:1 ratio, following the original setup of MessyTable.

4.2 Implementation details

Our multi-view object detection model is based on Keras-FasterRCNN, which is an implementation of single-view Faster R-CNN using Keras. The source codes for Keras-FasterRCNN and Keras can be found at https://github.com/you359/Keras-FasterRCNN and https://github.com/keras-team/keras, respectively. For our implementation, we choose VGG16 [38], which was pre-trained on ImageNet [39], as the backbone network. The anchor boxes used in our model have sizes of [128, 256, 512] and aspect ratios of [1:1, 1:2, 2:1]. Region proposal boxes are resized to 7×7 in the ROI pooling layer. We used dropout layers [40] with a drop probability of 0.5. We used the ADAM optimizer [41] with an initial learning rate of 0.00001 for the region proposal, view embedding, and classification networks. All of our experiments were conducted on a single NVIDIA 1080 Ti GPU.

4.3 Metrics

To evaluate the effectiveness of our model, we employ detection and re-identification metrics. For detection, we use MODA, MODP, and F1-score (F1). MODA takes into account both false positives and false negatives, while MODP measures the localization error of true positives [42]. F1 is a harmonic mean of recall and precision. We use a threshold of 0 to compute MODA and F1. Re-identification performance is evaluated using AP and FPR-95, as in [20]. AP is calculated as a weighted sum of precisions, counting the number of positive and negative matches at each threshold. FPR-95, commonly used in patch-based matching, is the false positive rate when recall is 95% [43], and complements AP.

4.4 Baselines

In this section, we explain the baselines used in our experiments. To the best of our knowledge, there are no existing studies on multi-view multi-class object detection without the ground plane assumption. Therefore, we employ heuristics or deep learning-based methods for single-view detection, re-identification, and multi-view classification. We then integrate the results from each step to generate the final results for multi-view detection, which serve as the baseline for our experiments. We note that the performance of single-view faster R-CNN, which forms the backbone of our model, represents the lower bound in our experiments.

To re-identify the objects detected in single views, we use ASNet and TripleNet, which are state-of-the-art methods for re-identification on the MessyTable dataset. ASNet uses neighboring information around a bounding box when the appearance features of a pair of boxes are dissimilar, while TripleNet is a feature extractor trained with triplet loss [22] that measures the feature similarity of a pair of boxes. In our experiments, we train ASNet and TripleNet on the ground truth labels and refine the similarity scores using epipolar geometry in the inference step for the fair comparison, following the methodology of FaceNet [22].

To evaluate the effect of ground plane assumption, we also utilize homographic projection as another method of re-identification. Homographic projection is a widely used technique in multi-view pedestrian detection [21] and tracking [44, 45], which is based on the assumption of a ground plane. It projects the coordinates of objects from each view onto a 2D ground plane and determines whether they represent the same object based on the distance between their projected locations. We calculate the similarity scores as the reciprocal of the distances between the projected locations.

Since ASNet, TripleNet and homographic projection generate only the similarity scores between two boxes, a method for the complete re-identification on the boxes across all views is required. Therefore, we build a graph, where the nodes represent the detection boxes, and the edges are weighted by the similarity scores of the corresponding nodes. We only include edges with similarity scores higher than 0.5 to ensure reliable re-identification. Next, a maximum bipartite graph matching is applied to identify the valid paths in the graph, where all nodes in a valid path are considered as the same instances. Finally, the paths with

Table 1 Detection and re-identification results of baselines and our method on MessyTable. Our method is most effective for detection and re-identification

Method	Detection			Re-ID	
	MODA↑	MODP↑	F1↑	AP↑	FPR-95↓
SVDet	0.35	0.7	0.66		
SVDet+Homograpy+Majority	0.04	0.69	0.48	0.187	0.907
SVDet+Homograpy+MVCNN	0.04	0.71	0.46	0.119	0.9
SVDet+ASNet+MVCNN	0.3	0.7	0.63	0.441	0.852
SVDet+TripleNet+MVCNN	0.3	0.7	0.63	0.482	0.852
SVDet+ASNet+Majority	0.29	0.69	0.63	0.535	0.865
SVDet+TripleNet+Majority	0.3	0.69	0.63	0.575	0.858
MVDet(Ours)	**0.51**	**0.71**	**0.7**	**0.834**	**0.548**
GT SVDet+TripleNet+MVCNN	0.73	0.99	0.86	0.783	0.71
GT SVDet+TripleNet+Majority	0.87	0.99	0.93	0.862	0.622
GT SVDet+GT ReID+MVCNN	0.87	1.0	0.93	1.0	0
GT SVDet+GT ReID+Majority	1.0	1.0	1.0	1.0	0

The values highlighted in bold indicate the highest performance among the models

Table 2 Detection and re-identification results on subsets of different scene complexity

Subsets	Method	Detection			Re-ID	
		MODA↑	MODP↑	F1↑	AP↑	FPR-95↓
Easy	SVDet	0.44	0.7	0.71		
	SVDet+TripleNet+Majority	0.41	0.7	0.69	0.735	0.814
	MVDet (Ours)	**0.62**	**0.72**	**0.78**	**0.898**	**0.505**
Medium	SVDet	0.41	0.7	0.7		
	SVDet+TripleNet+Majority	0.33	0.7	0.65	0.613	0.844
	MVDet (Ours)	**0.54**	**0.71**	**0.72**	**0.838**	**0.568**
Hard	SVDet	0.36	0.68	0.65		
	SVDet+TripleNet+Majority	0.19	0.68	0.56	0.415	0.892
	MVDet (Ours)	**0.39**	**0.68**	**0.61**	**0.757**	**0.692**

The values highlighted in bold indicate the highest performance among the models

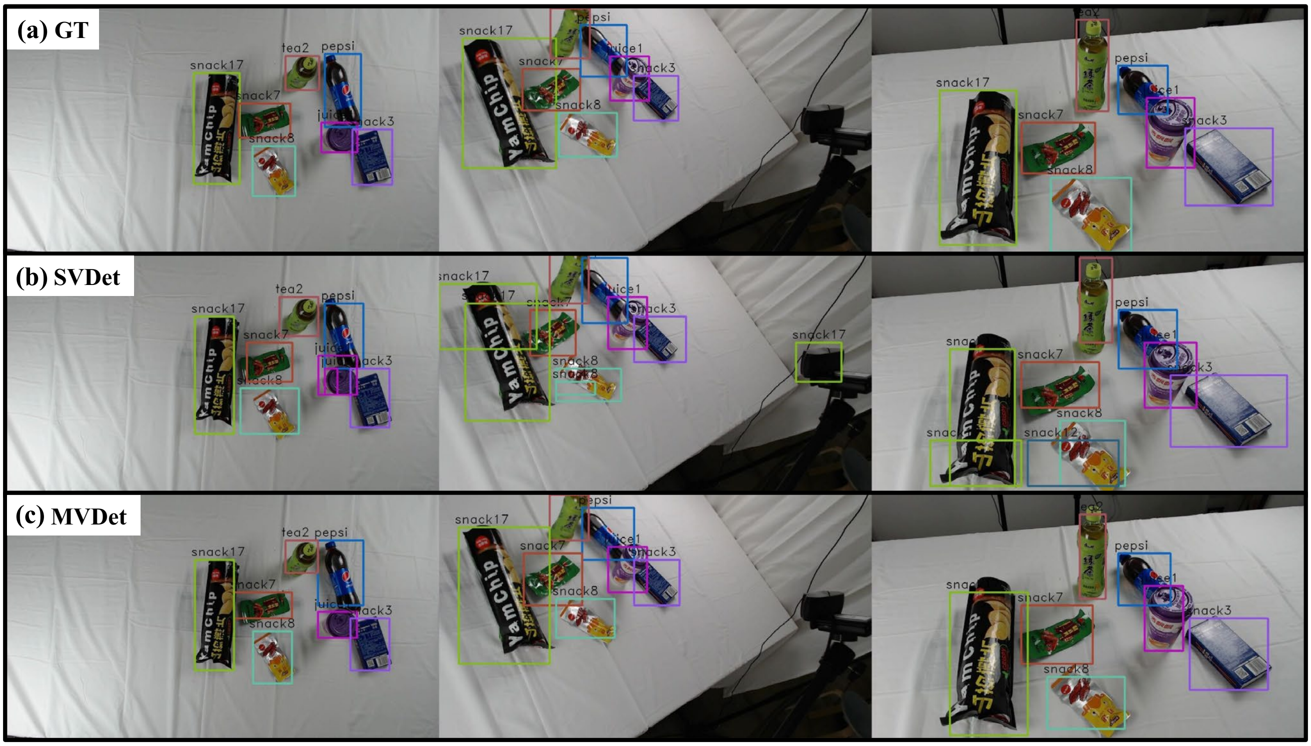

Fig. 2 (**a**) Ground truth and detection results of (**b**) SVDet and (**c**) MVDet. SVDet generates false positives due to the repeated detection of the same object and detection of the background region. In contrast, MVDet alleviates these issues and produces accurate detection results

the repeated nodes and subset paths are eliminated to avoid duplicate re-identification of single instances.

To assign a class label to the above re-identified instance, we use majority voting and MVCNN [36]. Majority voting method selects the class label for the re-identified instance based on the class that appears in the highest number of bounding boxes. The confidence score is calculated by averaging the detection scores of the boxes whose class is the majority class. If there is a tie between multiple classes, the final class is randomly selected from the tied classes.

MVCNN is a deep learning-based method that classifies multi-view images of a single object. However, in the original MVCNN, the number of input views must be fixed. Therefore, if a view is missing from the previous re-identification step, we compensate for the missing view by duplicating the other views and feeding them as inputs to MVCNN with the fixed number of views. Additionally, to enable background classification, which is not possible in the original MVCNN, we cut out patches without objects from the images and use them as background image samples during training. Finally, we also apply non-maximum suppression (NMS), which is also used in MVDet, to both the majority voting and MVCNN.

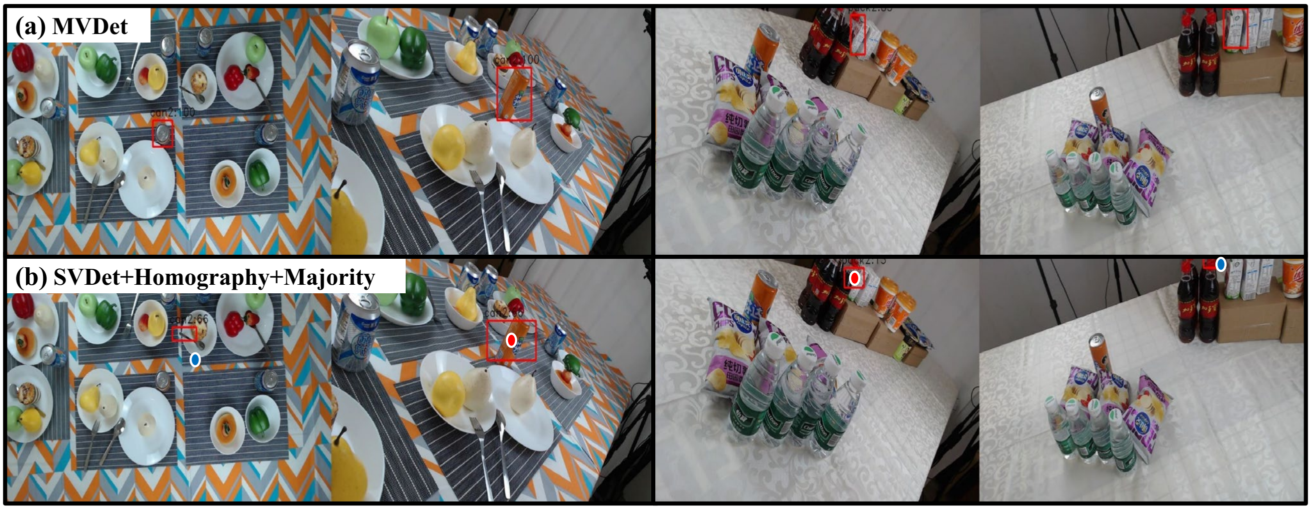

Fig. 3 First and second rows display the detection results of (**a**) MVDet and (**b**) SVDet+Homography+Majority, respectively. Due to varying viewpoints in the first two columns and objects positioned at the elevated surface in the last two columns, the accuracy of homolographic projection may decrease. The reason for this is that the center point of an object in one view, indicated by a red point, can be projected to a blue point in another view that is far from the original location of the same instance. Consequently, the object that is closer to the projected point may be incorrectly identified as the same object

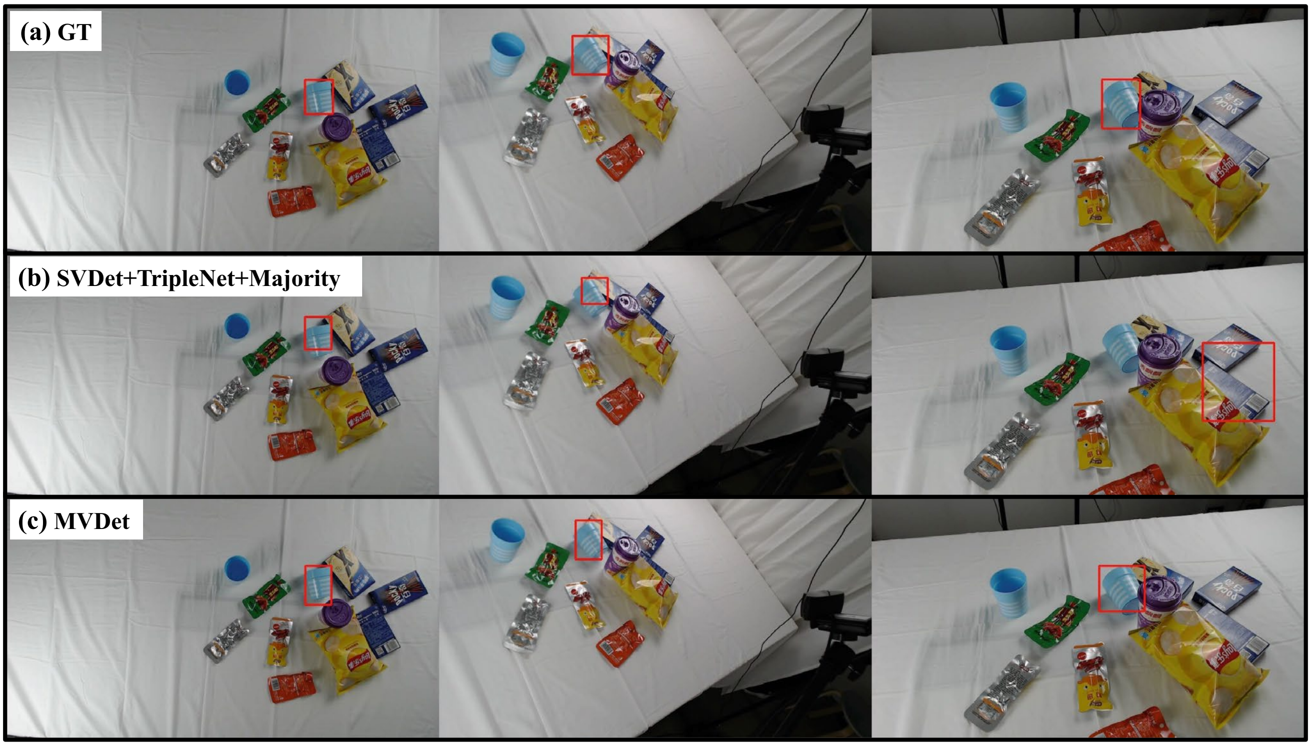

Fig. 4 **a** Ground truth and re-identification results of (**b**) SVDet+TripleNet+Majority and **c** MVDet. If the SVDet boxes deviates slightly from the object, SVDet+TripleNet+Majority generates a false re-identification

Table 1 demonstrates that our MVDet model achieves a MODA performance gain of +16% over single-view faster R-CNN (SVDet). The reason for this improvement is illustrated in Fig. 2, where we can see that MVDet successfully reduces many false positives that occur in SVDet. This is achieved by utilizing re-identification and classification methods to remove false region proposal boxes.

The results from Table 1 and Fig. 3 indicate that SVDet+Homography+Majority and SVDet+Homography+MVCNN have poor performance due to the presence of viewpoint variation and elevated surfaces. Homographic projection assumes that an object is located at the center of the occupied area on the ground. Therefore, the top-down view should represent the object with the center coordinates of the bounding box, while in other views, it should use the bottom center coordinates of the bounding box to represent the object. If multiple views with different coordinate matching methods are combined, re-identification performance deteriorates, as illustrated in Fig. 3. Additionally, objects not on the ground plane can hinder accurate homographic projection as they violate the ground plane assumption.

Also, MVDet surpasses the detection and re-identification performance of SVDet+TripleNet+Majority by +21% MODA and +25.9% AP, respectively, as shown in Table 1. It is worth noting that the detection accuracy of the separate multi-view detection models is even lower than that of SVDet due to the lack of robustness to false positives generated by SVDet. This is because the re-identification and classification networks are trained using error-free ground truth boxes, which are not robust to the false positives produced by SVDet. Therefore, the performance of GT SVDet+TripleNet+Majority, which assumes that the SVDet results are accurate, is much better than SVDet+TripleNet+Majority. In other words, the performance of re-identification and classification networks heavily relies on the accuracy of SVDet results. In contrast, MVDet uses region proposal boxes instead of ground truth boxes in the re-identification training process, making it robust against localization errors in the boxes, as shown in Fig. 4.

Furthermore, since the accuracy of the MVCNN model is highly dependent on the accuracy of the SVDet and ReID results, GT SVDet+GT ReID+MVCNN achieves 87% MODA, which is +14% higher than SVDet+TripleNet+MVCNN in Table 1. However, GT SVDet+GT TripleNet+Majority performs better than GT SVDet+GT TripleNet+MVCNN in Table 1 because MVCNN often misclassifies instances with minor localization errors. Therefore, the MVCNN model, which is trained on the ground truth boxes, performs well only when the accuracy of the SVDet and ReID results is guaranteed.

We conducted additional validation of our model on Easy, Medium, and Hard test sets, which are divided based on the scene complexity. The more difficult the test set, the more it contains similar objects and occlusions. As shown in Table 2, MVDet outperforms SVDet and SVDet+TripleNet+Majority on all test sets.

5 Conclusion

This paper addresses the problem of multi-view multi-class object detection that does not assume a ground plane. The proposed MVDet model performs region proposal, re-identification, and classification simultaneously in an end-to-end manner, using faster R-CNN and triplet loss. The model also employs an epipolar constraint-based re-identification algorithm to avoid the ground plane assumption. Experimental results on the MessyTable dataset demonstrate that MVDet outperforms both single-view detectors and separate multi-view detectors in terms of detection and re-identification accuracy. Overall, the proposed MVDet model presents a promising solution to the multi-view multi-class object detection problem in the absence of a ground plane assumption.

Data availibility Publicly available data are used.

Declarations

Conflict of interest The authors have no relevant financial or nonfinancial interests to disclose.

References

1. Hameed K, Chai D, Rassau A (2021) Class distribution-aware adaptive margins and cluster embedding for classification of fruit and vegetables at supermarket self-checkouts. Neurocomputing 461:292–309
2. Rigner A (2019) Ai-based machine vision for retail self-checkout system. Master's Theses in Mathematical Sciences
3. Mozaffari S, Al-Jarrah OY, Dianati M, Jennings P, Mouzakitis A (2020) Deep learning-based vehicle behavior prediction for

autonomous driving applications: a review. IEEE Trans Intel Transp Syst 23(1):33–47

4. Pierson HA, Gashler MS (2017) Deep learning in robotics: a review of recent research. Adv Robot 31(16):821–835
5. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In. In: Proceedings of the IEEE international conference on computer vision, pp. 2961–2969
6. Redmon J, Farhadi A (2018) Yolov3: an incremental improvement. arXiv preprint arXiv:1804.02767
7. Lin T-Y, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988
8. Noh J, Lee S, Kim B, Kim G (2018) Improving occlusion and hard negative handling for single-stage pedestrian detectors. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 966–974
9. Wang A, Sun Y, Kortylewski A, Yuille AL (2020) Robust object detection under occlusion with context-aware compositionalnets. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 12645–12654
10. Kortylewski A, Liu Q, Wang A, Sun Y, Yuille A (2021) Compositional convolutional neural networks: a robust and interpretable model for object recognition under occlusion. Int J Comput Vis 129(3):736–760
11. Song S, Xiao J (2014) Sliding shapes for 3d object detection in depth images. In: European conference on computer vision, Springer. pp. 634–651
12. Wang T, He X, Barnes N (2013) Learning structured Hough voting for joint object detection and occlusion reasoning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1790–1797
13. Qi CR, Liu W, Wu C, Su H, Guibas LJ (2018) 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
14. Ye M, Xu S, Cao T (2020) Hvnet: hybrid voxel network for lidar based 3d object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1631–1640
15. Zhou Y, Sun P, Zhang Y, Anguelov D, Gao J, Ouyang T, Guo J, Ngiam J, Vasudevan V (2020) End-to-end multi-view fusion for 3d object detection in lidar point clouds. In: Conference on Robot Learning, PMLR pp. 923–932
16. Roig G, Boix X, Shitrit HB, Fua P (2011) Conditional random fields for multi-camera object detection. In: 2011 International Conference on Computer Vision, IEEE. pp. 563–570
17. Baqué P, Fleuret F, Fua P (2017) Deep occlusion reasoning for multi-camera multi-target detection. In: Proceedings of the IEEE international conference on computer vision, pp. 271–279
18. Chavdarova T, Fleuret F (2017) Deep multi-camera people detection. In: 2017 16th IEEE international conference on machine learning and applications (ICMLA), IEEE. pp. 848–853
19. Nassar AS, D'aronco S, Lefèvre S, Wegner JD (2020) Geograph: Graph-based multi-view object detection with geometric cues end-to-end. In: European conference on computer vision, Springer. pp. 488–504
20. Cai Z, Zhang J, Ren D, Yu C, Zhao H, Yi S, Yeo CK, Change Loy C (2020) Messytable: instance association in multiple camera views. In: European conference on computer vision, Springer. pp. 1–16
21. Hou Y, Zheng L, Gould S (2020) Multiview detection with feature perspective transformation. In: European conference on computer vision, Springer. pp. 1–18.
22. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 815–823
23. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: towards realtime object detection with region proposal networks. arXiv preprint arXiv:1506.01497
24. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788
25. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC (2016) Ssd: single shot multibox detector. In: European Conference on Computer Vision, Springer. pp. 21–37
26. Tan M, Pang R, Le QV (2020) Efficientdet: scalable and efficient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10781–10790
27. Zhao L, Li X, Zhuang Y, Wang J (2017) Deeply-learned partaligned representations for person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp. 3219–3228
28. Wang G, Yuan Y, Chen X, Li J, Zhou X (2018) Learning discriminative features with multiple granularities for person re-identification. In: Proceedings of the 26th ACM international conference on multimedia, pp. 274–282
29. Sun Y, Zheng L, Yang Y, Tian Q, Wang S (2018) Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European conference on computer vision (ECCV), pp. 480–496
30. Zhao H, Tian M, Sun S, Shao J, Yan J, Yi S, Wang X, Tang X (2017) Spindle net: person re-identification with human body region guided feature decomposition and fusion. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1077–1085
31. Xiang Y, Choi W, Lin Y, Savarese S (2015) Data-driven 3d voxel patterns for object category recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1903–1911
32. Chen X, Kundu K, Zhu Y, Berneshawi AG, Ma H, Fidler S, Urtasun R (2015) 3d object proposals for accurate object class detection. Adv Neural Inf Process Syst. 28
33. Chen X, Kundu K, Zhang Z, Ma H, Fidler S, Urtasun R (2016) Monocular 3d object detection for autonomous driving. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2147–2156
34. Zia MZ, Stark M, Schiele B, Schindler K (2013) Detailed 3d representations for object recognition and modeling. IEEE Trans Pattern Anal Mach Intell 35(11):2608–2623
35. Zeeshan Zia M, Stark M, Schindler K (2014) Are cars just 3d boxes?-jointly estimating the 3d shape of multiple objects. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3678–3685
36. Su H, Maji S, Kalogerakis E, Learned-Miller E (2015) Multiview convolutional neural networks for 3d shape recognition. In: Proceedings of the IEEE international conference on computer vision, pp. 945–953
37. Nassar AS, Lefèvre S, Wegner JD (2019) Simultaneous multiview instance detection with learned geometric soft-constraints. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 6559–6568
38. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556

39. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE pp. 248–255
40. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958
41. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980
42. Kasturi R, Goldgof D, Soundararajan P, Manohar V, Garofolo J, Bowers R, Boonstra M, Korzhova V, Zhang J (2008) Framework for performance evaluation of face, text, and vehicle detection and tracking in video: data, metrics, and protocol. IEEE Trans Pattern Anal Mach Intell 31(2):319–336
43. Han X, Leung T, Jia Y, Sukthankar R, Berg AC (2015) Matchnet: unifying feature and metric learning for patch-based matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3279–3286
44. Xu Y, Liu X, Liu Y, Zhu S-C (2016) Multi-view people tracking via hierarchical trajectory composition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4256–4265
45. Xu Y, Liu X, Qin L, Zhu S-C (2017) Cross-view people tracking by scene-centered spatio-temporal parsing. In: Proceedings of the AAAI conference on artificial intelligence, vol. 31