Multiview Detection with Feature Perspective Transformation

Yunzhong Hou, Liang Zheng, and Stephen Gould

Australian National University
Australian Centre for Robotic Vision
{firstname.lastname}@anu.edu.au

Abstract. Incorporating multiple camera views for detection alleviates the impact of occlusions in crowded scenes. In a multiview system, we need to answer two important questions when dealing with ambiguities that arise from occlusions. First, how should we aggregate cues from the multiple views? Second, how should we aggregate unreliable 2D and 3D spatial information that has been tainted by occlusions? To address these questions, we propose a novel multiview detection system, MVDet. For multiview aggregation, existing methods combine anchor box features from the image plane, which potentially limits performance due to inaccurate anchor box shapes and sizes. In contrast, we take an anchor-free approach to aggregate multiview information by projecting feature maps onto the ground plane (bird's eye view). To resolve any remaining spatial ambiguity, we apply large kernel convolutions on the ground plane feature map and infer locations from detection peaks. Our entire model is end-to-end learnable and achieves 88.2% MODA on the standard Wildtrack dataset, outperforming the state-of-the-art by 14.1%. We also provide detailed analysis of MVDet on a newly introduced synthetic dataset, MultiviewX, which allows us to control the level of occlusion. Code and MultiviewX dataset are available at https://github.com/hou-yz/MVDet.

Keywords: multiview detection, anchor-free, feature perspective transformation, fully convolutional

1 Introduction

Occlusion is a fundamental issue that confronts many computer vision tasks. Specifically, in detection problems, occlusion introduces great difficulties and many methods have been proposed to address it. Some methods focus on the single view detection problem, *e.g.*, part-based detection [33,23,44], loss design [42,36], and learning non-maximum suppression [13]. Other methods jointly infer objects from multiple cues, *e.g.*, RGB-D [10,12,25], LIDAR point cloud [6], and multiple RGB camera views [7,3]. In this paper, we focus on pedestrian detection from multiple RGB camera views (multiview).

Multiview pedestrian detections usually have images from multiple synchronized and calibrated cameras as input [7,27,3]. These cameras focus on the same

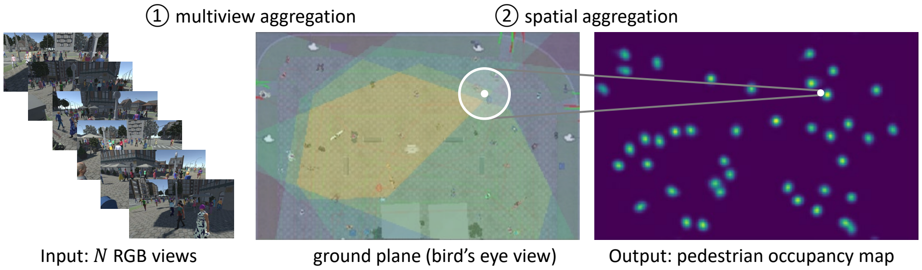

Fig. 1. Overview of the multiview pedestrian detection system. **Left**: the system takes synchronized frames from N cameras as input. **Middle**: the camera field-of-views overlaps on the ground plane, where the multiview cues can be aggregated. **Right**: the system outputs a pedestrian occupancy map (POM). There are two important questions here. **First**, how can we aggregate multiple cues. **Second**, how can we aggregate spatial neighbor information for joint consideration (large white circle), and make a comprehensive decision for pedestrian occupancy (small white circle).

area, and have overlapping field-of-view (see Fig. 1). Camera calibrations provide the matching between 2D image coordinate (u, v) and 3D world location (x, y, z). We refer to points with $z = 0$ in the 3D world as being on the ground plane (bird's eye view). For each point on the ground plane, based on 3D human width and height assumption, its corresponding bounding box in multiple views can be calculated via projection and then stored. Since the bounding boxes can be retrieved via table lookup, multiview pedestrian detection tasks usually evaluate pedestrian occupancy on the ground plane [7,3].

Addressing the ambiguities from occlusions and crowdedness is the main challenge for multiview pedestrian detection. Under occlusion, it is difficult to determine if a person exists in a certain location, or how many people exist and where they are. To solve this, one must focus on two important aspects of multiview detection: first, multiview aggregation and, second, spatial aggregation (Fig. 1). Aggregation of multiview information is essential since having multiple views is the main difference between monocular-view detection and multiview detection. Previously, for a given ground plane location, multiview systems usually choose an anchor-based multiview aggregation approach and represent certain ground plane location with multiview anchor box features [4,1,17]. However, researchers find the performance of anchor-based detection systems might be limited by inaccurate anchor boxes in monocular view systems [45,16,40], while multiview anchor boxes calculated from pre-defined human 3D height and width might also be inaccurate. Aggregation of spatial neighbors is also vital for occlusion reasoning. Previous methods [7,27,1] usually adopt conditional random field (CRF) or mean-field inference to jointly consider the spatial neighbors. These methods usually requires specific potential terms design in CRF or additional operations outside the CNN forward pass during inference.

In this paper, we propose a simple yet effective method, MVDet, that has heretofore not been explored in the literature for multiview detection. First, for *multiview aggregation*, as inaccurate anchor boxes can limit system performance [45,16,40], rather than anchor-based approach, MVDet choose an anchor-free approach. For it to work on multiview systems, MVDet projects the convolution feature map via perspective transformation and concatenates the multiple projected feature maps. Second, for *spatial aggregation*, to minimize human design and operations outside of CNN, instead of CRF or mean-field inference, MVDet adopts an fully convolutional solution. It applies (learned) convolutions on the aggregated ground plane feature map, and use the large receptive field to jointly consider ground plane neighboring locations. The proposed fully convolutional MVDet can be trained in an end-to-end manner.

We demonstrate the effectiveness of MVDet on two large scale datasets. On Wildtrack, a real-world dataset, MVDet achieved 88.2% MODA [15], a 14.1% increase over previous state-of-the-art. On MultiviewX, a synthetic dataset, MVDet also achieves competitive results under multiple levels of occlusions.

2 Related Work

Monocular view detection. Detection is one of the most important problems in computer vision. Neural network based methods like the R-CNN family [9,8,26] achieve great performance. On pedestrian detection, some researchers detect pedestrian bounding boxes through head-foot point detection [30] and center and scale detection [21]. Occlusion handling in pedestrian detection draws great attention from the research community. Part-based detectors are very popular [23,33,22,42] since the occluded people are only partially observable. Hosang *et al.* [13] learn the non-maximal suppression for occluded pedestrians. Repulsion loss [36] is proposed to repulse bounding boxes.

3D object understanding with multiple information sources. Incorporating multiple information sources, such as depth, point cloud, and other RGB camera views is studied for 3D object understanding. For multiple view 3D object classification, Su *et al.* [31] use maximum pooling to aggregate the features from different 2D views. For 3D object detection, aggregating information from RGB image and LIDAR point cloud are widely studied. Chen *et al.* [5] investigate 3D object detection with stereo image. Later, Chen *et al.* first generate 3D proposals and then aggregate multiview information for each proposal [6]. View aggregation for 3D anchors is studied in [17], where the researchers extract features for every 3D anchor from RGB camera and LIDAR bird's eye view via a table lookup of 3D anchor projection. Liang *et al.* [19] calculate the feature for each point from bird's eye view as multi-layer perceptron (MLP) output from camera view features of K nearest neighbor LIDAR points. Frustum PointNets [25] first generate 2D bounding boxes proposal from RGB image, then extrude them to 3D viewing frustums to aggregate information.

Multiview pedestrian detection. In multiview pedestrian detections, first, aggregating information from multiple RGB cameras is important. In [4,1], researchers fuse multiple information source for multiview 2D anchors. Given fixed assumption of human width and height, all ground plane locations (anchors) and their corresponding multiview 2D bounding boxes are first calculated and stored in a lookup table. Then, researchers in [4,1] use this lookup table to calculate the features for all ground plane anchors. In [7,38,27], single view detection results are fused instead. Second, in order to aggregate spatial neighbor information, mean-field inference [7,1] and conditional random field (CRF) [27,1] are exploited. For mean-field inference [7,1], ideal 2D images under certain occupancy are first estimated, and then compared with the real multiview inputs. In [7,1], the overall occupancy in the scenario is cast as an energy minimization problem and solved with CRF. Baque *et al.* [1] construct higher-order potentials as consistency between CNN estimations and generated ideal images, and train the CRF with CNN in a combined manner. Their CNN-CRF combined method achieved state-of-the-art performance on Wildtrack dataset [3].

Geometric transformation in deep learning. Geometric transformations such as affine transformation and perspective transformation can model many phenomena in computer vision, and can be explicitly calculated with a fixed set of parameters. Jaderberg *et al.* [14] propose Spatial Transformer Network that learns the affine transformation parameters for translation and rotation on the 2D RGB input image. Wu *et al.* [37] estimate the projection parameters and project 2D key points from the 3D skeleton. Yan *et al.* [39] translate one 3D volume to 2D silhouette via perspective transformation. Geometry-aware scene text detection is studied in [35] through estimating instance-level affine transformation. For cross-view image retrieval, Shi *et al.* [28] apply polar transformation to bring the representations closer in feature space.

3 Methodology

In this work, we focus on the occluded pedestrian detection problem in an multiview scenario and design MVDet for dealing with ambiguities. MVDet applys anchor-free *multiview aggregation* that alleviate influence from inaccurate anchor boxes in previous works [6,17,4,1], and fully convolutional *spatial aggregation* that does not rely on CRF or mean-field inference [7,27,1]. As shown in Fig. 2, MVDet takes multiple RGB images as input, and outputs the pedestrian occupancy map (POM) estimation. In the following sections, we will introduce the proposed multiview aggregation (Section 3.1), spatial aggregation (Section 3.2), and training and testing configurations (Section 3.3).

3.1 Multiview Aggregation

Multiview aggregation is a very important part of multiview systems. In this section, we explain the anchor-free aggregation method in MVDet that alleviate influence from inaccurate anchor boxes, and compare it with several alternatives.

Feature map extraction. In MVDet, first, given N images of shape $[H_i, W_i]$ as input (H_i and W_i denote the image height and width), the proposed architecture use a CNN to extract N C-channel feature maps (Fig. 2). Here, we choose ResNet-18 [11] for its strong performance and light-weight. This CNN

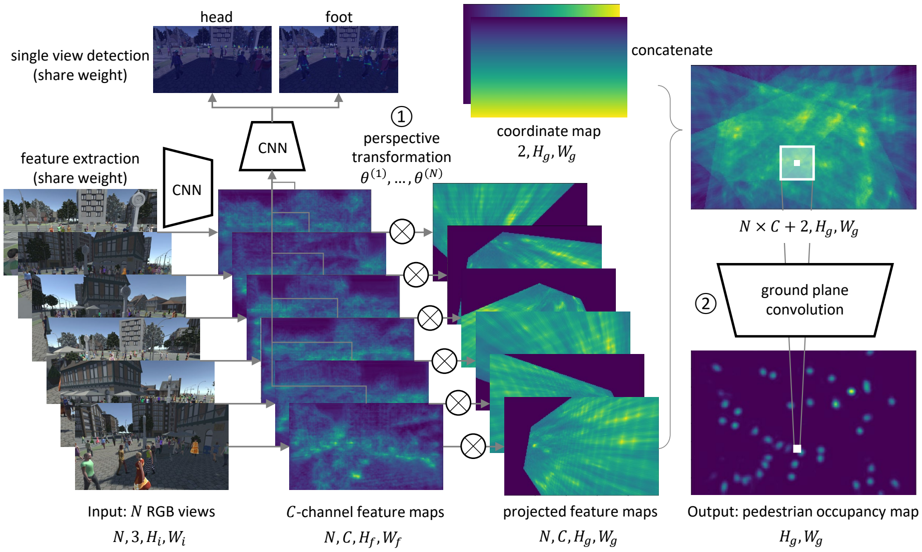

Fig. 2. MVDet architecture. First, given input images of shape $[3, H_i, W_i]$ from N cameras, the proposed network uses a CNN to extract C-channel feature maps for each input image. The CNN feature extractor here shares weight among N inputs. Next, we reshape the C-channel feature maps into a size of $[H_f, W_f]$, and run single view detection by detecting the head-foot pairs. Then, for **multiview aggregation** (circled 1), we take an anchor-free approach and use perspective transformation to project N feature maps according to corresponding camera calibrations $\theta^{(1)}, \ldots, \theta^{(N)}$, which results in N feature maps of shape $[C, H_g, W_g]$. For each ground plane location, we store its X-Y coordinates in a 2-channel coordinate map [20]. Through concatenating N projected feature maps with a coordinate map, we aggregate the ground plane feature map for the whole scenario at this moment (of shape $[N \times C + 2, H_g, W_g]$). At last, we apply large kernel convolutions on the ground plane feature map, so as to **aggregate spatial neighbor information** (circled 2) for the final occupancy decision.

calculates C-channel feature maps separately for N input images, while sharing weight among all calculations. In order to maintain a relatively high spatial resolution for the feature maps, we replace the last 3 strided convolutions with dilated convolutions [41]. Before projection, we resize N feature maps into a fixed size $[H_f, W_f]$ (H_f and W_f denote the feature map height and width). In each view, similar to [18,30], we then detect pedestrians as a pair of head-foot points with a shared weight single view detector.

Anchor-free feature aggregation. Previously, in detection tasks that have multiple cues, *e.g.*, 3D object detection and multiview pedestrian detection, anchor-based aggregation with anchor box features is commonly adopted [6,17,4,1]. Given assumption of 3D human height and width, for a ground plane location (red points in Fig. 3), one can create a lookup table for its corresponding multiview 2D anchor boxes (green boxes in Fig. 3) with projection. Then, one

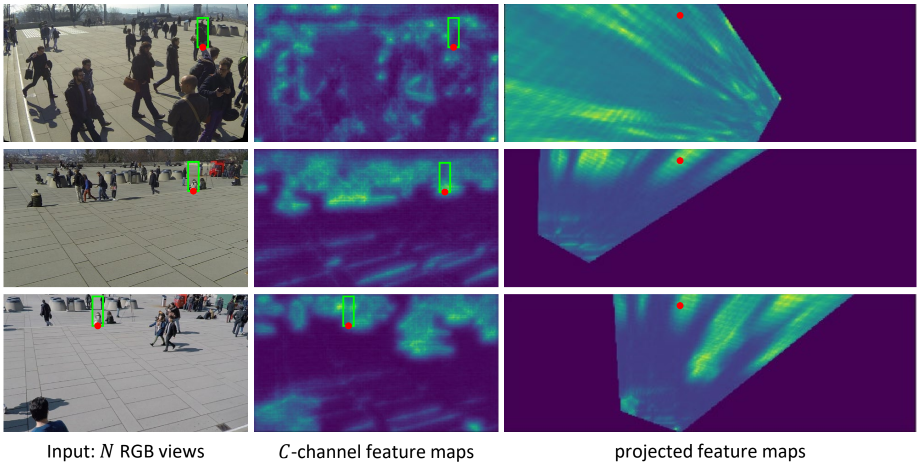

Fig. 3. Representing ground plane locations with feature maps projection or anchor boxes features. Red dots represent a certain ground plane location and its corresponding pixel in different views. Green bounding boxes refer to anchor boxes whose bottom center (human foot point) is at that ground plane location. As human targets might not be the same height as the default assumption (*e.g.*, sitting rather than standing), ROI-pooling for multiview anchor boxes might fail to provide the most accurate feature representation for that location. On the contrary, being anchor-free, for one ground plane location, feature map projection can retrieve accurate representations.

can use ROI-pooling [8] to represent the bounding box features and aggregate via concatenation to represent the corresponding ground plane location [4,1].

However, one potential problem is that the anchor boxes might *not* be the most accurate, which can potentially limit system performance [45,16,40]. As in Fig. 3, the lady in the white coat is sitting and only takes up half of the anchor box. In contrast, being anchor-free, the proposed feature maps projection method does not suffer from inaccurate anchor boxes. Also, even though the projection method does not pool feature from anchor boxes to represent 2D regions, each pixel in the feature map extract information from an adaptive region in its receptive field. As a result, ground plane feature maps constructed via anchor-free feature perspective transformation are more accurate, and still contains sufficient information from 2D images for detection.

Perspective transformation. We then project the feature maps with perspective transformation. Translating between 3D locations (x, y, z) and 2D image pixel coordinates (u, v) is done via the point-wise transformation

$$s \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = P_\theta \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} = A\,[R|\mathbf{t}] \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} & \theta_{14} \\ \theta_{21} & \theta_{22} & \theta_{23} & \theta_{24} \\ \theta_{31} & \theta_{32} & \theta_{33} & \theta_{34} \end{bmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}, \tag{1}$$

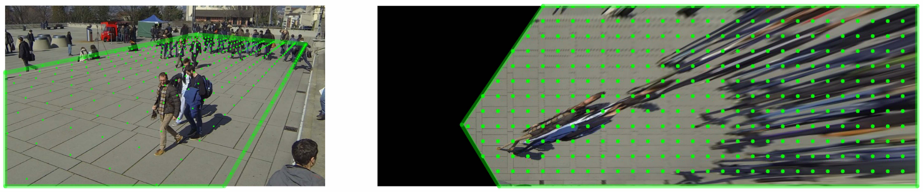

Fig. 4. Illustration of perspective transformation. Assuming all pixels are on the ground plane ($z = 0$), we can use a customized sampling grid to project a 2D image (left) to the ground plane (right). The remaining ground plane locations are padded with 0.

where s is a real-valued scaling factor, and P_θ is a 3×4 perspective transformation matrix. A is the 3×3 intrinsic parameter matrix. $[R|\mathbf{t}]$ is the 3×4 joint rotation-translation matrix, or extrinsic parameter matrix, where R specifies the rotation and $\mathbf{t}$ specifies the translation.

A point (pixel) from an image lies on a line in the 3D world. To determine exact 3D locations of image pixels, we consider a common reference frame: the ground plane, $z = 0$. For all 3D location $(x, y, 0)$ on this ground plane, the point-wise transformation can be written as

$$s \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = P_{\theta,0} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{14} \\ \theta_{21} & \theta_{22} & \theta_{24} \\ \theta_{31} & \theta_{32} & \theta_{34} \end{bmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}, \tag{2}$$

where $P_{\theta,0}$ denotes the 3×3 perspective transformation matrix that have the third column canceled from P_θ.

We quantize the ground plane locations into a grid of shape $[H_g, W_g]$, where H_g and W_g specifies ground plane grid height and width. For camera $n \in \{1, \ldots, N\}$ with calibration $\theta^{(n)}$, we can project the image onto the $z = 0$ ground plane through a customized sampling grid of shape $[H_g, W_g]$ based on Equation 2 (Fig. 4). The remaining (out-of-view) ground plane locations are padded with zero. We concatenate a 2-channel coordinate map [20] to specify the X-Y coordinates for ground plane locations (Fig. 2). Together with projected C-channel feature maps from N cameras, we have a $(N \times C + 2)$ channel ground plane feature map that is of shape $[H_g, W_g]$.

Different projection choices. For multiview aggregation, there are multiple choices for projection: we can project the RGB images, feature maps, or single view results (Fig. 5). First, RGB pixels on its own contains relatively little information, and much information is preserved in the spatial structures. However, projection breaks the spatial relationship between neighboring RGB pixels. As a result, this limits the performance of the multiview detector. Second, projecting the single view results (foot points) limits the information to be aggregated. In fact, in this setup, the system has no access to cues other than the single view detection results. Since single view results might not be accurate under occlusion (which is the reason for introducing multiple views), this setup can also limit the overall performance. In this paper, we propose to project

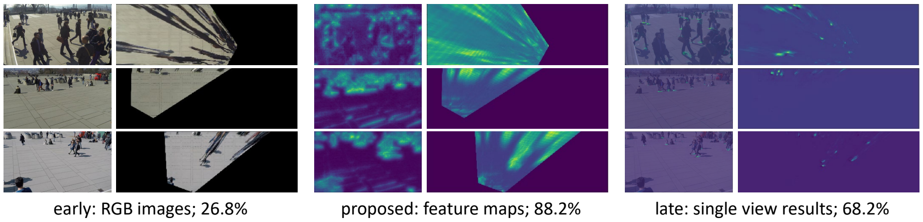
early: RGB images; 26.8%
proposed: feature maps; 88.2%
late: single view results; 68.2%

Fig. 5. Different network setups for multiview aggregation and their performance (MODA [15]). **Left**: early projection of RGB images breaks the spatial relationship between RGB pixels, which introduces great difficulties to the convolutions. **Right**: late projection of single view detection results (foot) limits the information to be aggregated. **Middle**: the proposed projection of feature maps not only are more robust to the pixel structure break (high-level semantic feature can represent information by itself, thus suffer less from structure break), but also contain more information.

the feature map. Compared to other choices, feature maps not only suffer less from the spatial structure break (since 2D spatial information have already been concentrated into individual pixels in feature maps), but also contain more information. As shown in Fig. 5, aggregation via feature maps projection achieves highest MODA [15] performance.

3.2 Spatial aggregation

In the previous section, we show that multiview information can be aggregated channel-wise through perspective transformation and concatenation. One remaining problem is how to aggregate information from spatial neighbors.

Occlusion are generated by human crowd within a certain area. To deal with the ambiguities, one can consider the certain area and the human crowd in that area jointly for an overall informed decision. Previously, CRFs and mean-field inference have been adopted to solve this problem. In this work, we propose an alternative with large kernel convolutions on the ground plane feature map. In fact, Zheng *et al.* [43] find that CNN can model some behavior and characteristics of CRFs. And Peng *et al.* [24] outperform CRFs with large kernel convolutions for semantic segmentation. We feed the $(N \times C + 2)$ channel ground plane feature map to convolution layers that have a relatively large receptive field, so as to jointly consider the ground plane neighbors. Here, we use three layers of dilated convolution for having minimal parameters while still keeping a larger ground plane receptive field. The last layer outputs an 1-channel $[H_g, W_g]$ pedestrian occupancy map (POM) $\tilde{\mathbf{g}}$ with no activation.

3.3 Training and Testing

In training, we train MVDet as a regression problem. Given ground truth pedestrian occupancy $\mathbf{g}$, similar to landmark detection [2], we use a Gaussian kernel $f(\cdot)$ to generate a "*soft*" ground truth target $f(\mathbf{g})$. In order to train the whole network, we use Euclidean distance $\|\cdot\|_2$ between network output $\tilde{\mathbf{g}}$ and "*soft*" target $f(\mathbf{g})$ as loss function,

$$\mathcal{L}_{\text{ground}} = \left\|\tilde{\mathbf{g}} - f(\mathbf{g})\right\|_2 . \tag{3}$$

We also include bounding box regression loss from N camera inputs as another supervision. The single view head-foot detection is also trained as a regression problem. For single view detection results $\tilde{\mathbf{s}}_{\text{head}}^{(n)}, \tilde{\mathbf{s}}_{\text{foot}}^{(n)}$ and the corresponding ground truth $\mathbf{s}_{\text{head}}^{(n)}, \mathbf{s}_{\text{foot}}^{(n)}$ in view $n \in \{1, ..., N\}$, the loss is computed as,

$$\mathcal{L}_{\text{single}}^{(n)} = \left\|\tilde{\mathbf{s}}_{\text{head}}^{(n)} - f\left(\mathbf{s}_{\text{head}}^{(n)}\right)\right\|_2 + \left\|\tilde{\mathbf{s}}_{\text{foot}}^{(n)} - f\left(\mathbf{s}_{\text{foot}}^{(n)}\right)\right\|_2 . \tag{4}$$

Combining ground plane loss $\mathcal{L}_{\text{ground}}$ and N single view losses $\mathcal{L}_{\text{single}}^{(n)}$, we have the overall loss for training MVDet,

$$\mathcal{L}_{\text{combined}} = \mathcal{L}_{\text{ground}} + \alpha \times \frac{1}{N}\sum_{n=1}^{N} \mathcal{L}_{\text{single}}^{(n)}, \tag{5}$$

where α is a hyper-parameter for singe view loss weight.

During testing, MVDet outputs a single-channel occupancy probability map $\tilde{\mathbf{g}}$. We filter the occupancy map with a minimum probability of 0.4, and then feed the ground plane location proposals to a standard NMS. This NMS has a Euclidean distance threshold of 0.5 meters, which is the same threshold for considering this location proposal as true positive in evaluation [3].

4 Experiment

4.1 Datasets

Datasets. We test on two multiview pedestrian detection datasets (Table 1).

The *Wildtrack* dataset includes 400 synchronized frames from 7 cameras, covering a 12 meters by 36 meters region. For annotation, the ground plane is quantized into a 480×1440 grid, where each grid cell is a 2.5-centimeter square. The 7 cameras capture images with a 1080×1920 resolution, and are annotated at 2 frames per second (fps). On average, there are 20 persons per frame in Wildtrack dataset and each locations in the scene is covered by 3.74 cameras.

The *MultiviewX* dataset is a new synthetic dataset collected for multiview pedestrian detection. We use Unity engine [34] to create the scenario. As for pedestrians, we use human models from PersonX [32]. MultiviewX dataset covers a slightly smaller area of 16 meters by 25 meters. Using the same 2.5-centimeter square grid cell, we quantize the ground plane into a 640×1000 grid. There are 6 cameras with overlapping field-of-view in MultiviewX dataset, each of which outputs a 1080×1920 resolution image. We also generate annotations for 400 frames in MultiviewX at 2 fps (same as Wildtrack). On average, 4.41 cameras are

Table 1. Datasets comparison for multiview pedestrian detection

	#camera	resolution	frames	area	crowdedness	avg. coverage
Wildtrack	7	1080×1920	400	12×36 m^2	20 person/frame	3.74 cameras
MultiviewX	6	1080×1920	400	16×25 m^2	40 person/frame	4.41 cameras

covering the same location. Being a synthetic dataset, there are various potential configurations for the scenario with free annotations. In the current setting, MultiviewX has 40 persons per frame, doubling the crowdedness in Wildtrack. If not specified, MultiviewX refers to this default setting.

Evaluation metrics. Following [3], we use the first 90% frames in both datasets for training, and the last 10% frames for testing. We report precision, recall, MODA, and MODP. MODP evaluates the localization precision, whereas MODA accounts for both the false positives and false negatives [15]. We use MODA as the primary performance indicator, as it considers both false positives and false negatives. A threshold of 0.5 meters is used to determine true positives.

4.2 Implementation Details

For memory usage concerns, we downsample the 1080×1920 RGB images to $H_i = 720, W_i = 1280$. We remove the last two layers (global average pooling; classification output) in ResNet-18 [11] for $C = 512$ channel feature extraction, and use dilated convolution to replace the strided convolution. This results in a $8\times$ downsample from the 720×1280 input. Before projection, we bilinearly interpolate the feature maps into shape of $H_f = 270, W_f = 480$. For Wildtrack and MultiviewX, the ground plane grid sizes are set as $H_g = 120, W_g = 360$ and $H_g = 160, W_g = 250$ with $4\times$ downsampling. In the ground plane grids, each cell represents a 10 centimeter square. For spatial aggregation, we use 3 convolutional layers with 3×3 kernels and dilation of $1, 2, 4$. This will increase the receptive field for each ground plane location (cell) to a 15×15 square cells, or a 1.5×1.5 square meters. In order to train MVDet, we use an SGD optimizer with a momentum of 0.5, L2-normalization of 5×10^{-4}. The weight α for single view loss is set to 1. We use the one-cycle learning rate scheduler [29] with the max learning rate set to 0.1, and train for 10 epochs with batch size set to 1. We finish all experiments on two RTX-2080Ti GPUs.

4.3 Method Comparisons

In Table 2, we compare multiview aggregation and spatial aggregation in different methods. For multiview aggregation, previous methods either project single view detection results [38,7] or use multiview anchor box features [4,1]. For spatial aggregation, clustering [38], mean-field inference [7,1], and CRF [1,27] are investigated. In order to compare against previous methods, we create the following variants for MVDet. To compare feature map aggregation methods (Section 3.1),

Table 2. Multiview aggregation and spatial aggregation in different methods

Method	Multiview aggregation	Spatial aggregation
RCNN & clustering [38]	detection results	clustering
POM-CNN [7]	detection results	mean-field inference
DeepMCD [4]	anchor box features	N/A
Deep-Occlusion [1]	anchor box features	CRF + mean-field inference
MVDet (project images)	RGB image pixels	large kernel convolution
MVDet (project results)	detection results	large kernel convolution
MVDet (w/o large kernel)	feature maps	N/A
MVDet	feature maps	large kernel convolution

Table 3. Performance comparison with state-of-the-art methods on multiview pedestrian detection datasets. * indicates that the results are from our implementation

	Wildtrack				MultiviewX			
Method	MODA	MODP	Prec.	Recall	MODA	MODP	Prec.	Recall
RCNN & clustering	11.3	18.4	68	43	18.7^*	46.4^*	63.5^*	43.9^*
POM-CNN	23.2	30.5	75	55	-	-	-	-
DeepMCD	67.8	64.2	85	82	70.0^*	73.0^*	85.7^*	83.3^*
Deep-Occlusion	74.1	53.8	**95**	80	75.2^*	54.7^*	$\mathbf{97.8}^*$	80.2^*
MVDet (project images)	26.8	45.6	84.2	33.0	19.5	51.0	84.4	24.0
MVDet (project results)	68.2	71.9	85.9	81.2	73.2	79.7	87.6	85.0
MVDet (w/o large kernel)	76.9	71.6	84.5	93.5	77.2	76.3	89.5	85.9
MVDet	**88.2**	**75.7**	94.7	**93.6**	**83.9**	**79.6**	96.8	**86.7**

we create “MVDet (w/o large kernel)”, which remove the large kernel convolutions. This variant is created as a direct comparison against DeepMCD [4], both of which do not include spatial aggregation. To compare different projection choices (Section 3.1), we include two variants that either project RGB image pixels “MVDet (project images)” or single view detection results “MVDet (project results)”. “MVDet (w/o large kernel)” also show the effectiveness of spatial aggregation. All aforementioned variants are anchor-free. All variants follow the same training protocol as MVDet.

4.4 Evaluation of MVDet

Comparison against state-of-the-art methods. In Table 3, we compare the performance of MVDet against multiple state-of-the-art methods on multiview pedestrian detection. Since there are no available codes for some of the methods, for a fair comparison on MultiviewX, we re-implement these methods to the best as we can. On Wildtrack dataset, MVDet achieves 88.2% MODA, a +14.1% increase over previous state-of-the-art. On MultiviewX dataset, MVDet achieves 83.9% MODA, an 8.7% increase over our implementation of Deep-Occlusion [1]. MVDet also achieves highest MODP and recall on both datasets, but slightly falls

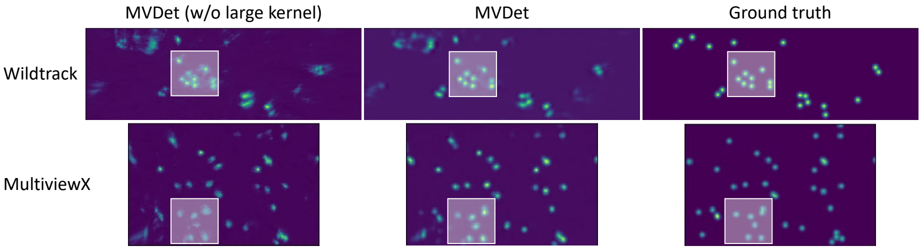

Fig. 6. Effectiveness of spatial aggregation via ground plane convolution. Compared to "MVDet (w/o large kernel)", MVDet outputs occupancy probabilities more similar to the ground truth, especially in highlighted areas.

behind Deep-Occlusion in terms of precision. It is worth mentioning that Deep-Occlusion outperforms MVDet in terms of precision, but falls behind in terms of recall. This shows that their CNN-CRF method is very good at suppressing the false positives, but sometimes has a tendency to miss a few targets.

Effectiveness of anchor-free multiview aggregation. Even without spatial aggregation, "MVDet (w/o large kernel)" achieves 76.9% MODA on Wildtrack dataset and 77.2% MODA on MultiviewX dataset. In fact, it slightly outperforms current state-of-the-art by +2.8% and +2.0% on two datasets. The high performance proves the effectiveness of our anchor-free aggregation via feature map projection. In Section 3.1, we hypothesize that inaccurate anchor boxes could possibly result in less accurate aggregated features and thus proposed an anchor-free approach. In Table 3, we prove the effectiveness of our anchor-free approach by comparing anchor-based DeepMCD [4] against anchor-free "MVDet (w/o large kernel)", both of which do not include spatial aggregation. The variant of MVDet outperforms DeepMCD by 9.1% on Wildtrack dataset, and 7.2% MODA on MultiviewX dataset, which demonstrates anchor-free feature maps projection can be a better choice for multiview aggregation in multiview pedestrian detection when the anchor boxes are not accurate.

Feature map projection brings less improvement over multiview anchor box features on MultiviewX dataset (+7.2% on MultiviewX compared to +9.1% on Wildtrack). This is because MultiviewX dataset has synthetic humans, whereas Wildtrack captures real-world pedestrians. Naturally, the variances of human height and width are higher in the real-world scenario, as synthetic humans are of very similar sizes. This suggests less accurate anchor boxes on average for the real-world dataset, Wildtrack. As a result, aggregation via feature map projection brings larger improvement on Wildtrack dataset.

Comparison between different projection choices. We claim that projecting the feature maps is a better choice than projecting the RGB images or single view results in Section 3.1. Projecting the RGB images breaks the spatial relationship between pixels, and a single RGB pixel represents little information. As a result, in Table 3, we find "MVDet (project images)" leads to largely inferior performance on both datasets (26.8% and 19.5% MODA). Although single

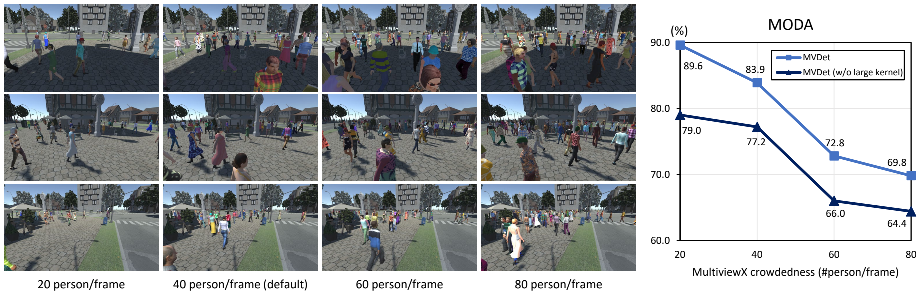

Fig. 7. MultiviewX dataset under different crowdedness configuration (left), and corresponding MVDet performance (right).

view results are robust to spatial patter break from projection, the information contained in them is limited. Due to crowdedness and occlusion, single view detection might lose many true positives. As such, clustering these projected single view results as in "RCNN & clustering" [38] are proven to be extremely difficult (11.3% and 18.7% MODA). Replacing the clustering with ground plane convolution "MVDet (project results)" increases the performance by a large margin (68.2% and 73.2% MODA), as it alleviates the problem of formulating 1-size clusters (clusters that have only one component, as the detections are missing from occlusion) and can be trained in an end-to-end manner. Still, the restricted information in detection results prevents the variant from higher performance.

Effectiveness of spatial aggregation via ground plane convolution. Spatial aggregation with large kernel convolutions brings forward a +11.3% MODA increase on Wildtrack dataset, and a +6.7% performance increase on MultiviewX dataset. In comparison, spatial aggregation with CRF and mean-field inference brings forward increases of +6.3% and +5.2% on the two datasets, going from DeepMCD to Deep-Occlusion. We do not assert superiority of either the CRF-based or CNN-based methods. We only argue that the proposed CNN-based method can effectively aggregate spatial neighbor information to address the ambiguities from crowdedness or occlusion. As shown in Fig. 6, large kernel convolutions manages to generate pedestrian occupancy probability maps that are more similar to the ground truth.

For spatial aggregation, both the proposed large kernel convolution and CRF bring less improvement on MultiviewX dataset. As mentioned in Table 1, even though there are fewer cameras in MultiviewX dataset, each ground plane location in MultiviewX dataset is covered by more cameras on average. Each location is covered by 4.41 cameras (field-of-view) on average for MultiviewX dataset, as opposed to 3.74 in Wildtrack. More camera coverage usually introduces more information and reduces the ambiguities, which also limits the performance increase from addressing ambiguities via spatial aggregation.

Influence of different crowdedness and occlusion levels. Being a synthetic dataset, there are multiple available configurations for MultiviewX. In

Fig. 7 (left), we show the camera views under multiple levels of crowdedness. As the crowdedness of the scenario increases, the occlusion also increases. In Fig. 7 (right), we show the MVDet performance under multiple levels of occlusions. As crowdedness and occlusions increase (more difficult), MODA of both MVDet and MVDet "MVDet (w/o large kernel)" decrease. In addition, performance increases from spatial aggregation also drop, due to the task being more challenging and heavy occlusion also affecting the spatial neighbors.

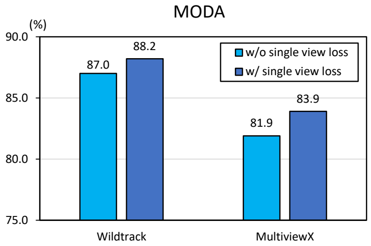

Fig. 8. MODA performance of MVDet with ($\alpha = 1$) or without ($\alpha = 0$) single view detection loss.

Influence of single view detection loss. In our default setting of MVDet, for the combined loss in Eq. 5, the ratio α is set to 1. In Fig. 8, we investigate the influence of removing the single view loss. Without single view detection loss, we find a -1.2% and a -2.0% performance loss on both datasets, which are still very competitive. In fact, we believe single view foot detection loss does not further benefit the system, as the foot points are already supervised on the ground plane. The head point detection loss, on the other hand, can produce heterogeneous supervision, thus further improving system performance. As discussed in Section 3.1 and Section 4.4, less accurate bounding box annotations limit the performance gain from single view loss on Wildtrack dataset.

5 Conclusion

In this paper, we investigate pedestrian detection in a crowded scene, through incorporating multiple camera views. Specifically, we focus on addressing the ambiguities that arise from occlusion with multiview aggregation and spatial aggregation, two core aspects of multiview pedestrian detection. For multiview aggregation, we directly project the feature maps. For spatial aggregation, the proposed convolution method reaches high performance. The proposed system, MVDet, achieves 88.2% MODA on Wildtrack dataset, outperforming previous state-of-the-art by 14.1%. On MultiviewX, a new synthetic dataset for multiview pedestrian detection, MVDet also achieves very competitive results. We believe the proposed MVDet can serve as a strong baseline for multiview pedestrian detection, encouraging further studies in related fields.

Acknowledgement

Dr. Liang Zheng is the recipient of Australian Research Council Discovery Early Career Award (DE200101283) funded by the Australian Government. The authors thank all anonymous reviewers and ACs for their constructive comments.

References

1. 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)
2. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In: arXiv preprint arXiv:1812.08008 (2018)
3. 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)
4. Chavdarova, T., et al.: Deep multi-camera people detection. In: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 848–853. IEEE (2017)
5. Chen, X., Kundu, K., Zhu, Y., Berneshawi, A.G., Ma, H., Fidler, S., Urtasun, R.: 3d object proposals for accurate object class detection. In: Advances in Neural Information Processing Systems. pp. 424–432 (2015)
6. Chen, X., Ma, H., Wan, J., Li, B., Xia, T.: Multi-view 3d object detection network for autonomous driving. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1907–1915 (2017)
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. Girshick, R.: Fast r-cnn object detection with caffe. Microsoft Research (2015)
9. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 580–587 (2014)
10. Gupta, S., Girshick, R., Arbeláez, P., Malik, J.: Learning rich features from rgb-d images for object detection and segmentation. In: European conference on computer vision. pp. 345–360. Springer (2014)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Hoffman, J., Gupta, S., Leong, J., Guadarrama, S., Darrell, T.: Cross-modal adaptation for rgb-d detection. In: 2016 IEEE International Conference on Robotics and Automation (ICRA). pp. 5032–5039. IEEE (2016)
13. Hosang, J., Benenson, R., Schiele, B.: Learning non-maximum suppression. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4507–4515 (2017)
14. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in neural information processing systems. pp. 2017–2025 (2015)
15. Kasturi, R., Goldgof, D., Soundararajan, P., Manohar, V., Garofolo, J., Bowers, R., Boonstra, M., Korzhova, V., Zhang, J.: Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(2), 319–336 (2008)
16. Kong, T., Sun, F., Liu, H., Jiang, Y., Shi, J.: Foveabox: Beyond anchor-based object detector. arXiv preprint arXiv:1904.03797 (2019)

17. Ku, J., Mozifian, M., Lee, J., Harakeh, A., Waslander, S.L.: Joint 3d proposal generation and object detection from view aggregation. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1–8. IEEE (2018)
18. Law, H., Deng, J.: Cornernet: Detecting objects as paired keypoints. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 734–750 (2018)
19. Liang, M., Yang, B., Wang, S., Urtasun, R.: Deep continuous fusion for multi-sensor 3d object detection. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 641–656 (2018)
20. Liu, R., Lehman, J., Molino, P., Such, F.P., Frank, E., Sergeev, A., Yosinski, J.: An intriguing failing of convolutional neural networks and the coordconv solution. In: Advances in Neural Information Processing Systems. pp. 9605–9616 (2018)
21. Liu, W., Liao, S., Ren, W., Hu, W., Yu, Y.: High-level semantic feature detection: A new perspective for pedestrian detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5187–5196 (2019)
22. Noh, J., Lee, S., Kim, B., Kim, G.: 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 (2018)
23. Ouyang, W., Zeng, X., Wang, X.: Partial occlusion handling in pedestrian detection with a deep model. IEEE Transactions on Circuits and Systems for Video Technology **26**(11), 2123–2137 (2015)
24. Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J.: Large kernel matters–improve semantic segmentation by global convolutional network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4353–4361 (2017)
25. 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)
26. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
27. 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)
28. Shi, Y., Liu, L., Yu, X., Li, H.: Spatial-aware feature aggregation for image based cross-view geo-localization. In: Advances in Neural Information Processing Systems. pp. 10090–10100 (2019)
29. Smith, L.N., Topin, N.: Super-convergence: Very fast training of neural networks using large learning rates. In: Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications. vol. 11006, p. 1100612. International Society for Optics and Photonics (2019)
30. Song, T., Sun, L., Xie, D., Sun, H., Pu, S.: Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 536–551 (2018)
31. Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E.: Multi-view convolutional neural networks for 3d shape recognition. In: Proceedings of the IEEE international conference on computer vision. pp. 945–953 (2015)
32. Sun, X., Zheng, L.: Dissecting person re-identification from the viewpoint of viewpoint. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 608–617 (2019)

33. Tian, Y., Luo, P., Wang, X., Tang, X.: Deep learning strong parts for pedestrian detection. In: Proceedings of the IEEE international conference on computer vision. pp. 1904–1912 (2015)
34. Unity: Unity technologies, https://unity.com/
35. Wang, F., Zhao, L., Li, X., Wang, X., Tao, D.: Geometry-aware scene text detection with instance transformation network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1381–1389 (2018)
36. Wang, X., Xiao, T., Jiang, Y., Shao, S., Sun, J., Shen, C.: Repulsion loss: Detecting pedestrians in a crowd. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7774–7783 (2018)
37. Wu, J., Xue, T., Lim, J.J., Tian, Y., Tenenbaum, J.B., Torralba, A., Freeman, W.T.: Single image 3d interpreter network. In: European Conference on Computer Vision. pp. 365–382. Springer (2016)
38. 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)
39. Yan, X., Yang, J., Yumer, E., Guo, Y., Lee, H.: Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In: Advances in neural information processing systems. pp. 1696–1704 (2016)
40. Yang, T., Zhang, X., Li, Z., Zhang, W., Sun, J.: Metaanchor: Learning to detect objects with customized anchors. In: Advances in Neural Information Processing Systems. pp. 320–330 (2018)
41. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
42. Zhang, S., Wen, L., Bian, X., Lei, Z., Li, S.Z.: Occlusion-aware r-cnn: detecting pedestrians in a crowd. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 637–653 (2018)
43. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.: Conditional random fields as recurrent neural networks. In: Proceedings of the IEEE international conference on computer vision. pp. 1529–1537 (2015)
44. Zhou, C., Yuan, J.: Multi-label learning of part detectors for heavily occluded pedestrian detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3486–3495 (2017)
45. Zhu, C., He, Y., Savvides, M.: Feature selective anchor-free module for single-shot object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 840–849 (2019)