Multi-view Target Transformation for Pedestrian Detection

Wei-Yu Lee, Ljubomir Jovanov, and Wilfried Philips
TELIN-IPI, Ghent University-imec, Gent, Belgium
{Weiyu.Lee, Ljubomir.Jovanov, Wilfried.Philips}@ugent.be

Abstract

Occlusion is one of the most challenging problems in single-view pedestrian detection. To alleviate the occlusion problem, multi-view systems have been exploited to fully acquire and recognize blocked targets. Most often, methods from the literature exploit perspective transformation to aggregate different sensing view angles of the scene, but projection distortion issues cause spatial structure break and prevent these methods from fully exploring the projected features. In this paper, we propose a novel approach, Multi-view Target Transformation (MVTT), to address the distortion problem inherent in multi-view aggregation by encoding the full target features and limiting the area of interest of the projected features. Experiment results show that the performance of our proposed method compares favorably against recent relevant methods on public datasets. The ablation studies also confirm the effectiveness of the proposed components.

1. Introduction

Despite the recent advances in object detection towards more accurate localization and more robust performance in various scenarios, occlusion is still one of the challenging problems. When occlusions occur in a single-view system, blocked objects cannot be fully acquired and recognized. One of the mainstream solutions for handling occlusions is introducing multiple perspectives to discover the occluded targets. Compared to the single-view setup, the multi-view system utilizes the information from different perspectives to discover the objects and improve the robustness to occlusions. In this paper, we focus on the multi-view pedestrian detection problem, using images captured by multiple cameras with various view angles as input to perform detection.

In previous multi-view solutions, perspective transformation was used to aggregate different sensing view angles of a scene. Geometric information is provided by synchronized and calibrated cameras. Hence, the targets in single-view 2D images can be projected to the 3D world plane for spatial aggregation. The existing methods overlap the projected bird's eye view (BEV) images to consider multiple perspectives simultaneously [11, 10] and predict the pedestrian occupancy. In addition, various perspective transformation methods have also been explored to improve the usage of the projected features [23, 18, 12].

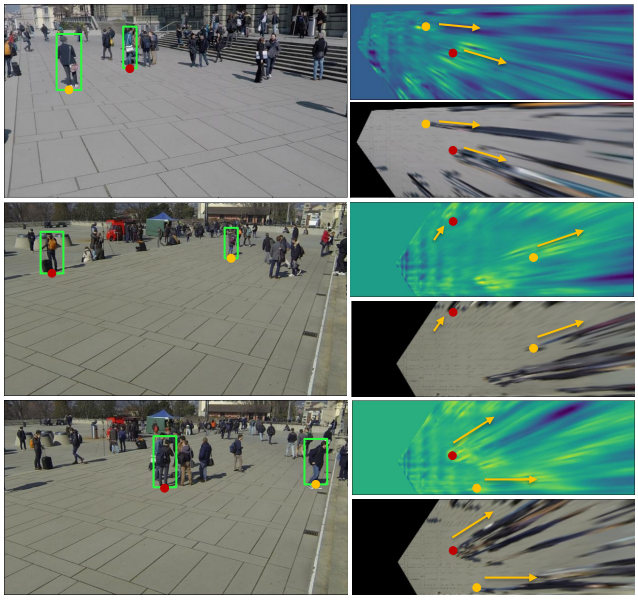


Figure 1. Feature projection examples. The left column show the original single-view images with selected pedestrian bounding boxes and foot points. The right column show the corresponding projected ground plane features and images. According to the different position of the pedestrians and cameras, the projected features would be spread over a large area toward different directions or compressed into a small grid of the ground plane. The projection distortion breaks the spatial structure and prevent the fully exploration of the extracted features.

However, due to the required computational load, most previous methods apply 2D projection to project the features to the ground plane (i.e. BEV) images, which usually introduces severe distortion issues. Depending on the positions of the pedestrians and cameras, the projected pedestrians could be rather different [10]. Various distortion patterns may cause spatial structure break and prevent the existing methods from fully exploring the projected features.

As shown in Fig. 1, some of the distorted targets are spread over a large area and some of them are compressed into a small grid of the ground plane. Hence, it is difficult for the traditional convolution kernel or even the transformer-based methods to effectively adapt to the features with such a large variance. As a result, the detection performance upper bound would be limited.

In this paper, we propose a novel and straightforward approach, Multi-view Target Transformation (MVTT), that alleviates the distortion problem inherent in multi-view aggregation by encoding the full target features and limiting the area of interest of the projected features. In order to enable the model to discover the projected features across different perspectives without learning through various distorted patterns, we utilize the single-view detection results of each perspective to extract and encode the full features of each pedestrian before the distortion happens. Next, we rely on extracted features to build an auxiliary feature map derived from the foot locations of the pedestrians. Because of the limited spatial size of the encoded features, the distortion of the projected auxiliary features would be also constrained. This auxiliary feature can be easily combined with the original image features for assisting the model to sense the entire targets by limited receptive fields.

It is worth of noting that, instead of proposing unique network architectures or loss functions, the main idea lies in the proposed novel feature transformation, which improves the capability of learning and recognizing the distorted features to existing multi-view detection solutions. In our experiments, we apply our proposed MVTT on three different previous models, and conduct the experiments on the benchmark datasets of Wildtrack [2] and MultiviewX [11]. We show that our method clearly improves the performance of existing multi-view detection models for recognizing pedestrians across multiple perspectives.

Our contributions can be summarized as follows:

- We introduce a novel transformation scheme, Multi-view Target Transformation (MVTT), that effectively tackles the projection distortion problems and enables the full utilization of the projected pedestrian features.
- Instead of proposing new network architectures or learning methods, our MVTT is a novel transformation module, applicable to most existing multi-view detection approaches that leverage the spatial aggregation on the same projection plane.
- In our experiments, we qualitatively and quantitatively verify the effectiveness of our method against recent relevant methods and achieve comparable or better detection results on the benchmark datasets of *Wildtrack* and *MultiviewX*.

2. Related Work

Single-view detection Single-view setup is the most commonly used scenario for object detection in computer vision. Previous research could be roughly separated into two branches: anchor-based and anchor-free methods. Anchor-based methods such as Faster R-CNN [20], YOLOv3 [19], and SSD [14] have achieved accurate detection results. On the other hand, in recent years, anchor-free approaches are also coming up with the simplified detection processes to breakthrough the performance limitation caused by the pre-defined anchors [30, 26, 28, 13, 5]. Different from the general object detection, pedestrian detection methods are developed to explore the physical attributes of the human body. For instance, the head-foot or center point detection were proposed to find the bounding box in [24, 15]. In addition, occlusion is also an important issue in pedestrian detection. Some researchers develop the part-based solutions to solve the partial occlusion problem [16, 17, 25, 29]. Nevertheless, even though the partial occlusion problem can be handled, crowded scenes or severe occlusion of pedestrians are still a challenging problem for the single-view setting due to the lack of essential information captured by the camera.

Multi-view detection Multiple camera setup is one of the mainstream solutions for the heavy occlusion. Synchronized and calibrated cameras with different perspectives are used to capture the same area and the multi-view detection system aggregates the images with overlapping field-of-view to perform pedestrian detection. Before the recent advances made by deep learning, researchers were focused on probabilistic modeling of objects [4, 22]. Then, for aggregating the information from multiple views, researchers have been using mean-field inference [6, 1] and conditional random field (CRF) [21, 1] to combine the single-view detection results. However, these methods usually require additional calculations or specific designs outside deep learning models. Hou et al. [10] propose a convolution-based end-to-end trainable method to consider the neighboring location for aggregating the information from multiple perspectives, which achieves significant improvement. Nevertheless, the limited receptive field and translation-invariant calculation still cannot suit to the projected feature maps well. In [10] a method was presented which relies on the improved deformable transformer to adapt to the different distortion patterns across multiple view features. Nevertheless, the spatial structure break caused by 2D projection still cannot be properly addressed.

Feature projection for spatial aggregation In order to utilize the neighboring information across different viewpoints, projecting the single-view feature maps onto the

ground plane is one of the mainstream solutions. Song et al. [23] introduce the stacked homography transformation to approximate 3D point detection. Feature maps are projected onto different height levels depending on different body parts of the pedestrian and then fused by relyuing on the soft selection to achieve a homography with higher accuracy. Furthermore, Qiu et al. [18] also propose a similar solution to project the feature map of each view to multiple parallel planes. However, most previous research on perspective projection has assumed that all pixels are on a certain height or on the ground plane ($z = 0$). In other words, the greater the distance between the projected pixel and the foot point, the greater the number of errors it contains [7]. Hence, overlapping the projected features from multiple perspectives might cause the ambiguity problem. Furthermore, various distortion patterns also make it difficult for the model to sense all features belonging to the pedestrians. In this paper, we propose a novel transformation to tackle this problem. Instead of approximating 3D projection for higher accuracy, we compress and encode pedestrian features before severe distortions happen, and build an auxiliary feature map as the complementary information for the detection.

3. Methodology

Instead of focusing on approximation of more accurate projection [23, 18] or better usage of projected features [10], we propose a novel feature transformation approach to leverage single-view detection results to alleviate the projection distortion issue and fully explore the pedestrian features in multi-view aggregation for better localization.

The objective of our proposed transformation method is to build an auxiliary feature map, named *meta*, for assisting spatial aggregation and detection on the ground plane. The proposed multi-view detection system consists of (1) single-view feature extractor, (2) perspective transformation, and (3) ground plane heat map estimation. As illustrated in Fig. 2, our proposed Multi-view Target Transformation (MVTT) uses the single-view image features F_s as an input and generates the transformed meta feature map M_f. Afterwards, F_s and M_f are concatenated to the corresponding camera view for the perspective transformation. Then, we can get the projected features maps $\tilde{F_{sf}}$. We overlap the N views and use the ground plane heat map generator $\mathrm{G_h}$ to estimate the occupancy heat map. Finally, the post-processing is applied to find the detected targets. In the following subsections, we will briefly describe multi-view aggregation algorithm and then focus on the proposed MVTT module and explain the details of each contribution.

3.1. Brief Review of Multi-view Aggregation

Multi-view feature projection A great deal of feature aggregation in multi-view detection problem focus on projecting feature maps from different single-views to the ground plane. We denote the set of input images from N camera views as $I_s = \{I_1, I_2, ..., I_N\}$ with $H \times W$ size and the corresponding extracted feature maps of the single-view images from the feature extractor as $F_s = \{F_1, F_2, ..., F_N\}$ with downsampled size $H_f \times W_f \times C$. Given the image plane coordinate (u, v) and ith camera view, we rely on the intrinsic parameters $A_i \in \mathbb{R}^{3\times3}$ and the extrinsic parameters $E_i = [R_i|t_i] \in \mathbb{R}^{3\times4}$ to calculate the perspective transformation matrix P_i. Similarly as in [11, 10], we assume the objects in the scene are with height $z = 0$. The extracted feature maps F_s with image coordinate (u, v) can be projected to the ground plane (x, y) as $\tilde{F}_s$ using:

$$\gamma \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = P_i \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} = A_i[R_i|t_i] \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} \tag{1}$$

where γ is the scale factor, and R_i and t_i are the rotation and translation matrix respectively.

After retrieving the projected feature maps, one can easily explore the features of a certain location across different perspectives on the same grid of the ground plane feature maps. Subsequently, the ground plane heat map generator $\mathrm{G_h}$ is able to consider the neighboring locations and estimates the occupancy heat map for detection results.

Projection distortion Although the feature projection allows the spatial aggregation across different perspectives, various distortion patterns caused by the perspective transformation also limit the upper bound of discrimination ability. Specifically, according to the position of the pedestrians and cameras, the projected features could be spread or compressed into different shapes and break the spatial structure, which prevents the extracted features to be fully exploited.

3.2. Multi-view Target Transformation

In this paper, we propose a novel feature transformation, Multi-view Target Transformation (MVTT), to tackle the distortion problem. Instead of focusing on approximating more accurate projection, our MVTT relies on the single-view predicted bounding boxes to build a auxiliary feature map for assisting the multi-view spatial aggregation.

Single-view meta feature representation Given the extracted feature maps of the single-view images F_s and the single-view predicted bounding boxes of each perspective $B_s = \{B_1, B_2, ..., B_N\}$, we first utilize ROI alignment [8] to extract the pedestrian features $F_p = \{F_{p,1}, F_{p,2}, ..., F_{p,N}\}$. For i-th camera view, we can collect $F_{p,i} = \mathrm{ROI}_{\mathrm{align}}(F_i, B_i)$. Specifically, for d-th detected pedestrian in a single-view image, we can get the feature with size (H_d, W_d, C), where (H_d, W_d) represents the

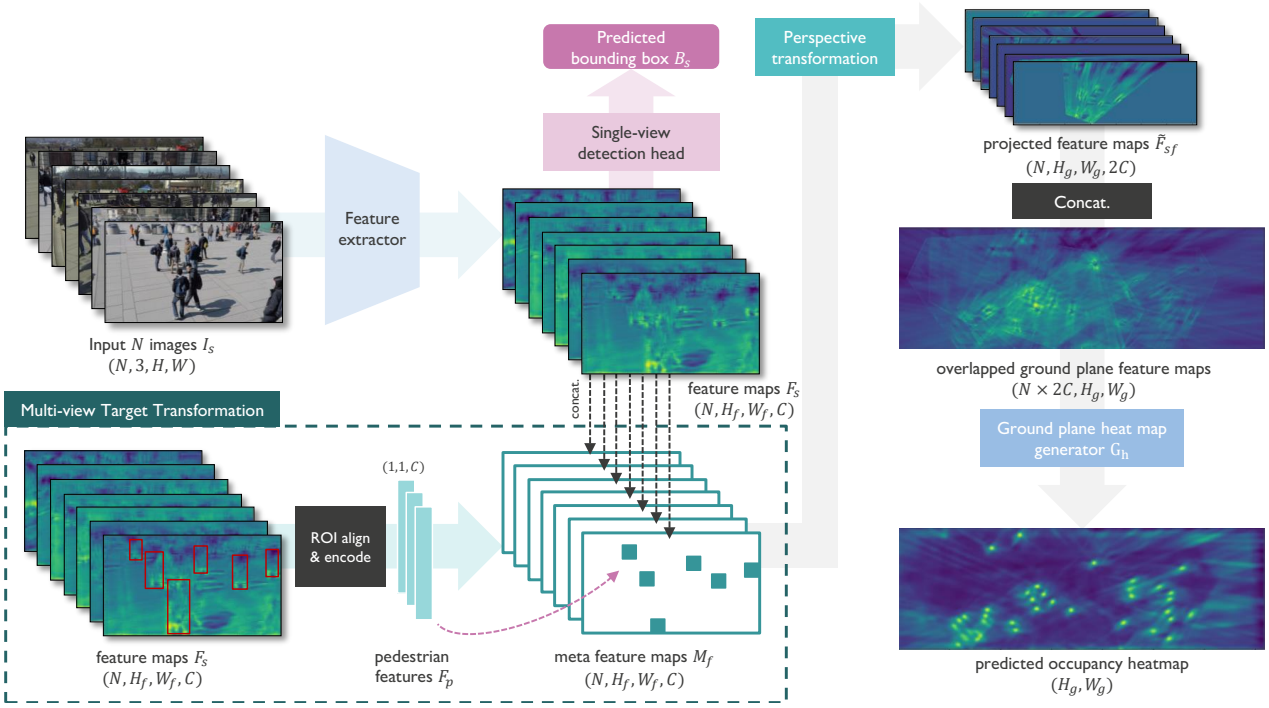


Figure 2. Overview of the proposed Multi-view Target Transformation (MVTT) module and the system pipeline. Given the input images from N camera views and the extracted single-view feature maps, our MVTT module leverage the single-view detection results to build a meta feature map, providing the auxiliary information for spatial aggregation after the perspective transformation.

height and width of the bounding box and C is the channel number. Next, we leverage the ROI alignment to downsample and unify the various bounding box sizes into (s, s, C), where s is the pooled size. The reason why we downsample the size is to make sure all the pedestrian features can have the same size for further process then.

After we extract the pooled features, we apply a fully connected layer as an encoder for each pedestrian feature $F^l_{p,i} \in \mathbb{R}^{s \times s \times C}$, where l indicate the l-th pedestrian in i-th view. Subsequently, the pedestrian feature would be encoded into a one dimension vector $\hat{F}^l_{p,i} \in \mathbb{R}^{1 \times 1 \times C}$. In other words, for each pedestrian, we can use a single vector to describe the attributes without the location information.

In order to fully explore the pedestrian features, we exploit the encoded features to build a meta feature map M_f as the auxiliary information to the image features F_s. First, we follow the size of F_s to create new tensors (N, H_f, W_f, C) filled with zeros. According to the bounding boxes of each single-view B_s, the foot points can be localized at the center of the bounding box bottom. As illustrated in Fig. 2, we insert the encoded pedestrian features $\hat{F}^l_{p,i}$ into the corresponding foot points to complete the meta feature maps for all the perspectives to clearly indicate the location of each pedestrian. The reason why we use the foot point to insert the encoded features is because we assume that the objects in the scene are with zero height, and the foot points, in most cases, on the images should intersect with the ground $(z = 0)$. Hence, we choose the foot point as the representative of the pedestrian to associate the encoded features.

Multi-view aggregation For multi-view aggregation, we first concatenate the meta feature maps with the corresponding single-view feature maps. Then, similarly as in [11] we project the features to the ground plane by applying the perspective transformation. Afterwards, projected feature maps $\tilde{F}_{sf}$ are fed into the ground plane heat map generator G_h for aggregating the information from spatial neighbors. Previously, convention convolution and deformable transformer are adopted [11, 10] as the generator, the limited receptive field and the spatial structure break prevent the existing methods to fully exploit the extracted features. In contrast to the existing methods, our proposed meta feature maps can preserve the full ROI features from the single-view detection results and limit the distortion in a relative small area (Fig. 3). Specifically, the foot point location can be clearly highlighted for the ground plane heat map generator with full pedestrian features, which significantly improves the detection performance.

4. Experiments

In this section, we evaluate the proposed method by conducting several experiments on the public Wildtrack [2] and MultiviewX [11] datasets to compare with the other multi-

view detection methods. Moreover, we also conduct ablation studies to demonstrate the impact of the proposed components. In all comparisons, we use the recomended default settings presented in [11, 10, 23] as well as the original implementation to conduct all the experiments.

4.1. Dataset and Implementation Details

We evaluate the proposed method on two public multi-view datasets, as shown in Table 1. The following sections provide the dataset and our implementation details.

Wildtrack dataset Wildtrack [2] is a real-world dataset captured using 7 synchronized and calibrated cameras with the overlapping field-of-view. The pedestrians within the are of 12 meters by 36 meters are captured and annotated at a ground plane with the resolution of 480×1440 grid, where each grid division represents a 2.5 centimeter square. The average number of pedestrians per frame is 20, and each location is covered by 3.74 cameras. For each single-view image, the resolution is 1080×1920, and there are total 400 frames in this dataset, where 360 frames are for training and the remaining 40 frames are used for evaluation.

MultiviewX dataset MultiviewX [11] is a synthetic dataset generated by the Unity engine. It contains view generated by 6 virtual cameras with overlapping field-of-view. The captured area is 16 meters by 25 meters, which is smaller than the Wildtrack dataset. For annotation, the ground plane is quantized into a grid containing 640×1000 fields, where each grid represents the same 2.5 centimeter square. The average number of pedestrians per frame is 40, while each location is covered by 4.41 cameras. For each single-view image, the resolution is 1080×1920, and there are 400 frames in total, same as in the Wildtrack dataset.

Implementation details As a transformation module that is applicable to the existing methods, we follow their network architectures to implement the feature extractor, ground plane heatmap generator, and the post-processing method to make a fair comparison. Subsequently, we plug our proposed Multi-view Target Transformation (MVTT) module after the feature extractor ResNet-18 [9]. We use ROI align module proposed in [8] to downsample the bounding box content into a 9×9 feature map with 128 channels, and the encoder is a single fully connected layer with output length 128. The non-maximum suppression threshold $K = 0.3$ for the single-view detection. For training, we use SGD optimizer with learning rate 0.1 and 0.15, momentum 0.5 and 0.9 for MVDet [11] and SHOT [23]. For MVDeTr [10], the optimizer is Adam with the learning rate $5e^{-4}$. In our experiments, all models are trained with batch size 1 on a single NVIDIA TITAN RTX. Please refer to the Supplementary Materials for more details.

4.2. Evaluation Metric

During evaluation, we use the same data split in [11] and follow the metrics used in [11, 10, 23], including Multiple Object Detection Accuracy (MODA), Multiple Object Detection Precision (MODP), precision, and recall rate. We evaluate the predicted ground plane occupancy map instead of considering intersection-over-union (IoU) of the bounding box. The distance between the predicted target location and the ground truth is used to determine true positives. As in [11, 10, 23], we use a threshold 0.5 meter in all the experiments. We use MODA as the primary metric because the false positives and false negatives are both considered [11].

4.3. Experiment Results

Evaluation on Wildtrack dataset As shown in Table 2, we evaluate our method by introducing the proposed module into the existing models to evaluate the performance and compare it with the original methods. Our method can improve the MODA performance 2.1% for MVDet [11], 2.9% for SHOT [23], and 2.6% for MVDetr [10]. The results clearly demonstrate that our proposed module can improve the detection performance of the existing methods.

Evaluation on MultiviewX dataset As illustrated in Table 2, we also conduct the same experiment on the MultiviewX dataset. Our method can improve the MODA performance 10.1% for MVDet [11], 5.9% for SHOT [23], and 1.3% for MVDetr [10]. This table not only shows that our proposed MVTT can significantly improve the detection performance but also demonstrates that our module is applicable to the existing methods that utilize spatial aggregation on the same projection plane.

Visual comparisons of projected features For validating the effectiveness of our proposed transformation, we compare the projected feature maps with the previous methods to see if the distortion problem has been addressed in Fig. 3. We observe that the distortion area can be effectively limited by our proposed MVTT module. When combined with the images features, meta features can clearly strengthen the location of the foot point, which would improve the ground plane heatmap generator in localizing the pedestrians. Furthermore, encoded meta features with limited projected area contain the full features from the single-view detection ROIs, which means that the ground plane heatmap generator do not have to adapt to the long-term dependency of the distorted images.

4.4. Ablation Study

In order to investigate the effect of our proposed components, we use MVDeTr [10] and our MVTT to conduct several experiments on the public Wildtrack [2] dataset.

Table 1. Comparisons of the two public multi-view pedestrian detection datasets.

Dataset	Resolution	Frames	Camera	Area	Ground Plane Grid Size	Crowdedness	Scene Type
Wildtrack	1080×1920	400	7	$12 \times 36m^2$	120×360	20 person/frame	Real-world
MultiviewX	1080×1920	400	6	$16 \times 25m^2$	120×250	40 person/frame	Synthetic scene

Table 2. Comparisons with the state-of-the-art methods on the Wildtrack and MultiviewX datasets. We plugged our proposed MVTT module into the existing methods to demonstrate the effectiveness of our method. * indicates the results are not including the additional data augmentations for fair comparisons.

Method	Wildtrack				MultiviewX			
	MODA	MODP	Precision	Recall	MODA	MODP	Precision	Recall
RCNN-2D/3D [27]	0.113	0.184	0.68	0.43	0.187	0.464	0.635	0.439
DeepMCD [3]	0.678	0.642	0.85	0.82	0.700	0.730	0.857	0.833
Deep Occlusion [1]	0.741	0.538	0.95	0.80	0.752	0.547	0.978	0.802
3DROM* [18]	0.912	0.769	0.959	0.953	0.900	0.837	0.975	0.924
MVDet [11]	0.882	0.757	0.947	0.936	0.839	0.796	0.968	0.867
SHOT [23]	0.902	0.765	0.961	0.940	0.883	0.820	0.966	0.915
MVDeTr [10]	0.915	**0.821**	0.974	0.940	0.937	0.913	**0.995**	0.942
MVDet + ours	0.903	0.819	**0.979**	0.917	0.940	0.926	0.994	0.946
SHOT + ours	0.931	0.805	0.967	0.951	0.942	0.922	0.989	0.924
MVDeTr + ours	**0.941**	0.813	0.976	**0.965**	**0.950**	**0.928**	0.994	**0.956**

Table 3. Comparisons with different size of extracted meta features. We change the dimension of the encoded pedestrian features to increase the spatial region on the image plane for comparisons.

Setting	MODA	MODP	Precision	Recall
MVDeTr + ours 1x1xC	**0.941**	**0.813**	0.976	0.965
MVDeTr + ours 3x3xC	0.924	0.805	0.968	0.956
MVDeTr + ours 5x5xC	0.907	0.805	0.980	0.925
MVDeTr + ours 7x7xC	0.892	0.801	0.980	0.921

Table 4. Comparisons with different combination methods between image and meta features. We conduct this experiment with three different combination methods and also compare the performance without combination.

Setting	MODA	MODP	Precision	Recall
MVDeTr *w/o meta feat.*	0.915	**0.821**	0.974	0.940
MVDeTr + ours *w/o image feat.*	0.741	0.712	0.991	0.727
MVDeTr + ours *Concatenate*	**0.941**	0.813	0.976	0.965
MVDeTr + ours *Addition*	0.891	0.804	0.961	0.934
MVDeTr + ours *Multiplication*	0.881	0.794	0.957	0.915

Effects of the extracted meta feature size on the ground plane For further analysis of the effect of our proposed transformation method, we conduct an experiment to compare the performance between different extracted meta feature sizes. Our default setting is using a one dimension vector with length $C = 128$. We keep the length constant and change the dimension of the encoded pedestrian features to increase the spatial region on the image plane (u, v) for comparisons. In Fig. 5, we visualize the different size of meta features on the image plane. We observe that the distortion area would be enlarged significantly if we increase the size of extracted meta features. Moreover, in Table 3, we also find that in the case when the size is increased, the detection performance would be also degraded. This observation leads to the same conclusion mentioned in [7], which states that farther the distance, the greater errors, which would cause the ambiguity problem during the detection. Hence, we choose to encode the meta feature into one dimensional vector and embed it in a single location to limit the distortion area, for a better performance.

Effects of the combination of image features and meta features Another effect we would like to analysis is the method for combining the image and meta features. We compare three different methods to find the best way to use

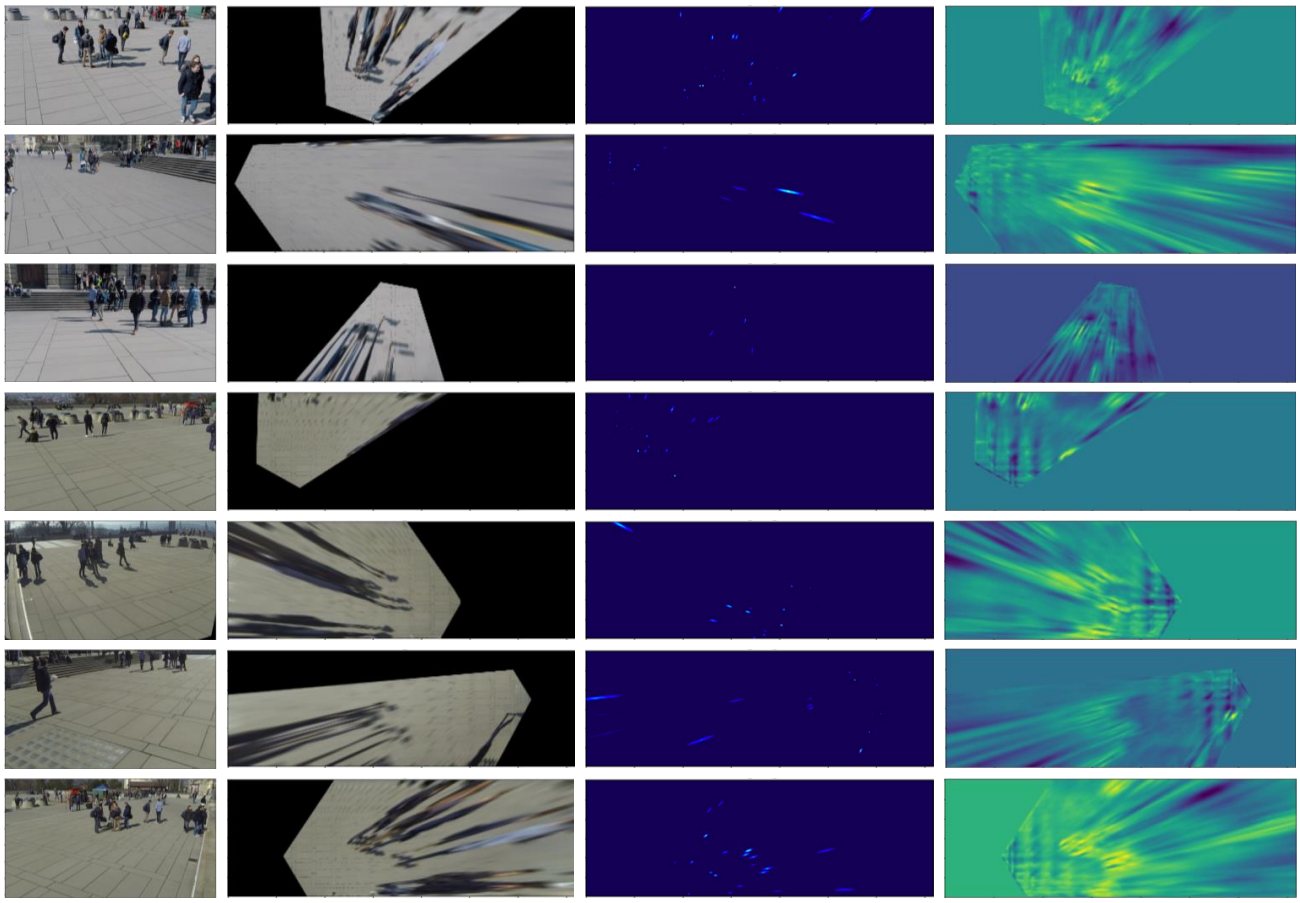


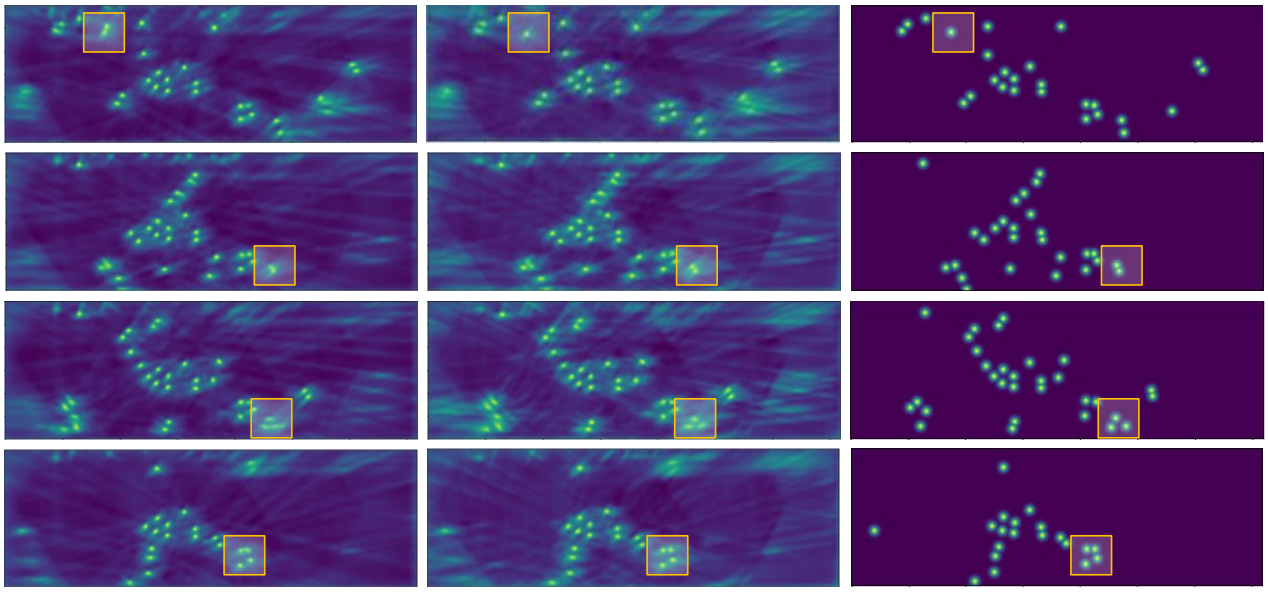
Figure 3. Visualization comparisons of projected features. (a) original single-view images from cameras. (b) projected single-view images. (c) projected proposed meta feature maps. (d) projected extracted image features. Compared to the distorted image features, our MVTT can effectively encode the full pedestrian features with limited the distortion area.

the auxiliary information from our meta features. In Table 4, the concatenation shows the superior performance than the others. Pixel-wised addition or multiplication cannot effectively enhance the original image features for detection. It is worth noting that when we remove the original image features and only use meta features for detection, the performance would be degraded significantly. This is because the limitation from the single-view detection results. Due to the occlusion or lighting condition issues, the predicted bounding boxes might not perfectly discover all the pedestrians in the single-view image. Therefore, the features of the undiscovered targets would not be projected to the ground plane, and therefore the essential features would be lost, which in turn leads to the limited detection performance.

Table 5. Comparisons with different number of bounding box from single-view detection. We use the NMS threshold K to control the bounding box number. Higher K represents more bounding boxes are involved.

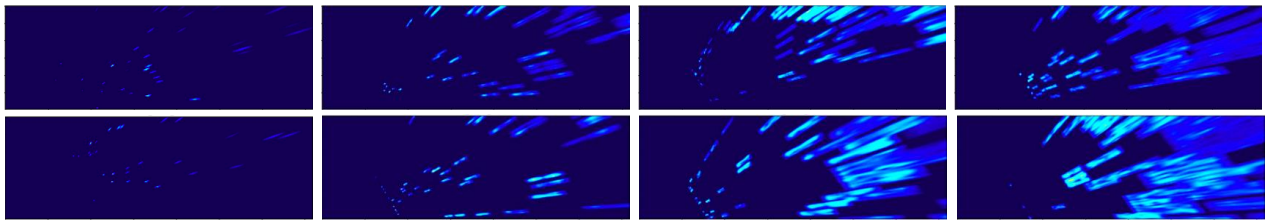
Setting	MODA	MODP	Precision	Recall
MVDeTr + ours $K = 0.1$	0.933	0.811	0.975	0.957
MVDeTr + ours $K = 0.3$	**0.941**	**0.813**	0.976	0.965
MVDeTr + ours $K = 0.5$	0.930	0.812	0.960	0.965
MVDeTr + ours $K = 0.7$	0.925	0.810	0.974	0.960

Effects of the single-view detection bounding box As the previous paragraph mentioned, the predicted bounding boxes of the single-view detection directly affect the usage of the meta features and the detection performance. Hence, we conduct another experiment to analyze the effect of increasing the number of the predicted bounding boxes. We adjust the threshold of the non-maximum suppression K to involve more candidates in a single-view image. When K



(a) MVDeTr (b) MVDeTr + ours (c) ground truth

Figure 4. Estimated occupancy maps comparison. We use MVDeTr [10] as baseline to see the improvement of our proposed method. The highlighted regions show that the proposed MVTT assist the model to output the occupancy map with higher quality.



(a) $1 \times 1 \times C$ (b) $3 \times 3 \times C$ (c) $5 \times 5 \times C$ (d) $7 \times 7 \times C$

Figure 5. Visualization of different size of meta features. We increase the spatial size of the encoded pedestrian features to visualize the distortion problem. Larger spatial sizes bring more overlapping areas between the independent features and cause ambiguity problem during spatial aggregation.

is increased, more bounding boxes would be involved. As we can see in Table 5, excessive number of bounding boxes cause possible ambiguity problem and degrade the detection performance. On the other hand, extreme less bounding boxes would also have lower performance because the high NMS threshold would easily make the bounding boxes be merged to the other one. Therefore, the targets with small bounding boxes would be lost during the ground plane detection. As a conclusion, a moderate number of candidates in a single-view image would be preferable.

5. Conclusions

In this paper, we propose a novel, low complexity approach that tackles the distortion problem caused by feature projection on multi-view pedestrian detection task. We introduce Multi-view Target Transformation (MVTT) module to leverage the single-view detection results to build a auxiliary feature map for assisting the system to localize the pedestrians. Our module successfully limits the spatial size of the pedestrian features, which prevents the spatial structure break and enables the spatial aggregation in a limited receptive field. Moreover, instead of proposing new network architectures or learning methods, our MVTT is applicable to most existing multi-view detection approaches that leverage the spatial aggregation on the same projection plane. The experiment results on the public Wildtrack and MultiviewX datasets confirm that our method performs favourably against recent relevant methods, and support the use of our proposed transformation for improved multi-view pedestrian detection. The ablation studies confirm the effectiveness of the proposed components.

Acknowledgements This work was funded by EU Horizon 2020 ECSEL JU research and innovation programme under grant agreement 876487 (NextPerception) and by the Flemish Government (AI Research Program).

References

[1] Pierre Baqué, François Fleuret, and Pascal Fua. Deep occlusion reasoning for multi-camera multi-target detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

[2] Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[3] Tatjana Chavdarova and François Fleuret. Deep multi-camera people detection. In *IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2017.

[4] Adam Coates and Andrew Y Ng. Multi-camera object detection for robotics. In *IEEE International Conference on Robotics and Automation*, 2010.

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

[6] Francois Fleuret, Jerome Berclaz, Richard Lengagne, and Pascal Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):267–282, 2007.

[7] Xin Gao, Yijin Xiong, Guoying Zhang, Hui Deng, and Kangkang Kou. Exploiting key points supervision and grouped feature fusion for multiview pedestrian detection. *Pattern Recognition*, 131:108866, 2022.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[10] Yunzhong Hou and Liang Zheng. Multiview detection with shadow transformer (and view-coherent data augmentation). In *Proceedings of the 29th ACM International Conference on Multimedia (MM'21)*, 2021.

[11] Yunzhong Hou, Liang Zheng, and Stephen Gould. Multiview detection with feature perspective transformation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[12] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[13] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[14] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[15] Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, and Yinan Yu. High-level semantic feature detection: A new perspective for pedestrian detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[16] Junhyug Noh, Soochan Lee, Beomsu Kim, and Gunhee Kim. Improving occlusion and hard negative handling for single-stage pedestrian detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

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

[18] Rui Qiu, Ming Xu, Yuyao Yan, Jeremy S Smith, and Xi Yang. 3d random occlusion and multi-layer projection for deep multi-camera pedestrian localization. *arXiv preprint arXiv:2207.10895*, 2022.

[19] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[20] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 2015.

[21] Gemma Roig, Xavier Boix, Horesh Ben Shitrit, and Pascal Fua. Conditional random fields for multi-camera object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2011.

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

[23] Liangchen Song, Jialian Wu, Ming Yang, Qian Zhang, Yuan Li, and Junsong Yuan. Stacked homography transformations for multi-view pedestrian detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

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

[25] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning strong parts for pedestrian detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015.

[26] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[27] Yuanlu Xu, Xiaobai Liu, Yang Liu, and Song-Chun Zhu. Multi-view people tracking via hierarchical trajectory composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[28] Tong Yang, Xiangyu Zhang, Zeming Li, Wenqiang Zhang, and Jian Sun. Metaanchor: Learning to detect objects with customized anchors. *Advances in Neural Information Processing Systems*, 2018.

[29] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Occlusion-aware r-cnn: detecting pedestrians in a crowd. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[30] Chenchen Zhu, Yihui He, and Marios Savvides. Feature selective anchor-free module for single-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.