Stacked Homography Transformations for Multi-View Pedestrian Detection

Liangchen Song[1], Jialian Wu[1], Ming Yang[2], Qian Zhang[2], Yuan Li[3], and Junsong Yuan[1]
[1]University at Buffalo [2]Horizon Robotics, Inc. [3]Google, Inc.
{lsong8,jialianw,jsyuan}@buffalo.edu

Abstract

Multi-view pedestrian detection aims to predict a bird's eye view (BEV) occupancy map from multiple camera views. This task is confronted with two challenges: how to establish the 3D correspondences from views to the BEV map and how to assemble occupancy information across views. In this paper, we propose a novel Stacked HOmography Transformations (SHOT) approach, which is motivated by approximating projections in 3D world coordinates via a stack of homographies. We first construct a stack of transformations for projecting views to the ground plane at different height levels. Then we design a soft selection module so that the network learns to predict the likelihood of the stack of transformations. Moreover, we provide an in-depth theoretical analysis on constructing SHOT and how well SHOT approximates projections in 3D world coordinates. SHOT is empirically verified to be capable of estimating accurate correspondences from individual views to the BEV map, leading to new state-of-the-art performance on standard evaluation benchmarks.

1. Introduction

Multi-view detection, a.k.a. multi-camera detection [20, 1], aims to detect objects from a set of synchronized images from different sensing viewpoints of a scene. Compared with the single view detection, multi-view object detection enables to aggregate the information across multiple viewpoints and infer 3D structures of the scene [8, 5], hence is generally robust to occlusions, which is a major challenge to single-view detection in crowded scenes. In this paper, as shown in Figure 1, we focus on detecting pedestrians from multi-view images, where the input is a batch of images from different viewpoints and the output is an occupancy map from the bird's eye view (BEV) of the plane.

Estimating the occupancy map from a set of multi-view images is challenging in two aspects. First, due to the change of view point, occlusions and ambiguities in object appearances often present in different views, therefore it is not a trivial problem to match the features of pedestrians accurately among input views. Second, even if the correspondences are adequately estimated, one view only provides an incomplete representation for the whole scene, bringing difficulties in assembling the knowledge of occupancy across all views. For example, due to occlusions, an area may be only visible in one view, so we have to identify that view and exclude the distractions from other views according to the pre-established correspondences.

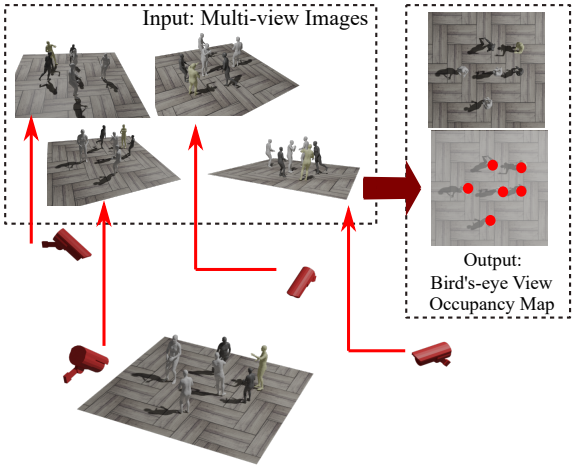
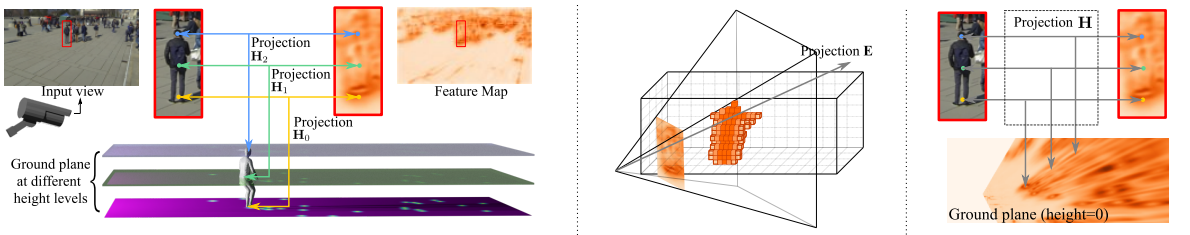


Figure 1. The task of multi-view pedestrian detection: Given a batch of synchronized images captured from different view angles, our goal is to predict an occupancy map of the scene.

To optimize the correspondence and feature extraction jointly in an end-to-end manner, recent work [13, 12] proposed to project the features extracted from 2D images to a shared space for aggregating information from all views, while keep the framework differentiable. However, previous works either project the features to large 3D grids [13], or only project the features to the ground plane [12]. The full 3D projection proposed by [13] is expensive since 3D convolutions are involved in dealing with the projected features. Meanwhile, 2D projection used in [12] is not accurate due to misalignments.

In this paper, for the first challenge of establishing the 3D correspondence, we propose to project the feature maps onto different height levels according to different semantic parts of the pedestrian. As illustrated in Figure 2, our mo-



high efficiency *high* accuracy
(a) approximate 3D projection with our SHOT

less efficiency *high* accuracy
(b) 3D projection

high efficiency *less* accuracy
(c) 2D projection

Figure 2. Illustration of different projection schemes: (a) the proposed stacked homography transformations (SHOT) approximates 3D projection with a stack of homographies; (b) 3D projection proposed in [13] project 2D feature points to 3D grids; (c) 2D projection proposed in [12] projects 2D feature onto the 2D ground plane. Our method achieves a better tradeoff between projection efficiency and accuracy than the other two schemes.

tivation is that each pixel should be projected to a ground plane at a proper height. To achieved this, we construct a stack of homography projections, which are $\mathbf{H}_0$, $\mathbf{H}_1$ and $\mathbf{H}_2$ in the figure.

For the second challenge, assembling occupancy information across views, we design a soft selection module to ensure the network differentiable, thus learning how to aggregate the occupancy information end-to-end. Specifically, for each pixel of the features extracted from individual views, we design a likelihood map prediction module to softly select projections from the stack of transformations. Since each pixel is projected with a stack of homography transformations, our method is named Stacked HOmography Transformations (SHOT).

Intuitively, SHOT can be viewed as an approximation of a 3D projection with a stack of homographies. Then we theoretically analyze the properties of SHOT in two aspects: (1) the requirements of acquiring the pre-computed homographies and (2) the requirements of serving as a 3D projection. SHOT achieves the state-of-the-art performance of 90.2% MODA on WILDTRACK and 88.3% MODA on MultiviewX, which outperforms the recent method [12] by 2% on WILDTRACK and 4.2% on MultiviewX, respectively. Moreover, we investigate the performance of SHOT under a practical yet challenging setting: both the scene and the camera locations are distinctly different between training and testing. The results validate the generalization ability of SHOT. To sum up, our contributions are as follows:

- We propose a novel stacked homography transformations (SHOT) to establish accurate 3D correspondences between individual input views and the BEV occupancy map.
- We theoretically analyze the geometry property of SHOT and demonstrate two properties: 1) The stack of transformations can be effectively constructed without knowing extrinsic parameters; 2) SHOT can project all human body parts to the same grid on the BEV map with proper hyperparameters.
- We conduct experiments on standard benchmarks and achieve new state-of-the-art results. Moreover, we investigate the performance of SHOT under a new challenging setting: training and testing involve different scenes and camera viewpoints.

2. Related work

Multi-view detection. The most challenging part of multi-view detection is gathering occupancy knowledge about objects or pedestrians from multiple views. Before the surge of deep neural networks, modeling the correspondence across cameras was mostly done by probabilistic modeling of objects [8, 5, 20, 18]. Rubino *et al*. [19] estimate a quadric (ellipsoid) in 3D from a set of 2D ellipses fitted to the object detection bounding boxes in multiple views. Baqué *et al*. [2] proposed an end-to-end trainable model based on CRF and use high-order CRF terms to represent potential occlusions. In [25], Xu *et al*. proposed to re-formulate the computation of correspondences among views as a problem of compositional structure optimization. In [3], the authors proposed a large scale dataset WILDTRACK and systematically review the performance of recent multi-view detection methods. Recently, Hou *et al*. [12] proposed the simple yet effective feature perspective transformation for the multi-view detection task, which was trained end-to-end and reached new SOTA.

Multiple projections for parallel planes. The proposed SHOT projects each view to multiple parallel planes with respect to the ground plane. The idea of multiple projections for parallel planes was made possible by the results published in [6]. Khan *et al*. [15] proposed Multiple Scene Planes for tracking occluded people. Similarly, Eshel *et al*. [7] proposed to detect head and apply the plane transformation. Also, similar idea has been successfully adopted in the field of multi-view crowd counting [27, 29, 28, 30]. Our method differs from previous methods for the reason that the proposed SHOT is combined with the soft selection module for pedestrian detection.

Geometry integrated deep networks. Multiple view ge-

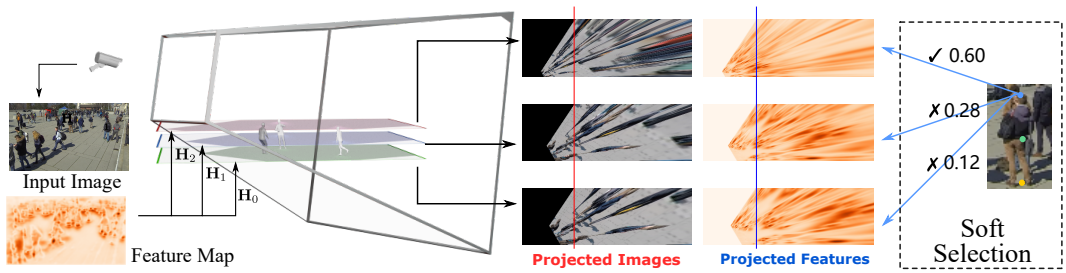


Figure 3. Illustration of the proposed SHOT. Each feature map will be projected with a stack of homographies and the projection results are selected softly to form a proper screen→BEV projection.

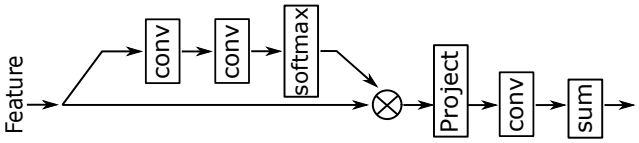


Figure 4. Network details about the soft selection module.

ometry is one of the cornerstones in computer vision [9]. Each projection in our work is essentially a homography, which describes the translation of a plane for the pin-hole camera model. In [24, 26], the authors used perspective projection to connect 2D estimations to 3D worlds. Shi *et al*. [21] proposed polar transformation to align an aerial image with a ground-view panorama. Perspective transformation is also widely used in pose estimation tasks [13, 11, 23], to infer the 3D positions of body joints. In [16], Nassar *et al*. proposed to learn the warping and detection together with geometric soft constraints. Besides, Roddick *et al*. [17] propose to project features on to the ground plane and predict BEV semantic segmentation maps. Our method differs from previous methods because we aim at learning to predict the likelihood of transformations to each pixel for the pedestrian detection task.

3. Preliminaries and notations

We denote the input images from N views as $(\boldsymbol{I}^1, \cdots, \boldsymbol{I}^N)$ and the extracted feature maps for each image as $(\boldsymbol{J}^1, \cdots, \boldsymbol{J}^N)$. Denote the intrinsic parameters of camera i (for view i) as $\mathbf{K}^i \in \mathbb{R}^{3\times3}$. Assume that for a point from camera i, its image coordinate is $(u, v)^T$ and world coordinate is $(X, Y, Z)^T$, then by the pinhole camera model we have

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} \sim \mathbf{K}^i[\mathbf{R}^i|\mathbf{t}^i] \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}, \tag{1}$$

where $[\mathbf{R}^i|\mathbf{t}^i] \in \mathbb{R}^{3\times4}$ denotes the extrinsic parameters of camera i. Following [12], we define ground plane d_0 as the $Z = 0$ plane in the world coordinates. From Equation (1) we can see that if $Z = 0$ the projection matrix $\mathbf{K}^i[\mathbf{R}^i|\mathbf{t}^i]$ can be converted to a 3×3 matrix, representing the projection from $Z = 0$ plane to the screen plane. That is, if we denote the extrinsic matrix as $\mathbf{E}^i = [\mathbf{R}^i|\mathbf{t}^i] = (\mathbf{e}^i_1, \mathbf{e}^i_2, \mathbf{e}^i_3, \mathbf{e}^i_4)$, where each $\mathbf{e}$ is a column vector, the matrix for the projection from $Z = 0$ plane to the screen plane is $\mathbf{K}^i\mathbf{E}^i_0$ where $\mathbf{E}^i_0 = (\mathbf{e}^i_1, \mathbf{e}^i_2, \mathbf{e}^i_4)$, as the world coordinates will be $\mathbf{e}^i_1 X + \mathbf{e}^i_2 Y + \mathbf{e}^i_4$.

Next, we denote the projective mapping from world coordinates to an occupancy map as $\mathbf{K}_g \in \mathbb{R}^{3\times3}$. We can treat $\mathbf{K}_g$ as the intrinsic parameters for the "ground camera", which quantizes the ground plane world coordinates into grids [12]. Now using $\mathbf{K}_g$, the transformation from image i to the ground occupancy map can be represented as $\mathbf{H}^i_0 = \mathbf{K}_g(\mathbf{E}^i_0)^{-1}(\mathbf{K}^i)^{-1}$, which is a homography matrix. Finally, we denote the size of input images as $H \times W$ and denote the output ground plane occupancy map as $H_g \times W_g$.

4. Proposed method

Our motivation is to approximate transformations in 3D world coordinates with a stack of homographies, such that we can project all body parts of one person to the same location on the BEV map. To achieve this goal, we propose stacked homography transformations (SHOT), which consists of two steps: construction of a stack of homographies and soft selection of the transformations.

4.1. Construction of a stack of homographies

Instead of directly estimating the 3D locations of pedestrians, we approximate the transformations in 3D world coordinates with a stack of homographies. Each homography in the stack is designed to project a view to a ground plane at a certain height. Specifically, for each view i, we define that there are $D + 1$ transformations in total, and Δz is the distance between each two target planes, which are parallel to the ground plane. The homographies are the projections from the screen plane to the $Z = \{k\Delta z|k = 0, \cdots, D\}$ planes in the world coordinates, where D indicates the size of the stack. In Figure 3, we demonstrate an example of $D = 3$ which projects the screen plane to $Z = \{0, \Delta z, 2\Delta z\}$ planes. As we can see from the figure, by selecting the target of projection we can align different body parts of one person to the same BEV position.

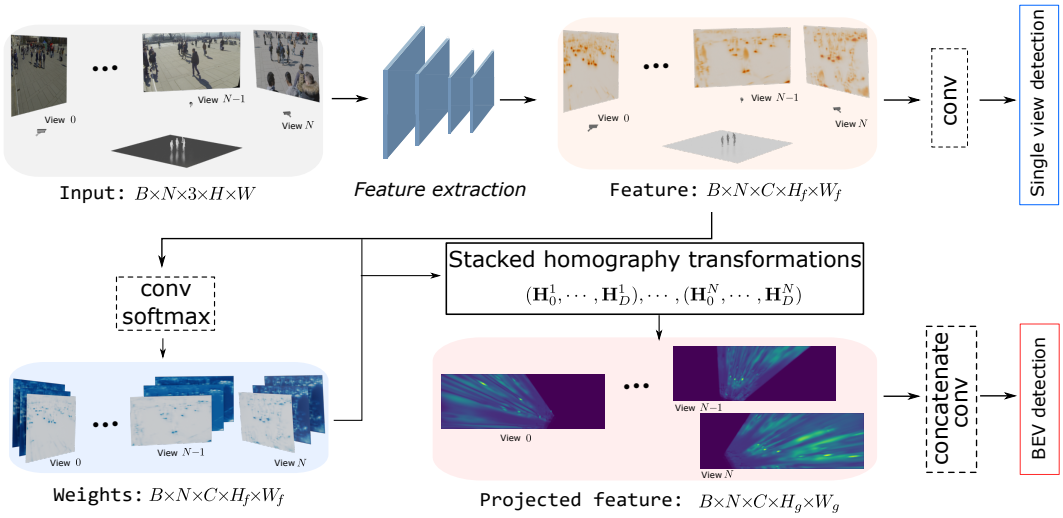


Figure 5. Overview of our proposed framework.

Calculating the homography matrices is straightforward if the extrinsic parameters are known. Following notations in Section 3, the homography $\mathbf{H}_k^i$ for projection from image i to $k\Delta z$ plane can be computed as

$$\mathbf{H}_k^i = \mathbf{K}_g(\mathbf{E}_k^i)^{-1}(\mathbf{K}^i)^{-1}, \text{where } \mathbf{E}_k^i = (\mathbf{e}_1^i, \mathbf{e}_2^i, \mathbf{e}_4^i + k\Delta z\mathbf{e}_3^i). \tag{2}$$

As we have N cameras and there are $D+1$ transformations pre-computed for each camera, the number of transformations is then in total $N(D+1)$, *i.e.*, $\{(\mathbf{H}_0^i,\cdots,\mathbf{H}_D^i)\}_{i=1}^{i=N}$. Here, note that we can still compute the homographies without knowing extrinsic parameters and further analysis to be discussed in Section 6.4.

4.2. Soft selection module

With a stack of homographies, we now face a new problem: Which homography should be used for which pixel? Since we need to keep the network end-to-end differentiable, we propose to generate a selection mask for each input view from the extracted features, *i.e.*, $(\boldsymbol{J}^1,\cdots,\boldsymbol{J}^N)$. The motivation of using the features as input is that the network can predict a proper likelihood score for transformations according to the semantic information presented by the features. For example, the network may recognize the head of a person and then assign the transformation that best fits the height of the head.

The computation graph of the soft selection module is shown in Figure 4. With the features $\boldsymbol{J}^i$ as inputs, we first estimate the selection likelihood scores on each pixel of the feature maps with softmax for all the transformations. Let the likelihood tensors be $\{g_k(\boldsymbol{J}^i)\}_{k=0}^{k=D}$, where $g_k()$ means the likelihood prediction for transformation k. Then homography matrices are applied accordingly and we have $\{\mathbf{H}_k^i \cdot \big(g_k(\boldsymbol{J}^i) \circ \boldsymbol{J}^i\big)\}_{k=0}^{k=D}$. After applying a convolutional layer to each feature, for each view we sum up all the features projected by the stack of homographies.

4.3. Overall framework

In Figure 5, we demonstrate the overall framework for multi-view pedestrian detection, which is inspired by the structure designed in [12]. Specifically, ResNet-18 [10] with dilation is chosen as the backbone and reduces the resolution by 8 times from the input. Then the feature maps are upsampled to $H_f \times W_f$ and sent to a classifier for detecting pedestrians in each image. The loss is $L_{\text{single}} = \sum_i \|\mathbf{s}_h^i - \mathbf{y}_h^i\|_2^2 + \|\mathbf{s}_f^i - \mathbf{y}_f^i\|_2^2$, where for image i, $\mathbf{s}_h^i$ and $\mathbf{s}_f^i$ are the classification map for head and foot, and $\mathbf{y}_h^i$ and $\mathbf{y}_f^i$ are the ground truth labels blurred with a Gaussian kernel for head and foot, respectively.

Next, the features are projected with the proposed SHOT. The features from all views are concatenated and sent to the classifier layer for pedestrian detection. After the final BEV detection results are generated, the loss is calculated with $L_{\text{ground}} = \|\mathbf{s}_g - \mathbf{y}_g\|_2^2$. Here, $\mathbf{s}_g$ and $\mathbf{y}_g$ are the prediction of the BEV occupancy map and the ground truth of the BEV occupancy map blurred with a Gaussian kernel, respectively. Finally, the total loss is the combination of the above two losses, that is, $L = L_{\text{single}} + L_{\text{ground}}$. All network modules are jointly optimized in an end-to-end manner: the convolutional backbone for feature extraction of the pedestrians, soft selection module for establishing the correspondences, and classifier layers for detecting pedestrians.

5. Properties of SHOT

In this section, we analyze the property of the proposed SHOT in two aspects: 1) Besides camera intrinsic parameters, what else should we know for constructing the homographies? 2) How do SHOT approximate the 3D point projection? Without loss of generality, we omit the superscript i for cameras in this section for simplicity.

5.1. Requirements for running SHOT

In Section 4.1, we illustrate our approach with the assumption of knowing the extrinsic parameters of a camera, which sometimes may not be practical. Meanwhile, observing that $\mathbf{H}_0$ is a homography, we can mark corresponding points on the ground easily, such as intersections of bricks on the pavement. Note that only 4 point correspondences are theoretically enough for estimating $\mathbf{H}_0$ [9]. A question then naturally arises: what extra information is needed for constructing a stack of homographies in our method? We first show that two extra annotations of the pedestrians with the same height are enough for acquiring a stack of transformations.

As demonstrated in Figure 6, for a pedestrian in image coordinates we denote the location of feet as $\boldsymbol{f}$ and the location of head as $\boldsymbol{h}$. Next, we denote the coordinates on the BEV occupancy map of this person as $\boldsymbol{o}$ correspondingly. Then formally we have the following proposition (full proof included in supplementary):

Proposition 1. *If $\mathbf{K}, \mathbf{K}_g$ and $\mathbf{H}_0$ are known, we can construct a stack of transformations $\{(\mathbf{H}_0, \cdots, \mathbf{H}_D)\}$ with only two extra annotations of pedestrians if they have the same height. Two extra annotations means two sets of points correspondences in the camera image and BEV image,* i.e., *$(\boldsymbol{f}_1, \boldsymbol{h}_1, \boldsymbol{o}_1)$ and $(\boldsymbol{f}_2, \boldsymbol{h}_2, \boldsymbol{o}_2)$.*

Proof. (Sketch) First observe that we can recover part of extrinsic parameters with $\mathbf{E}_0 = \mathbf{K}^{-1}\mathbf{H}_0^{-1}\mathbf{K}_g$, then if we define $\mathbf{E}_D = \mathbf{K}^{-1}\mathbf{H}_D^{-1}\mathbf{K}_g$, we have $\mathbf{E}_D = \mathbf{E}_0 + \Delta\mathbf{T}$ where $\Delta\mathbf{T}$ is

$$\begin{pmatrix} 0 & 0 & \Delta t_1 \\ 0 & 0 & \Delta t_2 \\ 0 & 0 & \Delta t_3 \end{pmatrix}. \tag{3}$$

To construct a set of transformations, we only need to know $\Delta t_1, \Delta t_2$ and Δt_3. Next, from one annotation $(\boldsymbol{f}, \boldsymbol{h}, \boldsymbol{o})$ we have the equations

$$\begin{cases} \mathbf{E}_0\mathbf{K}_g^{-1}\boldsymbol{o} \sim \mathbf{K}^{-1}\boldsymbol{f}, \\ \mathbf{E}_D\mathbf{K}_g^{-1}\boldsymbol{o} \sim \mathbf{K}^{-1}\boldsymbol{h}. \end{cases} \tag{4}$$

Then there are two equations for solving $\Delta\mathbf{T}$,

$$\begin{cases} \Delta t_1 - h_u \Delta t_3 = (h_u - f_u)(e_{31}o_x + e_{32}o_y + e_{33}), \\ \Delta t_2 - h_v \Delta t_3 = (h_v - f_v)(e_{31}o_x + e_{32}o_y + e_{33}), \end{cases} \tag{5}$$

where h_u, h_v, f_u, f_v and o_x, o_y are the first and second elements from $\mathbf{K}^{-1}\boldsymbol{h}$, $\mathbf{K}^{-1}\boldsymbol{f}$ and $\mathbf{K}_g^{-1}\boldsymbol{o}$, respectively. e_{31}, e_{32} and e_{33} are elements of the third row of $\mathbf{E}_0$. As each point provides two equations and there are three variables, we need two extra annotations for constructing the transformations. □

In the above proposition, the requirement of two pedestrians with the same height could be hard to meet. Fortunately, in practice we usually have a bunch of point pairs with noise, which reduce the impact of not satisfying the height constraint. In fact, we find that in our experiments we can even set $\Delta t_3 = 0$ while still can achieve good projection result. In other words, only one annotation of pedestrian shows promising results in our experiments. The reason may be that the gap of the scale between the whole scene and pedestrian height is large. We present results on a real image in Section 6.4 and calculate transformations with Equation (5) from one annotation.

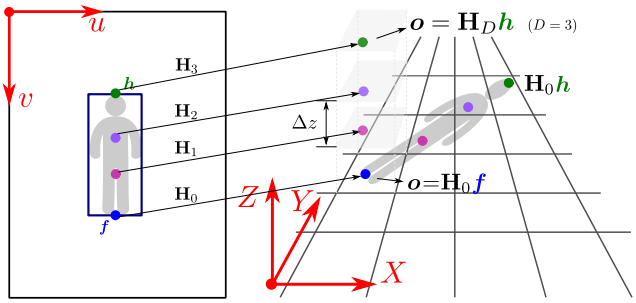


Figure 6. Geometry of SHOT.

5.2. Approximating 3D projection with SHOT

Since a homography matrix is a projection between planes but we need 3D point projections in multi-view detection, we now investigate how SHOT approximates a 3D point projection. The key observation is that there are two discretizations in the whole framework: First, a stack of homographies is a discretization of Z axis of the world coordinates; Second, a BEV map is a discretization of the ground plane. The two discretizations play different roles. The discretization of the Z axis can be viewed as quantizing the input occupancy map while the BEV map can be viewed as quantizing the output occupancy map. Therefore theoretically we can align the two discretizations such that all points of an object can be projected on the same grid. In Figure 6, we show an example of SHOT with $D = 3$, which is able to project all points into one grid on BEV. Formally, we have the following proposition for the alignment

Proposition 2. *If we set $D = \lceil \|\mathbf{H}_0(\boldsymbol{f} - \boldsymbol{h})\| \rceil$ ($\lceil \cdot \rceil$ is the ceiling function) and $D\Delta z\mathbf{e}_3$ is a solution for $(\Delta t_1, \Delta t_2, \Delta t_3)$ in Equation (5), all points with the same Z value can be projected to the same grid on BEV map.*

Proof. Since $\mathbf{H}_0$ is a homography, vector $\boldsymbol{f} - \boldsymbol{h}$ will be projected as another vector on the BEV occupancy map. Observe that the number of occupied pixels of $\mathbf{H}_0(\boldsymbol{f} - \boldsymbol{h})$ reflects the number of grids needed to be aligned, thus the number of transformations should be $D = \lceil \|\mathbf{H}_0(\boldsymbol{f} - \boldsymbol{h})\| \rceil$. Next, if projection $\mathbf{H}_D$ can transform the topmost point to the same occupancy pixel, then all points on $\boldsymbol{f} - \boldsymbol{h}$ can be transformed to the same pixel. From the proposition introduced last section, transforming the topmost point means

Method	WILDTRACK				MultiviewX			
	MODA	MODP	Precision	Recall	MODA	MODP	Precision	Recall
RCNN & clustering [25]	11.3	18.4	68	43	18.7	46.4	63.5	43.9
POM-CNN [8]	23.2	30.5	75	55	-	-	-	-
DeepMCD [4]	67.8	64.2	85	82	70.0	73.0	85.7	83.3
Deep-Occlusion [2]	74.1	53.8	95	80	75.2	54.7	**97.8**	80.2
MVDet [12]	88.2	75.7	94.7	93.6	83.9	79.6	96.8	86.7
Volumetric [13]*	88.6	73.8	95.3	93.2	84.2	80.3	97.5	86.4
Ours	**90.2**	**76.5**	**96.1**	**94.0**	**88.3**	**82.0**	96.6	**91.5**

Table 1. Comparison to state-of-the-art methods. Our results are averaged on 5 repeated runs. * Volumetric is our re-implementation and all other results of comparison methods are quoted from [12].

Setting	MODA	MODP	Prcn	Rcll
project images				
MVDet [12]	19.5	51.0	84.4	24.0
Ours	**32.3**	**73.2**	**92.4**	**35.1**
project results				
MVDet [12]	73.2	**79.7**	87.6	85.0
Ours	**77.0**	79.6	**89.8**	**86.8**
w/o large kernel				
MVDet [12]	77.2 (6.7↓)	76.3 (3.3↓)	89.5 (7.3↓)	85.9 (0.8↓)
Ours	**86.1** (2.2↓)	**81.8** (0.2↓)	**91.9** (4.7↓)	**94.4** (2.9↑)

Table 2. Applying SHOT under different settings. Results on MultiviewX is reported.The settings *project images* and *project results* are applying perspective projection to images and results. The setting *w/o large kernel* means not using large kernel in the final occupancy map classification layers, which is used for aggregating knowledge across views in [12]. The numbers in parentheses are the performance drop compared to the setting with large kernel.

that $D\Delta z\mathbf{e}_3$ is a solution for $(\Delta t_1, \Delta t_2, \Delta t_3)$ in Equation (5). □

The above theoretical analysis proves that the proposed SHOT can be constructed effectively with very few annotations, which demonstrates the suitability for pedestrian detection task. Also, with a proper D and Δz, SHOT is capable of perfectly projecting all points of interest to the same grid on the BEV map, which demonstrates the broad applicability.

6. Experiments

To evaluate our proposed SHOT, we first conduct experiments following the evaluation protocols used in [3, 12]. Then, we conduct various ablation studies and present visualization to validate the effectiveness of our method[1].

6.1. Datasets

WILDTRACK [3]. In this dataset, pedestrians on 12 meters by 36 meters region are captured from 7 cameras. The size of images is 1080×1920 and annotated 2 frames per second. There are 400 images for the scene and the total number of images is 2,800 since there are 7 cameras. On average, each frame captures 23.8 persons and each person is seen in 30.41 frames.

MultiviewX [12]. This dataset is a synthetic dataset generated with Unity engine and human models from PersonX [22]. Same as WILDTRACK, 400 frames are of size 1080×1920 and annotated 2 frames per second. The ground plane is of size $16 \times 25m^2$, slightly smaller than WILDTRACK. Unlike WILDTRACK, 6 cameras are used in this dataset and there are 40 persons in each frame.

6.2. Implementation details and metrics

Our implementation is based on the released code of [12]. Specifically, as [12], the input image size (H, W) is set to $(720, 1280)$ and the size of output feature for each view (H_f, W_f) is set to $(270, 480)$. For training, we use the SGD optimizer with learning rate 0.15 and momentum 0.9. D is set to 4. Δz is set to 10 on WILDTRACK and 0.1 on MultiviewX. In our experiments, all networks are trained with batch size 1 on two Titan XP GPUs.

Evaluation metrics. We use the data split in [12] and follow the metrics proposed in [14]. Four metrics are reported: Multiple Object Detection Accuracy (MODA), Multiple Object Detection Precision (MODP), precision and recall. To calculate the metrics, false positives (FP), false negatives (FN) and true positives (TP) are first computed. Then, MODA is computed by $1 - \frac{\text{FP}+\text{FN}}{N}$, where N is the number of ground truth pedestrians. MODP is computed by $\frac{\sum 1-d[d<t]/t}{\text{TP}}$, where d is the distances from a detection to its ground truth and t is the threshold set to 20. MODP tells us the precision of detection. Finally, precision is computed by $\frac{\text{TP}}{\text{FP}+\text{TP}}$ and recall is computed by $\frac{\text{TP}}{N}$. For all metrics, we report the percentage.

6.3. Comparison to state-of-the-art methods

We compare SHOT to the state-of-the-art methods in multiple aspects. We present quantitative comparisons on standard validation benchmarks and then demonstrate the effectiveness of our proposed SHOT.

[1] Our code is avaliable in the supplementary.

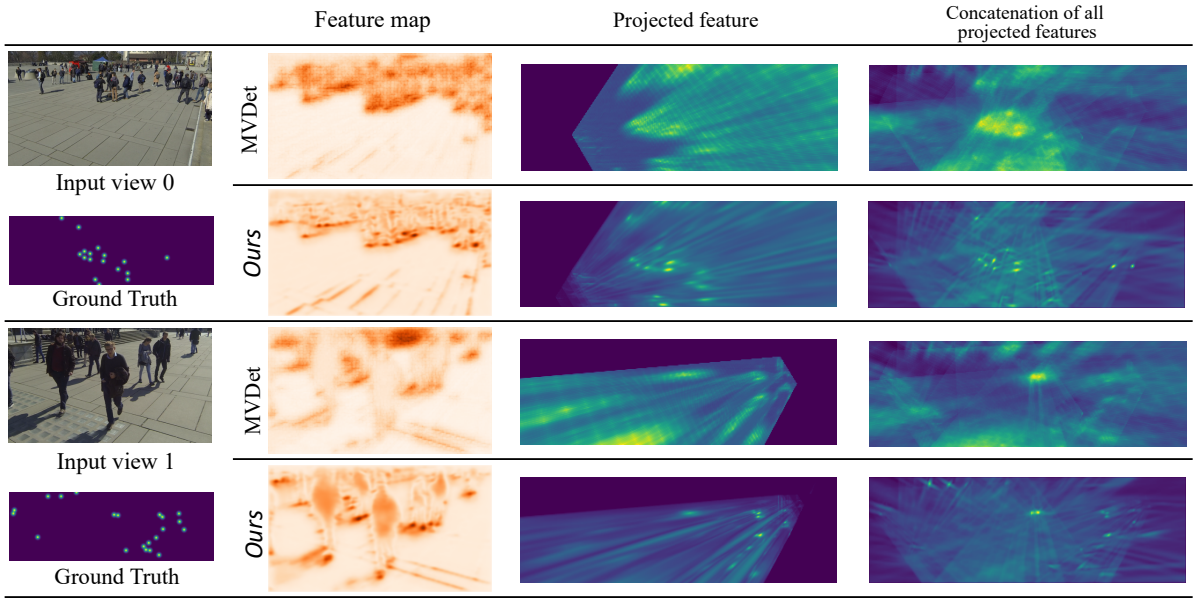


Figure 7. Visual comparison of extracted feature maps and projected feature maps. First column: one example view and current ground truth BEV occupancy map. Second column: the feature maps extracted for the demonstrated view. Third column: Feature maps of the view after projecting to BEV map. Fourth column: Concatenation of projected features from all views.

On standard multi-view detection benchmarks. As the first experiment, we report the performance of our method and compare it with the state-of-the-art methods in Table 1. The comparison shows that our method outperforms all competing methods significantly. More precisely, on WILDTRACK, our method is better than others for all the metrics. For MultiviewX, our method outperforms the others by a large margin (*4.4%* higher than the current best) for the MODA metric. Besides, our recall is remarkably better than others (*4.8%* higher than the current best), while the precision remains as good as the best one.

Applying SHOT under different settings. In [12], the authors first studied three different projection schemes: image level, result level, and feature level. They found that projecting the intermediate feature is the most effective scheme, which is the final result reported in Table 1. Then they found that using a large kernel in the final occupancy classification layer is beneficial for aggregating the knowledge across views. As a comparison, we conduct experiments to validate the effectiveness of SHOT accordingly. First, we observe that applying SHOT to the other two schemes is also beneficial. From the results shown in Table 2, we can see that our SHOT module consistently improves the detection performance. Moreover, under the setting *w/o large kernel*, it is clear that our method is less affected by using small kernels when performing BEV occupancy map classification. The reason is that SHOT is capable of aligning the features, therefore relaxing the requirement on the final occupancy map classification layers from aggregating occupancy knowledge across views.

Visual comparisons of projected feature maps. To further validate our claim that SHOT can directly help to gather knowledge across views, we visualize the intermediate feature maps of our method. In Figure 7, we present two randomly selected samples from the test set of WILDTRACK. For visualizing features, we normalize the feature tensor along the channel dimension. The first point we can observe is that the features after projection and concatenation (3rd and 4th column) are very different between MVDet and ours. The features projected with our method are more focused while features from MVDet are more blurred, which demonstrates that SHOT can indeed help align the features on the BEV map. The second observation is that by comparing the extracted image feature (2nd column), our method also focuses on other human body parts. For example, from the last row, we can observe that the response on the human body is higher than the counterpart, indicating that features from the human body are useful for classification on the BEV map. To sum up, all the above results validate our motivation and claim that SHOT can help align the feature on the BEV map.

6.4. Analysis

Previous results verify the effectiveness of our method. In this section, we analyze our method in terms of explainability and applicability.

Transformation selection module. As introduced in Section 4.2, we predict the likelihood for each pixel of softly selecting a specific homography from the stacked homography transformations. It is important to analyze whether the likelihood prediction module $g()$ is functioning as expected. In Figure 8, we demonstrate the selection likelihood values for each pixel. Observe that the color on the

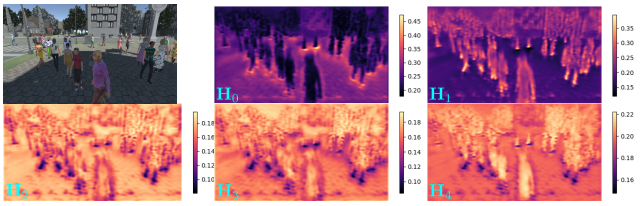
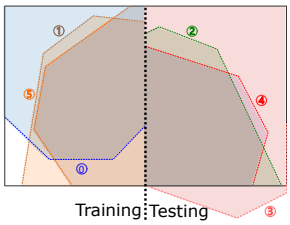


Figure 8. Visualization of estimated selection likelihood for each homography transformation.

D	1	2	3	4	5	6
MODA	85.5	86.2	87.2	88.3	88.6	OOM

Table 3. Performance on MultiviewX with different D values and each D means that $D+1$ homographies are used. OOM means out of memory.



	MVDet	*Ours*
MODA	33.0	49.1
MODP	76.5	77.0
Prcn	64.5	73.3
Rcll	73.4	77.1

Table 4. Performance comparison under the setting shown in the left image. The cameras and the ground plane are different between training and testing. The numbers in the left image are the camera ids.

human body becomes darker with a higher level transformation, indicating that the likelihood of selecting the higher level transformations increases on body parts. The changes in the likelihood of different body parts prove that the likelihood prediction module works expectedly.

The number D of stacked homography transformations. Recall that the stacked homography transformations can be viewed as a discretization of the Z axis in the world coordinates, so the number D represents how fine the discretization is. Although we introduce the condition for achieving perfect alignment in Section 5.2, the alignment is different for feature maps since the receptive field is usually large for networks. Still, the foreseeable impact of D is that as D increases the performance gets better until D reaches a certain value. The results are shown in Table 3, from which we can observe that the change of performance meets our expectations. Specifically, with a larger D, the performance keeps improving with a diminishing gain. The performance gap between $D=4$ and 5 is getting quite small.

Input views are changed when testing. In this paper, we follow the setting that the scene and the training and testing views are the same. While this setting has its practical application scenarios such as surveillance cameras. A more practical and challenging setting is to train the model once and deploy to a different scene. Therefore, we suggest to evaluate in a new setting: the scene and the camera locations are different between training and testing. To investigate the performance under this setting, we create a new training-testing split on MultiviewX, by using cameras 0, 1, and 5 for training and camera 2, 3, and 4 for testing. Also, we split the ground plane into two parts from the middle vertically, then the left part is used for training while the right part is used for testing. In Table 4, we present a visualization of our setting and performance comparison. Our method again outperforms the baseline method MVDet in this much more challenging setting.

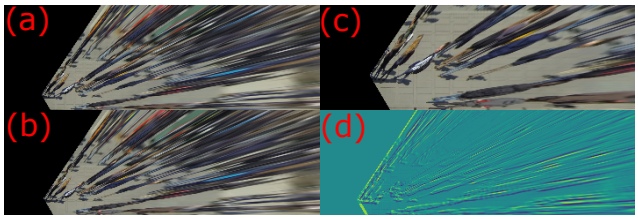


Figure 9. Computing $\mathbf{H}_D$ from $\mathbf{H}_0$ with one pedestrian annotation. (a) Projection with $\mathbf{H}_D$ computed with one annotation. (b) Projection with $\mathbf{H}_D$ computed with extrinsic parameters. (c) Projection with $\mathbf{H}_0$. (d) Difference between (a) and (b).

Computing transformations without extrinsic parameters. In Section 5, we mention that in practice we can construct SHOT with only one pedestrian label from the ground homography. Here we present an example from WILDTRACK in Figure 9. As can be observed from the images, there is very little difference between (a) and (b), thus computing a stack of transformations will not restrict the applicability.

7. Conclusion

In this paper, we propose the stacked homography transformations (SHOT) as an approximation to the 3D point projection. SHOT consists of two steps: first constructing the transformations, then soft selecting transformations. We theoretically analyze the requirements of applying SHOT and how SHOT approximate 3D projections in the framework. On standard benchmarks, our method reaches new state-of-the-art with notable gain. Moreover, extensive analysis validates our claims and motivations.

Acknowledgements

This work is supported in part by a gift grant from Horizon Robotics and the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C0124. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). The authors thank Xuan Gong, Runze Li, Jiemin Fang and the reviewers for the comments and support.

References

[1] Hamid Aghajan and Andrea Cavallaro. *Multi-camera networks: principles and applications*. Academic press, 2009. 1

[2] Pierre Baqué, François Fleuret, and Pascal Fua. Deep occlusion reasoning for multi-camera multi-target detection. In *Int. Conf. Comput. Vis.*, pages 271–279, 2017. 2, 6

[3] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5030–5039, 2018. 2, 6

[4] Tatjana Chavdarova and François Fleuret. Deep multi-camera people detection. In *IEEE International Conference on Machine Learning and Applications*, pages 848–853. IEEE, 2017. 6

[5] Adam Coates and Andrew Y Ng. Multi-camera object detection for robotics. In *IEEE International Conference on Robotics and Automation*, pages 412–419. IEEE, 2010. 1, 2

[6] Antonio Criminisi, Ian Reid, and Andrew Zisserman. Single view metrology. *Int. J. Comput. Vis.*, 40(2):123–148, 2000. 2

[7] Ran Eshel and Yael Moses. Tracking in a dense crowd using multiple cameras. *Int. J. Comput. Vis.*, 88(1):129–143, 2010. 2

[8] Francois Fleuret, Jerome Berclaz, Richard Lengagne, and Pascal Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(2):267–282, 2007. 1, 2, 6

[9] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 3, 5

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 4

[11] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7779–7788, 2020. 3

[12] Yunzhong Hou, Liang Zheng, and Stephen Gould. Multiview detection with feature perspective transformation. *Eur. Conf. Comput. Vis.*, 2020. 1, 2, 3, 4, 6, 7

[13] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Int. Conf. Comput. Vis.*, pages 7718–7727, 2019. 1, 2, 3, 6

[14] Rangachar Kasturi, Dmitry Goldgof, Padmanabhan Soundararajan, Vasant Manohar, John Garofolo, Rachel Bowers, Matthew Boonstra, Valentina Korzhova, and Jing Zhang. 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, 2008. 6

[15] Saad M Khan and Mubarak Shah. Tracking multiple occluding people by localizing on multiple scene planes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(3):505–519, 2008. 2

[16] Ahmed Samy Nassar, Sébastien Lefèvre, and Jan Dirk Wegner. Simultaneous multi-view instance detection with learned geometric soft-constraints. In *Int. Conf. Comput. Vis.*, pages 6559–6568, 2019. 3

[17] Thomas Roddick and Roberto Cipolla. Predicting semantic map representations from images using pyramid occupancy networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11138–11147, 2020. 3

[18] Gemma Roig, Xavier Boix, Horesh Ben Shitrit, and Pascal Fua. Conditional random fields for multi-camera object detection. In *Int. Conf. Comput. Vis.*, pages 563–570. IEEE, 2011. 2

[19] Cosimo Rubino, Marco Crocco, and Alessio Del Bue. 3d object localisation from multi-view image detections. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1281–1294, 2017. 2

[20] 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. 1, 2

[21] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. In *Adv. Neural Inform. Process. Syst.*, pages 10090–10100, 2019. 3

[22] Xiaoxiao Sun and Liang Zheng. Dissecting person re-identification from the viewpoint of viewpoint. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 6

[23] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. Eur. Conf. Comput. Vis., 2020. 3

[24] Jiajun Wu, Tianfan Xue, Joseph J Lim, Yuandong Tian, Joshua B Tenenbaum, Antonio Torralba, and William T Freeman. Single image 3d interpreter network. In *Eur. Conf. Comput. Vis.*, pages 365–382. Springer, 2016. 3

[25] Yuanlu Xu, Xiaobai Liu, Yang Liu, and Song-Chun Zhu. Multi-view people tracking via hierarchical trajectory composition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4256–4265, 2016. 2, 6

[26] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *Adv. Neural Inform. Process. Syst.*, pages 1696–1704, 2016. 3

[27] Qi Zhang and Antoni B Chan. Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8297–8306, 2019. 2

[28] Qi Zhang and Antoni B Chan. 3d crowd counting via multi-view fusion with 3d gaussian kernels. In *AAAI*, volume 34, pages 12837–12844, 2020. 2

[29] Qi Zhang and Antoni B Chan. Wide-area crowd counting: Multi-view fusion networks for counting in large scenes. *arXiv preprint arXiv:2012.00946*, 2020. 2

[30] Qi Zhang, Wei Lin, and Antoni B Chan. Cross-view cross-scene multi-view crowd counting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 557–567, 2021. 2