From a Bird's Eye View to See: Joint Camera and Subject Registration without the Camera Calibration

Zekun Qian[1], Ruize Han[2,3†], Wei Feng[1], Song Wang[4]
[1]College of Intelligence and Computing, Tianjin University
[2]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[3]City University of Hong Kong [4]University of South Carolina
{clarkqian, han_ruize, wfeng}@tju.edu.cn, songwang@cec.sc.edu

Abstract

We tackle a new problem of multi-view camera and subject registration in the bird's eye view (BEV) without pre-given camera calibration, which promotes the multi-view subject registration problem to a new calibration-free stage. This greatly alleviates the limitation in many practical applications. However, this is a very challenging problem since its only input is several RGB images from different first-person views (FPVs), without the BEV image and the calibration of the FPVs, while the output is a unified plane aggregated from all views with the positions and orientations of both the subjects and cameras in a BEV. For this purpose, we propose an end-to-end framework solving camera and subject registration together by taking advantage of their mutual dependence, whose main idea is as below: i) creating a subject view-transform module (VTM) to project each pedestrian from FPV to a virtual BEV, ii) deriving a multi-view geometry-based spatial alignment module (SAM) to estimate the relative camera pose in a unified BEV, iii) selecting and refining the subject and camera registration results within the unified BEV. We collect a new large-scale synthetic dataset with rich annotations for training and evaluation. Additionally, we also collect a real dataset for cross-domain evaluation. The experimental results show the remarkable effectiveness of our method. The code and proposed datasets are available at BEVSee.

1. Introduction

There are just three problems in computer vision: registration, registration, and registration.

– Takeo Kanade

Registration is an important task in computer vision. In this work, we study a new and challenging problem of *camera and person registration in the BEV without camera calibration*. Specifically, as shown in Figure 1, given the multi-view images for a multi-person scene, we aim to generate the position and orientation of every person (referred to as subject in this paper) and camera in BEV.

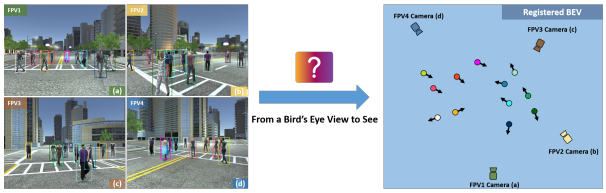

Figure 1. An illustration of the multi-view camera and subject registration problem.

This problem is practical for the multi-view camera multi-human scene analysis, which has many applications such as video surveillance, social scene understanding, *etc*. In this case, the bird's eye view, also called the top view, is a good way to observe the whole scene. In BEV, we can obtain the global spatial layout and trajectories of all the persons in the scene without mutual occlusion, which is very useful in many typical scenarios including the automatic driving [15, 42, 43, 46, 57], outdoor human detection [4, 13] and complementary-view crowd analysis [21, 22].

A popular research problem related to this task is the multi-view human detection [4, 13, 29, 30, 53], which projects the subjects detected from each view to their locations on the ground plane and then generate an occupancy map in the bird's eye view. Note that, these methods all *require the pre-given camera calibration parameters* among the multi-view cameras as input, which, however, limits the applications of the method in many scenes. Another series of research focuses on the complementary-view multi-human analysis using a top-view camera (*e.g.*, on a UAV) and several first-person-view cameras [20–23]. The main limitation of these methods is that the usage of *a top-view camera carried by a drone is not easy to deploy*. Similarly, BEV detection in the automatic driving area also relies on the given camera calibration or depth sensor, *e.g.*, LiDAR.

†Corresponding author.

Given the above reasons, in this work, we propose to study a more practical yet challenging problem that achieves not only the subject registration (human localization and face orientation estimation), but also the camera registration (camera localization and view direction estimation) in BEV. Different from previous works, we register the subjects in BEV but *without* a real bird's-eye-view image, where we generate a virtual BEV. Moreover, we do not use the camera calibration as input, but we need to generate a camera registration result, *i.e.*, the camera location and view direction estimation (can be regarded as a weak version of the camera calibration) as output. This makes this problem very difficult given the *very limited input information and multiple output results*.

A straightforward idea for this problem may be the local descriptors-based methods for multi-view camera pose estimation [27, 48, 54]. However, these algorithms can not handle the proposed problem since they need the input images to have enough overlapped area with textural information. However, in our problem, the overlap may be very limited given the large view difference (even on the opposite side), and the scene (dominant by humans) is unfavorable for local descriptor extraction. In this work, we propose a novel framework to address this problem. Our basic idea is the *registration of the camera and subject are interdependent and complementary*. We *alternately achieve the camera and subject registrations* to make them help each other. Specifically, the subject distribution in the real 3D world is fixed, which presents variously in the 2D image given different camera poses. This way, we first restore the subject 3D localization in BEV from the respective camera, and then leverage the prior of the unified subject spatial distribution in the 3D world to estimate the relative camera pose (from subject to camera registration), and finally based on the camera registration to further refine the subject registration in the BEV (from camera to subject registration).

Based on the above insights, we propose a joint framework to simultaneously achieve the subject and camera registration in the BEV. Specifically, in each side view, we first apply a view-transform subject detection module (VTM) to obtain the subject detection results in the BEV. We then propose a computing-geometry-based spatial alignment module (SAM) to estimate the relative pose of the multiple cameras in the BEV, in which we also apply a self-supervised multi-view human association strategy to obtain the cross-view human corresponding among the multiple views. With the camera pose estimation from SAM, in the final registration module, we use a camera pose selection strategy to obtain the camera registration and subject fusion scheme to get unified subject registration in the BEV. We summarize the main contributions in this work:

❶ To the best of our knowledge, this is the first work to study the camera and subject registration for the multi-view multi-human scene, in which we alternately achieve the camera and human registration results in a unified BEV. This work breaks the limitations of using pre-given camera calibration or real BEV images in previous works.

❷ We propose a novel solution for this problem, in which we integrate the deep network-based VTM and a multi-view geometry-based SAM. This framework integrates both the generalization of the deep network for the human localization task and the stability of the classical geometry for the camera pose estimation task.

❸ We build a new large-scale synthetic dataset for the proposed problem. Extensive experimental results on this dataset show the superiority of the proposed method and the effectiveness of the key modules. Furthermore, the cross-domain study on the real dataset verifies the generalization of our method.

2. Related Work

Multi-view object detection seems like the most related work to this paper, which aims to aggregate information about the same object from different views. The main difficulty of this problem is solving the serious occlusion problem. Recent approaches for this work are mainly based on the pre-given camera calibration to project the objects detected from each view to their locations on the real-world ground plane, and then generate an occupancy map in the BEV. In [4, 13], researchers estimate the positions of pedestrians on the ground plane with corresponding anchor box features. In [29, 30, 53], feature perspective transformation is employed to project all view features into a shared plane without any anchor. A couple of datasets have been developed for multi-view pedestrian detection. One is created in a virtual environment, while the other is captured from the real world. These datasets are proposed in [14, 30], respectively. Besides the multi-view detection, some recent related works have been employed in different fields, *e.g.*, cross-view human association and tracking, multi-view 3D human pose estimation [18, 50, 56], *etc*. Note that, these works all use fixed cameras and require the prior camera calibration as input. Differently, in this work, we not only do not need the pre-given camera calibration but also provide the camera registration results in the BEV as output.

3D object detection in autonomous driving aims to detect objects in traffic scenes. Existing solutions can be broadly classified into three categories. The first category only relies on monocular images, where the localization is directly estimated from monocular images without any depth sensor. For instance, general objects are modeled as 3D boxes for localization [15, 36, 38, 43, 57, 58]. For 3D pedestrian detection works, the 3D skeleton of pedestrians is extracted for localization instead of 3D boxes [7, 8, 28]. The second category of methods is based on multi-view images. These methods [16, 31, 33, 37, 39, 59] use camera

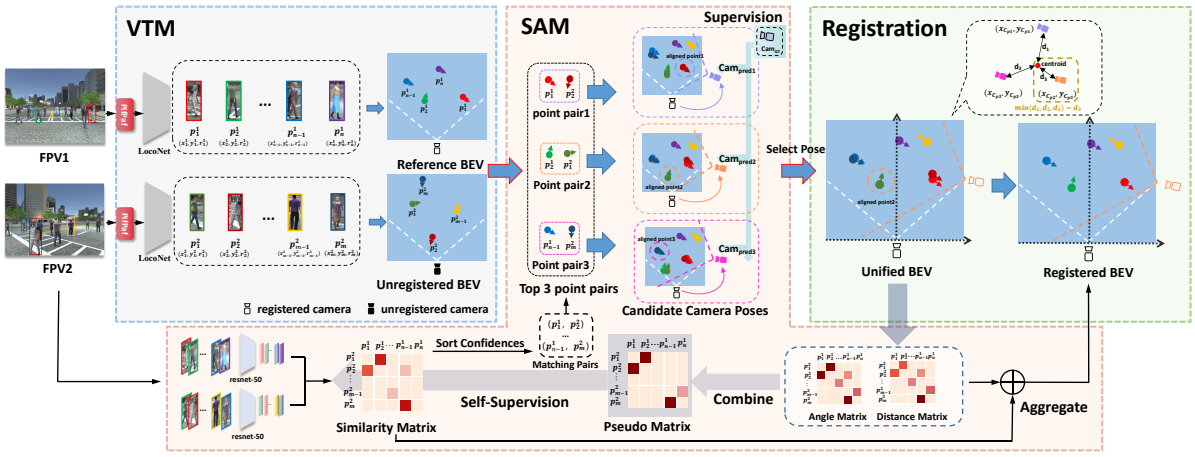

Figure 2. Framework of the proposed method, which can be divided into three parts, *i.e.*, VTM, SAM and Registration. We use hollow camera icons to represent registered cameras and filled camera icons to represent unregistered cameras.

calibration to align different viewpoints and estimate image depth to construct a unified BEV, which can be used to achieve 3D object detection. The third category of methods [5, 6, 10, 12, 17, 44, 60–63] utilizes depth sensors such as LiDAR to capture the 3D point cloud of the entire scene, based on which they achieve 3D object detection. Each of these three categories has its limitations. The monocular view cannot effectively handle occlusions and capture global information. The second and third categories rely on the camera calibration and depth sensor, which are costly and have limited applicability. In contrast to the above mentioned methods, our method not only use the aggregated information from multi-view images but also do not need additional data from camera calibration and depth sensors.

Camera pose estimation is a related problem to this work, which is a fundamental problem in computer vision. In the long history of its exploration, many methods have been proposed. Conventional methods to solve this problem used to be helped with some extra measuring devices. In [40], the laser rangefinders are used to combine cameras from different views. In [9, 19], the visual sensors are applied to bridge the huge differences between different fields of views (FOV). In [11, 49], some structure from motion (SfM) methods are proposed to track the movement of objects from different views. The core for recent vision-based methods [27, 48] is to find and match the feature points from different views, which, however, are not very useful in this work given the large FOV difference.

Bird's-eye-view visual analysis. Recently, some works have proposed to associate the top view (BEV) with the first-person views for collaborative analysis. In [51, 52], such idea is employed to locate the first-person-view camera in BEV aerial images in a large field, which is used for GEO-localization. Later, some related works [20–23, 25, 26] focus on the localization of humans, which aims to associate and track the multiple humans by the spatial reasoning based on the pre-acquired detection. Another series of works [1–3] use graph matching-based methods to locate camera wearers by combining information from FPVs and BEV. The main difference between the previous works and this work is that they require a BEV image (*e.g.*, captured by a UAV) as input, which is not practical in many real applications.

3. Proposed Method

3.1. Overview

We first give an overview of the proposed method mainly containing three stages, as shown in Figure 2. 1) Given multiple images simultaneously captured from different views for a multi-human scene, we apply a view-transform subject detection module (VTM) to get the position and the face orientation estimation of each person in the BEV (Section 3.2). 2) We then apply a geometric transformation based spatial alignment module (SAM) to estimate the relative camera pose candidates in the BEV (Section 3.3). 3) We next use a centroid distance based candidate selection strategy to choose the final camera pose estimation result (camera registration) from the candidates obtained by the SAM. For the subject registration task, we take both spatial and appearance information to aggregate the same person in the BEV for multi-view subject registration (Section 3.4). Besides, with the subject registration results, we propose a backward training strategy to learn subject similarity for human association in SAM using a self-supervised manner (Section 3.5).

3.2. View-Transform Detection Module (VTM)

For the input of multiple images captured in a multi-human scene, we first get the subject position and face orientation of each person in the BEV. For this purpose, we develop a **LocoNet** using a lightweight FC-based structure with three

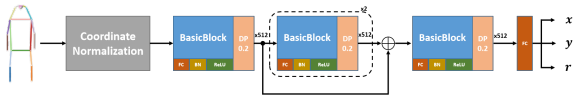

Figure 3. The structure of LocoNet. Here $\times 512$ means the 512 feature channels, Fc means fully connected layer, BN means batch normalization layer, ReLU is an activation function, DP0.2 is the dropout layer with ratio 0.2.

heads. Before that, we first apply an existing human pose detector called PifPaf[34] to predict the 2D skeleton joints of each person from the original RGB image, which will be inputted into the LocoNet, whose structure is shown in Figure 3. At the end of LocoNet, we use a human 3D localization head composed of simple MLP layers to predict the position and face orientation of each person. The process of the network can be represented as

$$\mathbf{p}_i^v \triangleq (x_i^v, y_i^v, r_i^v) = \text{LocoNet}(\mathbf{k}_i^v), \tag{1}$$

where v denotes the v-th view and i denotes the i-th person in v-th view, $\mathbf{k}_i^v$ is the 2D skeleton joints belonging to the person i in view v. $\mathbf{p}_i^v$ is the output prediction of LocoNet, which is composed of (x_i^v, y_i^v) representing the subject position in the BEV and r_i^v representing the face orientation.

3.3. Spatial Alignment Module (SAM)

We then show the relative camera pose estimation (in the BEV) via the subject localization alignment. For convenience, we first present the case of two views. Our basic idea is that the human position and face orientation are unique in the real-world 3D coordinate system, which can be used for aligning the cameras to generate multiple 2D images. In the BEV maps with human position and face orientation generated from different first-person view (FPV) images, we can obtain the camera pose in the BEV by aligning the corresponding human position and facing orientation (as aligned points) as shown in SAM of Figure 2.

For this purpose, the first step is to find the same subject from different views. We identify the subjects in the input images through the human appearance features, and the corresponding subjects in the BEV are then matched across different views. We use a ResNet-50 network to extract the feature of every person and apply Euclidean distance and sigmoid function to create a similarity matrix ($\mathbf{M}_{\text{pred}}$), which indicates the subject similarities among the subjects from two views. Then we sort the similarities of each subject pair and select the top-K pairs as the *matching pairs*.

After that, we apply the geometric transformation to align two BEVs (containing all subjects and cameras on them), which are denoted as a reference BEV map and an unregistered one. Specifically, for a pair of matching points, we apply a geometric transformation, as shown in Figure 4, to rotate and move the camera position and orientation in the unregistered BEV to that in the reference BEV.

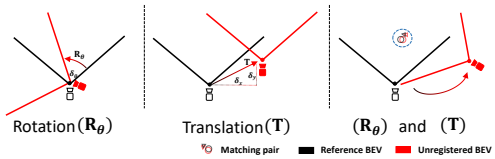

Figure 4. An illustration of the rotation and translation transformation for two BEVs with a matching pair.

We calculate the relative pose (in the BEV) between two points in a matching pair as discussed above, which are denoted as $\mathbf{p}_{\text{ref}} = (x_{\text{ref}}, y_{\text{ref}}, r_{\text{ref}})$ and $\mathbf{p}_{\text{unr}} = (x_{\text{unr}}, y_{\text{unr}}, r_{\text{unr}})$. Specifically, the relative pose transformation between them can be formulated as

$$\begin{pmatrix} x_{\text{ref}} \\ y_{\text{ref}} \\ 1 \end{pmatrix} = \mathbf{T}\mathbf{R}_\theta \begin{pmatrix} x_{\text{unr}} \\ y_{\text{unr}} \\ 1 \end{pmatrix}, \tag{2}$$
$$r_{\text{ref}} = \delta_\theta + r_{\text{unr}},$$

where $\mathbf{R}_\theta = \begin{pmatrix} \cos\delta_\theta & -\sin\delta_\theta & 0 \\ \sin\delta_\theta & \cos\delta_\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}, \mathbf{T} = \begin{pmatrix} 1 & 0 & \delta_x \\ 0 & 1 & \delta_y \\ 0 & 0 & 1 \end{pmatrix}$.

We denote $\mathbf{R}_\theta$ as the rotation matrix with a rotation angle δ_θ, and $\mathbf{T}$ is the translation matrix, in which (δ_x, δ_y) is the translation vector.

The corresponding matching pair can be aligned after applying this transformation. This means the transformation matrix is just the relative camera pose between the two cameras in the BEV. Note that, this relative camera pose only contains three degrees of freedom, *i.e.*, the translation and rotation projected into the BEV plane. This way, we can obtain the relative camera pose $(\delta_x, \delta_y, \delta_\theta)$ by solving the above Eq. (2) and get

$$\begin{cases} \delta_x &= x_{\text{ref}} - x_{\text{unr}}\cos\delta_\theta + y_{\text{unr}}\sin\delta_\theta \\ \delta_y &= y_{\text{ref}} - x_{\text{unr}}\sin\delta_\theta - y_{\text{unr}}\cos\delta_\theta \\ \delta_\theta &= r_{\text{ref}} - r_{\text{unr}} \end{cases}. \tag{3}$$

As discussed above, we use K point pairs to estimate the relative pose, each of which can generate a relative pose estimation result as shown in SAM in Figure 2. We use the camera pose estimation loss function for training the LocoNet as

$$\mathcal{L}_{\text{Cam}} = \sum_{k=1}^{K}(\|(\delta_x^k, \delta_y^k) - (\delta_x^{\text{gt}}, \delta_y^{\text{gt}})\| + \|\delta_\theta^k - \delta_\theta^{\text{gt}}\|), \tag{4}$$

where $(\delta_x^k, \delta_y^k, \delta_\theta^k)$ is the k-th candidate relative camera pose estimation generated from the k-th point pair, and $(\delta_x^{\text{gt}}, \delta_y^{\text{gt}}, \delta_\theta^{\text{gt}})$ is the ground-truth camera pose. Note that, we apply the supervision on the camera position, *i.e.*, δ_x, δ_y, and the view direction, *i.e.*, δ_θ in our method. However, the camera view direction is very hard to measure and annotate in real-world applications. In the experiments, we show that our method is not very sensitive to the supervision of δ_θ.

3.4. Camera and Subject Registration

Camera Registration. Based on the relative camera pose $(\delta_x^k, \delta_y^k, \delta_\theta^k)$ obtained in Section 3.3, we have got K candidates of relative camera pose estimation between the reference and unregistered BEVs. Here we denote the camera pose on the reference BEV as (0, 0, 0). We then get a camera pose of the unregistered BEV on the coordinate system of the reference BEV as

$$\mathbf{c}^k = (c_x^k, c_y^k, c_\theta^k) = (0,0,0) + (\delta_x^k, \delta_y^k, \delta_\theta^k) = (\delta_x^k, \delta_y^k, \delta_\theta^k), \tag{5}$$

which denotes the k-th candidate camera pose from $(\delta_x^k, \delta_y^k, \delta_\theta^k)$, as shown in Figure 5.

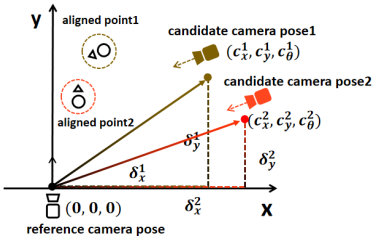

Figure 5. Candidate camera poses in the coordinate system.

The next step is to find the selected camera pose from the K candidates, *i.e.*, achieving the camera registration task. We calculate $\overline{c_x} = \frac{\sum_{k=1}^K c_x^k}{K}, \overline{c_y} = \frac{\sum_{k=1}^K c_y^k}{K}$, where (c_x^k, c_y^k) is the candidate position of the unregistered camera, $(\overline{c_x}, \overline{c_y})$ is the centroid point. We then compute the distance of each candidate position to the centroid point as

$$d_{\text{centroid}}^k = \|(c_x^k, c_y^k) - (\overline{c_x}, \overline{c_y})\|, \tag{6}$$

The candidate with the minimum distance will be selected, which is used to register the unregistered BEV into the reference BEV, then we can get a unified BEV as shown in the left part of Registration in Figure 2.

Subject Registration. With the camera registration result, we can register the camera position and its view direction, together with the subject localization and face orientation of the unregistered BEV, into the reference BEV. Note that, for multiple views, we select one as the reference BEV and others as the unregistered BEVs, all of which can be registered into the reference BEV, respectively. The next step is aggregating the same person from different views in the unified BEV, which can be achieved by two steps, *i.e.*, subject matching and fusion.

1) *Subject Matching.* To match the subjects from multiple views, we create a person spatial distance matrix $\mathbf{M}_{\text{dis}}$ and an angle difference matrix $\mathbf{M}_{\text{ang}}$ in the unified BEV, which measure the distance and angle differences of all persons from different views. We then combine it with similarity matrix $\mathbf{M}_{\text{pred}}$ provided in Section 3.3. We first employ three thresholds as filters to select potential matching subject pairs, whereby only pairs that fall within the distance and angle thresholds and surpass the similarity threshold will be identified as the same subject. Besides, we further consider two constraints for accurate matching. The first one is cycle consistency [32], which means the connection of the same subject from all views should form a loop. The second one is uniqueness, which means one subject should not be connected to more than one subject in another view.

For the above constraints, first, we use a classical data structure, *i.e.*, union-find, to aggregate the transitive relations, which makes all the subjects with direct and indirect connections in a union of union-find to be clustered as a sub-graph, as shown in Figure 6(b), which solves the problem of cycle consistency for all the subject connected as a loop. Second, we define the problem as a hierarchical maximum spanning subgraph problem, the layer-by-layer (view-by-view) spanning constrains that a subject is connected at most one node in each view to avoid the uniqueness conflict, as shown in Figure 6(c). To solve this problem, we propose an algorithm referenced from the Prim algorithm [45]. We provide more details and the algorithm flow of the above strategy in the *supplementary material*.

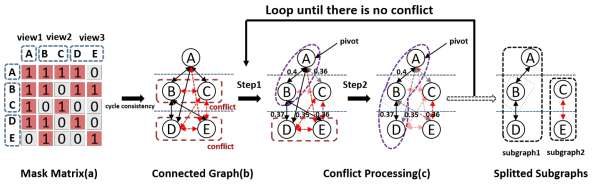

Figure 6. Solving the cycle consistency and uniqueness.

2) *Subject Fusion.* For the subjects from multiple views to be regarded as the same person using the above subject matching method, we then estimate the final registration result of a subject using the same strategy in Eq. (6). The position and orientation of the subject with the minimum centroid distance will be retained, and other same subject will be removed. Especially, if there are only two views, we use the mean position and orientation between two subjects as the fused result. Finally, we can get the unified BEV with the camera and subject registration from multiple views.

3.5. Self-supervision for Subject Association

Based on the above subject registration results, we further consider to use them for supervising the appearance-based subject association network with a back-propagation strategy. As shown in the bottom of Figure 2, we propose to train the appearance feature extraction network, *e.g.*, ResNet-50, for similarity matrix calculating in a self-supervised manner, to make full use of the spatial information from Section 3.4. Specifically, we inversely normalize each row of the spatial distance matrix $\mathbf{M}_{\text{dis}}$ and angle difference matrix $\mathbf{M}_{\text{ang}}$ discussed above, to get the spatial-aware similarity matrixes as $\mathbf{M}_{\text{spatial}} = \alpha\overline{\mathbf{M}_{\text{dis}}} + (1-\alpha)\overline{\mathbf{M}_{\text{ang}}}$, where α is the hyper-parameter, $\overline{\mathbf{M}_{\text{dis}}}$ and the $\overline{\mathbf{M}_{\text{ang}}}$ are the nor-

malized similarity matrix obtained from $\mathbf{M}_{\text{dis}}$ and $\mathbf{M}_{\text{ang}}$, respectively. After that, we apply a self-supervised loss to train the appearance feature extraction network as below

$$\mathcal{L}_{\text{App}} = \|\mathbf{M}_{\text{pred}} - \mathbf{M}_{\text{spatial}}\|. \quad (7)$$

3.6. Implementation Details

We pretrain the LocoNet using the camera location and view direction labels in our other synthetic data with MSE loss and use the pretrained model of ResNet-50 in [24]. We use the camera pose loss function and the self-supervised appearance learning loss function as the total loss function $\mathcal{L} = \mathcal{L}_{\text{Cam}} + \mathcal{L}_{\text{App}}$. We set the number of candidate K as 3 in Section 3.3 and the similarity matrix threshold as 0.25, distance threshold as 2.0 m and angle threshold as 15° in Section 3.4. We set the hyper-parameters of the pseudo matrix in Section 3.5 as $\alpha = 0.5$. We use a pair of FPVs to train our framework. In the inference stage, the number of FPVs is not limited, in which one FPV will be selected as the reference view and others can be registered in that reference view. We use Pytorch as our main framework and the work runs on the server with RTX 3090 GPU.

4. Experiments

4.1. Proposed Dataset

To our best knowledge, there is no available dataset that can be used for the task in this work, which requires the multi-view relative camera poses, the 3D position and the face orientation of each person. Even with expensive hardware, it is also very hard to obtain accurate annotations of them in the real world. So we consider using the modeling engine to create a synthetic dataset.

• *Flexible data controlling.* We use a 3D game development Unity 3D [47] to build a city scene and apply open-source 3D human model library PersonX [55] containing more than 1,000 different persons to generate subjects in the scene. Through the flexible development platform, we can create various scenes to simulate the real world. The cover area is set as $25m \times 25m$, in which all the objects are simulated to the real environment with a scaling.

• *Diverse subject settings.* For data diversity, the number of subjects in each frame is different, where the range of subjects in the scene is from 10 to 25, containing 5-20 people walking free and 5 camera-wears. Further, we generate every frame by random function, which means camera registration and subject registration are various in each frame.

• *Large scale.* We create two *C*amera *S*ubject *R*egistration *D*atasets, *i.e.*, CSRD-II and CSRD-V, which contain two views and five views, respectively. In total, CSRD-II includes 2,000 pairs of images, with 1,000 for training and another 1,000 for testing. CSRD-V includes 1,000 groups of images, in which each group contains 5 synchronous images. CSRD-V is only used for testing in our experiments.

• *Rich and accurate annotations.* Our annotations contain the position (in meters) and face orientation of each subject in the BEV, as well as the camera poses. Besides, we also provide the bounding box with the unified ID number of each subject in all views. More details about the proposed datasets can be obtained in the *supplementary material*.

4.2. Setup

Evaluation Metrics. Metric-I: We first evaluate the accuracy of the *camera registration* results, including the position and orientation results in the BEV. For the position, we calculate the distance between the predicted and ground-truth positions. Then we count the average error (*Cam.Pos.Avg*) and the percentages of the error within a list of a certain distance, including 0.5, 1, and 1.5 meters. Similarly, we calculate the angle error in average (*Cam.Ori.Avg*) and percentages of degree errors within certain ranges, including 5, 10 and 15 degrees. **Metric-II**: We also evaluate the *subject registration* results. It is similar to the metric-I, which evaluates the position distance and orientation error of the subjects. **Metric-III**: We finally evaluate the multi-view multiple human association (identification) results. We use precision, recall, and F_1 scores as the metrics.

Comparison Methods. As discussed above, there is no method that can directly handle the proposed problem. We include the following comparison methods for the *camera registration* task: *DMHA* [23] achieves the task of camera registration by using a real BEV image. We provide the FPV images and the corresponding BEV image to DMHA. *SIFT [41] + KNN* and other deep-learning-based methods [35, 48, 54, 65] are local descriptor (key point) matching based methods, which are combined with the classical camera pose estimation method with the matched key points for relative camera estimation. For the second task of *subject registration*, we include the following three methods. *Monoloco++* [8] is a network predict 3D-localization and face orientation of each person in the view. We concatenate it with our geometric transformation and subject fusion methods for comparison. *MVDet* and *MVDetr* [29, 30] are used for multi-view object detection with the camera calibrations. We provide more details about the implementation details of the comparison methods in the *supplementary material*.

4.3. Comparative Results

Camera Registration Results. We first evaluate the camera registration results on CSRD-II as shown at the top half in Table 1. We can first see that all the comparison methods provide very poor results. Among them, we provide the ground-truth BEV image to DMHA, which is used to find the camera wearer from the BEV instead of our position regression. The key point matching based methods almost fail because of the huge view differences. Monoloco++ gener-

Table 1. Camera registration results. The top half is comparison experiments, the bottom half is ablation study, in which 'Cam.Pos.Avg' and 'Cam.Ori.Avg' present the average error in meters of the camera position and the orientation error in degrees in BEV, 'Cam.Pos@d' represents the percentage of distance error within d meters and 'Cam.Ori.@r' represents the percentage of angle error within r degrees.

Methods	Cam.Pos.Avg	Cam.Ori.Avg	Cam.Pos@0.5	Cam.Pos.@1	Cam.Pos.@1.5	Cam.Ori.@5	Cam.Ori.@10	Cam.Ori.@15
Monoloco++ [8]	3.00	21.84	7.60%	21.60%	36.40%	17.50%	34.60%	47.10%
DMHA [23]	5.99	47.43	46.50%	47.60%	48.60%	46.20%	50.00%	53.60%
SIFT [41]	7.11	144.46	1.26%	2.34%	3.60%	4.80%	8.20%	11.10%
LoFTR [54]	11.50	90.11	0.70%	1.20%	1.70%	3.70%	6.50%	8.50%
SuperGlue [48]	11.17	89.74	0.60%	1.10%	1.50%	3.70%	6.50%	8.60%
CVNet [35]	11.38	115.10	0.88%	1.25%	1.75%	3.10%	5.5%	7.40%
R2Former [65]	13.55	102.52	0.35%	0.47%	0.83%	3.90%	7.20%	9.50%
Max	2.27	15.22	20.00%	42.30%	59.60%	33.90%	60.30%	76.00%
Random	1.91	12.62	21.60%	47.30%	65.00%	37.50%	65.80%	81.20%
w/o pre-train	6.98	33.02	0.50%	1.40%	3.20%	10.20%	20.90%	29.50%
w/o GT δ_θ	0.93	5.91	37.80%	71.80%	85.60%	59.10%	85.60%	94.30%
Ours	0.89	5.78	42.20%	72.40%	88.40%	59.50%	86.50%	94.80%

Table 2. Subject registration results. The expression of metrics of subject here is in the same way as Table 1.

Methods	Sub.Pos.Avg	Sub.Ori.Avg	Sub.Pos.@0.5	Sub.Pos.@1	Sub.Pos.@1.5	Sub.Ori.@5	Sub.Ori.@10	Sub.Ori.@15
Monoloco++ [8]	1.32	32.50	26.05%	61.47%	77.65%	13.21%	26.05%	38.17%
MVDetr [29]	2.41	-	11.18%	29.54%	46.07%	-	-	-
MVDet [30]	2.44	-	11.28%	29.19%	45.65%	-	-	-
w/o pre-train	6.35	89.29	1.62%	6.62%	11.41%	2.29%	4.74%	6.97%
w/o GT δ_θ	0.83	16.36	41.15%	77.89%	89.31%	32.30%	56.79%	72.77%
Max	1.27	21.56	37.39%	72.38%	82.87%	30.46%	54.95%	69.13%
Random	1.06	17.19	39.19%	74.62%	85.07%	33.61%	59.01%	73.39%
Ours	0.75	14.67	43.23%	81.43%	92.12%	35.07%	63.24%	79.15%

ates a relatively acceptable result since it's equipped with the proposed geometric transformation methods. For our method, the mean distance error is only 0.89 meters, less than 1. The most remarkable thing is the accuracy under camera angle error ≤ 15 degrees is more than 94%, even ≤ 5 degrees is up to 59%, and the mean error is less than 6 degrees. This is promising for many real-world applications.

Subject Registration Results. We also evaluate the subject registration in CSRD-II using Metric-II as shown in Table 2. Even MVDet and MVDetr take the camera calibration as prior, our method achieves much superior results in all metrics. At the same time, our method keeps the average distance error within 0.8 meters and the average orientation error within 15 degrees.

4.4. Ablation Study

• w/o pre-train.: Removing the pre-training of LocoNet.
• w/o GT δ_θ: Removing the supervision of the camera orientation in Eq. (4).
• Max/Random: In the candidate camera selection strategy, we choose the max confidence pair or choose randomly instead of our method in Eq. (6).

As shown at the bottom half of Table 1, the ablation study, *i.e.*, ' w/o pre-train', verifies the necessity of the pretrained LocoNet in VTM. We can also see from the next row that, when removing the camera orientation supervision in SAM, *i.e.*, 'w/o GT δ_θ', the performance only drops a little. This demonstrates that our method *is not heavily dependent on* the camera orientation supervision, which is not easy to obtain in the real world. For camera pose selection in Registration module, we can see that no matter whether using the strategy of the max confidence one or the random one, which, not considering the spatial-aware selecting strategy, both provide a relatively poor performance than our centroid strategy. We also conduct the ablation study on the subject registration task, as shown in Table 2. Similar to the above results, we can see the effectiveness of the pretrained LocoNet in VTM, camera orientation supervision in SAM, and the centroid strategy in Registration module.

We further evaluate the results of multi-view human association in CSRD-II, which can verify *the effectiveness of the proposed backward self-supervised training strategy in SAM*. As shown in Table 3, the baseline is the ResNet-50 model pre-trained on the person Re-ID dataset named Market-1501 [64], on which we apply the self-supervised training strategy as discussed in Section 3.5. 'w GT re-id' denotes that we provide the ground-truth assignment matrix to supervise the result of the similarity matrix. We can see from Table 3 that our self-supervision strategy improves the F_1 score from the original 66.78% to 85.98% with a large margin. We can also see that our results are very close to

Table 4. Multi-view camera and subject registration, and multi-view subject association results.

Methods	Cam.Pos.Avg	Cam.Ori.Avg	Cam.Pos.@1	Cam.Ori.@10	Sub.Pos.Avg	Sub.Ori.Avg	Sub.Pos.@1	Sub.Ori.@10	F_1
Pair-wise	1.06	6.96	62.71%	79.61%	0.75	14.67	80.76%	58.81%	83.85%
Multi-view w/o constraints	1.06	6.93	63.55%	80.60%	1.10	15.64	63.78%	50.47%	85.64%
Multi-view w constraints	1.06	6.93	63.55%	80.60%	0.94	13.45	70.57%	57.73%	86.12%

the result of supervised training, with only a small gap of 0.45%. The results of the association verify the effectiveness of the proposed self-supervision strategy.

Table 3. Cross-view subject association results.

Methods	Precision	Recall	F_1
Baseline [24]	57.48%	82.98%	66.78%
Ours	79.33%	95.45%	85.98%
w GT re-id (oracle)	77.97%	98.18%	86.43%

We further evaluate the proposed method on the scenes using multiple cameras for camera and subject registration on CSRD-V, as shown in Table 4. The first row shows the results that we split the 5 views into $C_5^2 = 10$ pair-wise views, on which we apply the proposed method for two views as above. The second and third rows are the results of multi-view subject and camera registration without or with the constraints during subject matching in Section 3.4. We can see that using the proposed constraints effectively improves the results on subject registration and multi-view subject association, which demonstrates *the effectiveness of the subject registration strategy in the Registration module*. With respect to the results of two views registration, we can see that even though the results are slightly worse in 5 views, the overall results are still very impressive, which demonstrates the stability of our method in multiple views.

4.5. In-depth Analysis

Real-world Dataset Evaluation. We propose a large-scale real-world *evaluation dataset* CSRD-R, to test the performance of the cross-domain of our method, which includes 15,490 frames and five different scenes. There are 1,500 synchronous frame groups for the two-view scene, 830 synchronous frame groups for the three-view scene, and 2,500 synchronous frame groups for the four-view scene. In addition to the first-person views provided by the wearing cameras, we also capture the real BEV using a UAV. For all the first-person-view and BEV videos in the dataset, we annotate the bounding boxes for each subject and label the unified ID for same subject in all views. Next, we conducted cross-domain experiments where we train our model on the synthetic dataset CSRD-II and performed a cross-domain evaluation on the real dataset, CSRD-R.

As shown in Table 5, we provide the detection performance of our method. Note that, considering the gap between the BEV generated by our method and the real BEV, we define the new detection metric on CSRD-R, which is provided in the *supplementary material*. The results demonstrate the effectiveness of our method on real data and its reliable cross-domain generalization ability.

Table 5. Results on CSRD-R for different numbers of views.

	Two Views	Three Views	Four Views
Ours	82.50%	85.07%	86.31%

Qualitative Analysis. Figure 7 shows a case, in which we can see that the prediction of both camera and subject registration can achieve a good coincidence with the ground truth, thanks to the high accuracy of our method. We also provide the real-world case, as shown in Figure 8. Note that, we directly apply our method to the real-world case without any additional annotation. We can see that, except for a wrong fusion coming from the incorrect matching, the prediction of the subject and camera distributions are very close to the real BEV. This demonstrates the robustness and generalization of the proposed method. More visualization results with special conditions and analyses on sensitivity and complexity are available in the *supplementary material*.

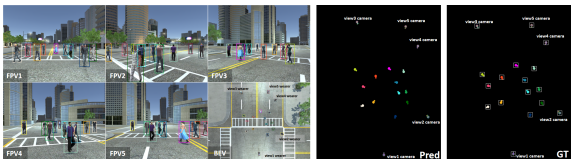

Figure 7. Qualitative case analysis. We add a white rectangle around every ground-truth subject.

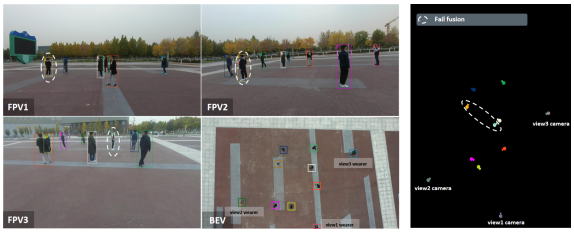

Figure 8. Real-world case study.

5. Conclusion

In this paper, we have studied a new problem of multi-view camera and subject registration tasks in BEV without camera calibrations. For this problem, we develop a new approach that can simultaneously handle these two tasks. Specifically, the proposed method uses an end-to-end framework, which makes full use of deep network based appearance information and multiple view geometry based spatial knowledge to complement each other's advantages. We also create new synthetic and real-world datasets with various settings and rich annotations. Experimental results show the superior performance of our method.

Acknowledgment. This work was supported in part by the NSFC under Grants 62072334, U1803264.

References

[1] Shervin Ardeshir and Ali Borji. Ego2top: Matching viewers in egocentric and top-view videos. In *Proceedings of the European Conference on Computer Vision*, pages 253–268, 2016. 3

[2] Shervin Ardeshir and Ali Borji. Egocentric meets top-view. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(6):1353–1366, 2018.

[3] Shervin Ardeshir and Ali Borji. Integrating egocentric videos in top-view surveillance videos: Joint identification and temporal alignment. In *Proceedings of the European Conference on Computer Vision*, pages 285–300, 2018. 3

[4] Pierre Baqué, François Fleuret, and Pascal Fua. Deep occlusion reasoning for multi-camera multi-target detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 271–279, 2017. 1, 2

[5] Erkan Baser, Venkateshwaran Balasubramanian, Prarthana Bhattacharyya, and Krzysztof Czarnecki. Fantrack: 3D multi-object tracking with feature association network. In *Proceedings of the IEEE Intelligent Vehicles Symposium*, pages 1426–1433, 2019. 3

[6] Nuri Benbarka, Jona Schröder, and Andreas Zell. Score refinement for confidence-based 3D multi-object tracking. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 8083–8090, 2021. 3

[7] Lorenzo Bertoni, Sven Kreiss, and Alexandre Alahi. Monoloco: Monocular 3D pedestrian localization and uncertainty estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6861–6871, 2019. 2

[8] Lorenzo Bertoni, Sven Kreiss, and Alexandre Alahi. Perceiving humans: from monocular 3D localization to social distancing. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):7401–7418, 2021. 2, 6, 7

[9] Tolga Birdal, Emrah Bala, Tolga Eren, and Slobodan Ilic. Online inspection of 3D parts via a locally overlapping camera network. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 1–10, 2016. 3

[10] Guillem Brasó and Laura Leal-Taixé. Learning a neural solver for multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6247–6257, 2020. 3

[11] Andrea Censi, Antonio Franchi, Luca Marchionni, and Giuseppe Oriolo. Simultaneous calibration of odometry and sensor parameters for mobile robots. *IEEE Transactions on Robotics*, 29(2):475–492, 2013. 3

[12] Mohamed Chaabane, Peter Zhang, J Ross Beveridge, and Stephen O'Hara. Deft: Detection embeddings for tracking. *arXiv preprint arXiv:2102.02267*, 2021. 3

[13] Tatjana Chavdarova and François Fleuret. Deep multi-camera people detection. In *Proceedings of the IEEE International Conference on Machine Learning and Applications*, pages 848–853, 2017. 1, 2

[14] 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*, pages 5030–5039, 2018. 2

[15] Hansheng Chen, Yuyao Huang, Wei Tian, Zhong Gao, and Lu Xiong. Monorun: Monocular 3D object detection by reconstruction and uncertainty propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10379–10388, 2021. 1, 2

[16] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, and Feng Zhao. Graph-detr3d: Rethinking overlapping regions for multi-view 3D object detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5999–6008, 2022. 2

[17] Hsu-kuang Chiu, Jie Li, Rareş Ambruş, and Jeannette Bohg. Probabilistic 3D multi-modal, multi-object tracking for autonomous driving. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 14227–14233, 2021. 3

[18] Junting Dong, Qi Fang, Wen Jiang, Yurou Yang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3D pose estimation and tracking from multiple views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6981–6992, 2021. 2

[19] Shuai Dong, Xinxing Shao, Xin Kang, Fujun Yang, and Xiaoyuan He. Extrinsic calibration of a non-overlapping camera network based on close-range photogrammetry. *Applied optics*, 55(23):6363–6370, 2016. 3

[20] Ruize Han, Yujun Zhang, Wei Feng, Chenxing Gong, Xiaoyu Zhang, Jiewen Zhao, Liang Wan, and Song Wang. Multiple human association between top and horizontal views by matching subjects' spatial distributions. *arXiv preprint arXiv:1907.11458*, 2019. 1, 3

[21] Ruize Han, Jiewen Zhao, Wei Feng, Yiyang Gan, Liang Wan, and Song Wang. Complementary-view co-interest person detection. In *Proceedings of the ACM International Conference on Multimedia*, pages 2746–2754, 2020. 1

[22] Ruize Han, Wei Feng, Yujun Zhang, Jiewen Zhao, and Song Wang. Multiple human association and tracking from egocentric and complementary top views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5225–5242, 2021. 1

[23] Ruize Han, Yiyang Gan, Jiacheng Li, Feifan Wang, Wei Feng, and Song Wang. Connecting the Complementary-View Videos: Joint Camera Identification and Subject Association. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2416–2425, 2022. 1, 3, 6, 7

[24] Ruize Han, Yun Wang, Haomin Yan, Wei Feng, and Song Wang. Multi-view multi-human association with deep assignment network. *IEEE Transactions on Image Processing*, 31:1830–1840, 2022. 6, 8

[25] Ruize Han, Yiyang Gan, Likai Wang, Nan Li, Wei Feng, and Song Wang. Relating view directions of complementary-view mobile cameras via the human shadow. *International Journal of Computer Vision*, 131(5):1106–1121, 2023. 3

[26] Ruize Han, Wei Feng, Feifan Wang, Zekun Qian, Haomin Yan, and Song Wang. Benchmarking the complementary-

view multi-human association and tracking. *International Journal of Computer Vision*, 132(1):118–136, 2024. 3

[27] Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14141–14152, 2021. 2, 3

[28] Jun Hayakawa and Behzad Dariush. Recognition and 3D localization of pedestrian actions from monocular video. In *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*, pages 1–7, 2020. 2

[29] Yunzhong Hou and Liang Zheng. Multiview detection with shadow transformer (and view-coherent data augmentation). In *Proceedings of the ACM International Conference on Multimedia*, pages 1673–1682, 2021. 1, 2, 6, 7

[30] Yunzhong Hou, Liang Zheng, and Stephen Gould. Multiview detection with feature perspective transformation. In *Proceedings of the European Conference on Computer Vision*, pages 1–18, 2020. 1, 2, 6, 7

[31] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3D object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 2

[32] Qi-Xing Huang and Leonidas Guibas. Consistent shape maps via semidefinite programming. In *Computer graphics forum*, pages 177–186, 2013. 5

[33] Yanqin Jiang, Li Zhang, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, and Yu-Gang Jiang. Polarformer: Multi-camera 3D object detection with polar transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1042–1050, 2023. 2

[34] S. Kreiss, L. Bertoni, and A. Alahi. PifPaf: Composite Fields for Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11977–11986, 2019. 4

[35] Seongwon Lee, Hongje Seong, Suhyeon Lee, and Euntai Kim. Correlation verification for image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5374–5384, 2022. 6, 7

[36] Peiliang Li, Tong Qin, et al. Stereo vision-based semantic 3D object and ego-motion tracking for autonomous driving. In *Proceedings of the European Conference on Computer Vision*, pages 646–661, 2018. 2

[37] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3D object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1477–1485, 2023. 2

[38] Zhenyu Li, Zehui Chen, Ang Li, Liangji Fang, Qinhong Jiang, Xianming Liu, and Junjun Jiang. Unsupervised domain adaptation for monocular 3D object detection via self-training. In *Proceedings of the European Conference on Computer Vision*, pages 245–262, 2022. 2

[39] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *Proceedings of the European Conference on Computer Vision*, pages 1–18, 2022. 2

[40] Zhen Liu, Fengjiao Li, and Guangjun Zhang. An external parameter calibration method for multiple cameras based on laser rangefinder. *Measurement*, 47:954–962, 2014. 3

[41] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 6, 7

[42] Shujie Luo, Hang Dai, Ling Shao, and Yong Ding. M3dssd: Monocular 3D single stage object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6145–6154, 2021. 1

[43] Xinzhu Ma, Yinmin Zhang, Dan Xu, Dongzhan Zhou, Shuai Yi, Haojie Li, and Wanli Ouyang. Delving into localization errors for monocular 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4721–4730, 2021. 1, 2

[44] Ziqi Pang, Zhichao Li, and Naiyan Wang. Simpletrack: Understanding and rethinking 3D multi-object tracking. In *Proceedings of the European Conference on Computer Vision*, pages 680–696, 2023. 3

[45] Robert Clay Prim. Shortest connection networks and some generalizations. *The Bell System Technical Journal*, 36(6): 1389–1401, 1957. 5

[46] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8555–8564, 2021. 1

[47] John Riccitiello. John riccitiello sets out to identify the engine of growth for unity technologies (interview). *VentureBeat. Interview with Dean Takahashi. Retrieved January*, 18 (3), 2015. 6

[48] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4938–4947, 2020. 2, 3, 6, 7

[49] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016. 3

[50] Matthew Shere, Hansung Kim, and Adrian Hilton. 3D human pose estimation from multi person stereo 360 scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2019. 2

[51] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[52] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am I looking at? joint location and orientation estimation by cross-view matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4064–4072, 2020. 3

[53] 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*, pages 6049–6057, 2021. 1, 2

[54] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8922–8931, 2021. 2, 6, 7

[55] Xiaoxiao Sun and Liang Zheng. Dissecting person re-identification from the viewpoint of viewpoint. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 608–617, 2020. 6

[56] Xiao Tan, Zhigang Wang, Minyue Jiang, Xipeng Yang, Jian Wang, Yuan Gao, Xiangbo Su, Xiaoqing Ye, Yuchen Yuan, Dongliang He, et al. Multi-camera vehicle tracking and re-identification based on visual and spatial-temporal features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 275–284, 2019. 2

[57] Li Wang, Liang Du, Xiaoqing Ye, Yanwei Fu, Guodong Guo, Xiangyang Xue, Jianfeng Feng, and Li Zhang. Depth-conditioned dynamic message propagation for monocular 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 454–463, 2021. 1, 2

[58] Tai Wang, Jiangmiao Pang, and Dahua Lin. Monocular 3D object detection with depth from motion. In *Proceedings of the European Conference on Computer Vision*, pages 386–403. Springer, 2022. 2

[59] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3D object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022. 2

[60] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. 3D multi-object tracking: A baseline and new evaluation metrics. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 10359–10366, 2020. 3

[61] Xinshuo Weng, Yongxin Wang, Yunze Man, and Kris Kitani. Gnn3dmot: Graph neural network for 3D multi-object tracking with multi-feature learning. *arXiv preprint arXiv:2006.07327*, 2020.

[62] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3D object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11784–11793, 2021.

[63] Jan-Nico Zaech, Alexander Liniger, Dengxin Dai, Martin Danelljan, and Luc Van Gool. Learnable online graph representations for 3D multi-object tracking. *IEEE Robotics and Automation Letters*, 7(2):5103–5110, 2022. 3

[64] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1116–1124, 2015. 7

[65] Sijie Zhu, Linjie Yang, Chen Chen, Mubarak Shah, Xiaohui Shen, and Heng Wang. R2former: Unified retrieval and reranking transformer for place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19370–19380, 2023. 6, 7