Multiple Human Association and Tracking From Egocentric and Complementary Top Views

Ruize Han, Wei Feng, *Member, IEEE*, Yujun Zhang, Jiewen Zhao, and Song Wang, *Senior Member, IEEE*

Abstract—Crowded scene surveillance can significantly benefit from combining egocentric-view and its complementary top-view cameras. A typical setting is an egocentric-view camera, e.g., a wearable camera on the ground capturing rich local details, and a top-view camera, e.g., a drone-mounted one from high altitude providing a global picture of the scene. To collaboratively analyze such complementary-view videos, an important task is to associate and track multiple people across views and over time, which is challenging and differs from classical human tracking, since we need to not only track multiple subjects in each video, but also identify the same subjects across the two complementary views. This paper formulates it as a constrained mixed integer programming problem, wherein a major challenge is how to effectively measure subjects similarity over time in each video and across two views. Although appearance and motion consistencies well apply to over-time association, they are not good at connecting two highly different complementary views. To this end, we present a spatial distribution based approach to reliable cross-view subject association. We also build a dataset to benchmark this new challenging task. Extensive experiments verify the effectiveness of our method.

Index Terms—Crowded scene surveillance, top view, horizontal view, complementary view, human association, tracking, wearable cameras, video surveillance, egocentric perception

1 Introduction

Video surveillance is widely applicable to many civil and military scenarios. Classical video surveillance using fixed cameras has the limitation of covering only pre-specified area from fixed view angles. Mobile camera technology provides a new perspective to address this limitation. In particular, cameras mounted to drones in the air can provide birds-eye top views of a group of subjects on the ground, and wearable cameras, e.g., GoPro, mounted over the head of a wearer (one of the subjects on the ground), can provide egocentric horizontal views of the same group of subjects. In this paper, we leverage such top- and horizontal-view videos taken by respective moving cameras for better collaborative video surveillance, as shown in Fig. 1a. A typical application scenario is for outdoor law enforcement: a flying drone with a mounted camera, together with several officials with helmet cameras, can form such a mobile-camera network with which we can perform collaborative tracking, localization, and individual/group activity recognition, for the covered subjects.

Such top- and horizontal-view videos actually provide highly *complementary* information that is very important for surveillance. As shown in Figs. 1b and 1c, with a high altitude and a birds-eye view vertical to the ground, a top-view camera can provide a *global picture* of the whole scene and spatial distribution of the subjects, but could not capture detailed appearance of individual ones, while horizontal-view cameras on the ground can better capture *local details* of subjects appearance by moving cameras closer to the subjects of interest. Furthermore, mutual occlusions are common for crowded subjects in horizontal-view videos, but very rare in a top-view video. For simplicity, in this paper, we consider the mobile-camera network of one top-view camera and one horizontal-view camera for collaborative complementary-view analysis, i.e., we jointly analyze a top-view video and a horizontal-view video, both taken at the same time.

While the above complementary-view mobile-camera network is applicable to various video-surveillance scenarios, in this paper we focus on the fundamental problem of video-based human tracking. Different from traditional human tracking based on frame-by-frame *over-time* subject association along a single video, in this work we need not only *over-time* subject association along the top- or horizontal-view video, but also *cross-view* subject association between the two complementary views, to achieve a reliable collaborative multiple-human tracking. As shown in Figs. 1d and 1e, such collaborative tracking can capture global spatial distribution and trajectories of all the subjects in the top view as well as their local detailed appearance and activities in the horizontal view.

To achieve this goal, we need to first build data similarity models for over-time and cross-view subject associations,

- *Ruize Han, Wei Feng, Yujun Zhang, and Jiewen Zhao are with the School of Computer Science and Technology, College of Intelligence and Computing, Tianjin University, Tianjin 300350, China, and also with the Key Research Center for Surface Inspection and Analysis of Cultural Relics, SACH, Tianjin 300350, China. E-mail: {han_ruize, wfeng, yujunzhang, zhaojw}@tju.edu.cn.*
- *Song Wang is with the Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29201 USA, and also with the School of Computer Science and Technology, College of Intelligence and Computing, Tianjin University, Tianjin 300350, China. E-mail: songwang@cec.sc.edu.*

Manuscript received 29 Feb. 2020; revised 9 Feb. 2021; accepted 17 Mar. 2021. Date of publication 2 Apr. 2021; date of current version 4 Aug. 2022. (Corresponding authors: Wei Feng and Song Wang.) Recommended for acceptance by H. Kjellstrom. Digital Object Identifier no. 10.1109/TPAMI.2021.3070562

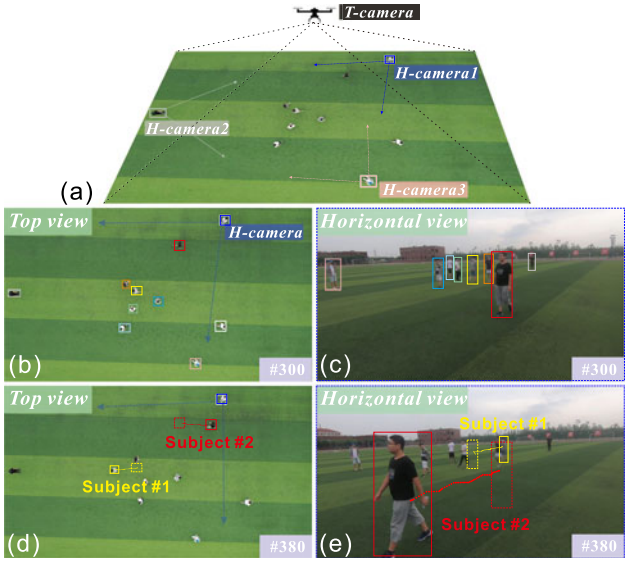


Fig. 1. An illustration of the proposed complementary-view subject association and collaborative tracking. (a) Mobile-camera network that produces both top- and horizontal-view videos. Here the top-view video is taken by a camera mounted to a drone from a high altitude with view direction roughly vertical to the ground and the horizontal-view videos are taken by GoPro worn by wearers among the subjects on the ground. (b) and (c) illustrate a sample cross-view subject association: identical-color boxes indicate the same subject across the top view (b) and the horizontal view (c). The location of horizontal-view camera that produces (c) is indicated by a blue box in (b). (d) and (e) illustrate the trajectories of two subjects in top and horizontal views, respectively.

respectively, and then develop an optimization scheme to integrate the over-time and cross-view associations for final collaborative tracking. While data similarity for over-time association on each video can be modeled by measuring the consistency of subject appearance and motion features frame by frame as in traditional multiple human tracking, cross-view similarity modeling and subject association are more challenging, since subject appearance and motion features are usually highly inconsistent between the top and horizontal views. As shown in Fig. 1b, the top-view direction is roughly vertical to the ground from a high altitude. In the top view, only the head and shoulders of subjects are visible and such limited appearance is barely useful to distinguish and associate subjects between the top and horizontal views. Given such a large view difference, it is also very difficult to use classical key point detection/matching and multi-view geometry theory to build cross-view spatial correspondence. In addition, subject motions in these two views are inconsistent and cannot be related, since both cameras are moving without calibration. In practice, even human eyes may not be able to identify the same person across the top and horizontal views only based on the appearance and motion. Besides, integrating over-time and cross-view associations for final collaborative multi-subject tracking is not trivial either, given inevitable errors in both over-time and cross-view associations.

In this paper, we propose a new joint optimization formulation to integrate the over-time and cross-view subject associations for collaborative tracking. Specifically, we split the video in each view into short clips with equal length, and extract tracklets in each clip that we refer to as *single-view tracklets*. The tracklets from adjacent clips in both top- and horizontal-views are used to establish the over-time and cross-view subject association, resulting in the *cross-view short trajectories*. We then solve the multi-clip cross-complementary-view data association as a constrained mixed integer programming problem. We finally stitch the short trajectories over time to get the *cross-view long trajectories* as the collaborative tracking results.

In this formulation, over-time subject association is modeled by classical appearance and motion consistency based similarity. For cross-complementary-view association, we propose a new frame-level similarity measure by comparing the subjects' spatial distributions across the top- and horizontal-views. Specifically, in each frame, from the horizontal view, we detect all subjects and estimate their spatial distribution and rough depths using the positions and sizes, respectively. In the corresponding top-view frame, we traverse all detected subjects and possible view directions to localize the horizontal-view camera (wearer), as well as its view orientation. For each traversed location and orientation, we estimate the spatial distribution of all visible subjects. We finally define a robust matching cost between the subjects' spatial distributions, to decide the horizontal-view camera location and view orientation, with which we can reliably associate multiple subjects across the top- and horizontal-views. Extensive experiments verify that our approach can effectively associate and track multiple subjects across the complementary top- and horizontal-views.

The main contributions of this paper are:

- We propose a new mobile-camera network setting, combining top- and horizontal-view videos for collaborative human tracking, where the top-view is from a high altitude with an approximate vertical-to-ground view angle to capture the global picture and trajectories of the subjects.
- We develop a new effective geometry-based algorithm for cross-complementary-view subject association by identifying and matching the subjects' spatial distributions, and build a joint optimization scheme to integrate the over-time and cross-view associations for reliable collaborative multi-subject tracking.
- We build a new dataset of top- and horizontal-view video pairs for benchmark evaluation. We have released the dataset to the public.[1]

The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 elaborates the proposed method. Section 4 reports the experimental results, followed by a further discussion in Section 5 and a brief conclusion in Section 6. This paper is a substantial extension from a preliminary conference version [1] with a number of major changes. First, we add spatial distribution based cross-view human association algorithm in Section 3, which is a key component for effectively measuring cross-view subject similarity (Section 3.2). Second, a new appearance reasoning for cross-view subject association is introduced in Section 3.3. Finally, we also significantly extend experimental comparisons and analyses in Section 4, and discuss the potential applications and limitations of our method in Section 5.

1. https://github.com/RuizeHan/CVMHT

2 Related Work

Single-View Multi-Object Tracking. Multiple object tracking (MOT) on a single view or a single video has been studied for a long time. In general, existing MOT methods can be divided into offline and online methods. The former treats MOT as an optimization problem of global data association [2], [3], [4] over the video and is limited to offline applications. The latter only uses the information on the current and previous frames [5], [6], [7], [8] when tracking objects in the next frame, thus is suitable for real-time applications. Online MOT methods, which, to some extent, can be viewed as associating multiple tracklets provided by single object tracking (SOT) [9], [10], [11], [12], usually have difficulties in handling long-term occlusions or mis-detections. To achieve reliable association, many appearance and motion features are used in MOT. Color histogram is a popular feature representation for appearance in MOT [3], [13], [14]. Recently, deeply learned appearance features are also used in tracking [6], [15], [16]. Both linear and nonlinear models have been used for representing motion features in MOT. While linear motion models assume constant velocity of moving objects across frames [14], [17], [18], [19], nonlinear motion models may lead to more accurate predictions [20], [21]. In this paper, we also use appearance and motion consistency for over-time association as in these existing MOT methods. The difference lies in that we further introduce cross-view association for collaborative tracking in top- and horizontal-view videos.

Multi-View Multi-Object Tracking. Multi-view MOT combines multiple videos taken from different view angles for object tracking and similar to the proposed work, it requires models for both over-time and cross-view associations. Many new problem formulations and solutions have been proposed for multi-view MOT in recent years. For example, Fleuret *et al.* [22] propose a generative model with dynamic programming for multi-object tracking. Wu *et al.* [23] formulate the problem as multidimensional assignment, which can be solved by a greedy randomized search algorithm. Liu *et al.* [24] introduce a unified probabilistic framework and develop a cluster Markov Chain based method for cross-view human tracking. Prior works also studied multi-view object tracking [25], [26] by exploring the subject appearance/motion consistency across different horizontal-view angles, e.g., frontal, back and side views, where no top-view videos are involved. Multiple-view information other than appearance and motion, such as geometrical relations [27] and spatial reconstructions [28] has also been used for association and MOT in many existing works. More recently, Tang *et al.* [29] jointly use the appearance model, geometry and pose information for associating tracklets across different views. Similarly, Liu *et al.* [30] integrate more semantic attributes including moving speed, accessories and activities to complement appearance features for associating subjects across cameras. However, multi-views used in these methods are usually horizontal views with different view angles. This way, most of these multi-view MOT methods can still use appearance and motion feature matching to infer the cross-view association. In this paper, we aim to track and associate subjects across top and horizontal views, in which neither the appearance/motion features nor other attributes, e.g., pose and activity information, are very useful.

Combining Top and Horizontal Views. More related to our work is a series of recent works by Ardeshir and Borji [31], [32], [33] on building association between top-view and horizontal-view cameras. In [31], [32], by jointly handling a set of egocentric (first-person) horizontal-view videos and a top-view video, a graph-matching based algorithm is developed to locate all the horizontal-view camera wearers in the top-view video. In [33], the problem is extended to locate not only the camera wearers, but also other subjects across the horizontal- and top-view videos. However, this series of works are based on two assumptions: 1) the top-view camera bears certain slope view angle and lower altitude to enable the partial visibility of human body and the use of appearance matching for cross-view association. Such setting may reduce the top-view camera's field of view and coverage, as well as introducing mutual occlusions between subjects, thus affects its applicability in practice; 2) the looking-at direction of the horizontal-view camera is the same as the moving direction of the camera wearer, such that optical flow can be used to help identify the camera wearer's view direction in the top view. This is not always true in practice, since a horizontal-view camera wearer may turn head when she/he walks, leading to inconsistency between her/his moving direction and view direction. In this paper, we remove these two assumptions and leverage the spatial distribution of subjects for cross-view subject association. We propose a new spatial-distribution-aware matching function to associate subjects across a top-view frame and its corresponding horizontal-view frame in Section 3.2. We also consider multi-frame information for spatial-temporal subject association and tracking in Section 3.3.

Recently, in the autonomous driving research, the LiDAR based bird's eye view (BEV) and RGB camera view are jointly utilized for 3D object detection and tracking, which shows certain similarity to the proposed collaborative analysis of top- and horizontal-view videos. Yang *et al.* propose a proposal-free single-stage detector for 3D object detection [34] and further use high-definition (HD) maps as priors to enhance the robustness of 3D object detectors [35]. Chen *et al.* propose a sensory-fusion framework that takes both LIDAR point clouds and RGB images as input to predict the 3D bounding boxes [36]. Liang *et al.* design an end-to-end architecture to fuse the feature maps of image and point clouds [37]. Besides, Zhang *et al.* [38] propose a multi-modality multi-object tracking (mmMOT) framework, and Kuang *et al.* [39] further propose a general multi-modality cascaded fusion framework for robust 3D multi-object tracking. These works differ from the proposed work in two aspects: 1) the bird's eye view (BEV) is obtained by a top-view projection of the point clouds obtained by LiDAR. It is not a real top view as in our problem since it can not capture horizontally occluded subjects in LiDAR imaging. 2) These methods usually take multi-modality data, e.g., RGB images and point clouds, as input and try to construct their effective representations for 3D object detection/tracking, while our method only uses the complementary-view RGB images for collaborative association and tracking.

Different from these prior works, this paper, to the best of our knowledge, for the first time studies complementary-view multiple subject association and tracking. We investigate both the over-time and cross-view subject associations

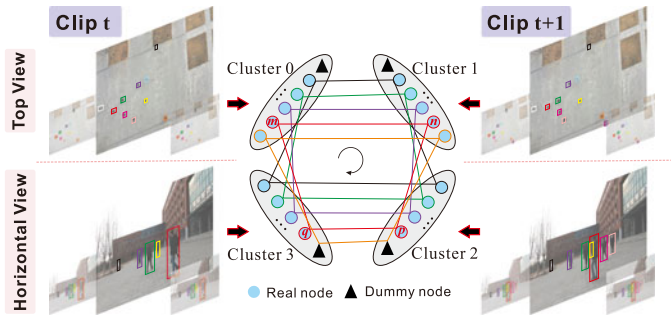


Fig. 2. An illustration of subject associations between consecutive clips and across the two views. Solid triangle in each cluster represents its dummy node.

for collaborative tracking by combining a top-view video, taken by a camera with vertical-to-the-ground view angle and high altitude, and a horizontal-view video, taken by camera worn by one subject on the ground.

3 Complementary-View Multiple Human Association and Tracking

In this section, we formulate the proposed complementary-view multiple human association and tracking (CVMHAT) within a joint optimization scheme in Section 3.1, which integrates cross-view and over-time tracklet-level subject association, as elaborated in Sections 3.3 and 3.4, respectively. Section 3.2 specifically studies cross-complementary-view association at frame level, that is a key component of our method. Section 3.5 finally summarizes the proposed tracking framework.

3.1 Problem Formulation

Given a pair of temporally-synchronized videos that are taken from the top view and horizontal view, respectively, we first synchronously split these two videos into short clips with a fixed length, e.g., 10 frames. In each video clip, we extract a set of subject tracklets using a simple strategy based on spatial overlap and appearance similarity, with implementation details given in Section 4.1. We refer to the resulting tracklets as *single-view tracklets* since they are extracted from the top-view and horizontal-view videos, respectively. We then conduct cross-view over-time data association for these single-view tracklets. More specifically, as shown in Fig. 2, the single-view tracklets from two adjacent clips in two views are fed into corresponding cluster as nodes. We then establish the subject associations between clips and across views by solving a joint optimization function. We refer to the generated tracking trajectories between two clips and across two views as *cross-view short trajectories*. Finally, we stitch the short trajectories considering the frame overlap over time and obtain the *cross-view long trajectories* as the final cross-view tracking result.

The single-view tracklets from two adjacent clips in two views constitute four clusters 0, 1, 2, and 3, as shown in Fig. 2. Each (real) node in a cluster represents a single-view tracklet in respective clip/view. We define e^i_{mn} as the binary variable for the edge between the node m in cluster i and the node n in cluster $(i+1)|4$, where $|$ is the modulo operation: $e^i_{mn}=1$ indicates that the corresponding two tracklets are associated in tracking while $e^i_{mn}=0$ indicates not. For the four clusters shown in Fig. 2, cluster $(i+1)|4$ is adjacent to cluster i in a clockwise order. This way, edge connection is only considered between tracklets from the same view or from the same clip. Besides, the edge connection is intended to be single-direction and we define c^i_{mn} as the weight of edge e^i_{mn}. Edge weight reflects the similarity of two tracklets over time or across views and we will elaborate on its definitions later.

The association among these four clusters can be formulated as maximizing an energy function

$$\arg\max_{\mathbf{e},\mathbf{d}} \quad \sum_{i=0}^{3}\left(\sum_{n=1}^{N_{i'}}\sum_{m=1}^{N_i} c^i_{mn}e^i_{mn} + c_0 d^i\right), \tag{1}$$

where N_i denotes the number of real nodes (tracklets) in cluster i, and $i'=(i+1)|4$. Besides real nodes, we also add a dummy node [18] for each cluster, which can be connected to multiple nodes (both real and dummy nodes) in other clusters.[2] The variable d^i counts the number of incoming and outgoing edges connected to the dummy node in cluster i, which can be any nonnegative integer value.

This is a mixed integer programming (MIP) problem and we further consider three constraints for $i=0,1,2,3$:

Constraint 1 limits that there is at most one edge between 1) nodes in cluster i and a real node in cluster i', 2) a real node in cluster i and nodes in cluster i', where $i'=(i+1)|4$

$$\sum_{m=1}^{N_i} e^i_{mn} \le 1, \quad \sum_{n=1}^{N_{i'}} e^i_{mn} \le 1. \tag{2}$$

Constraint 2 ensures that resulting edge connections form loops among four clusters

$$e^i_{mn} + e^{i'}_{np} + e^{i''}_{pq} \le 2 + e^{i'''}_{qm}, \tag{3}$$

where $i'=(i+1)|4$, $i''=(i'+1)|4$, and $i'''=(i''+1)|4$.

Constraint 3 ensures that the same number of K edges are selected for association between two adjacent clusters

$$\sum_{q=1}^{N_{i'''}}\sum_{m=1}^{N_i} e^{i'''}_{qm} + \sum_{m=1}^{N_i}\sum_{n=1}^{N_{i'}} e^i_{mn} + d^i = 2K. \tag{4}$$

Considering the above three constraints, we solve the integer programming to obtain edge sets between different clusters, which provide the desired subject association and tracking results.

In the following, we show the subject association reasonings between two tracklets from the complementary views and adjacent clips, respectively. For that, we define the cross-view and over-time tracklet similarity in Sections 3.3 and 3.4, respectively. Here the cross-view subject similarity measurement is a new challenging problem in this work, for which we propose a spatial-aware cross-view subject association mechanism in Section 3.2.

2. We use dummy nodes to handle the cases of misdetection, occlusion and out of view. For example, a match between a real node in cluster 1 and dummy node in cluster 2 indicates the tracked subject underlying the real node in cluster 1 is not detected/tracked in cluster 2.

3.2 Spatial-Aware Cross-View Subject Association

In this section, we first elaborate on the spatial distribution based approach for cross-view subject association at frame level, i.e., the association between two temporally-synchronized frames from the top- and horizontal-view videos, respectively [40]. Then, we discuss its extension to two sequences and get the cross-view association results frame by frame, which is a key component in cross-view tracklet similarity measurement in Section 3.3.

3.2.1 Definition

Given a top-view image and a horizontal-view image that are taken by respective cameras at the same time, we detect all subjects (humans), each in form of a bounding box, on both images by a human detector [41]. Let $\mathcal{T} = \{T_m\}_{m=1}^{M}$ be the collection of M subjects detected on the top-view image, with T_m being the mth detected subject. Similarly, let $\mathcal{H} = \{H_q\}_{q=1}^{Q}$ be the collection of Q subjects detected on the horizontal-view image, with H_q being the qth detected subject. The goal of cross-view subject association is to identify all the matched subjects between $\mathcal{T}$ and $\mathcal{H}$ that indicate the same persons.

We address this problem by exploring the spatial distributions of the detected subjects in both views. More specifically, from each detected subject T_m in the top view, we infer a vector $\mathbf{V}_m^{\text{top}} = (x_m^{\text{top}}, y_m^{\text{top}})$ that reflects its relative position to the horizontal-view camera (wearer) on the ground. Then for each detected subject H_q in the horizontal view, we also infer a vector $\mathbf{V}_q^{\text{hor}} = (x_q^{\text{hor}}, y_q^{\text{hor}})$ to reflect its relative position to the horizontal-view camera on the ground. We associate the subjects detected in the two views by seeking matchings between two vector sets $\mathbf{V}^{\text{top}}(\mathcal{T}, \theta, O) = \{\mathbf{V}_m^{\text{top}}\}_{m=1}^{M}$ and $\mathbf{V}^{\text{hor}}(\mathcal{H}) = \{\mathbf{V}_q^{\text{hor}}\}_{q=1}^{Q}$, where O and θ are the location and view angle of the horizontal-view camera (wearer) in the top-view image and they are unknown priorly. Finally, we define a matching cost function Φ to measure the dissimilarity between the two vector sets and optimize this function for finding the matching subjects between the two views, as well as the camera location O, and camera view angle θ. In the following, we elaborate on each step of the proposed method.

3.2.2 Spatial Distribution Representation

First, we discuss how to derive $\mathbf{V}^{\text{top}}$ and $\mathbf{V}^{\text{hor}}$. Note that, the matching of the subjects, and the view orientation θ and horizontal-view camera location O in the top view are coupled and difficult to be solved together. In this case, we first assume that the camera location O and its view orientation θ are given. This way, we can compute its field of view (FOV) in the top-view image and all the detected subjects' relative positions to the horizontal-view camera on the ground. Horizontal-view image is egocentric and we can compute the detected subjects' relative positions to the camera based on the subjects' positions and sizes on the horizontal-view image.

Top-View Representation. As shown in Fig. 3a, in the top-view image we can easily compute the left and right boundaries of the field of view of the horizontal-view camera, denoted by $\vec{L}$, $\vec{R}$, respectively, based on the given camera location O and its view orientation θ. For a subject located at point P in the field of top view, we estimate its relative position to the horizontal-view camera by using two geometry parameters $\hat{x}$ and $\hat{y}$, where $\hat{x}$ is the (signed) distance to the horizontal-view camera along the (camera) right direction $\vec{V}$, as shown in Fig. 3a and $\hat{y}$ is the depth. From the pinhole camera model, we can calculate them by

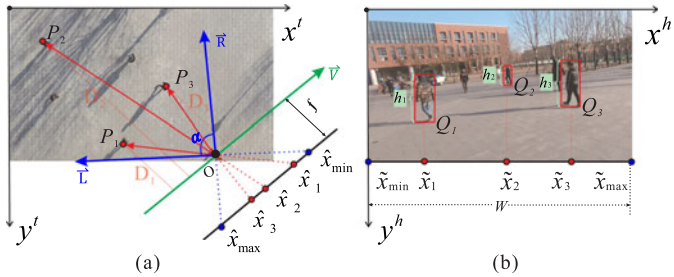


Fig. 3. An illustration of the vector representation of subjects in (a) top view and (b) horizontal view.

$$\begin{cases} \hat{x} = f \cdot \cot\langle \vec{OP}, \vec{V} \rangle \\ \hat{y} = |\vec{OP}| \cdot \sin\langle \vec{OP}, \vec{V} \rangle, \end{cases} \tag{5}$$

where $\langle \cdot, \cdot \rangle$ indicates the angle between two directions and f is an interior parameters of horizontal-view camera.

We then consider the range of $\hat{x}$. From Fig. 3a, we have

$$\begin{cases} \hat{x}_{\min} = f \cdot \cot\langle \vec{L}, \vec{V} \rangle = f \cdot \cot(\frac{\pi+\alpha}{2}) \\ \hat{x}_{\max} = f \cdot \cot\langle \vec{R}, \vec{V} \rangle = f \cdot \cot(\frac{\pi-\alpha}{2}), \end{cases} \tag{6}$$

where $\alpha \in [0, \pi]$ is the given field-of-view angle of the horizontal-view camera as indicated in Fig. 3a. From Eq. (6), we have $\hat{x}_{\max} = -\hat{x}_{\min} > 0$.

To enable the matching to the vector representation from the horizontal view, we further normalize the value range of $\hat{x}$ to $[-1, 1]$, i.e.,

$$\begin{cases} x^{\text{top}} = \frac{\hat{x}}{f \cdot \cot(\frac{\pi-\alpha}{2})} \\ y^{\text{top}} = \hat{y}. \end{cases} \tag{7}$$

With this normalization, we actually do not need the actual value of f in the proposed method.

Let O_k^{top}, $k \in \mathcal{K} \subset \{1, 2, \ldots, M\}$ be the subset of detected subjects in the field of view in the top-view image. We can find the vector representation for all of them and sort them in terms of x^{top} values in an ascending order. We then stack them together as

$$\mathbf{V}^{\text{top}} = (\mathbf{x}^{\text{top}}, \mathbf{y}^{\text{top}}) \in \mathbb{R}^{|\mathcal{K}| \times 2}, \tag{8}$$

where $|\mathcal{K}|$ is the size of $\mathcal{K}$, and $\mathbf{x}^{\text{top}}$ and $\mathbf{y}^{\text{top}}$ are the column-wise vectors of all the x^{top} and y^{top} values of the subjects in the field of view, respectively.

Horizontal-View Representation. For each subject in the horizontal-view image, we also compute a vector representation to make it consistent to the top-view vector representation, i.e., x-value reflects the distance to the horizontal-view camera along the right direction and y-value reflects the depth to the horizontal-view camera. As shown in Fig. 3b, in the horizontal-view image, let $(\tilde{x}, \tilde{y})$ and h be the location and height of a detected subject, respectively. If we take the top-left corner of the image as the origin of the coordinate, $\tilde{x} - \frac{W}{2}$, with

Fig. 4. An illustration of mutual occlusion in the horizontal view (b) and its corresponding top view (a). (c) Damping function used for reducing searching space.

W being the width of the horizontal-view image, is actually the subject's distance to the horizontal-view camera along the right direction. To facilitate the matching to the top-view vectors, we normalize its value range to $[-1, 1]$ by

$$\begin{cases} x^{\text{hor}} = \frac{\tilde{x} - \frac{W}{2}}{\frac{W}{2}} \\ y^{\text{hor}} = \frac{1}{h}, \end{cases} \tag{9}$$

where we simply take the inverse of the subject height as an estimate of its depth to the horizontal-view camera.

For all Q detected subjects in the horizontal-view image, we can find their vector representations and sort them in terms of x^{hor} values in an ascending order. We then stack them together as

$$\mathbf{V}^{\text{hor}} = (\mathbf{x}^{\text{hor}}, \mathbf{y}^{\text{hor}}) \in \mathbb{R}^{Q \times 2}, \tag{10}$$

where $\mathbf{x}^{\text{hor}}$ and $\mathbf{y}^{\text{hor}}$ are the column-wise vectors of all the x^{hor} and y^{hor} values of the Q subjects detected in the horizontal-view image, respectively.

Occlusion Scenario. An occlusion in the horizontal-view image indicates that two subjects and the horizontal-view camera are collinear, as shown by P_1 and P_2 in Fig. 4a. In this case, the subject with larger depth, i.e., P_2, is not visible in the horizontal view and we simply ignore this occluded subject in vector representation of $\mathbf{V}^{\text{top}}$. In practice, we set a tolerance threshold τ and if $\langle \overline{OP_1}, \overline{OP_2} \rangle < \tau$, we ignore the one with larger depth.

3.2.3 Spatial Distribution Matching

In this section we associate the subjects across two views by matching the vectors between the two vector sets $\mathbf{V}^{\text{top}}$ and $\mathbf{V}^{\text{hor}}$. Since the x values of both vector sets have been normalized to the range of $[-1, 1]$, they can be directly compared. However, the y values in these two vector sets are of different scale and not directly comparable: y^{top} values are computed in the top-view image while y^{hor} values are computed in the horizontal-view image. It is non-trivial to normalize them into a same scale – y^{hor} is a rough depth estimation and is very sensitive to subject detection errors and the height difference of the subjects.

We first find reliable subset matchings between $\mathbf{x}^{\text{top}}$ and $\mathbf{x}^{\text{hor}}$ and use them to estimate the scale difference between their corresponding y values. More specifically, we find a scaling factor μ to scale y^{top} values to make them comparable to the y^{hor} values. For this purpose, we use a RANSAC-alike strategy [42]: for each element x^{top} in $\mathbf{V}^{\text{top}}$, we find the nearest x^{hor} in $\mathbf{V}^{\text{hor}}$. If $|x^{\text{top}} - x^{\text{hor}}|$ is less than a very small threshold value, we consider x^{top} and x^{hor} a matched pair and take the ratio of their corresponding y values and the average of this ratio over all the matched pairs is taken as the scaling factor μ.

With the scaling factor μ, we match $\mathbf{V}^{\text{top}}$ and $\mathbf{V}^{\text{hor}}$ via dynamic programming (DP). Specifically, we define a dissimilarity matrix $\mathbf{D}$ of dimension $|\mathcal{K}| \times Q$, where D_{ij} is the dissimilarity between $\mathbf{V}_i^{\text{top}}$ and $\mathbf{V}_j^{\text{hor}}$, defined as

$$D_{ij} = \lambda |x_i^{\text{top}} - x_j^{\text{hor}}| + |\mu y_i^{\text{top}} - y_j^{\text{hor}}|, \tag{11}$$

where $\lambda > 0$ is a balance factor. Given that $\mathbf{x}^{\text{top}}$ and $\mathbf{x}^{\text{hor}}$ are both ascending sequences, we use dynamic programming algorithm to search for a monotonic path in $\mathbf{D}$ from $D_{1,1}$ to $D_{|\mathcal{K}|,Q}$ to build the matching between $\mathbf{V}^{\text{top}}$ and $\mathbf{V}^{\text{hor}}$ with minimum total dissimilarities. If a vector $\mathbf{V}^{\text{top}}$ matches to multiple vectors in $\mathbf{V}^{\text{hor}}$, we only keep the one with the smallest dissimilarity given in Eq. (11). After that, we check if a vector $\mathbf{V}^{\text{hor}}$ matches to multiple vectors in $\mathbf{V}^{\text{top}}$ and we keep the one with the smallest dissimilarity. These two-step operations will guarantee the resulting matching is one-on-one and we denote κ to be the number of final matched pairs. Denote the resulting matched vector subsets to be $\mathbf{V}_*^{\text{top}} = (\mathbf{x}_*^{\text{top}}, \mathbf{y}_*^{\text{top}})$ and $\mathbf{V}_*^{\text{hor}} = (\mathbf{x}_*^{\text{hor}}, \mathbf{y}_*^{\text{hor}})$, both of dimension $\kappa \times 2$. We define a matching cost between $\mathbf{V}^{\text{top}}$ and $\mathbf{V}^{\text{hor}}$ as

$$\Phi(\mathbf{V}^{\text{top}}, \mathbf{V}^{\text{hor}}) = \frac{1}{\kappa} \rho^{\frac{L}{\kappa}} (\lambda \|\mathbf{x}_*^{\text{top}} - \mathbf{x}_*^{\text{hor}}\|_1 + \|\mu \mathbf{y}_*^{\text{top}} - \mathbf{y}_*^{\text{hor}}\|_1), \tag{12}$$

where $\rho > 1$ is a pre-specified factor and $L = \max(|\mathcal{K}|, Q)$. In this matching cost, the term $\rho^{\frac{L}{\kappa}}$ encourages the inclusion of more vector pairs into the final matching, which is important when we use this matching cost to search for optimal camera location O and view orientation θ, which will be discussed in the following.

3.2.4 Spatial-Aware Association Framework

In practice, both the top-view and horizontal-view cameras generate videos and we need to perform cross-view subject association between them over time. By assuming consistent frame rate and temporal synchronization between them, we can simply apply the above image-based subject association algorithm between two videos frame by frame.

Complexity of Association Algorithm. In calculating the matching cost of Eq. (12), we need to know the horizontal-view camera location O and its view orientation θ to compute the vector $\mathbf{V}^{\text{top}}$. However, we do not know O and θ priorly. In the first frame, we try all possible values for O and θ and then select the ones that lead to the minimum matching cost Φ. The matching with such minimum cost provides us the cross-view subject association. Assume the number of subjects detected in the top and horizontal views are M and Q, respectively. The complexity of the dynamic programming (DP) based vector matching is $\mathrm{O}(MQ)$. Based on the matching cost, we try all possible values for O and θ and then select the ones that lead to the minimum matching cost. For view orientation θ, we sample the range $[0, 2\pi)$ uniformly with an interval of $\Delta\theta$. For the horizontal-view camera location O, we try every subject in the top view. This way, the whole complexity of our algorithm is $\mathrm{O}(M^2 Q) \cdot \frac{2\pi}{\Delta\theta}$.

Searching Space Reduction Strategy. The above association method searches O and θ in an exhaustive way. To reduce the

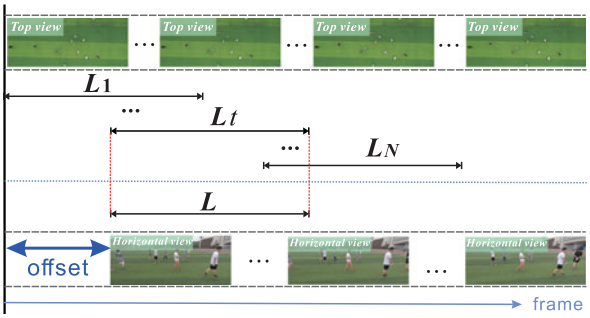


Fig. 5. An illustration of the proposed temporal synchronization strategy.

searching space of θ, we employ a searching decrement strategy. Specifically, with computed θ_{t-1} on frame $t-1$, we limit the camera view orientation angle on frame t to be in the range of $[\theta_{t-1} - \frac{1}{2}S_t, \theta_{t-1} + \frac{1}{2}S_t]$ instead of exhaustive search over $[0, 2\pi)$, where S_t is the size of the search space. We use a damping function for the search space, i.e., $S_t = 2\pi \frac{r^{t-1}+u}{1+u}$, where we set $r = 0.95$ and $u = 0.1$ and the function is illustrated in Fig. 4c. We can see that the search range S_t is large at the beginning and then keeps decreasing over time. This setting can improve the algorithm robustness against possible errors in early frames as well as reduce the searching space in the subsequent frames. We also reduce the searching time for the horizontal-view camera location O. Actually, for the online tracking task, we only need to exhaustively search O in the top view at the initialization stage. Then, we can leverage the top-view subject online tracking result of O to quickly locate it in the subsequent frames.

Non-Synchronized Problem. Due to latency and instability in signal transmission, the assumption of perfect synchronization between two-view videos may not stand in real world. We can use the proposed spatial-aware association to handle this issue. As shown in Fig. 5, given the input top-view and horizontal-view videos with possible temporal offset, we take a clip L from one video, e.g., the horizontal-view video in Fig. 5. Then, on the other video, we temporally slide the window defined by L equally to both ends to generate a set of N clips $L_1, L_2, \ldots, L_N$, each of which has the same length as L. After that, we compute the cost between clips L_i and L, $i = 1, 2, \ldots, N$, by calculating their matching cost frame by frame as defined in Eq. (12) and then averaging over all the frames in each clip. As shown in Fig. 5, we finally select the clip L_t with the lowest average matching cost and take its offset as the one between the two videos. We will show the experimental analysis on the video pairs with temporal offset in Section 4.2.2.

3.3 Cross-View Tracklet Association for CVMHAT

In this section, we define the edge weight in the problem formulation namely Eq. (1) in Section 3.1, i.e., the data similarity between tracklets across top and horizontal views. For that, in Section 3.3.1, we first use the spatial reasoning based on the spatial-aware subject association approach proposed in Section 3.2. Besides, although the appearance-based features are not very useful for cross-view subject association as discussed above, they are still supplementary to the spatial reasoning and can improve the cross-view association to some extent. So, we also consider the appearance reasoning in Section 3.3.2.

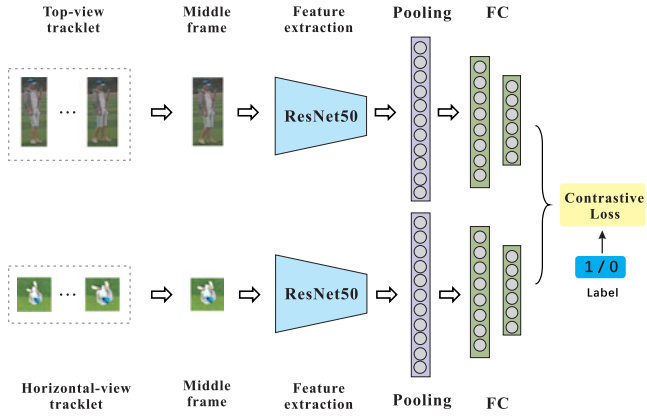


Fig. 6. Siamese neural network structure for measuring the cross-view appearance similarity.

3.3.1 Spatial Reasoning

Given a pair of synchronized clips from the top view and horizontal view, respectively. We first calculate the subject association results frame by frame using the approach in Section 3.2. As discussed above, let $\mathcal{T} = \{T_m\}_{m=1}^{M}$ be the collection of M subjects detected on a frame in the top view, and $\mathcal{H} = \{H_q\}_{q=1}^{Q}$ be the collection of Q subjects detected on the corresponding frame in the horizontal view. The proposed cross-view subject association tries to identify all the matched subjects between $\mathcal{T}$ and $\mathcal{H}$. The association results obtained in Section 3.2 can be expressed by a function $\mathrm{A}(T_m, H_q) = 1$ if T_m, H_q represent the same person, while $\mathrm{A}(T_m, H_q) = 0$ otherwise. Given two single-view tracklets $\mathbf{T}_m, \mathbf{H}_q$ (or $\mathbf{T}_n, \mathbf{H}_p$) extracted from the clip t (or clip $t+1$) in two views, respectively, we identify their temporally overlapped L frames, on each of which we have subjects T_m^l and H_q^l, $l = 1, \ldots, L$. The spatial-aware similarity between $\mathbf{T}_m$ and $\mathbf{H}_q$ can be calculated by

$$\hat{c}_{qm}^{i} = \frac{\sum_{l=1}^{L} \mathrm{A}(T_m^l, H_q^l)}{\max(|\mathbf{T}_m|, |\mathbf{H}_q|)}, \quad i = 3, \tag{13}$$

where $|\mathbf{T}_m|$ and $|\mathbf{H}_q|$ denotes the length of the tracklet $\mathbf{T}_m$ and $\mathbf{H}_q$, respectively. Here $i = 3$ indicates the set of edges connecting the nodes from Cluster 3 to Cluster 0 in Fig. 2 that reflects cross-view association. Accordingly, we can compute the similarity $\hat{c}_{np}^{i}$ $(i = 1)$ for the tracklets $\mathbf{T}_n$ and $\mathbf{H}_p$ at clip $t+1$.

3.3.2 Appearance Reasoning

For the appearance reasoning, we used a two-branch Siamese network [43], [44], as shown in Fig. 6, to measure the appearance similarity between two tracklets, where each branch follows the network structure that has been successfully used for feature extraction in the challenging person re-identification task [45]. Specifically, given a pair of tracklets from complementary views, we only take the involved bounding boxes in their middle frames as the input of the two branches, respectively, since the appearance variation of a subject along a 10-frame short clip is usually very small. Each bounding box is resized to 256×128 and ResNet-50 [46] is adopted as backbone, following by a pooling layer and then two fully connected (FC) layers with the dimension of 1,024 and 128 respectively. The first FC layer is

followed by the batch normalization, ReLU and dropout layers. We use the pair-based contrastive loss function [47] for training this network. By calculating the euclidean distance of the two branches' outputs, we can obtain the appearance similarity $\overline{c}^i_{qm}$ for $i = 3$ ($\overline{c}^i_{np}$ for $i = 1$) for the cross-view tracklets $\mathbf{T}_m$ and $\mathbf{H}_q$ ($\mathbf{T}_n$ and $\mathbf{H}_p$).

3.4 Over-Time Tracklet Association for CVMHAT

To measure the similarity of single-view tracklets across different clips, we use the classical appearance and motion features as the representation.

3.4.1 Appearance Consistency

For the appearance consistency, we first compute the histogram for all the subjects of a single-view tracklet. Then the median of all the histogram is selected as the appearance descriptor of the tracklet [18]. Let $\varphi(\mathbf{T}_m)$ and $\varphi(\mathbf{T}_n)$ be the color histogram based appearance descriptors of the tracklets $\mathbf{T}_m$ and $\mathbf{T}_n$ ($\mathbf{H}_q$ and $\mathbf{H}_p$ for the horizontal view), respectively. We use Histogram Intersection [48] as their appearance similarity

$$\hat{c}^i_{mn} = \mathrm{K}(\varphi(\mathbf{T}_m), \varphi(\mathbf{T}_n)), \quad i = 0, \tag{14}$$

where $\mathrm{K}(\cdot)$ denotes a kernel function [48] taking value in [0,1]. Similarly, we can get the appearance similarity for the tracklets $\mathbf{H}_q$ and $\mathbf{H}_p$ namely $\hat{c}^i_{pq}$ for $i = 2$. Here, $i = 0, 2$ indicate the two sets of edges in Fig. 2 that reflect the over-time association.

3.4.2 Motion Consistency

We also consider the motion consistency for single-view over-time association. We use the constant velocity motion model to predict motion consistency as in most previous multiple object tracking methods. Given two tracklets $\mathbf{T}_m$ and $\mathbf{T}_n$ from two adjacent clips in the top view (or $\mathbf{H}_q$ and $\mathbf{H}_p$ for the horizontal view), we calculate the deviation error between the two tracklets as

$$\delta = \sigma(\delta_f + \delta_b), \tag{15}$$

where σ is a scaling factor, and δ_f and δ_b denote the forward and backward deviation errors respectively, that can be computed by the constant velocity motion model as in [14]. To measure the difference between $\mathbf{T}_m$ and $\mathbf{T}_n$, we convert the error into a similarity value by

$$\overline{c}^i_{mn} = e^{-\delta}, \quad i = 0, \tag{16}$$

where $\overline{c}^i_{mn}$ takes the value in [0,1]. Similarly, we compute the similarity $\overline{c}^i_{pq}$ ($i = 2$) for the tracklets $\mathbf{H}_q$ and $\mathbf{H}_p$.

3.5 Overall Tracking Framework

Given two cross-view tracklets $\mathbf{T}_m, \mathbf{H}_q$, we can use the above methods to get the spatial similarity score $\hat{c}^i_{qm}$ and appearance similarity score $\overline{c}^i_{qm}$ for ($i = 3$), respectively, as discussed in Section 3.3. We use linear combination to get the final edge weight c^i_{qm} as

$$c^i_{qm} = w_1 \hat{c}^i_{qm} + (1 - w_1)\overline{c}^i_{qm}, \quad i = 3, \tag{17}$$

where w_1 is a pre-set parameter. Similarly, we compute the edge weight c^i_{np} for $i = 1$.

Besides, given two single-view tracklets $\mathbf{T}_m, \mathbf{T}_n$ from different clips, we calculate the edge weight c^i_{mn} ($i = 0$) by

$$c^i_{mn} = w_2 \hat{c}^i_{mn} + (1 - w_2)\overline{c}^i_{mn}, \quad i = 0, \tag{18}$$

where w_2 is a pre-set parameter, and $\hat{c}^i_{mn}$, $\overline{c}^i_{mn}$ are the tracklet similarity scores calculated by the appearance color histogram features in Eq. (14) and the motion features in Eq. (16), respectively. Similarly, we compute the edge weight c^i_{pq} for $i = 2$.

The proposed multiple subject tracking method is summarized in Algorithm 1. After getting the short trajectories, we stitch two over-time adjacent trajectories by considering the consistency of overlapping frames between them. By stitching one after another, we obtain the final tracking results of the whole video.

Algorithm 1. Complementary-View Multiple Human Association and Tracking (CVMHAT)

Input: V_T, V_H: Top-view and horizontal-view videos; Pre-set hyper-parameters.
Output: Tracked subject bounding boxes with ID numbers.

```
Split the cross-view videos into T clips respectively.
for t = 1 : T do
    Detect the subjects then extract the single-view tracklets
      T_m^t, H_q^t, T_n^{t+1}, H_p^{t+1} in clip t and t + 1.
    Calculate the tracklets similarity scores then compute the
      edge weight c by Eqs. (17) and (18).
    Solve for e by Eq. (1) to get cross-view short trajectories.
    if t = 1 then
        if e_qm = 1 then
            Assign the same ID numbers to the bounding boxes
              in the tracklets T_m^t, H_q^t.
        else
            Assign the incremental ID (the maximum of current
              ID numbers plus one) to the other ones.
    else
        if e_np = 1 then
            Assign the ID number of T_n^{t+1} to H_p^{t+1} .
        else if e_mn = 1 (e_pq = 1) then
            Assign the ID number of T_m^t (H_q^t) to T_n^{t+1} (H_p^{t+1}),
              respectively.
        else
            Assign the incremental ID to the other ones.
return Bounding boxes with ID numbers
```

4 Experimental Results

In this section, we first present the dataset collection and experimental setup. Then, we independently evaluate the proposed cross-view subject association algorithm described in Section 3.2. After that, we report the extensive experimental results of tracking on the complementary top- and horizontal-view videos.

4.1 Setup

Dataset Collection. We do not find publicly available dataset with temporally synchronized top- and horizontal-view

videos with ground-truth labeling for evaluating the performance of the proposed cross-view multiple human association and tracking. Therefore, we collect a new dataset by flying a drone with a camera to take top-view videos and mounting GoPro over the head of persons on the ground to take the horizontal-view videos. More specifically, we arrange two to three subjects, but not all the subjects on the ground, to wear cameras to record the horizontal-view videos. Videos are taken at five sites with different backgrounds. The subjects are free to move or stop in the scene without any specific instructions and there may be random mutual occlusions between subjects in the horizontal view. We manually synchronize the top- and horizontal-view videos such that corresponding frames between the two views are taken at the same time. We then cut out 15 pairs of sequences with length from 600 to 1,200 frames, and in total 23,400 frames as our dataset. The number of subjects varies from 3 to 10 in horizontal-view videos and 6 to 14 in top-view videos. For both views, the image resolution is $2,688 \times 1,512$. We manually annotate the subjects in the forms of rectangular bounding boxes and ID numbers: the same subject in each view and across the two views are labeled with the same ID number. Note that, this manual labeling is very labor intensive – we need scrutiny to identify subjects in the top-view videos (see Fig. 1 for an example).

Evaluation Metrics. We first define the metrics for evaluating the cross-view subject association performance. We use the cross-view subject association precision score, i.e., CVIDP and recall score, i.e., CVIDR for performance evaluation. Then, the cross-view ID F_1 measure (CVIDF_1) can be defined as

$$\mathrm{CVIDF}_1 = \frac{2\mathrm{CVIDP} \times \mathrm{CVIDR}}{\mathrm{CVIDP} + \mathrm{CVIDR}}. \tag{19}$$

We further define the cross-view matching accuracy (CVMA) as

$$\mathrm{CVMA} = 1 - \left(\frac{\sum_t \mathrm{m}_t + \mathrm{fp}_t + 2\mathrm{mme}_t}{\sum_t \mathrm{g}_t}\right), \tag{20}$$

where m_t, fp_t, mme_t are the numbers of misses, false positives, and mismatch pairs of cross-view subject matchings at time t, which follows the definitions in multi-object tracking accuracy (MOTA) [49], and g_t is the total number of subjects in both the top and horizontal views at time t. The proposed metrics are different from the IDF_1 and MOTA metrics used in the DukeMTMC benchmark [50] for multi-view MOT performance evaluation. The metrics in DukeMTMC reflect the predicted ID consistency of the same person along the time across multiple cameras. Differently, the proposed CVIDF_1 and CVMA metrics reflect the level of cross-view human ID matching for all the detected subjects at a time. We also apply standard MOT metrics for evaluating the tracking performance [49], including multi-object tracking precision (MOTP) and multi-object tracking accuracy (MOTA). One key task of the proposed collaborative MOT is to identify/track the same subject across two views. Therefore, we also use four ID-based performance metrics, i.e., ID precision (IDP), ID recall (IDR), and ID F_1 measure (IDF_1).

Implementation Details. We implement the main program in Matlab on a desktop computer with an Intel Core i9 3.6 GHz CPU, and the Siamese network for cross-view appearance similarity measurement is implemented on GPU. The mixed integer programming problem is solved by the MIP solver – *cplex*.

• *Model Training Details.* We use the general YOLOv3 [41] detector to detect subjects in the form of bounding boxes in both top- and horizontal-view videos. For top-view subject detection, we fine-tune the network using 600 top-view human images. For training the Siamese based network, given a subject detected in the top-view frame, we regard it paired with its corresponding subject in horizontal view as a positive sample, and paired with other subjects as negative training samples. In the experiments, we use the pre-trained person re-identification model [45] in our Siamese network. Note that all the training data have no overlap with our test dataset.

• *Tracklet Extraction.* We split the videos into short clips with a fixed length of 10 frames. In each clip, we extract the subject tracklets using the criteria of bounding box overlap and color similarity. Specifically, we first calculate the subject (in the form of bounding boxes) intersection over union (IOU) between two adjacent frames. Detected subjects with good IOU between two adjacent frames, are connected. Besides, to alleviate possible mismatches in crowd scenarios with occlusions, we further integrate the color histogram based appearance features to measure the similarity between connected bounding boxes in the horizontal view. Specifically, we connect a pair of bounding boxes if their IOU and color-histogram similarity are larger their respective preset thresholds (0.5 and 0.3 in experiments). In the top view, we only use the bounding box IOU as the similarity since there is no mutual occlusion and color features of different subjects are not very discriminative. Finally, we repeatedly use the dynamic programming (DP) to achieve a set of optimal sequences of connected bounding boxes across the clip, i.e., those with highest total similarities, as tracklets.

• *Parameter Selection.*We set the tolerance threshold $\tau = 2^\circ$ in Section 3.2.2. In the following, without specific claim, we use the setting of $\Delta\theta = 1^\circ$ with search range of S_t as described in Section 3.2.4. The weighting parameters c_0 in Eq. (1), and w_1, w_2 in Eqs. (17) and (18) are set to 0.2, 0.3 and 0.6, respectively. The pre-specified parameters ρ and λ in Eq. (12) are set to 15 and 0.003 respectively, and the parameters σ in Eq. (15) is set to 0.05.

Baseline Methods. For the cross-view association problem, we did not find available methods for subject association between the top and horizontal views. As discussed above, the appearance and motion features are not very useful for the proposed top- and horizontal-view matching. However, as we all know, appearance and motion are two backbone modalities for video object matching. Hence, we use the state-of-the-art multiple object tracker of DMAN [6] that uses the appearance-motion-based features to build the over-time subject association. We help DMAN by manually associating the subjects between the top and horizontal views on the frames when each subject first appears in the video. We then track all the subjects in each video by DMAN, respectively, and finally use the tracking results to propagate the subject association to later frames.

For the tracking problem, we first choose three multiple object tracking (MOT) methods, i.e., GMMCP [18], MDP [5],

and DMAN [6] as the comparison methods. To validate the effectiveness of the MIP based optimization method, we replace it with the widely used spectral clustering based algorithm [19], and compare the tracking performance. Besides, we also include the prior version of this work namely CVMHAT [1] for comparison.

• *GMMCP* uses the color histogram based appearance feature and the constant-velocity-based motion features for subject association, which is the same as our over-time association and can be regarded as our baseline method.

• Both *MDP* and *DMAN* are online MOT methods, where DMAN learns deep appearance features for association. All the comparison trackers are implemented to track on the top-view and horizontal-view videos separately, initialized with the ground-truth subjects and ID labels in the first frame. Note that, this initialization is an extra favor to these comparison methods since the proposed method takes the association result generated by Eq. (1) in the first frame as the initialization. Other than that, we use the same subject detector for all the methods, including the proposed method and the comparison methods.

• *Clustering*. As shown in Fig. 2, given the nodes from different clips and views, we take the similarity scores from Eqs. (17) and (18) and use the obtained similarity matrix as the input of spectral clustering. We can achieve the desired cross-view cross-clip subject association by clustering the detections of the same subject from different views and clips into one group. In spectral clustering, we examine the number of subjects in each involved clip and then select the largest one as the number of groups.

In practice, similar to cross-view association, we did not find other directly related methods that can handle the proposed complementary-view multiple human tracking. All prior seemingly related works for cross-view multiple human tracking, e.g., [25], only handle multiple horizontal views or top views with sloped view angles, and use the human pose/appearance features for subject association. These features are barely useful given our top-view video is taken from a vertical-to-ground view direction.

4.2 Cross-View Human Association Results

4.2.1 Results

In this section, to evaluate the proposed cross-view subject association method in Section 3.2, we apply it on all 15 pairs of videos in our dataset and compute the average performance on each frame. On each pair of corresponding frames between two views, we can compute the precision and recall in association by comparing to the ground truth. From Table 1, we can see that the average precision and recall scores of our method over all frame pairs are 65.4 and 64.8 percent respectively. DMAN produces a satisfactory precision score but a poor recall score since MOT performance is usually not satisfactory in horizontal view due to mutual occlusions and ID switches. Besides, we evaluate our method by tracking the horizontal-view camera O after manual assignment in the first frame. Specifically, we manually assign the horizontal-view camera location in the top view at the first frame and use the tracking results to locate it in the subsequent frames. We can see that the performance increases significantly as shown in the last row of Table 1. However, we couldn't manually assign the camera location O in practice. In the following of this subsection, we evaluate the subject association as an independent algorithm by automatically searching O frame by frame without manual assignment or tracking.

TABLE 1
A Comparison of Cross-View Association Results From Different Methods

Method	CVIDP	CVIDR	CVIDF_1	CVMA
DMAN	63.2	34.5	44.6	43.2
Ours	65.4	64.8	65.1	62.8
Ours w O	72.0	71.1	71.5	67.8

We use the average precision, recall, F_1 score and matching accuracy, i.e., CVIDP (%), CVIDR (%), CVIDF_1 (%), and CVMA (%), over all the frame pairs, respectively.

4.2.2 Detailed Analysis

Vector Representations. We compare the association results using different vector representation methods as shown in Table 2. The first row denotes that we represent the subjects in the two views by using one-dimensional vectors $\mathbf{x}^{\text{top}}$ and $\mathbf{x}^{\text{hor}}$ respectively. The second row denotes that we represent the subjects in the two views by one-dimensional vectors $\mathbf{y}^{\text{top}}$ and $\mathbf{y}^{\text{hor}}$, respectively, which are simply normalized to the range [0,1] to make them comparable. The third row denotes that we combine the one-dimensional vectors for the first and second rows to represent each view, which differs from the proposed method (the 4th row of Table 2). By comparing the results in the third and fourth rows, we can see that the use of RANSAC strategy for estimating the scaling factor μ does improve the final association performance. The results in the first and second rows show that using only one dimension of the proposed vector representation cannot achieve performance as good as the proposed method that combines both. We can also see that $\mathbf{x}^{\text{top}}$ and $\mathbf{x}^{\text{hor}}$ provide more accurate information than $\mathbf{y}^{\text{top}}$ and $\mathbf{y}^{\text{hor}}$ when used for cross-view subject association.

Qualitative Analysis. It is a common situation that many subjects in two views are not the same persons. In this case, the shared subjects may only count for a small proportion in both views. Two examples are shown in the top row of Fig. 7, where associated subjects bear same number labels. In the left, we show a case where many subjects in the top view are not in the field of view of the horizontal-view camera. In the right, we show a case where many subjects in the horizontal view are too far from the horizontal-view camera and not covered by the top-view camera. We can see that the proposed method can handle these two cases very well,

TABLE 2
A Comparison of Cross-View Association Results Using Different Representations (%)

Vector	CVIDP	CVIDR	CVIDF_1	CVMA
$\mathbf{x} : \mathbf{x}^{\text{top,hor}}$	40.5	39.7	40.1	42.2
$\bar{\mathbf{y}} : \bar{\mathbf{y}}^{\text{top,hor}}$	22.8	15.5	18.5	25.0
$\mathbf{x} + \bar{\mathbf{y}}$	53.9	54.0	53.9	53.2
Ours	**65.4**	**64.8**	**65.1**	**62.8**

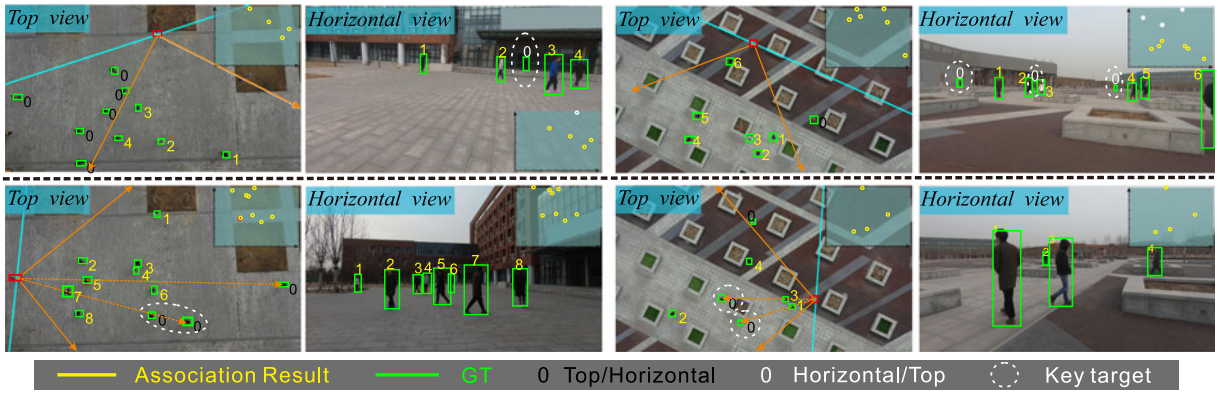


Fig. 7. Illustration of association results. Row 1:Two sample results with large number of unshared subjects between two views. Row 2: Two sample results on image pairs with occlusions. Vector sets $\mathbf{V}^{top}$ and $\mathbf{V}^{hor}$ are shown in the top-right corner of every image. The subjects with black '0' label are visible in the top view but invisible in the horizontal view, and with white '0' are visible in the horizontal view but invisible in the top view.

by exploring the spatial distribution of the shared subjects. Sample results on image pairs with occlusions (in the horizontal view) are shown in the second row of Fig. 7, where associated subjects bear same number labels. We can see that occlusions occur more often when 1) the subjects are crowded, and 2) one subject is very close to the horizontal-view camera. We can also see that the proposed method can handle the occlusion scenarios effectively.

Complexity and Speed Analysis. As shown in Table 3, we compare the exhaustive and the improved searching strategies described in Section 3.2.4 in terms of the size of searching space, running speed and corresponding association performance. We can see that when the searching space Θ is reduced from $[0, 2\pi)$ to S_t, the running speed improves significantly at the cost of a little association performance drop. Further, with the temporal tracking of O in the top view, the searching space is actually reduced to $MQ \cdot \frac{S_t}{\Delta\theta}$. This also leads to better performance because the temporal tracking of O in the top view is robust, which facilitates the temporal consistency of O. Note that, in our experiments, the orders of magnitude of $\frac{S_t}{\Delta\theta}$, M and Q are between 10^0 to 10^1, which make the size of searching space always acceptable. From Table 3, we can see that the running speed of our method, with online tracking for Θ and O, can reach 44.8 fps on CPU.

Robustness Analysis. We evaluate the association performance under different temporal offsets between the two videos. As shown in Table 4, we manually misplace the two videos with an offset of $5, 10, 15$ and 20 frames respectively and the results show that the proposed cross-view association method can tolerate the misplacement up to 5 frames, with performance degradation less than 4 percent. However, the association performance decreases significantly when the offset further increases, e.g., to 10, 15 and 20 frames. We apply the proposed synchronization strategy in Section 3.2.4 to our dataset to evaluate its effectiveness. In our experiments, the clip length is set to 10 frames, the step length of sliding window is set to 1 frame, and the number of sliding clips N is set to 30. The results are shown in Table 5, where 'Initial' indicates the offset we manually set for the input video pairs and 'Result' indicates the offset after the use of the proposed synchronization strategy, averaged over the whole dataset. We can see that, even if the initial offset is as large as 20 frames, the proposed strategy can still reduce it to less than 5 frames, which can be tolerated by the proposed method.

TABLE 3
Complexity and Speed Analysis of the Cross-View Association Algorithm Using Different Searching Strategies

Search.	Exhaustive	Reduce Θ	Reduce $\Theta + O$
Space	$M^2Q \cdot \frac{2\pi}{\Delta\theta}$	$M^2Q \cdot \frac{S_t}{\Delta\theta}$	$MQ \cdot \frac{S_t}{\Delta\theta}$
Time (sec)	0.320	0.038	0.022
Speed (fps)	3.1	26.3	44.8
$CVIDF_1$	67.2%	65.1%	71.5%
CVMA	64.1%	62.8%	67.8%

4.3 Complementary-View Human Tracking Results

4.3.1 Results

In this section, we evaluate the proposed tracking method on our dataset. We evaluate the MOT results on each video separately using the standard MOT metrics. We show the results of different trackers in Table 6 (left 5 columns). We can see that although using the same features as GMMCP for overtime data association, our method outperforms the baseline method GMMCP by a wide margin in the ID-related metrics. The proposed method achieves the comparable performance with the state-of-the-art DMAN tracker that combines many

TABLE 4
Cross-View Association Results Under Different Temporal Offsets (%)

Offset	CVIDP	CVIDR	$CVIDF_1$	CVMA
0 frames	65.4	64.8	65.1	62.8
5 frames	62.8	61.8	62.3	60.4
10 frames	55.9	54.6	55.2	54.4
15 frames	52.3	50.5	51.4	51.4
20 frames	46.8	45.2	46.0	46.9

TABLE 5
Temporal Synchronization Results Under Different Initial Offsets (Frames)

Initial	Result	Initial	Result	Initial	Result
10	3.40	15	3.87	20	4.47

TABLE 6
A Comparison of Cross-View Association and Tracking Results (%) of Different Methods

Method	IDP	IDR	IDF_1	MOTP	MOTA	$CVIDF_1$	CVMA
GMMCP [18]	49.4	50.7	50.1	**75.2**	79.3	17.9	27.2
MDP [5]	65.8	68.4	67.1	**75.2**	84.9	33.7	38.1
DMAN [6]	72.3	77.2	74.7	75.1	82.4	44.6	43.2
CVMHT [1]	77.6	76.6	77.1	74.9	84.2	84.0	78.3
Clustering	66.1	66.1	66.1	75.0	83.9	85.4	81.3
Ours	**80.3**	**80.5**	**80.4**	75.0	**85.3**	**86.9**	**83.9**

IDP, IDR, IDF_1, MOTP, and MOTA are standard MOT metrics. $CVIDF_1$ and CVMA are the proposed new metrics for evaluating the cross-view association.

TABLE 7
A Comparison of Tracking Results (%) From Different Methods on the Subsets of Top-View Videos and Horizontal-View Videos, Respectively

Method	Top view					Horizontal view				
	IDP	IDR	IDF_1	MOTP	MOTA	IDP	IDR	IDF_1	MOTP	MOTA
GMMCP [18]	50.7	50.7	50.7	69.3	76.5	47.7	50.8	49.2	83.6	83.4
MDP [5]	76.2	77.8	77.0	69.6	**86.3**	50.8	54.3	52.5	83.6	82.7
DMAN [6]	**85.1**	**87.4**	**86.2**	**70.0**	85.8	54.8	61.9	58.1	82.7	77.4
CVMHT [1]	79.5	80.9	80.2	69.2	84.0	74.3	70.1	72.2	**83.8**	84.5
Clustering	66.9	67.6	67.2	69.2	83.3	65.0	63.8	64.4	**83.8**	84.8
Ours	82.4	83.8	83.1	69.2	84.0	**75.7**	**77.2**	**76.4**	**83.8**	**87.2**

complex and advanced features. We believe our leading margin over DMAN can be further enlarged by adopting more advanced features, which, however, introduces more computation load and speed sacrifice. The bottom two rows in Table 6 show the comparison of our method using the MIP optimization and spectral-clustering algorithms, respectively. We can see that using MIP leads to better tracking performance than using spectral clustering, partly because we consider more constraints, e.g., the cyclic consistency constraint in Eq. (3), in the MIP optimization.

Moreover, we divide the dataset into top-view and horizontal-view videos and evaluate the MOT performance, respectively. As shown in Table 7, we can first find that the tracking results in top view shows better performance than those in horizontal view. This is due to the mutual occlusions which frequently appear in the horizontal view but are rare in top view. Compared with other methods, the proposed method achieves better tracking results in horizontal view with the assistance of tracking in top view. As for top-view tracking, the proposed method outperforms the baseline method GMMCP by a large margin with the cross-view joint optimization for subject association. The result verifies that the top view and horizontal view are complementary in improving the tracking accuracy.

Besides the standard MOT metrics, we also compare the cross-view MOT performance using $CVIDF_1$ and CVMA as shown in Table 6 (right 2 columns). While the selected comparison methods can only handle the single-view tracking, we provide the ground-truth ID of each subject on the first frame of both views. This way, the tracking on each view actually propagate the subject IDs to later frames and from the IDs, we can match the subjects across views over time. We can see that all the three MOT trackers, i.e., GMMCP, MDP, DMAN, produce very poor results because the cross-view subject matching will fail once a target is lost in tracking in any one view. Our method produces the acceptable $CVIDF_1$ and CVMA results of 86.9 and 83.9 percent, respectively. Note that, the cross-view association results, i.e., $CVIDF_1$ and CVMA scores, generated by the optimization of Eq. (1) are better than the frame-level results shown in Table 1. This demonstrates that the spatial-temporal joint optimization can further improve the cross-view subject association. To better evaluate the cross-view MOT, we show the average $CVIDF_1$ and CVMA scores over time[3] in Fig. 8. We can see that the performance of all the single-view trackers show a downward trend, while our method shows relatively steady scores without much performance decrease over time.

Actually, MOT is expected to maintain the subject ID after the initialization on the first frame. We evaluate the ID accuracy over time based on the consistency of current and initial ID numbers. As shown in Fig. 9, DMAN performs the best in the top view. However, in the horizontal view, our approach gets better performance than other trackers as the frame number increases. This is because our approach can re-identify the horizontal-view subjects by associating to the tracked subjects in the top view.

4.3.2 Ablation Studies

Parameters Selection. There are six free parameters in the proposed method: c_0 in Eq. (1), w_1 in Eq. (17), and w_2 in Eq. (18), σ in Eq. (15), λ in Eq. (12) and the tolerance threshold τ in Section 3.2.2. We select different values for them and examine their influence to the final tracking performance. Table 8

3. In this experiment, we only consider the first M frames of each video, where M is the minimum length of all the videos. This way, we can compute the average $CVIDF_1$ and CVMA scores of all the videos frame by frame.

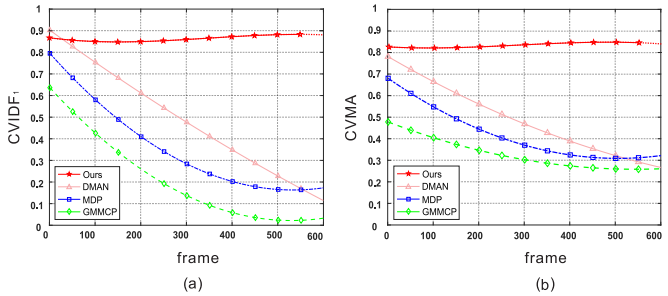


Fig. 8. Cross-view subject association results of $CVIDF_1$ score (a) and CVMA score (b) over time.

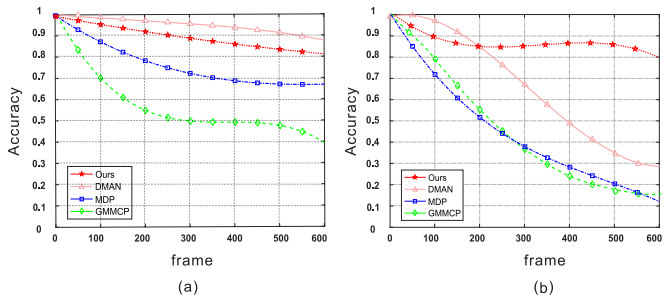


Fig. 9. ID consistency accuracy in top (a) and horizontal view (b).

reports the results by varying one parameter while fixing the other five. We can see that final tracking performance, including the standard MOT metrics of IDF_1 and MOTA, and the cross-view MOT metrics of $CVIDF_1$ and CVMA, are not very sensitive to the six parameters.

Features for Similarity Measurements. We study the influence of using different similarity measures. As shown in Table 9, 'w/o ot-App' and 'w/o Motion' denote the proposed method without the appearance and motion features in single-view over-time subject association, respectively, and 'w/o cv-App' and 'w/o Spatial' denote the proposed method without the appearance and spatial features in cross-view association, respectively. The results in the first and second rows show that using only appearance or motion features in over-time association cannot achieve performance as good as the proposed method that combines both features, although using any one of the two types of features can still produce good results. By comparing the results in the third and fourth rows with the last row, we can see that the proposed method using only appearance features in cross-view association leads to very poor tracking results. This verifies that the appearance features are not very useful for subject association between top and horizontal views. Fortunately, the proposed spatial distribution based method is effective for cross-view subject association.

TABLE 9
A Comparison of Association and Tracking Results by Using Different Features

Features	IDF_1	MOTA	$CVIDF_1$	CVMA
w/o ot-App	76.3	85.2	86.5	83.7
w/o Motion	77.0	85.2	86.9	83.9
w/o cv-App	79.2	85.3	86.6	83.7
w/o Spatial	59.4	77.7	20.3	25.5
Ours	80.4	85.3	86.9	83.9

'ot-App' and 'Motion' denote the appearance and motion features used in over-time association. 'cv-App' and 'Spatial' denote the appearance and spatial features used in cross-view association (%).

TABLE 10
Tracklet Extraction Results Using Different Features (%)

Method	Horizontal view			Top view		
	P	R	F_1	P	Recall	F_1
IOU	97.02	96.15	96.58	99.89	98.85	99.37
IOU + Color	98.46	97.96	98.21	99.89	93.19	96.42

Tracklet Extraction. We evaluate the performance of the tracklet generation result using precision, recall and F_1 score, which is shown in Table 10. A tracklet is taken as a true positive if and only if it consistently connects the bounding boxes of the same person across a clip. We use the bounding box intersection over union (IOU) and color similarity to extract the tracklet as described in Section 4.1. From the first row, we see that IOU can provide a good guidance for tracklet generation, especially for the top-view video clip where there is no mutual occlusion. In the horizontal view, the performance is improved when we combine the color similarity with IOU for tracklet generation. In the top view, however, the performance gets worse when considering color similarity as well, mainly because the color features of different subjects are not discriminative in top view, as shown in Fig. 1. Hence, we use only bounding-box IOU for the top view and the combination of bounding-box IOU and color similarity for the horizontal view in tracklet extraction.

4.3.3 Detailed Analysis

Speed Analysis. As shown in Table 11, we record the running time taken by each component of the proposed method. In

TABLE 8
A Comparison of Association and Tracking Results (%) by Varying Values of c_0, w_1, w_2, σ, λ, and τ

c_0	IDF_1	MOTA	$CVIDF_1$	CVMA	w_1	IDF_1	MOTA	$CVIDF_1$	CVMA	w_2	IDF_1	MOTA	$CVIDF_1$	CVMA
0.1	80.0	84.9	85.5	82.6	0.2	80.2	85.3	86.9	83.9	0.5	79.2	85.3	86.9	84.0
0.2	80.4	85.3	86.9	83.9	0.3	80.4	85.3	86.9	83.9	0.6	80.4	85.3	86.9	83.9
0.3	79.3	85.3	87.0	83.9	0.4	80.3	85.4	86.1	83.4	0.7	80.1	85.3	86.8	83.9
σ	IDF_1	MOTA	$CVIDF_1$	CVMA	λ	IDF_1	MOTA	$CVIDF_1$	CVMA	τ	IDF_1	MOTA	$CVIDF_1$	CVMA
0.1	79.0	85.3	86.8	83.9	0.0015	80.5	85.2	86.8	83.9	1	79.4	85.4	86.7	84.0
0.05	80.4	85.3	86.9	83.9	0.003	80.4	85.3	86.9	83.9	2	80.4	85.3	86.9	83.9
0.025	80.1	85.3	86.9	83.8	0.006	80.4	85.3	86.3	83.7	3	78.8	84.6	85.2	81.7

TABLE 11
Time Performance (Sec/Frame) of Each Component

Component	tracklet	ot-sim	cv-sim	solution
Time	0.106	0.132	0.084	0.048
Proportion	28.6%	35.7%	22.7%	13.0%

this table, 'tracklet' denotes the single-view tracklet construction, 'ot-sim' and 'cv-sim' denote the single-view over-time and cross-view data similarity computation, respectively, 'solution' denotes the step of solving the optimization problem. We can find that the single-view similarity computation takes 35.7 percent of the total running time, which is similar to the combined time taken by cross-view similarity computation and final optimization. So compared to classical motion and appearance similarity computation used in single-view tracklet construction, the proposed cross-view similarity computation and optimization are quite efficient.

Robustness Analysis. We further study the influence of initialization noise on the tracking performance when using the proposed method. Specifically, in the initialization stage, we replace the cross-view association result generated by Eq. (1) with a random assignment between the top and horizontal views. As shown in the second row of Table 12, both the online tracking performance, i.e., IDF_1, MOTA, and the cross-view association performance, i.e., $CVIDF_1$, CVMA, are quite robust to the initialization noise. The main reason is that our method jointly performs the cross-view association and over-time tracking, with which the mismatches between the top and horizontal views may be identified and corrected at later frames. As shown in the last three rows in Table 12, even if we uniformly select 1, 5 and 10 percent frames, instead of just the first frame, to apply the above random assignment noise, the proposed method can still achieve reasonable tracking performance by correcting matching errors over time.

Occlusion and Out of View. In horizontal-view videos, it is common to have subjects with mutual occlusion and being out-of-view. In this case, existing online trackers, e.g., DMAN can not associate the long-term lost subjects when they reappear. Two examples are shown in Fig. 10. The top two rows show the case of mutual occlusions. From top view frame #180, we can find that two subjects (ID 2 and 3) are occluded by others in the horizontal view, and DMAN could not identify them correctly when they reappear at frame #210. Our method can track these occluded subjects with consistent ID. Similarly, let's look at subject 4 that goes out of view at frame #165 in the horizontal view. We see that this subject is reassigned to a new ID by DMAN. Our method preserves the original ID of this subject, that is consistent to its ID in the top view.

TABLE 12
A Comparison of Association and Tracking Results (%) Under Different Cross-View Association Noise

Init_Noise	IDF_1	MOTA	$CVIDF_1$	CVMA
w/o noise	80.4	85.3	86.9	83.9
1st frame	79.9	85.1	85.1	82.4
1%	79.9	80.3	82.6	73.1
5%	78.0	74.4	76.2	63.7
10%	76.7	72.5	72.7	60.0

4.4 Results on KITTI Benchmark

Overview. As discussed in Section 2, the LiDAR-based bird's eye view (BEV) and the RGB-camera front view utilized for 3D object detection and tracking in autonomous driving also provide a kind of top and horizontal views, respectively, as shown in Fig. 11. Although the same subject shows totally different appearance in such two views, in this section we conduct experiments to show that the proposed method can still build cross-view association based on the spatial distribution of subjects.

Setup on KITTI. We first evaluate the proposed cross-view association method on the object detection training dataset of KITTI that contains 7,481 image pairs from both the LiDAR bird's eye view (top view) and RGB camera view

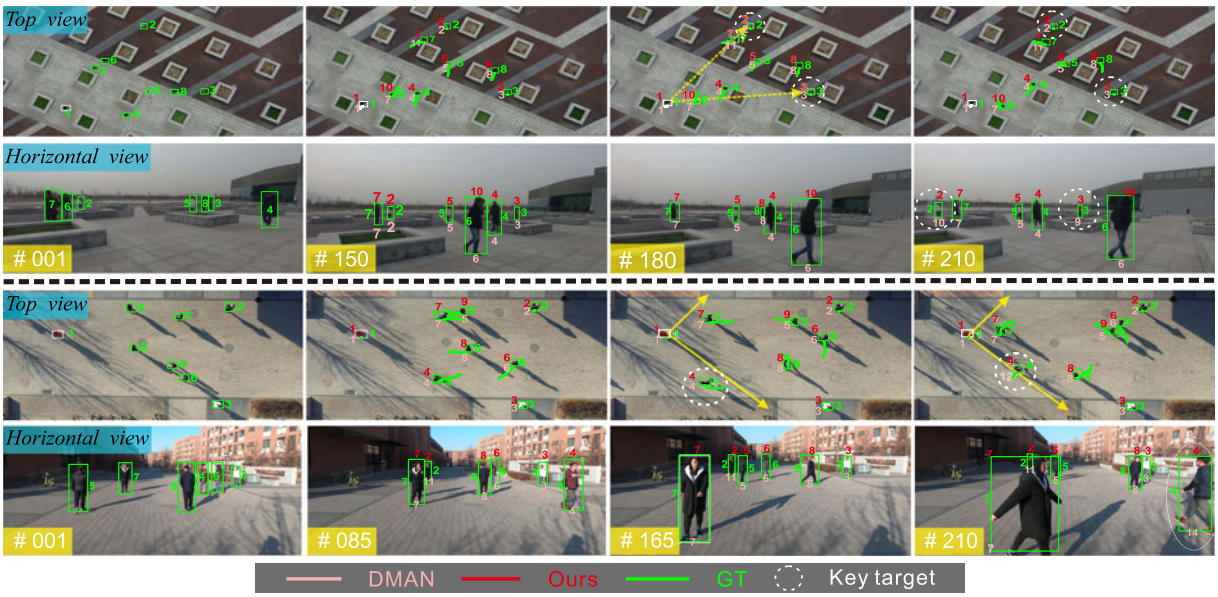


Fig. 10. Case analysis of long-term occlusion (top) and out-of-view (bottom) scenarios.

Fig. 11. LiDAR bird's eye view (top view) and RGB camera view (horizontal view) in KITTI dataset.

(horizontal view). We further evaluate our tracking method on the video-based object tracking dataset of KITTI, which contains 21 videos and the corresponding ground-truth annotated bounding boxes and labels.

Implementation Details. We use the benchmark subject bounding boxes as subject detections. Since the horizontal camera is not associated with a subject in KITTI data, we extend the strategy in Section 3.2.4 to automatically search for the horizontal-view camera location O in the top-view image, where the searching space consists of M equidistantly sampled candidate camera locations on the left edge of the top-view image, e.g., points $O_i, i = 1, 2, \ldots, M$ in Fig. 11. We set M as 5 in the experiments for a small searching space. On KITTI, we only use motion consistency for over-time subject association in the top view and the spatial reasoning for cross-view subject association.

Results and Discussions. As shown in Table 13, the proposed method, when using $\mathbf{x}$ and $\mathbf{y}$ vectors jointly, can achieve better performance than using one of them individually. For this dataset, the location of O can actually be derived because the cameras are calibrated. We also adapt our method by using the accurate O without searching and the resulting performance is shown in the bottom row of Table 13. All these results are promising and indicate the possible applicability of the proposed method on such datasets in autonomous driving. Table 14 shows the tracking performance of the proposed method and two comparison methods on the horizontal-view videos. The MOTP scores are high for all three methods because we use the annotated bounding boxes for association and tracking. In terms of the comprehensive MOTA metric, we can see that the proposed method outperforms the comparison method GMMCP, which uses the same features as in the proposed method in this experiment. The proposed method also outperforms the state-of-the-art tracking method DMAN. These results on KITTI benchmark are consistent to the comparison results on our dataset shown in Table 6.

TABLE 13
A Comparison of Cross-View Association Results on KITTI Detection Dataset Using Different Representations

Method	CVIDP	CVIDR	CVIDF$_1$	CVMA
w only $\mathbf{x}$	53.3	51.9	52.6	51.9
w only $\mathbf{y}$	52.5	45.0	48.5	45.0
Ours ($\mathbf{x}$ + $\mathbf{y}$)	73.1	72.8	72.9	72.8
Ours w O	95.5	95.2	95.4	95.2

TABLE 14
A Comparison of Tracking Results From Different Methods on KITTI Tracking Dataset

Method	IDP	IDR	IDF$_1$	MOTP	MOTA
GMMCP	33.0	33.4	33.2	96.8	57.6
DMAN	72.0	49.3	58.5	90.8	62.3
Ours	79.2	60.7	68.7	94.5	67.9

5 Discussion

As discussed above, the proposed method is designed under some specific assumptions. In this section, we provide more discussions about the limitations and feasibility of the proposed method to more general and practical settings.

5.1 General Setting

We assume the ideal top and horizontal views in our problem setting, i.e., the top view is almost vertical to the ground and the horizontal view has no pitching and rolling. When deviation from the ideal setting is not significant, the proposed method can tolerate it – actually the proposed dataset in the experiments are collected in real environment and the setting is usually not ideal. In practice, we can also estimate and compensate the disparity from the ideal setting before applying the proposed method. Note that, in the ideal setting, the roll of the top-view camera and the yaw of the horizontal-view camera are parallel to the ground and these two angles are related by searching for the optimal horizontal-view angle θ in the top-view image, as described in Section 3.2.4. In the following, we describe the general setting by considering the pitch and yaw of the top-view camera, and the pitch and roll of the horizontal camera.

Non-Vertical Top View. In this case, there is a non-zero angle β between the top-camera view and the exact vertical view, as shown in Fig. 12a. We can estimate a homography transform to map the captured top-view image to a real one under the ideal setting, i.e., exact vertical view. This homography transform takes the form of $\mathbf{H}^{\text{top}} = \mathbf{K}_1\mathbf{R}^{\text{top}}\mathbf{K}_1^{-1}$, where $\mathbf{K}_1$ is the intrinsic matrix of the top-view camera, $\mathbf{R}^{\text{top}}$ is the rotation matrix between the top-view camera and the one in the ideal setting [51]. The rotation matrix $\mathbf{R}^{\text{top}}$ can be calculated from the inclined angles β_x and β_y derived from angle β [51]. In the proposed method, we treat each subject in the top-view image as a point. Therefore, we can directly apply the homography transform to these points and get the subjects' locations in the ideal top view, with which we can estimate the top-view representation as described in Section 3.2.2 for the proposed cross-view association and tracking. Note that, this only works when the the ground is leveled, since homography transform only maps a plane to another. In addition, the disparity angle β cannot be too large. Otherwise, the top view may not cover all subjects and mutual occlusion may also appear and prevent the detection of some subjects in the top view.

Non-Parallel Horizontal View. When there is a non-zero pitch angle γ between the horizontal view and the ground, as shown Fig. 12b, we can also apply homography transform $\mathbf{H}^{\text{hor}} = \mathbf{K}_2\mathbf{R}^{\text{hor}}\mathbf{K}_2^{-1}$, where $\mathbf{K}_2$ is the horizontal-view-camera intrinsic matrix and $\mathbf{R}^{\text{hor}}$ can be calculated from γ,

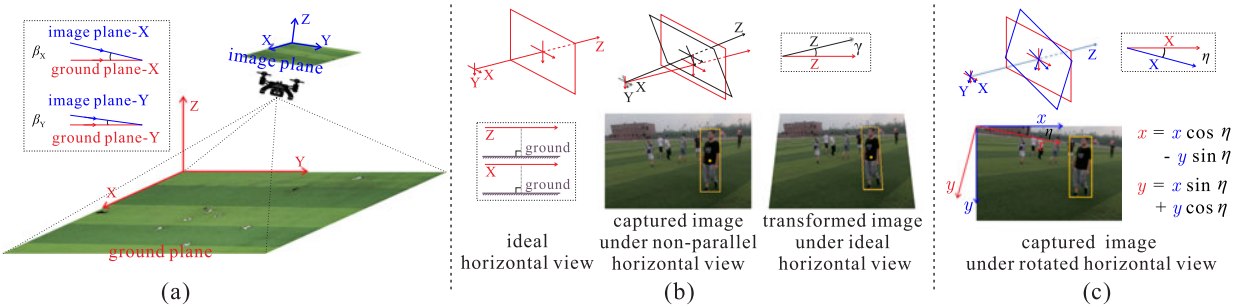


Fig. 12. General settings of our method. (a) Top view with an incline angle. (b) Horizontal view with pitching. (c) Horizontal view with rolling.

to map the captured image to its horizontal-view version with zero pitch angle. In this case a rectangular bounding box of the subject in the horizontal view may be mapped to a trapezoid after the homography transform. This is not a problem for our method since we only need the central point and height of the subject as described in Section 3.2.2 – we can directly apply $\mathbf{H}^{\text{hor}}$ to the central point while height can be transformed by multiplying a factor of $\cos\gamma$. Then we use the transformed central points and heights to compute the horizontal-view representation as discussed in Section 3.2.2. In practice, the pitch angle γ cannot be too large. Otherwise subjects may be cropped out of the field of view.

Rotated Horizontal View. When the horizontal-view camera has non-zero roll angle η as shown in Fig. 12c, we only need to rotate the captured horizontal-view image by the same angle η. If the subjects are detected in the captured image in the form of bounding boxes with the center location $C = (x, y)$ and the height h, the corresponding center location and height of the subject in the ideal horizontal view can be calculated as $C' = (x\cos\eta - y\sin\eta, x\sin\eta + y\cos\eta)$ and $h' = \frac{h}{\cos\eta}$, respectively. The transformed central points and heights can then be used to derive the horizontal-view representation as discussed in Section 3.2.2. The roll angle can be large since it often does not cause information loss for the covered subjects, although in practice the rolling angle of wearable cameras is usually limited.

5.2 Limitation Analysis

Despite the above discussions of the applicability to more general settings, our method still has some limitations.

Non-Ideal Camera Settings. As discussed above, when the disparity from the ideal camera setting is large, i.e., non-vertical top view and/or pitching and rolling of the horizontal-view camera, we may need to transform the captured images before applying the proposed method. These transforms require more camera parameters measured by other sensors equipped in the cameras. In particular, the disparity angle of the top-view camera and the pitching angle of the horizontal-view camera cannot be too large. The former may prevent the detection of the subject in the top view. The latter may crop subjects out of the view. Both of them may seriously affect the performance of the proposed association and tracking methods.

Number of Subjects. The cross-view subject association and tracking performance, to some extent, are related to the number of subjects in the scene. The association performance gets worse when the number of subjects is too high or too low in the horizontal-view video. When there are too many subjects, the crowdedness in the horizontal view may prevent the accurate detection of subjects. When there are too few subjects, the constructed vector representation is not sufficiently discriminative to locate the camera location O and camera-view orientation θ.

5.3 Potential Applications

The proposed multiple human association and tracking builds a new setting for multi-view video/image analysis, which has many potential applications.

Cross-View Person Identification (CVPI). CVPI [52], [53], using multiple wearable cameras, is challenging due to frequent subject interactions and occlusions. In this case, a top view provides a global picture of all subjects and can act as a bridge to associate subjects in all horizontal views.

Complementary-View Human Activity Analysis. Based on the human association and tracking results, we can synchronously monitor the subjects of interest from different views, which has many applications, such as multi-view multi-human activity recognition [54], important person identification in crowded scene [55]. This can significantly enhance the outdoor video surveillance capability on anomaly detection and scene understanding.

Sport Video Analysis. In many sports scenarios, e.g., basketball or soccer games, players move very fast and may be difficult to track in a horizontal-view video. By adding a top-view camera from a high altitude, together with the proposed complementary association and tracking algorithms, we can better track multiple players in the horizontal view and conduct more advanced sport scene analysis.

Autonomous Driving. The bird's eye view and RGB camera view are widely used for 3D object detection and tracking in autonomous driving, which are similar to the top and horizontal views studied in this paper. The experimental results on KITTI (Section 4.4) show the potential use of the proposed method to these applications. Specifically, our method can provide more cues, e.g., the geometrical features, for 3D object detection, association and tracking.

6 Conclusion

In this paper, we have studied a new problem of associating and tracking multiple subjects in the complementary top- and horizontal-view videos. We formulate such collaborative tracking as a joint optimization problem in the spatiotemporal domain that can be solved by a constrained mixed integer programming algorithm. Within each view,

we associate subjects over time using appearance and motion features. Across the top and horizontal views, where classical appearance and motion features are not consistent, we associate the subjects by further considering their spatial distributions. Specifically, we proposed a new geometry-based method to represent and match the subjects' spatial distributions across the top and horizontal views. For benchmark evaluation, we built a new complementary-view video dataset, consisting of temporally synchronized video pairs from top and horizontal views, as well as manually labeled ground-truth subjects association/tracking results. Extensive experimental results on this dataset verified that the proposed approach can effectively and collaboratively associate and track multiple subjects across spatial and temporal domains, by capturing both their global spatial distributions and trajectories as well as their local appearance details. In the future, we are interested in extending our formulation to multiple top-view and horizontal-view cameras, and exploring more application scenarios of the proposed complementary-view human association and tracking in real world.

ACKNOWLEDGMENTS

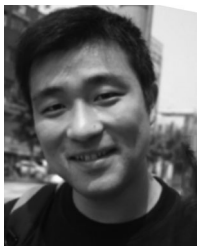
The authors would like to thank Nan Li, Zicheng Niu, Yiyang Gan, Tingliang Feng, Xiaotao Liu, Haomin Yan, Yibo Shi, and Qian Zhang for the daily academic communication for this paper. They would also like to thank Sibo Wang, Shuai Wang, Chenxing Gong, Xiaoyu Zhang, Yun Wang, and Haibo Li for their kind assistance in the collection and annotation of our dataset. This work was supported by the NSFC under Grants U1803264, 62072334, 61672376, and 61671325.

REFERENCES

[1] R. Han *et al.*, "Complementary-view multiple human tracking," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, pp. 10917–10924, 2020.
[2] L. Yuan, H. Chang, and R. Nevatia, "Learning to associate: Hybridboosted multi-target tracker for crowded scene," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 2953–2960.
[3] S. Tang, B. Andres, M. Andriluka, and B. Schiele, "Multi-person tracking by multicut and deep matching," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 100–111.
[4] A. Maksai and P. Fua, "Eliminating exposure bias and loss-evaluation mismatch in multiple object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4634–4643.
[5] Y. Xiang, A. Alahi, and S. Savarese, "Learning to track: Online multi-object tracking by decision making," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4705–4713.
[6] J. Zhu, H. Yang, N. Liu, M. Kim, W. Zhang, and M. Yang, "Online multi-object tracking with dual matching attention networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 379–396.
[7] J. Xu, Y. Cao, Z. Zhang, and H. Hu, "Spatial-temporal relation networks for multi-object tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3987–3997.
[8] P. Bergmann, T. Meinhardt, and L. Leal-Taixe, "Tracking without bells and whistles," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 941–951.
[9] R. Han, Q. Guo, and W. Feng, "Content-related spatial regularization for visual object tracking," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2018, pp. 1–6.
[10] W. Feng, R. Han, Q. Guo, J. Zhu, and S. Wang, "Dynamic saliency-aware regularization for correlation filter-based object tracking," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3232–3245, Jul. 2019.
[11] Q. Guo, R. Han, W. Feng, Z. Chen, and L. Wan, "Selective spatial regularization by reinforcement learned decision making for object tracking," *IEEE Trans. Image Process.*, vol. 29, pp. 2999–3013, 2020.
[12] R. Han, W. Feng, and S. Wang, "Fast learning of spatially regularized and content aware correlation filter for visual tracking," *IEEE Trans. Image Process.*, vol. 29, pp. 7128–7140, 2020.
[13] H. Izadinia, I. Saleemi, W. Li, and M. Shah, "MP^{2}T: Multiple people multiple parts tracker," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 100–114.
[14] A. R. Zamir, A. Dehghan, and M. Shah, "GMCP-Tracker: Global multi-object tracking using generalized minimum clique graphs," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 343–356.
[15] L. Lealtaixe, C. Cantonferrer, and K. Schindler, "Learning by tracking: Siamese CNN for robust target association," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 418–425.
[16] Q. Chu, W. Ouyang, H. Li, X. Wang, B. Liu, and N. Yu, "Online multi-object tracking using CNN-based single object tracker with spatial-temporal attention mechanism," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4846–4855.
[17] A. Milan, S. Roth, and K. Schindler, "Continuous energy minimization for multitarget tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 58–72, Jan. 2014.
[18] A. Dehghan, S. M. Assari, and M. Shah, "GMMCP Tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2015. pp. 4091–4099.
[19] E. Ristani and C. Tomasi, "Features for multi-target multi-camera tracking and re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6036–6046.
[20] B. Yang and R. Nevatia, "Multi-target tracking by online learning of non-linear motion patterns and robust appearance models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1918–1925.
[21] B. Yang and R. Nevatia, "An online learned CRF model for multi-target tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2034–2041.
[22] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 267–282, Feb. 2008.
[23] Z. Wu, N. I. Hristov, T. L. Hedrick, T. H. Kunz, and M. Betke, "Tracking a large number of objects from multiple views," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 1546–1553.
[24] X. Liu, "Multi-view 3D human tracking in crowded scenes," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 3553–3559.
[25] Y. Xu, X. Liu, Y. Liu, and S. Zhu, "Multi-view people tracking via hierarchical trajectory composition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4256–4265.
[26] Y. Xu, X. Liu, L. Qin, and S. Zhu, "Cross-view people tracking by scene-centered spatio-temporal parsing," in *Proc. AAAI Conf. Artif. Intell.*, 2017, 4299–4305.
[27] M. Ayazoglu, B. Li, C. Dicle, M. Sznaier, and O. I. Camps, "Dynamic subspace-based coordinated multicamera tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2462–2469.
[28] M. Hofmann, D. Wolf, and G. Rigoll, "Hypergraphs for joint multi-view reconstruction and multi-object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3650–3657.
[29] Z. Tang, R. Gu, and J.-N. Hwang, "Joint multi-view people tracking and pose estimation for 3D scene reconstruction," in Proc. *IEEE Int. Conf. Multimedia .Expo*, 2018, pp. 1–6.
[30] X. Liu, Y. Xu, L. Zhu, and Y. Mu, "A stochastic attribute grammar for robust cross-view human tracking," *IEEE Trans. Circuits. Syst. Video Technol.*, vol. 28, no. 10, pp. 2884–2895, Oct. 2018.
[31] S. Ardeshir and A. Borji, "Ego2Top: Matching viewers in egocentric and top-view videos," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 253–268.
[32] S. Ardeshir and A. Borji, "Egocentric meets top-view," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 6, pp. 1353–1366, Jun. 2019.
[33] S. Ardeshir and A. Borji, "Integrating egocentric videos in top-view surveillance videos: Joint identification and temporal alignment," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 300–317.
[34] B. Yang, W. Luo, and R. Urtasun, "Pixor: Real-time 3D object detection from point clouds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7652–7660.
[35] B. Yang, M. Liang, and R. Urtasun, "HDNet: Exploiting HD maps for 3D object detection," in *Proc. Conf. Robot Learn.*, 2018, pp. 146–155.
[36] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3D object detection network for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6526–6534.
[37] M. Liang, B. Yang, S. Wang, and R. Urtasun, "Deep continuous fusion for multi-sensor 3D object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp 663–678.

[38] W. Zhang, H. Zhou, S. Sun, Z. Wang, J. Shi, and C. C. Loy, "Robust multi-modality multi-object tracking," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 2365–2374.
[39] H. Kuang, X. Liu, J. Zhang, and Z. Fang, "Multi-modality cascaded fusion technology for autonomous driving," in *Proc. Int. Conf. Robot. Automat. Sci.*, 2020, pp. 44–49.
[40] R. Han *et al.* "Multiple human association between top and horizontal views by matching subjects' spatial distributions," in 2019, *arXiv:1907.11458*.
[41] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.
[42] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.
[43] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. Int. Conf. Mach. Learn. Workshop*, 2015.
[44] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, and S. Wang, "Learning dynamic Siamese network for visual object tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1781–1789.
[45] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5157–5166.
[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[47] R. Hadsell, S. Chopra, and Y. Lecun, "Dimensionality reduction by learning an invariant mapping," in Proc. *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 1735–1742.
[48] K. Grauman, "The pyramid match kernal: Discriminative classification with sets of image features," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2005, pp.pp. 1458–1465.
[49] L. Lealtaixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "MOTChallenge 2015: Towards a benchmark for multi-target tracking," in 2015, *arXiv:1504.01942*.
[50] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp 17–35.
[51] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110, 2004.
[52] K. Zheng, X. Fan, Y. Lin, H. Guo, and S. Wang, "Learning view-invariant features for person identification in temporally synchronized videos taken by wearable cameras," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2877–2885.
[53] G. Liang, X. Lan, X. Chen, K. Zheng, S. Wang, and N. Zheng, "Cross-view person identification based on confidence-weighted human pose matching," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 3821–3835, Aug. 2019.
[54] J. Zhao, R. Han, Y. Gan, L. Wan, W. Feng, and S. Wang, "Human identification and interaction detection in cross-view multi-person videos with wearable cameras," in *Proc. ACM Multimedia*, 2020, pp. 2608–2616.
[55] R. Han, J. Zhao, W. Feng, Y. Gan, L. Wan, and S. Wang, "Complementary-view co-interest person detection," in *Proc. ACM Multimedia*, 2020, pp. 2746–2754.

Ruize Han received the BS degree in mathematics and applied mathematics from the Hebei University of Technology, China, in 2016, and the ME degree, in 2019, in computer technology from Tianjin University, China, where he is currently working toward the PhD degree with the College of Intelligence and Computing. His major research interests include visual intelligence, specifically including multi-camera video collaborative analysis, visual object tracking, and solving preventive conservation problems of cultural heritages via artificial intelligence.



Wei Feng (Member, IEEE) received the PhD degree in computer science from the City University of Hong Kong, in 2008. From 2008 to 2010, he was a research fellow with The Chinese University of Hong Kong and the City University of Hong Kong. He is currently a professor with the School of Computer Science and Technology, College of Computing and Intelligence, Tianjin University, China. His major research interests include active robotic vision and visual intelligence, specifically including active camera relocalization and lighting recurrence, general Markov Random Fields modeling, energy minimization, active 3D scene perception, SLAM, generic pattern recognition, and solving preventive conservation problems of cultural heritages via computer vision and machine learning. He is an associate editor for *the Neurocomputing* and the *Journal of Ambient Intelligence and Humanized Computing*.

Yujun Zhang received the BE degree from the School of Software Engineering, Shanxi University, China, in 2017, and the ME degree, in 2020, in computer technology from Tianjin University, China, where she is currently working toward the PhD degree. Her research interests include visual object tracking, lighting recurrence, 3D scene perception, and reconstruction.

Jiewen Zhao received the BE and ME degrees from the School of Software Engineering, Tianjin University, China, in 2018 and 2021, respectively. His research focuses on multi-camera video collaborativ analysis, specially for the multi-human visual tracking and activity understanding.

Song Wang (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Illinois at UrbanaChampaign (UIUC), Champaign, IL, USA, in 2002. From 1998 to 2002, he was a research assistant with the Image Formation and Processing Group, Beckman Institute, UIUC. In 2002, he joined the Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, USA, where he is currently a professor. His current research interests include computer vision, image processing, and machine learning. He is currently the publicity or the web portal chair of the Technical Committee of Pattern Analysis and Machine Intelligence of the IEEE Computer Society, an associate editor for the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, the *Pattern Recognition Letters*, and the *Electronics Letters*. He is also a member of the *IEEE Computer Society*.