International Journal of Computer Vision (2024) 132:118–136
https://doi.org/10.1007/s11263-023-01857-z

Benchmarking the Complementary-View Multi-human Association and Tracking

Ruize Han[1,2] · Wei Feng[1] · Feifan Wang[1] · Zekun Qian[1] · Haomin Yan[1] · Song Wang[3]

Received: 3 March 2022 / Accepted: 13 July 2023 / Published online: 23 August 2023

Abstract
Using multiple moving cameras with different and time-varying views can significantly expand the capability of multiple human tracking in larger areas and with various perspectives. In particular, the use of moving cameras of complementary top and horizontal views can facilitate multi-human detection and tracking from both global and local perspectives. As a new challenging problem that draws more and more attention in recent years, one main issue is the lack of a comprehensive dataset for credible performance evaluation. In this paper, we present such a new dataset consisting of videos synchronously recorded by drone and wearable cameras, with high-quality annotations of the covered subjects and their cross-frame and cross-view associations. We also propose a pertinent baseline algorithm for multi-view multiple human tracking and evaluate it on this new dataset against the annotated ground truths. Experimental results verify the usefulness of the new dataset and the effectiveness of the proposed baseline algorithm.

Keywords Complementary view · Multi-human association · Multi-human tracking · Benchmark

1 Introduction

Multiple object tracking (MOT), especially multiple human tracking, is an essential task in visual intelligence and of fundamental importance in video surveillance (Luiten et al., 2020; Zhang et al., 2021; Zhu et al., 2022). Most of existing works on MOT, including the datasets and the methods, are focused on the videos captured by single-view or fixed cameras. They suffer from the problems of limited area coverage, frequent mutual occlusions, and short tracking durations. The use of multiple moving cameras, such as mobile phone cameras, the wearable cameras, and the drone-mounted cameras can well address these problems. In this paper, we explore the multiple human tracking on the videos taken by multiple moving cameras.

Furthermore, we are specifically interested in the videos taken by moving cameras from both top views, e.g., those mounted to the drone with view direction largely perpendicular to the ground, and various horizontal views, e.g., those worn by walking people on the ground. As shown in Fig. 1c–f, *top-view cameras from a high altitude can capture a global distribution of all or most of subjects on the ground* while horizontal-view cameras can approach individual subjects to observe more appearance details. Multiple human tracking on such *complementary views is highly desirable in*

Communicated by Matteo Poggi.

✉ Wei Feng
wfeng@tju.edu.cn

Ruize Han
han_ruize@tju.edu.cn

Feifan Wang
wff@tju.edu.cn

Zekun Qian
clarkqian@tju.edu.cn

Haomin Yan
yan_hm@tju.edu.cn

Song Wang
songwang@cec.sc.edu

[1] School of Computer Science and Technology, College of Intelligence and Computing, Tianjin University, Tianjin 300350, China

[2] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

[3] Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA

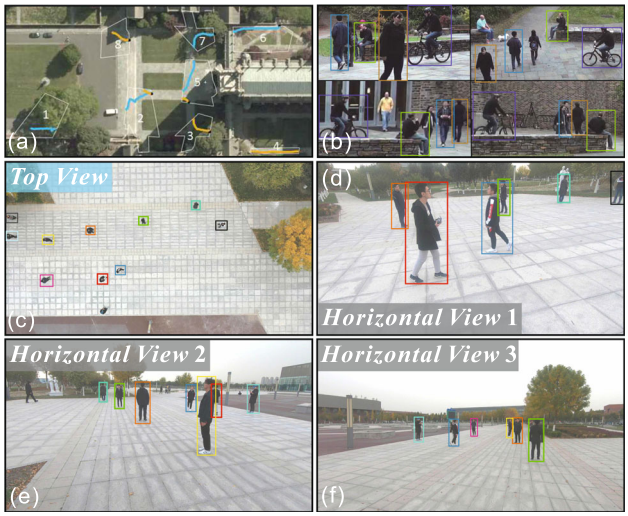


Fig. 1 **a** Plan view showing the location and field of view (FOV) of the cameras in DukeMTMC (Ristani et al., 2016). **b** Sampled multi-view frames from the UC CAMPUS dataset (Xu et al., 2016). **c**, **d** Sample frames from our new collected dataset, which are synchronously taken by one top-view camera (**c**) and three horizontal-view cameras (**d**–**f**), respectively, with identical-color boxes indicating the same subject across multiple views (Color figure online)

surveillance by providing information from both global and local perspectives. With the above setting, a typical practical use case is the air-ground collaborative moving camera system for outdoor law enforcement. Imagine a scene where we apply a flying drone with a mounted camera, together with several officials with helmet cameras, to form a moving camera system. Then, we can perform the collaborative human detection, association and tracking, to obtain the global picture of the people crowd and clear trajectories, from the top view in the air, and observe the details, e.g., pose, actions, of some specific subjects, from the horizontal view on the ground. This breaks through the limitation of the traditional surveillance system that requires the pre-installed cameras and has the inflexible coverage area.

With the above advantages, the complementary-view moving camera setting also makes the cross-view subject association and tracking a very difficult problem, e.g., the top view of a human only contains the head top and two shoulders without much useful appearance information to match its horizontal counterparts, the joint optimization of spatial-temporal subject association from uncertain cameras. A couple of prior works (Ardeshir & Borji, 2018, b) leverage a non-vertical inclined angle for the top view to facilitate the appearance-based association between complementary views. Another prior work (Han et al., 2020a, 2022a) tries to address both the association and tracking between the real vertical top and horizontal views by exploring the spatial distribution of the same group of subjects in different views. But it only handles the case of two cameras: one for top view and the other for horizontal view. In this work, we consider the more general setting of one top-view camera and multiple horizontal-view cameras—those horizontal cameras may also show large view differences, as shown Fig. 1d–f.

To attract more attentions to the **C**omplementary-**v**iew **M**ulti-**H**uman **A**ssociation and **T**racking (CvMHAT) task, one urgent problem is a public dataset that can be accessed and used by different researchers to quantitatively evaluate the developed algorithms. In this paper, we collect a new dataset where videos are taken by synchronous cameras mounted to drones and worn by people, as shown in Fig. 1c–f. The new dataset is composed of a real-world dataset and a synthetic dataset, which both contain 30 video groups, in total 200 videos adding up to more than 100 min. For annotation, we not only annotate the bounding box of each subject on each frame, but also an identity for each annotated subject such that the same subject in different frames and different views bears the same identity number.

Compared to existing datasets for MOT, besides using the multiple cameras that have been studied in several works (Ristani et al., 2016; Xu et al., 2016, 2017), CvMHAT has two peculiarities: (1) We use the *moving cameras* instead of fixed cameras. Compared to fixed cameras as in Fig. 1a, the moving cameras, e.g., GoPro, phone cameras, can flexibly cover larger areas and are particularly useful in the outdoor video surveillance at locations without pre-installed cameras, e.g., occasional gathering and open-air concert. (2) We leverage the *complementary-view* cameras rather than the congeneric-view ones. Specifically, we apply a top view taken by the drone-mounted cameras as show in Fig. 1c. Compared to the congeneric-view cameras, as shown in Fig. 1b, with the similar coverage areas and observation scales in most previous works, the complementary-top-horizontal views have several unique advantages. First, the top view can globally record the spatial distribution and temporal trajectories of all the subjects in the scene, which is the easiest form of MOT with respect to the perspective of trajectory tracking performance. Meanwhile, the multiple horizontal views can closely observe the local appearance and pose of selected subjects. With the cross-view observation and association, the multi-view human appearance or pose can be fused for addressing many higher-level video applications, e.g., multi-view human action recognition (Zhao et al., 2020) and 3D pose estimation (Dong et al., 2022) in various scenarios, e.g., outdoor party/show and outdoor law enforcement. Second, the top view can be taken as a central pivot of the moving camera network to better coordinate the collaborative analysis of the videos taken by the horizontal-view cameras without calibration, which, meanwhile, can actively observe the scene of interest to obtain more needed finer-grained information.

Based on the new dataset, we propose a baseline method for CvMHAT. Specifically, we first formulate the CvMHAT

as a generalized maximum (multi) clique problem (Zamir et al., 2012), in which we measure the cross-view and over-time subject similarities using appearance, motion features and the spatial reasonings. With these similarities, for problem solution, we propose an efficient Alternating Direction Method of Multipliers (ADMM) alike algorithm for optimization and produce the desired CvMHAT results. Experimental results show that the proposed method with relatively simple features can provide much better results than previous MOT algorithms with sophisticated deep features on our problem. The main contributions of this paper are: ❶ We build a new dataset of complementary top- and horizontal-view videos taken by multiple moving cameras. The dataset exhibits good varieties and consists of both real-world and synthetic data. We also provide new metrics for evaluating CvMHAT performance. ❷ We develop an effective baseline algorithm for CvMHAT, and conduct a serious of experiments to evaluate several existing related approaches and the proposed method on CvMHAT dataset. We report their performance and show the usefulness of this new dataset and the effectiveness of the proposed baseline method. ❸ We have released the video dataset, the annotations, the evaluation toolkit, as well as the source code and trained networks of the baseline algorithm, to the public at https://github.com/RuizeHan/MHATB, which may help attract more researchers to work on the important but challenging CvMHAT problem.

2 Related Work

Multiple Object Tracking (MOT) is a classical problem in computer vision. An important issue in MOT is data association over time—MOT can be treated as the problem of associating the object bounding boxes provided by a detector. Many different features have been used, while the most commonly used are appearance and motion features. The former includes many hand-crafted appearance features such as color histograms (Zamir et al., 2012; Dehghan et al., 2015) and recent deep network based appearance features (Leal-taixe et al., 2016; Chu et al., 2017; Zhu et al., 2018). The latter includes both linear motion models, which assume the target to have a linear movement with constant velocity for a period of time (Zamir et al., 2012; Dehghan et al., 2015; Ristani & Tomasi, 2018) and nonlinear motion models, which can better capture irregular movements and produce more accurate motion predictions (Yang & Nevatia, 2012a, b). Following the above tracking by detection, also called as separate detection and embedding (SDE) paradigm, classical DeepSort (Wojke et al., 2017) and the state-of-the-art StrongSort++ (Du et al., 2022) obtain the remarkable performance on MOT task. Rcent works also try to achieve object detection and tracking simultaneously using an end-to-end joint detection and tracking (JDT) paradigm (Wu et al., 2021). For example, Tracktor++ (Bergmann et al., 2019a) applies the regression and classification modules on a object detector for the tracking task. CenterTrack (Zhou et al., 2020) takes a pair of images as input, which localizes objects and predicts their associations with the previous frame. FairMOT (Zhang et al., 2021) presents a bunch of detailed designs for balancing the detection and Re-ID tasks, which are critical to achieving good tracking results. Bytetrack (Zhang et al., 2022) proposes to make use of the useful information from the low-confidence human detection bounding boxes. More recently, some trackers based on graph neural networks (GNN), e.g., Brasó and Leal-Taixé (2020) or transformers, e.g., Meinhardt et al. (2022); Ma et al. (2022); Zhou et al. (2022) are developed and obtain the promising performance.

Calibrated-Camera Based Video Analysis Most existing multi-view video analysis focus on the videos taken by fixed cameras, which can be mechanically calibrated in advance. For the multi-view MOT, Xu et al. (2016, 2017) leverage the given camera calibration and back-project each 2D bounding box onto the 3D ground as the unified motion features across different views. Another group of works (Kuo et al., 2010; Ristani et al., 2016; Ristani & Tomasi, 2018) aim to track and re-identify the humans in a large field, e.g., a campus, using many cameras installed at many sites with little or no field of view overlap. With prior calibrations, global trajectories of multiple targets can be computed across different cameras. Differently, in this paper, we employ multiple moving cameras with indeterminate motions and they cannot be priorly calibrated.

Multiple Moving-Camera Based Video Analysis Recently, a series of works focus on the multiple moving-camera based video analysis, e.g., cross-view person identification (Zheng et al., 2017; Liang et al., 2018, 2019), human activity detection (Zhao et al., 2020; Lin et al., 2015; Zheng et al., 2014), cross-view camera wearer identification (Ardeshir & Borji, 2018, 2016, 2018b), and multi-human association and tracking (Han et al., 2019, 2020a). Most of them use the egocentric-view videos taken by multiple wearable cameras. However, the multi-views in these works are mainly the congeneric horizontal views with different view orientations (Zheng et al., 2017; Liang et al., 2018; Zhao et al., 2020; Lin et al., 2015). Currently, several works propose to study the collaborative video analysis by combining top and horizontal views. For example, in Ardeshir and Borji (2018, 2016, 2018b), horizontal-view camera wearers are identified in the top view and then the same subjects between the top and horizontal views are associated. However, they still use the cameras with inclined angles and not very high altitude for the top view. In (Han et al., 2020a), MHAT with a top view largely vertical to the ground was addressed by combining different features, which is further extended to cross-view detection of important persons (Han et al., 2020b). However, both of them only consider two cameras—one for top view

Table 1 Comparison with the existing MOT datasets from different aspects, including the number of video groups (# Gro.) and videos, overall length, frame rate, number of subjects and cameras, camera settings, scene type and publication venues

Dataset	# Gro	# Vid	Length (min)	FPS	# ID	# Cam	Calib	Overlap	Scene	Publish
MOT 15	–	22	17	7–30	–	1	–	–	Both	arXiv 15
MOT 17	–	14	8	14–30	–	1	–	–	Both	arXiv 16
EPFL	5	30	96	25	7	4	✓	✓	Indoor	PAMI 08
PETS2009	1	8	8	7	30	4–8	✓	✓	Outdoor	IEEEW 09
USC Campus	1	3	75	30	146	8	✗	✗	Outdoor	ECCV 10
Dana36	–	36	–	–	21	36	✓	✗	Both	AVSS 12
CamNeT	1	8	240	25	50	5–8	✗	✗	Both	WACV 15
DukeMTMC	1	8	680	60	2834	8	✓	✗	Outdoor	ECCV 16
UC CAMPUS	4	16	50	30	25	3–4	✓	✓	Outdoor	CVPR 16
PPL-DA	4	12	48	30	25	4	✓	✓	Outdoor	AAAI 17
MHT	15	30	13	30	30	2	✗	✓	Outdoor	AAAI 20
CvMHAT-R (ours)	30	100	56	30	60	2–5	✗	✓	Outdoor	–
CvMHAT-S (ours)	30	100	48	30	100	2–5	✗	✓	Outdoor	–
CvMHAT (ours)	60	200	104	30	160	2–5	✗	✓	Outdoor	–

and the other for horizontal view. Differently, the proposed dataset and developed baseline algorithm can be used for the videos taken by one top-view and multiple horizontal-view cameras.

Multi-view MOT Datasets Several works focus on the research of multi-view multiple object tracking (MVMOT). Among them, the datasets in Kuo et al. (2010), Zhang et al. (2015), Ristani et al. (2016) are proposed to handle the human tracking and re-identification problem in a large non-overlapped field using multiple surveillance cameras. The cameras used in such datasets are pre-installed with limited and fixed fields of view (FOV). Another series of works focus on the collaborative analysis of a crowded scene using multiple cameras with overlapped area coverage (Fleuret et al., 2008; Ferryman, 2009; Xu et al., 2016). For example, early works propose to use multiple cameras to track the people at indoor (Fleuret et al., 2008) and outdoor (Ferryman, 2009) scenes, respectively. Recently, Xu et al. (2016, 2017) propose a multi-view multi-object tracking dataset which include multiple people with varied poses and different actions. The dataset is annotated with the trajectory of every person inside the scene (tracking), as well as their cross-view ID consistencies (association). As discussed above, in Han et al. (2020a) a relatively small-scale videos from two complementary-view cameras, one for top view and the other for horizontal view, are collected and used for CvMHAT evaluation. The data size in Han et al. (2020a) is also quite small compared to this work, which also does not provide the comprehensive annotations and metrics. In this paper, we construct a new dataset with much larger scale including one top view and multiple horizontal views, all necessary annotations and a baseline algorithm, for more general CvMHAT evaluation.

3 CvMHAT Benchmark

3.1 Problem Definition

Given a set of videos taken by temporally-synchronized complementary views, the desired output of CvMHAT is the *over-time* trajectory and the *cross-view* identification of each subject. Without loss of generality, we denote the input videos with the length of T as $\mathcal{T} = \{1, ..., T\}$ that are taken from multiple n time-varying views as $\mathcal{V} = \{v_1, ..., v_n\}$, CvMHAT problem aims to obtain the trajectory of each subject in all views along the video, i.e., the multi-view over-time subject association among all views and along each video. More specifically, the output of CvMHAT is the detected human bounding box $B^{v,t}$ in all frames $t \in \mathcal{T}$ of each view $v \in \mathcal{V}$, with the unified and unique ID number for each person.

3.2 CvMHAT Dataset

We build the CvMHAT dataset, which is composed of a real-world sub-dataset (**CvMHAT-R**) and a synthetic sub-dataset (**CvMHAT-S**). The former contains the videos taken by the drone and wearable cameras in the real-world scenes. The latter contains the synthetic videos generated by virtual 3D modeling. As shown in Table 1, we in detail compare our dataset with two representative single-view datasets (MOT 15,17) and several existing multi-view MOT datasets. For the data scale, the overall length of proposed dataset is 104 min with the frame rate of 30 fps. Different from the other datasets that collect long-time videos in fixed scenes, e.g., DukeMTMC contains 8 videos with the per duration of 85 min continuously taken by 8 cameras, CvMHAT contains

60 video groups (i.e., several continuous videos taken from multiple cameras at one site) and are taken in various outdoor scenes. Each video group contains 2–5 videos synchronously taken by multiple complementary-view cameras at the same multi-person scene.

3.2.1 Real-World Dataset

Data Collection To capture the complementary-view videos, we use a bird's eye view camera equipped on a flying Unmanned Aerial Vehicle (DJI Mavic 2) with a high altitude (e.g., 20–30 ms) to collect the top-view videos. We use the GoPro (HERO 8) mounting over the head of persons to take the horizontal-view videos. The output video has a resolution of 1920 × 1080 and a frame rate of 30 fps. As shown at the top of Fig. 2, the videos are taken at ten outdoor scenarios with different backgrounds, including the square, playground, park, garden, the entrance of teaching building and canteen, etc. In each scenario, there are 10–15 subjects who walk, sit or stand as they want, together with human-human interactions. We arrange some subjects to wear the GoPro cameras overhead for collecting the horizontal-view videos in which the mutual occlusions are common. The number of camera wearers in each scenario ranges in 1–4, and the recorded horizontal-view video(s) together with the corresponding top-view video make up a group of complementary-view videos.

Data Annotation and Statistics The top of Table 2 shows the statistics of the dataset according to the number of horizontal-view videos, i.e., 1 T(op) + X H(orizontal) views, including the amount of video groups and videos, and the number of frames and bounding boxes. We collect 30 groups of videos, in total 100 videos, in CvMHAT-R dataset. We manually synchronize these videos in each group such that corresponding frames of them are taken at the same time. The length of each video is from 300 to 1500 frames with the average length of 910.5. In total, CvMHAT-R dataset contains 91,050 frames.

The bounding boxes of all the subjects are annotated by outsourcing to a professional company. The subjects are annotated in the forms of rectangular bounding boxes and ID numbers: the same subject across different views in a video group are labeled with the same ID number. Following the previous works (Gan et al., 2021; Han et al., 2022a), we annotate one frame for each five frames, on each of which we manually annotate all the human bounding boxes and the corresponding over-time and cross-view unified IDs. For more accurate labels, we are not with the help of the auto-annotation system. All the annotations are double checked by the company and ourselves. Then we use the interpolation method to obtain the annotations for all the frames. The annotation generates 644,301 bounding boxes in total. Note that, with such a large scale, the annotation for the dataset is quite labor intensive given the difficulty in identifying persons in the top-view videos.

Table 2 Statistics of the CvMHAT-R and CvMHAT-S

	# Gro	# Vid	Len	# Frm	# Box
CvMHAT-R					
1 T + 1 H	10	20	760	15,200	111,483
1 T + 2 H	5	15	750	11,250	78,233
1 T + 3 H	10	40	740	29,600	213,359
1 T + 4 H	5	25	1440	36,000	241,226
Full	30	100	910.5	91,050	644,301
CvMHAT-S					
1 T + 1 H	10	20	1000	20,000	181,470
1 T + 2 H	5	15	1000	15,000	152,340
1 T + 3 H	10	40	1000	40,000	546,730
1 T + 4 H	5	25	1000	25,000	215,995
Full	30	100	1000	100,000	1,096,535

3.2.2 Synthetic Dataset

Data Collection Considering the high cost of the real-world data with burdensome data collection and annotation, we further build a synthetic video dataset CvMHAT-S by simulating the CvMHAT setting. It also has the advantage that we can control and record the accurate experimental settings during data capturing, e.g., the camera pose, human 3D location, etc., that are not easy to obtain in the real-world data. If the performance evaluation on synthetic data can reflect that on the real-world dataset, we can use the former for algorithm testing and evaluation, thus saving cost greatly. The effectiveness of the synthetic data for MOT has been verified by many previous datasets, e.g., Virtual KITTI (Fabbri et al., 2021), MOTSynth (Gaidon et al., 2016) and JTA (Fabbri et al., 2018) datasets, etc. To the best of our knowledge, this is the first synthetic dataset for *multi-view* MOT task.

We adopt the famous 3D modeling engine *Unity* to build the scenarios in CvMHAT-S dataset. We use the open source toolkit PersonX (Sun & Zheng, 2020) to generate the humans in our dataset. Similar to the real-world CvMHAT-R dataset, we set a top-view camera to look vertically down to the ground from a high altitude. We simultaneously set the cameras to be mounted to the head of several subjects in the scene as the horizontal-view camera wearers. The camera wearer may stand still, rotate his/her head (together with the camera) or walk freely in the scene. There are also 1 (top-view) + X (horizontal-view) cameras ($X = 1, 2, 3, 4$) with overlapped area coverage in CvMHAT-S dataset. As shown in the bottom of Fig. 2, we select six different scenarios, e.g., playground, running track, park, etc., and five groups of videos are generated in each scenario. The number of sub-

Fig. 2 Example video groups in the CvMHAT-R (top) and CvMHAT-S (bottom) datasets

jects in each scenario is set as 10–20 and all the subjects are controlled to walk freely in the scene. Note that, we do not require all the subjects to appear in all the views—we only set the FOV of the camera to cover most of the subjects. To ensure the reality of the synthetic dataset, we build the scene and human following their scales in the real world, including the human height, moving speed, camera coverage, etc. The altitude of the top-view is set as 15–20 ms in our dataset and the moving speed of the human is set to 1–1.5 ms per second. The synthesized video has a resolution of 1024 × 768 and a frame rate of 30 fps.

Data Annotation and Statistics In each video of CvMHAT-S dataset, the subjects are selected from 100 different subjects with individual ID number. Benefiting from the simulation environment, we can automatically obtain the bounding box and label of each subject without manual annotation. Specifically, the unique ID can be generated along with the subject and the same subject across all views in a video group is labeled with the same ID. We render one subject each time without the other subjects and background disturbance and apply a simple image cropping to get the bounding box, where occlusions in horizontal-view videos do not affect the correct bounding-box annotation. By repeating the above operation, we get the automatically generated annotations for all subjects including the bounding boxes and IDs for each frame in the video. Note that, the labels for synthetic dataset are more accurate compared to the manual annotations for real dataset. For example, it is hard to annotate the real bounding box of a subject mostly occluded by others on real-world image, which can be achieved in virtual environment.

As shown in the bottom of Table 2, CvMHAT-S dataset contains 30 video groups, in total 100 videos, the same as CvMHAT-R dataset. We set the video length as 1000 frames for all videos and in total generate 100,000 frames with over one million human bounding boxes.

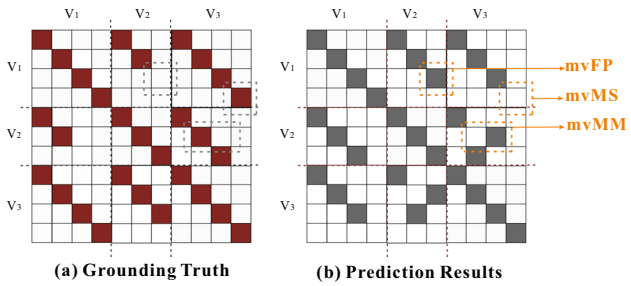


Fig. 3 Illustration of the multi-view association metrics, V_i indicates the view of i-th camera

3.3 Metrics for CvMHAT

Single-view Tracking Metric We take the widely used MOT metrics for evaluating the single-view tracking performance as in MOT Challenge (Leal-Taixé et al., 2015), including the multiple object tracking precision (MOTP) and multiple object tracking accuracy (MOTA) in CLEAR MOT Metrics (Bernardin & Stiefelhagen, 2008). A key task of MOT is to identify and track the same subject along the video, which is also very important in our problem. We also include the ID switches (IDS), fragment (FM) in Bernardin and Stiefelhagen (2008) and the ID-based metrics (Ristani et al., 2016): ID precision (IDP), recall (IDR), and F_1 measure (IDF_1).

Multi-view Association Metric In this work, besides temporal tracking, we also focus on the frame-by-frame multi-human cross-view association. We associate all the subjects

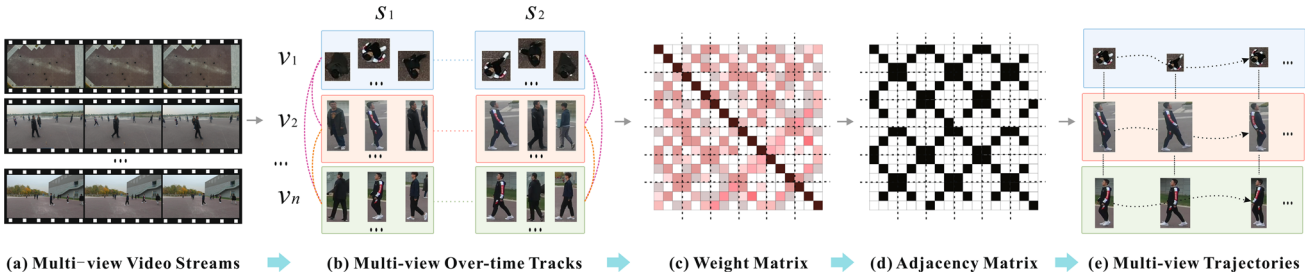


Fig. 4 Framework of the proposed baseline algorithm for CvMHAT. We first split the multi-view videos into shot clips, and in each clip we can obtain the single-view tracks. The main problem in this work is to associate the tracks from adjacent segments in all views. We formulate this task as a joint optimization problem, the solution of which can be used to form the final multi-view trajectories

appearing in different views (by giving the same ID) during the tracking, to obtain a multi-perspective observation of each subject. This way, we propose the new metrics for especially evaluating the multi-view association results, which are not generated and evaluated in the previous MOT problems. Specifically, we define new metrics multi-view ID precision (mvIDP), ID recall (mvIDR), ID F_1 score (mvIDF$_1$) and multi-object matching accuracy (mvMHAA). Specifically, mvIDP and mvIDR denote the multi-view subject association precision and recall, respectively. Given the subject IDs in all n views, we take two views each time and compute the pairwise subject matching performance. Based on mvIDP and mvIDR, the mvID F_1 is defined as

$$\text{mvIDF}_1 = \frac{2 \times \text{mvIDP} \times \text{mvIDR}}{\text{mvIDP} + \text{mvIDR}}. \tag{1}$$

We then define three metrics, i.e., mvMS, mvFP, mvMM to calculate the numbers of missed matches, false positives, and mismatch pairs for multi-view subject association. Specifically, for n views with in total N subjects, we first get ground-truth and predicted matching matrix with the dimension of $N \times N$ for all the views. An example is shown in Fig. 3. The metric mvMS counts the number of missed ground-truth matching pairs, the mvFP counts the number of falsely detected matching pairs, and the mvMM metric counts the number of mismatches. After that, we define multi-view multi-human association accuracy (mvMHAA) as

$$\text{mvMHAA} = 1 - \frac{\sum_t (\text{mvMS}_t + \text{mvFP}_t + 2\text{mvMM}_t)}{\sum_t N_t}, \tag{2}$$

where $\text{mvMS}_t, \text{mvFP}_t, \text{mvMM}_t$ are the missed matches, false positives, and mismatch pairs at time t, respectively. N_t is the total number of subjects for all views at time t.

4 The Proposed Baseline Method

Existing multi-camera MOT methods mainly study the over-time human tracking across the cameras but do not consider the frame-by-frame multi-human cross-view association. Also, existing methods mainly use the appearance features for association across the cameras with similar viewing angles. However, in our problem, we use top view, i.e., roughly vertical to the ground with a high altitude, as shown in Fig. 1c. This way, each subject is largely a small dark region and the appearance is not very useful for the human association between top and horizontal views. Therefore, existing methods can not directly handle the proposed problem.

For addressing the above problems, in this work, we formulate the CvMHAT as a classical generalized maximum (multi-) clique problem. In this formulation, we construct the spatial-temporal subject affinity matrix to build the subject correspondence both over time and across views, in which we apply the spatial reasoning for the affinity measurement between the top and horizontal views. We also apply the structural constraint conditions in our formulation, and the ADMM alike algorithm for efficient solution. The generated (multi) cliques form the spatial-temporal association relations in CvMHAT, which also consider the prior contained in the constraints. Overall, the proposed method is a simple and effective baseline given its appropriate input/output, structural constraints, and the reliable and efficient solution, which is presented in detail in the following.

4.1 Formulation

For the CvMHAT problem, the desired output is the *over-time* trajectories and the *cross-view* identification of each subject. To achieve this, we first track the subjects over time for generating the single-view tracks and we also associate them of the same subject across views to obtain the multi-view trajectories. In this work, we formulate the above task as a generalized maximum (multi) clique problem (Zamir et al., 2012).

Specifically, with the input videos that are taken from multiple moving cameras $\mathcal{V} = \{v_1, \ldots, v_n\}$. We first synchronously split these videos into short (temporal) segments with a fixed length, e.g., 10 frames, as shown in Fig. 4a, b. In each video clip at view v_n and segment s_i, we extract a set of *tracks* for each subject using a simple strategy based on spatial overlap, i.e., we connect two subjects between two adjacent frames if their intersection over union (IoU) of bounding boxes is larger than a threshold, e.g., a commonly used threshold 0.5 (Dehghan et al., 2015; Han et al., 2022a). All we need is then is to associate the tracks from different views and segments. Without loss of generality, we consider the multi-view over-time track association among all views, i.e., v_1, ..., v_n and between two adjacent segments, i.e., s_1, s_2, as shown in Fig. 4. We handle this task as a joint optimization problem. Specifically, we denote $x_i^{v,s}$ as a subject track in segment $s \in \{s_1, s_2\}$ of view $v \in \mathcal{V}$, where $i \in \mathcal{I}$ denotes the index, i.e., i-th track among all the tracks in all n views across two segments. We establish a graph $\mathcal{G} = \{\mathcal{N}, \mathbf{A}\}$ with nodes $\mathcal{N} = \{x_i^{v,s} | i \in \mathcal{I}\}$, and adjacency matrix $\mathbf{A} = \{a_{i,j} | i, j \in \mathcal{I}\}$. Then, the multi-view track association can be formulated as the following generalized maximum (multi) clique problem

$$\begin{aligned} \mathbf{A}^* &= \arg\max_{\mathbf{A}} \ \langle \mathbf{W}, \mathbf{A} \rangle \\ &= \arg\max_{a_{ij}} \ \sum_{i,j \in \mathcal{I}} w_{ij} a_{ij}, \\ \text{s.t.} \quad & \mathbf{A} \in \mathcal{S}, \end{aligned} \tag{3}$$

where $\langle \cdot \rangle$ denotes the matrix inner product operation. We use the adjacency matrix $\mathbf{A} = [a_{ij}]_{i,j} \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{I}|}$ to represent the association relation over graph $\mathcal{G}$ with each element $a_{i,j} \in \{0, 1\}$ representing the connectivity from node (with the index) i to node j, where $a_{i,j} = 1$ denotes that nodes i and j represent the same person in different views or segments. The weight matrix $\mathbf{W} = [w_{ij}]_{i,j} \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{I}|}$ is composed of $w_{i,j} \in [0, 1]$ representing the affinity score from node i to j. Here $\mathcal{S}$ denotes the internal constraint conditions that $\mathbf{A}$ should satisfy, which is considered as follows.

1) Symmetric-Similarity Constraint The adjacency matrix $\mathbf{A}$ should be a symmetric matrix, i.e.,

$$\begin{aligned} &\mathbf{A} = \mathbf{A}^{\mathrm{T}}, \\ \text{with} \quad &\mathbf{A}_{mm} = \mathbf{I}. \end{aligned} \tag{4}$$

This is not hard to get since if we have $a_{ij} = 1$, i.e., subject i and j denote the same person, we naturally obtain $a_{ji} = 1$ and vice versa. Specifically for the sub-matrix at diagonal block in $\mathbf{A}$, we have $\mathbf{A}_{mm} = \mathbf{I}$, as the identity matrix, which is shown in Fig. 4d.

2) Cycle-Consistency Constraint For a perfect association, the same person appearing in different clips should be connected as a cycle. Specifically, we denote $\mathcal{U}$ as the set of subjects in all $2n$ clips (in n views across two segments). For each two different clips m and n, we have $\mathbf{A}_{mn} = \mathbf{A}_m^u \, \mathbf{A}_u^n$, where $\mathbf{A}_m^u \in \mathbb{R}^{N_m \times |\mathcal{U}|}$ denotes the binary permutation matrix between the N_m subjects in clip m and the human set $\mathcal{U}$. For all clips, we concatenate their permutation matrices row by row and get $\mathbf{A}^u \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{U}|}$. Following previous works (Dong et al., 2019, 2021), the cyclic-consistency constraint requires that

$$\mathbf{A} = \mathbf{A}^u (\mathbf{A}^u)^{\mathrm{T}}, \tag{5}$$

which implies that all the concatenations among the subjects in different views form $|\mathcal{U}|$ cycles representing the $|\mathcal{U}|$ human identifications. Note, the number of nodes in each cycle is no more than $2n$.

3) Zero–One Constraint This is a mandatory constraint for assignment problem that a_{ij} should be a binary value, i.e.,

$$a_{ij} \in \{0, 1\}, \ 1 \leq i, j \leq |\mathcal{I}|. \tag{6}$$

4.2 Spatial-Temporal Affinity Measurement

We then consider the matrix $\mathbf{W}$ contains the affinity scores between two tracks.

• *Appearance similarity* To measure the similarity between two tracks, we extract the appearance feature using an efficient re-id network structure (He et al., 2020). Following Han et al. (2022a), by calculating the Euclidean distance of the output features, we obtain the appearance similarity between two over-time or cross-view tracks. The appearance feature is *only* used for the track-similarity measurement in horizontal views, but not across top and horizontal views. This is because the appearance in the top view and horizontal view has huge difference thus the appearance features are not useful here. We will show more details about this in the later experiments.

• *Motion similarity* We further apply the motion similarity for over-time track similarity measurement. We use the constant velocity motion model to predict motion similarity as in most previous MOT methods. Given two tracks from two adjacent clips in the single view, we first calculate the forward and backward motion propagation errors using the constant velocity motion model as in Zamir et al. (2012) and then generate the motion aware over-time track affinity.

• *Spatial reasoning* For the subject similarity between the top and horizontal views, both the appearance and motion are not very useful given the large view difference. We instead use a spatial distribution based method (Han et al., 2019) to address this problem, which represents each subject as a feature vector using their spatial layout and matches the vectors of all subjects in the horizontal-view and top view by a cost function for cross-view association. For a pair of

Table 3 Overview of the feature selection

Part	Appearance	Motion	Spatial
Cross-view top-hor	✗	✗	✓
Cross-view hor-hor	✓	✗	✗
Over-time top	✗	✓	✗
Over-time hor	✓	✓	✗

video clips from the top view and horizontal view, respectively, we first employ the spatial distribution based algorithm to get the subject association results of the subjects frame by frame. The result of cross-view subject association is to identify all the matched subjects between two views that indicate the same persons. With the frame-level subject matching results, we then calculate the spatial-reasoning-aware track similarity between the tracks from the clips in top and horizontal views. This is achieved through a voting strategy by the frames involved in the tracks with the frame-by-frame cross-view association (Han et al., 2019). Therefore, we obtain the similarity for two tracks across the complementary views.

We summarize where the appearance/motion/spatial features are used for similarity measurement in Table 3.

4.3 Optimization

Combining all the similarity measurements, we generate the weight matrix $\mathbf{W}$ as shown in Fig. 4c, where the weights for each pair of clips are normalized into [0,1]. We then discuss the solution of the constraint optimization problem in Eq. (3) to get the adjacency matrix $\mathbf{A}$. We first consider the constraints $\mathcal{S}$ in Eq. (3), in which Eqs. (4), (6) are the explicit constraints on $\mathbf{A}$, while Eq. (5) is implicit since $\mathbf{A}^u$ is unknown. We then provide the derivation of the constraint transformation.

First, Eq. (5) requires the matrix $\mathbf{A}$ can be factorized as $\mathbf{A}^u(\mathbf{A}^u)^T$. According to the theory of matrices (Gantmakher, 1959), from the above constraint we can get that $\mathbf{A}$ should satisfy that

$$\mathbf{A} \succeq 0,\ \text{rank}(\mathbf{A}) \leq |\mathcal{U}|, \tag{7}$$

where the former denotes $\mathbf{A}$ is a positive-semidefinite matrix and the latter requires its rank is less than $|\mathcal{U}|$, which counts the number of unique persons in the scene but is unknown in practice. Inspired by Han et al. (2022b), we further transform this constraint by using the nuclear norm $\|\mathbf{A}\|_*$.

Specifically, we denote the singular value of $\mathbf{A}$ as $g_q, q = \{1, 2, ..., Q\}$. From the basic properties of matrix we get

$$\begin{aligned} \|\mathbf{A}\|_* &= \sum_q g_q = \|\mathbf{g}\|_1, \\ \text{rank}(\mathbf{A}) &= \text{count}(g_q \neq 0) = \|\mathbf{g}\|_0, \end{aligned} \tag{8}$$

where $\mathbf{g}$ is the vector of singular values and the rank of $\mathbf{A}$ can be computed by the count of nonzero elements in $\mathbf{g}$. ❶ We denote e_q as the eigenvalue of $\mathbf{A}$, and have $g_q = |e_q| \geq e_q$ since $\mathbf{A}$ is a real symmetric matrix, where the symmetry is constrained by Eq. (4). Then we can get $\|\mathbf{A}\|_* = \sum_q g_q \geq \sum_q e_q$. This way, minimizing the nuclear norm $\|\mathbf{A}\|_*$ will push g_q, i.e., $|e_q|$ and e_q to close to each other, which is equivalent to $e_q \geq 0$, i.e., $\mathbf{A}$ is positive semidefinite in Eq. (7). ❷ For the second low-rank constraint in Eq. (7), from Eq. (8) we get it is equivalent to minimize the l_0 norm of $\mathbf{g}$. We know that the l_1 norm is commonly used as the optimal convex approximation of l_0 norm. This way, minimizing $\|\mathbf{A}\|_*$, i.e., the l_1 norm of $\mathbf{g}$ according to Eq. (8), also compels the low-rank constraint.

This way, from Eq. (7), we can get the following optimization problem

$$\begin{aligned} \mathbf{A}^* = \arg\min_{\mathbf{A}} \ & -\langle \mathbf{W}, \mathbf{A} \rangle + \lambda \|\mathbf{A}\|_* \\ \text{s.t.} \ & \mathbf{A} \in \bar{\mathcal{S}}, \end{aligned} \tag{9}$$

where λ is a pre-set parameter and $\bar{\mathcal{S}}$ denotes the constraints in Eqs. (4) (6), which we will consider latter. To solve the optimization problem with nuclear norm term, we employ the Augmented Lagrangian Method (ALM) algorithm (Boyd et al., 2011). We first construct the Augmented Lagrange formulation as

$$\begin{aligned} L_\rho(\mathbf{A}, \mathbf{B}, \mathbf{H}) = & -\langle \mathbf{W}, \mathbf{A} \rangle + \lambda \|\mathbf{B}\|_* \\ & + \langle \mathbf{H}, \mathbf{A} - \mathbf{B} \rangle + \frac{\rho}{2} \|\mathbf{A} - \mathbf{B}\|_2^2 \\ \text{s.t.} \ & \mathbf{A} \in \bar{\mathcal{S}}, \end{aligned} \tag{10}$$

where we introduce an auxiliary variable $\mathbf{B}$ requiring $\mathbf{A} = \mathbf{B}$, and $\mathbf{H}$ and $\rho > 0$ are the Lagrange multiplier and penalty factor, respectively. The problem can be solved iteratively using the ADMM (Boyd et al., 2011) technique. Each variable has closed form solution by optimizing the following three sub-problems

$$\begin{cases} \mathbf{A}^{(k+1)} = \arg\min_{\mathbf{A}} L_\rho(\mathbf{A}^{(k)}, \mathbf{B}^{(k)}, \mathbf{H}^{(k)}) \\ \mathbf{B}^{(k+1)} = \arg\min_{\mathbf{B}} L_\rho(\mathbf{A}^{(k+1)}, \mathbf{B}^{(k)}, \mathbf{H}^{(k)}) \\ \mathbf{H}^{(k+1)} = \mathbf{H}^{(k)} + \rho(\mathbf{A}^{(k+1)} - \mathbf{B}^{(k+1)}) \end{cases}. \tag{11}$$

Sub-problem for $\mathbf{A}$: We first extract the terms involving the variable $\mathbf{A}$ in Eq. (10) and get

$$L_\rho(\mathbf{A}) = -\langle \mathbf{W}, \mathbf{A} \rangle + \langle \mathbf{H}, \mathbf{A} - \mathbf{B} \rangle + \frac{\rho}{2} \|\mathbf{A} - \mathbf{B}\|_2^2. \tag{12}$$

By solving $\frac{\partial L_\rho(\mathbf{A})}{\partial \mathbf{A}} = 0$, the optimal solution for $\mathbf{A}$ can be obtained by

$$\mathbf{A}^* = \frac{1}{\rho}(\mathbf{W} - \mathbf{H}) + \mathbf{B}. \tag{13}$$

Sub-problem for $\mathbf{B}$: Similarly, we extract the terms involving $\mathbf{B}$ and get

$$L_\rho(\mathbf{B}) = \lambda\|\mathbf{B}\|_* + \langle \mathbf{H}, \mathbf{A} - \mathbf{B}\rangle + \frac{\rho}{2}\|\mathbf{A} - \mathbf{B}\|_2^2, \tag{14}$$

from which we can get the following equation by the formula deformation and irrelevant item removal

$$L'_\rho(\mathbf{B}) = \frac{\lambda}{\rho}\|\mathbf{B}\|_* + \frac{1}{2}\|\mathbf{B} - (\mathbf{A} + \frac{1}{\rho}\mathbf{H})\|_2^2, \tag{15}$$

where we denote $\Pi \triangleq \mathbf{A} + \frac{1}{\rho}\mathbf{H}$.

Following the Singular Value Thresholding (SVT) Algorithm in Cai et al. (2018), the optimal solution for $\mathbf{B}$ can be obtained by

$$\mathbf{B}^* = \mathcal{D}_\tau(\Pi) = \mathbf{U}\mathcal{D}_\tau(\Sigma)\mathbf{V}^{\mathrm{T}}, \tag{16}$$

where $\mathcal{D}_\tau(\Sigma) = \mathrm{diag}(\{\sigma_h - \tau\}_+)$, $\tau = \frac{\lambda}{\rho}$, $\{\cdot\}_+$ denotes to maintain the positive values and make others as 0. Here $\mathcal{D}_\tau(\Pi)$ in the SVT algorithm can be taken as applying a soft-threshold rule on Σ obtained by singular value decomposition (SVD) of Π as

$$\Pi = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}}, \quad \Sigma = \mathrm{diag}(\{\sigma_h\}_{1\le h\le H}), \tag{17}$$

where $\mathbf{U}$ and $\mathbf{V}$ denote the left and right singular vectors respectively, σ_h is the singular value. More details about the above deduction can be found in Cai et al. (2018).

Sub-problem for $\mathbf{H}$: We finally solve the sub-problem for $\mathbf{H}$ as

$$\mathbf{H}^* = \mathbf{H} + \rho(\mathbf{A} - \mathbf{B}). \tag{18}$$

4.4 The Framework

We present the overall algorithm of the proposed method in Algorithm 1, where we iteratively optimize the energy function with ADMM by K iterations, e.g., 75 iterations, for convergence, and use a threshold β (empirically set as 0.7) to obtain the binary assignment matrix. The ratio parameters of appearance and motion similarities in over-time horizontal-view association are 0.3 and 0.7, respectively. In the next experiment section, we will justify the usefulness of this baseline method and the high challenge of the proposed CvMHAT problem and dataset. Benefiting from the complementary-view videos, we apply the spatial-temporal collaborative subject association strategy for MHAT. This makes use of the information from different views (especially the top views) while obtaining trajectories for each view, for handling occlusion in horizontal view. Specifically, as shown in Algorithm 1, we show how to assign the ID for the track T_v^{t+1} (top view) and H_p^{t+1} (horizontal view) at segment $t+1$. For the temporal association, i.e., the single view tracking, given the association matrix, if its element $\bar{a}_{uv} = 1$, we associate tracks T_u^t and T_v^{t+1} by transferring the ID. We also consider the spatial (cross-view) association for the tracks without temporal association. At segment $t+1$, if $\bar{a}_{vq} = 1$, we assign the ID of the top-view track T_v^{t+1} to the horizontal-view track T_q^{t+1}. Let's explain it by an example. In a horizontal view v_1, we assume a person P firstly appears at time t_1, then disappears at t_2, and re-appears at t_3. In this case, at t_1, we use Algorithm 1 to assign a new ID $\#D$ to P. At t_2, there is no track for $\#D$, i.e., the subject P is not in view v_1, but in the top view the subject P continuously appears. At t_3, we can use the complementary-view subject association results to renewedly track P in view v_1, by assigning

Algorithm 1: Complementary-view MHAT:

Input: Complementary-view videos from multiple moving cameras.
Output: Subject trajectories with cross-view unified ID numbers.

Synchronously split the n videos, one from top view and $n-1$ from horizontal views, into S segments.
for $s = 1 : S$ **do**
  Implement human bounding box detection.
  Get the single-clip short tracks in each view and segments s and $s+1$.
  Calculte the track similarity score as in Sec. 4.2, and compute the weight matrix $\mathbf{W}$ in Eq. (3).
  // Solve $\mathbf{A}$ in Eq. (3) using ADMM algorithm.
  Construct the Augmented Lagrange formulation as Eq. (10)
  for $k = 1 : K$ **do**
    $\mathbf{A} = \frac{1}{\rho}(\mathbf{W} - \mathbf{H}) + \mathbf{B}$ in Eq. (13)
    $\mathbf{B} = \mathbf{U}\mathcal{D}_\tau(\Sigma)\mathbf{V}^*$ in Eq. (16)
    $\mathbf{H} = \mathbf{H} + \rho(\mathbf{A} - \mathbf{B})$ in Eq. (18)
    // To sitisfy the symmetry constraints in $\bar{\mathcal{S}}$
    $\mathbf{A} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^{\mathrm{T}})$
  // To obtain the zero-one association matrix
  Do binarization on $\mathbf{A}$ with a threshold β and get $\bar{\mathbf{A}}$.
  Get cross-view cross-segment middle trajectories from $\bar{\mathbf{A}}$.
  // Spatial-temporal subject association strategy **if** $\bar{a}_{uv} = 1$ ($\bar{a}_{pq} = 1$) **then**
    Assign the ID number of T_u^t (H_p^t) to T_v^{t+1} (H_q^{t+1}), respectively.
  else if $\bar{a}_{vq} = 1$ **then**
    Assign the ID number of T_v^{t+1} to H_q^{t+1} .
  else
    Assign the incremental ID to other subjects.
  Stitch the middle trajectories along the whole video.
return Multi-view long trajectories with unified IDs.

the ID #D from the top view to the re-appeared subject P. For the traditional MOT, it is hard to continuously track P if it disappears for a long time—P is usually assigned with a new ID when it re-appears.

5 Experimental Results

5.1 Setup

Comparison Methods We evaluate some methods on the CvMHAT dataset to verify the usefulness of the dataset. Actually, given the large view difference between the top and horizontal views and the multi-overlapped-view MOT setting, we did not find existing methods with code that can directly handle the proposed CvMHAT problem. Therefore, we select several famous single-view MOT methods for comparison.. The first category is separate detection and embedding (SDE) paradigm, i.e., MDP (Xiang et al., 2015), DMAN (Zhu et al., 2018), Tacktor++ (Bergmann et al., 2019b), and StrongSort++ (Du et al., 2022). Among them, MDP and DMAN are single object tracking (SOT) based MOT trackers, where MDP uses a Markov decision making process for data association and DMAN further learns appearance features for similarity measurement with a well-designed deep neural network. Tacktor++ achieves the tracking by using an object detector framework with the regression and classification branches. StrongSort++ is a state-of-the-art tracker based on the DeepSort framework. The second category is joint detection and tracking (JDT) paradigm. We select three famous trackers, i.e., CenterTrack (Zhou et al., 2020), Trackformer (Meinhardt et al., 2022), ByteTrack (Zhang et al., 2022) for comparison. The above methods aim to jointly achieve the detection and tracking tasks. Note, all the comparison methods are implemented on the single-view videos separately.

We also include two multi-view multi-human tracking methods BIPCC (Ristani et al., 2016) and DeepCC (Ristani & Tomasi, 2018) for comparison, which handle the human tracking using multiple cameras covering different areas. Besides, we include the method MHT (Han et al., 2020a) for comparison, which is only used for *two-view* video pairs thus cannot be directly applied to our dataset with multiple horizontal views. This way, we divide our dataset into top-horizontal-view video pairs and evaluate this method on each pair.

Experimental Details Following the previous works (Han et al., 2020a), for fair comparison, we use the same subject detection results, generated by the famous YOLOv3 (Redmon et al., 2016), for both the proposed method and the comparison methods. We have also tried the recent version YOLOX (Ge et al., 2021) as the detector in our method, and report the corresponding human association and tracking results in the following section. For top-view subject detection, we fine-tune the pre-trained network using extra 800 top-view images that are not in our CvMHAT dataset. In the ADMM algorithm, we set the parameters λ and ρ in Eq. (10) as 60 and 70, respectively. We fix the parameters for all experiments without fine-tuning them on each dataset. We will also show the stability analysis of parameters in the later experiment section.

5.2 Results on CvMHAT Dataset

We first evaluate the single-view MOT performance (over all views) using the standard MOT metrics on the real-world CvMHAT-R dataset, as shown in the top-left block of Table 4. We can see that the proposed CvMHAT baseline method outperforms the comparison methods, i.e., MDP, DMAN and Tracktor++, and gets the comparable results with the state-of-the-art MOT method StrongSort++, on the ID related metrics, i.e., IDP, IDR and IDF_1. This can be explained from that the multi-view joint optimization provides more constraints to the temporal ID consistency. The proposed method also achieves the comparable performance on MOTA with the state-of-the-art comparison trackers. It should be added that MDP outperforms Tracktor++ in this dataset. This may be due to the following two reasons. First, MDP is a classical Markov Decision Process based method, which do not use the deep appearance features. Tracktor++ is a deep learning based method using the pre-trained appearance model. However, the cross-domain gap between its training set and the proposed dataset may make the model not very robust. Note that, DMAN is based on MDP by integrating new deep features. The slightly lower IDF_1 score of DMAN than MDP can also verify the above point. Second, Tracktor++ uses the framework with joint detection and tracking. For fair comparison, we directly provide the public human detection without the correction in Tracktor++. This may also decrease the tracking performance.

Besides, we further evaluate the multi-view MOT performance using the proposed multi-view association metrics as shown in the top-right block of Table 4. For the comparison methods that only handle the single-view tracking, we additionally provide the ground-truth ID matches of the subjects among all views when they first appear in each view. 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, even with such additional information, all the single-view tracking methods, e.g., MDP, DMAN and Tracktor++, including the state-of-the-art MOT method StrongSort++, produce poor performance for the multi-view association task. This is because the cross-view subject matching fails once a subject is lost in one view. Previous multi-view multi-human tracking algorithms, e.g., BIPCC and DeepCC, also produce

Table 4 Comparative results of different methods on CvMHAT-R dataset

Box	Method	IDP	IDR	IDF_1	IDS	FM	MOTP	MOTA	mvIDP	mvIDR	$mvIDF_1$	mvMHAA
Detector	MDP	61.4	62.1	61.7	16,838	8356	77.5	68.1	43.3	23.1	30.2	39.1
	DMAN	60.5	62.5	61.5	12,013	9174	77.3	68.6	53.3	22.8	31.9	39.5
	Tracktor++	53.1	51.2	52.1	21,709	10,001	76.4	69.5	32.4	13.9	19.5	32.9
	StrongSort++	61.2	**67.2**	**64.1**	**1769**	8340	67.3	67.3	56.1	34.0	42.3	43.9
	BIPCC	34.9	33.6	34.2	9533	9360	78.1	**71.2**	9.1	3.4	5.0	25.4
	DeepCC	29.2	21.8	25.0	9092	**8250**	80.0	55.7	8.0	1.5	2.5	25.4
	MHT	54.7	54.1	53.5	4239	20,141	71.7	49.3	58.6	47.8	52.7	51.7
	Ours	**65.4**	62.6	64.0	3875	9797	78.1	70.7	**72.7**	**62.7**	**67.3**	**73.2**
	Ours-X	64.6	60.1	62.3	3530	9495	**81.2**	70.9	72.5	61.2	66.3	**73.2**
Annotation	MDP	73.3	71.8	72.5	24,867	4604	91.0	90.9	48.2	30.1	37.1	46.3
	DMAN	66.1	66.6	66.4	17,695	4276	93.9	89.7	51.6	27.3	35.7	41.0
	Tracktor++	62.4	59.9	61.1	24,285	5233	76.7	84.0	35.7	16.1	22.2	33.9
	StrongSort++	77.7	**80.6**	79.1	**1350**	3394	**99.0**	94.7	63.5	46.4	53.6	57.4
	BIPCC	52.6	50.5	51.5	4874	3900	97.6	**95.3**	10.2	5.0	6.7	29.0
	DeepCC	41.3	33.7	28.5	4315	**3792**	98.2	68.3	10.0	2.8	4.4	27.8
	MHT	74.5	70.7	72.6	2699	4401	97.2	87.7	75.0	68.0	71.3	74.8
	Ours	**81.5**	78.2	**79.8**	2499	4413	97.6	94.1	**75.2**	**72.8**	**74.0**	**85.7**

Bold values indicate the best performance for each metric
IDP↑ (%), IDR↑ (%), IDF_1↑ (%), IDS↓, FM↓, MOTP↑ (%), MOTA↑ (%) are standard MOT metrics. mvIDP↑ (%), mvIDR↑ (%), $mvIDF_1$↑ (%), and mvMHAA↑ (%) are the new metrics for evaluating the cross-view MOT

poor association results on our benchmark. This is because that these works aim to handle the *long-term* human tracking problem using multiple cameras covering different areas, which is different from our setting of the *multi-perspective* human association and tracking in the same scene using multiple complementary-view cameras. In our problem, we use a global top view, i.e., roughly vertical to the ground from a high altitude, that is different from most previous works using only horizontal views or slope-angled views, e.g., those in DuckMTMC (Ristani et al., 2016). Our top-view camera only captures the top of each subject's head and shoulders from high altitude, as shown in Fig. 1c. This way, *appearance is not very useful for the human association between top and horizontal views*. In addition, without the camera calibration, the motion feature is also inconsistent across two kinds of views. Since such methods (Ristani et al., 2016; Ristani & Tomasi, 2018) rely solely on appearance and motion features for cross-camera human association, they can not handle our cross-top-horizontal-view subject association in the proposed CvMHAT problem. The proposed method achieves an acceptable association performance by considering both the over-time and cross-view subject association. For our method with different detectors, we can see that YOLOX has a more accurate detection results, thus the corresponding 'Ours-X' providing a higher 'MOTP' and 'MOTA' scores. Although there is performance difference, the gap is not large. This also demonstrates that the proposed method is not very sensitive to the detection algorithms.

Further, to eliminate the effects brought by the false detections, we use the annotated bounding boxes as detections for tracking and the result is shown in the bottom of Table 4. We can still observe better performance on the ID related metrics from the proposed method, including the IDF_1, $mvIDF_1$ scores. Note that, the MOTP and MOTA scores are generally high for all methods because of the provided ground-truth bounding boxes. The comparison method MHT performs well with the bounding boxes of annotation but not well enough when using the detector, which demonstrate that this method is sensitive to the accuracy of detection.

We also compare with a series of most recent MOT methods, which are based on the joint detection and tracking (JDT) paradigm. We select three famous trackers, i.e., CenterTrack, Trackformer, ByteTrack for comparison. Note that, the above methods aim to jointly achieve the detection and tracking tasks, which mainly use the self-generated human bounding boxes in their framework. Even we providing the input detection, they also filter the boxes by their own detection results. This way, we directly provide the videos without detection for these methods. However, these methods can not handle the top-view videos in our dataset given the domain gap. To compare with them as much as possible, we evaluate the results in the horizontal-view videos in our dataset, as shown in Table 5. We can see that Bytetrack produces the overall good performance on all metrics because of the algorithm robustness. Also, Bytetrack provides the best IDS and FM scores and Trackformer provides the best MOTP score with

Table 5 Comparative results of different methods on CvMHAT-R (horizontal view) dataset

Method	IDP	IDR	IDF_1	IDS	FM	MOTP	MOTA	mvIDP	mvIDR	$mvIDF_1$	mvMHAA
CenterTrack	37.7	38.5	38.1	4598	7086	81.9	69.2	31.9	10.2	15.4	34.0
Trackformer	37.9	42.0	39.9	2363	4620	**84.1**	69.4	29.4	13.3	18.3	33.9
ByteTrack	55.4	**66.1**	60.3	**1209**	**4607**	82.0	69.1	55.4	66.1	60.3	46.3
Ours	**66.7**	64.5	**65.6**	3147	5459	82.8	**75.9**	**80.7**	**69.5**	**74.6**	**67.8**

Bold values indicate the best performance for each metric

Table 6 Comparative results of different methods on CvMHAT-S dataset

Method	IDP	IDR	IDF_1	IDS	FM	MOTP	MOTA	mvIDP	mvIDR	$mvIDF_1$	mvMHAA
MDP	61.5	47.0	53.3	35,629	16,072	88.6	70.2	32.4	14.9	20.4	29.2
DMAN	59.8	49.3	54.1	26,045	15,685	87.3	69.1	42.6	12.6	19.5	27.8
Tracktor++	48.0	35.4	40.8	30,661	25,577	73.4	62.2	29.1	9.1	13.9	26.0
StrongSort++	69.5	76.0	72.6	**1874**	18,324	**97.1**	88.6	62.4	38.0	47.2	45.8
MHT	66.8	50.6	57.6	7336	**8217**	91.1	74.8	67.0	33.0	44.2	39.7
Ours	**89.3**	**83.8**	**86.5**	54,409	11,815	93.0	**93.3**	**87.0**	**77.6**	**82.0**	**83.2**

Bold values indicate the best performance for each metric

the precise detection results. For other metrics, the proposed method outperforms the comparison methods, and especially with a large margin on the multi-view association task.

Table 6 shows the performance evaluation on our synthetic CvMHAT-S dataset, on which we use the automatically generated human bounding boxes within the dataset. We do not include the results of BIPCC and DeepCC (Ristani et al., 2016; Ristani & Tomasi, 2018) on CvMHAT-S, because the videos in which containing larger number of subjects make the solving of the Binary Integer Program in Ristani et al. (2016); Ristani and Tomasi (2018) difficult to implement. We can see that our method outperforms other algorithms obviously. From above results on both CvMHAT-R and CvMHAT-S datasets, we can see that the proposed benchmark is challenging and has the potential for further development. We can also see that synthetic data is also challenging, which can reflect the performance on real data to some extent.

5.3 Experimental Analysis

5.3.1 Subset Evaluation

As shown in Table 7, we evaluate the single-view tracking performance on the top- and horizontal-view videos, respectively. Experiments are implemented on CvMHAT-R dataset with the human bounding boxes generated by the detector. We can first see that, for the previous works, the overall performance in top view is better than that in horizontal views. This is mainly caused by frequent mutual occlusions in the horizontal views. We can further see that, in top view, the overall performance of our method is worse than the comparison methods, without some well-designed tracking techniques and tricks in previous works. On the contrary, the proposed method achieves better tracking performance, e.g., IDP, IDR, and IDF_1, in horizontal view using the joint optimization with the top view. This shows that the top view can provide complementary information for improving the tracking in the horizontal views, which verifies *the importance to study the problem of applying multiple and complementary views for collaborative tracking*.

5.3.2 Ablation Study

We conduct the ablation study of the proposed method. First, we discuss about the track extracting in each segment. Specifically, the segment is very short (i.e., 10 frames in our experiment), in which we use the bounding box intersection over union (IOU) to extract the track. We evaluate the performance of the track generation result using precision, recall and F_1 score, which is shown in the Table 8. A track is taken as a true positive if and only if it consistently connects the bounding boxes of the same person across a segment. We can see that the performance of single-view track generation is good enough. However, the performance of the overall tracking, as shown in Table 4, is relatively inferior. This is mainly because the tracking scenarios in our dataset is not easy with frequent occlusions, especially for the horizontal views, as shown in Fig. 2. Given the irregular movement and frequent occlusions of the subjects, the key challenge

Table 7 Comparative results of different methods on the subsets of top-view videos and horizontal-view videos, respectively

Method	Top view							Horizontal view						
	IDP	IDR	IDF_1	IDS	FM	MOTP	MOTA	IDP	IDR	IDF_1	IDS	FM	MOTP	MOTA
MDP	**77.9**	75.5	**76.7**	2011	3776	68.8	64.8	52.1	54.1	53.1	14,827	4580	82.3	70.1
DMAN	77.2	**75.8**	76.5	844	4163	68.5	**66.2**	51.3	54.6	52.9	11,169	5011	82.1	70.1
Tracktor++	73.7	70.3	72.0	2606	5439	63.4	61.5	41.0	39.7	40.3	19,103	**4562**	**83.4**	74.4
StrongSort++	76.1	**75.8**	76.0	**397**	**3592**	69.0	64.5	53.5	62.0	57.4	1372	4748	82.9	69.0
Ours	63.2	59.5	61.3	728	4338	**69.3**	62.1	**66.7**	**64.5**	**65.6**	**3147**	5459	82.8	**75.9**

Bold values indicate the best performance for each metric

is continuously to track the subject for a long time, i.e., to associate the track over time and across views in our framework. For this purpose, the most common scene is to associate the track between two adjacent segments, which is handled by the association matrix. Another scene is to associate the track between the nonadjacent segments (long-term occlusion). We addressed this problem through a spatial-temporal subject association strategy in Algorithm 1.

We then discuss the usefulness of the feature and solution in our method. We remove the appearance and motion features in single-view subject similarity measurement, respectively. The results are shown in Table 9. We can see that, although with some performance decline, the proposed method is robust. This is benefited from the collaborative framework using multiple views, which is not highly depended on the sophisticated deep features. We can not remove the spatial aware feature because it is necessary in our framework, which will be discussed in Sect. 5.3.3. To verify the proposed problem formulation and optimization method, we also compare our method with a clustering based method used in Ristani and Tomasi (2018); Ng et al. (2002). Specifically, given the weight matrix $\mathbf{W}$, we optimize Eq. (3) by the spectral clustering algorithm, which makes positive correlation within the same cluster and the negative one among different clusters. If a clustering group contains more than two subjects from the same view, we preserve only one with the highest (average) similarity with other subjects in this group. We then construct the adjacency matrix $\mathbf{A}$ according to the clustering. Similarly, we further use a self-tuning spectral clustering (Zelnik-Manor & Perona, 2004) that can automatically estimate the number of groups for comparison. We can see that the performance of these methods is not good enough, since they can not include the constraints of the proposed formulation in Eq. (3). The comparative results demonstrate the effectiveness of the problem formulation and optimization in our method.

Table 8 Single-view track extraction results using different features (%)

Metric	Horizontal view			Top view		
	P	R	F_1	P	R	F_1
Track	93.22	93.13	93.18	99.47	98.33	98.90

Table 9 Comparative study of our method. 'App.' and 'Motion' denote the appearance and motion features in single-view subject similarity measurement

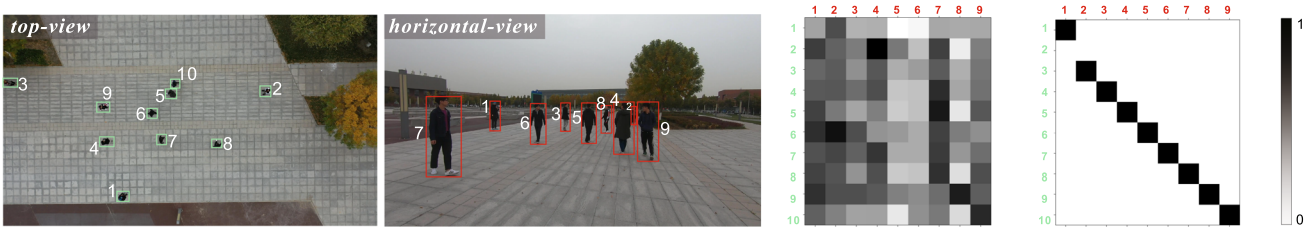
Methods	IDF_1	MOTA	$mvIDF_1$	mvMHAA
w/o App	63.7	70.7	67.1	72.8
w/o Motion	63.3	70.5	66.5	72.6
w Cluster	57.0	56.4	54.8	55.2
w Self-tune	51.8	68.6	65.2	69.0
Ours	**64.0**	**70.7**	**67.3**	**73.2**

Bold values indicate the best performance for each metric
'Cluster.' and 'Self-tune.' denote that we use a clustering or Self-tuning clustering algorithm for optimization (%)

5.3.3 Feature Usefulness Analysis

We clarify that the existing methods relying solely on *appearance and motion features can not handle our cross-top-horizontal-view subject association* in the proposed CvMHAT problem. The key challenge lies in the limited feature representation ability, for which we show some visualized results in the following.

Appearance Feature Most tracking methods use appearance features for association. In our problem, we use top view, i.e., roughly vertical to the ground with a high altitude, which is different from previous works that use cameras with similar slope-angled views, e.g., those in DuckMTMC (Ristani et al., 2016). Note, in top view, each subject is largely a small dark region, as shown following Fig. 5a, in which *appearance is not useful for human association between top and horizontal views*. Specifically, we use the state-of-the-art human re-identification method to extract the appearance feature of the subjects in top and horizontal views, with which



(a) *Top-view image* (b) *Horizontal-view image* (c) *Similarity matrix* (d) *GT matching matrix*

Fig. 5 A simple of the complementary-view images (**a**, **b**). Illustration of the similarity matrix and matching matrix (**c**, **d**)

(a) *Top view* (b) *Horizontal view* 1 (c) *Horizontal view* 2 (d) *Horizontal view* 3

Fig. 6 Illustration of the optical flow (generated by FlowNet2.0) in two views

we compute the similarity matrix. As shown in the following Fig. 5c, we compute the appearance similarity matrix between all the subjects in top view and those in horizontal view, and show the corresponding ground-truth matching results in Fig. 5d, where the darker denotes more similar. By comparing these two matrices, we can see that the appearance feature used in previous methods is inapplicable for the cross-complementary-view subject association.

Motion Feature The *motion features are also actually not suitable for the cross-complementary-view association* for two reasons: (1) with roughly 90° cross-view difference while without the camera calibration, the extracted subject motion features, like optical flow, from top and horizontal views, are not much related, e.g., *subject1* in different views as shown in Fig. 6; (2) the unconstrained camera motions (caused by the camera wearers' random movement) reflected in top and horizontal views are unmatched, which makes it more difficult to match the motion features in different views, e.g., *horizontal viewer1*∼3 in Fig. 6.

Differently, in the proposed baseline method, the appearance and motion features are used for the *single-view over-time* but not for the cross-top-horizontal-view subject association, for which we apply the spatial reasoning as discussed above.

5.3.4 Qualitative Analysis

We show the illustration of the complementary-view multi-human association and tracking results in Fig. 7. For clarity, we show the tracking trajectory in the top view and the multi-view subject IDs in all views. We can see that the temporal trajectory of each subject can be distinctly recorded in the top-view video, while the multiple horizontal-view cameras observe the subjects from all-around perspectives with the local details. This also demonstrates *the potential of the proposed CvMHAT for many applications, e.g., the outdoor video surveillance*, which aims to capture both global distribution and local appearance details of the involved people.

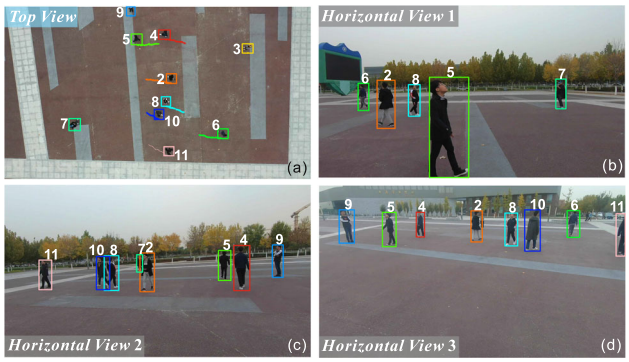


Fig. 7 Illustration of the CvMHAT results

5.3.5 Algorithm Speed Analysis

In this section, we further analyze the time consuming of our baseline method. We also compare the algorithm speeds of the proposed and other comparative methods.

As shown in Table 10, we record the running time of each component in the proposed baseline method. In this table, 'track' denotes the track generation, 'app.', 'motion' and 'spatial' denote the data similarity computation using different features, and 'optimize' denotes the solving of the optimization problem. We can see that the computation time is mainly taken by the track generation and appearance-based similarity measurement. The final optimization only takes

Table 10 Running time (second) of each component in our method

Part	Track	App	Motion	Spatial	Optimize
Time	0.125	0.087	0.011	0.012	0.002
Ratio (%)	52.7	36.7	4.6	5.1	0.9

Table 11 Speed comparison of different methods (fps)

Tracker	MDP	DMAN	Tracktor++	MHT	Ours
Speed	1.08	1.94	1.19	1.74	4.20

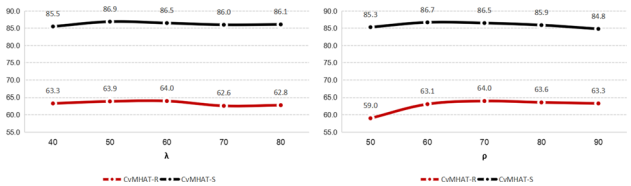


Fig. 8 Illustration of the performance (IDF_1 score) variation tendencies of our method with different parameters λ and ρ on CvMHAT-R and CvMHAT-S datasets

0.9% of total time. This denotes the efficiency of the proposed problem formulation and optimization.

We also compare the overall speed of our method with the comparison MOT trackers in Table 11. We can find that our method runs faster than all the comparison methods with a speedup of 2–4 times. Note that, our main program framework, except the neural network for appearance extraction, runs on a desktop computer using CPU only, which can be faster with the GPU acceleration.

5.3.6 Usefulness of the Synthetic Dataset

1) *Parameter Selection for Our Model* To verify the usefulness of the proposed synthetic dataset as far as possible, we provide the experimental analysis for parameter selection in our model. Specifically, in the real-world application of our model, we can not obtain the ground-truth results to fine-tune the parameters in our model for the best performance. But the ground-truth results in synthetic data are easy to obtain. This way, if the parameters selected on the synthetic data perform well enough, the synthetic data is useful for our model. This way, we select different parameter settings and test the corresponding performance on the synthetic and real-world data, respectively. As shown in Fig. 8, we investigate the performance of our method by changing the parameters λ and ρ in Eq. (10), on the real-world CvMHAT-R and synthetic CvMHAT-S datasets, respectively. We can see that the variation tendencies of the performance on two datasets are basically consistent. Note that, the impact of the parameter adjustment to the proposed method is not very significant. This is because that the parameter variation range here is not very large and the proposed method is not very sensitive to the parameter selection. The above experiments, to some extent, verify that the synthetic dataset can help our method for selecting the appropriate parameters to be applied to real-world data.

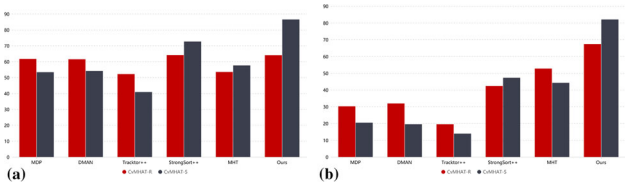


Fig. 9 IDF_1 score (**a**) and $mvIDF_1$ score (**b**) for all the compared methods on CvMHAT-S and CvMHAT-R datasets

2) *Pre-training Data for Deep Learning Algorithms* We also conduct the experiments to show the effectiveness of using synthetic data to obtain better performance on real data. We therefore select a deep learning method to train the model with and without the synthetic data and test it on real data to verify how much the synthetic data helps. Specifically, we selective a self-supervised learning method (Gan et al., 2021), which can be used for the multi-view multi-human association and tracking. However, this method only focuses on the first-person-view videos. This way, we select the multi-horizontal-view videos in our dataset to conduct the experiments. Specifically, we first directly test the model on the real-world dataset, whose results are shown in the first row in Table 12. We then re-train the model using our synthetic data and show the results in the second row. We can see that the overall results are not very good, since this is a self-supervised method and there exits a domain gap. By comparing the results, We can still see that the training on the synthetic data is proved to help the performance improvement.

3) *Testing Results for all Compared Methods* We further clarify that, even only considering the testing stage, the synthetic data is useful to some extent. In Fig. 9, we show the statistics of the performance for all the compared methods, on the synthetic data (CvMHAT-S) and the real-world data (CvMHAT-R), respectively. We can see that, these methods show the basically coincident relative performance on these two datasets. In practice, for a real-world scene, we can quickly build a synthetic scene following the real one and collect the data, on which we can test the performance of some methods. The results can reflect those on the real-world data, which, however, is not easy to obtain and annotate.

6 Applications and Future Work

All-Around Surveillance System CvMHAT can be regarded as a foundation for building the air-ground-synergetic video

Table 12 Comparative results of the method in Gan et al. (2021) on the multi-horizontal-view videos in CvMHAT-R dataset

Method	IDP	IDR	IDF_1	IDS	FM	MOTP	MOTA	mvIDP	mvIDR	$mvIDF_1$	mvMHAA
w/o train	33.0	37.8	35.2	31,008	17,952	83.0	53.2	20.2	12.2	15.2	15.9
w train. on CvMHAT-S	43.1	46.2	44.6	13,281	9633	84.5	69.6	30.8	17.3	22.2	32.6

surveillance system. Based on this, we can obtain the global picture of the people crowd and clear trajectory of each subject from the top view in the air. We can simultaneously observe the details, e.g., pose, actions, of some specific subjects, from the horizontal view on the ground.

Multi-view Action Recognition/Person Localization With multi-view MHAT as the basis for human scene analysis, it can achieve the multi-view collaborative human action recognition. As shown in Han et al. (2022b), a simple multi-view integration strategy can help the action recognition task in a crowded scene. This is because the multiple cameras are more likely to capture the better FOV for recognizing the human actions. Similarly, as discussed in Han et al. (2020b), we can use the horizontal-view camera for the interested activity perception and the top view as integration for the co-interest person localization.

Other Potential Applications The proposed problem may have other potential applications in the future, such as helping the visually impaired people for route navigation and obstacle avoidance, and developing the rough 3D reconstruction and mapping for a large-scale scene.

In the future, based on this benchmark, we aim to develop more effective features or techniques for the cross-view subject association, especially for the cross-top-horizontal views, e.g., those based on deep learning methods. Also, more joint learning paradigms and frameworks for the cross-view and over-time subject association and tracking are desirable. We believe these can very hopefully benefit the read-world applications in video surveillance, sport analysis, etc.

7 Conclusion

Complementary-view multiple human association and tracking (CvMHAT) is a relative new and challenging task. In this paper, we have built a new CvMHAT benchmark for this task, which contains both a real-world and a synthetic video dataset. Compared to existing datasets, the CvMHAT benchmark adopts the moving cameras from one top view and multiple horizontal views for data collection, with all the required annotations, including subject bounding boxes on each frame and consistent cross-frame and cross-view subject ID numbers. We further proposed a simple and effective baseline method. Experimental results verified the usefulness of this new dataset and effectiveness of the proposed baseline method. We have released all the data and code to the public, which we hope to bring convenience to other researchers to work on this emerging topic.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grants U1803264, 62072334. The authors would like to thank their team members, i.e., especially Jiewen Zhao for his technical assistance on implementation, and Liqiang Yin, Yiyang Gan, Yun Wang, Jiacheng Li, Sibo Wang, Shuai Wang, Songmiao Wang, and Likai Wang for their kind assistance in the collection and annotation of this dataset.

References

Ardeshir, S., & Borji, A. (2016). Ego2top: Matching viewers in egocentric and top-view videos. In *European conference on computer vision*.

Ardeshir, S., & Borji, A. (2018). Egocentric meets top-view. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(6), 1353–1366.

Ardeshir, S., & Borji, A. (2018b). Integrating egocentric videos in top-view surveillance videos: Joint identification and temporal alignment. In *European conference on computer vision*.

Bergmann, P., Meinhardt, T., Leal-Taixe, L. (2019a). Tracking without bells and whistles. In *IEEE international conference on computer vision*.

Bergmann, P., Meinhardt, T., Leal-Taixé, L. (2019b). Tracking without bells and whistles. In *IEEE international conference on computer vision*.

Bernardin, K., & Stiefelhagen, R. (2008). Evaluating multiple object tracking performance. *Journal on Image and Video Processing*.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J. (2011).

Brasó, G., Leal-Taixé, L. (2020). Learning a neural solver for multiple object tracking. In *IEEE conference on computer vision and pattern recognition*.

Cai, J. F., Candès, E., Shen, Z. (2018). A singular value thresholding algorithm for matrix completio. *SIAM Journal on Optimization*.

Chu, Q., Ouyang, W., Li, H., Wang, X., Liu, B., Yu, N. (2017). Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism. In *IEEE international conference on computer vision*.

Dehghan, A., Assari, S. M., Shah, M. (2015). GMMCP tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *IEEE conference on computer vision and pattern recognition*.

Dong, J., Jiang, W., Huang, Q., Bao, H., Zhou, X. (2019). Fast and robust multi-person 3d pose estimation from multiple views. In *IEEE conference on computer vision and pattern recognition*.

Dong, J., Fang, Q., Jiang, W., Yang, Y., Huang, Q., Bao, H., Zhou, X. (2021). Fast and robust multi-person 3d pose estimation and tracking from multiple views. *IEEE transactions on pattern analysis and machine intelligence*.

Dong, J., Fang, Q., Jiang, W., Yang, Y., Huang, Q., Bao, H., & Zhou, X. (2022). Fast and robust multi-person 3d pose estimation and track-

ing from multiple views. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(10), 6981–6992.
Du, Y., Song, Y., Yang, B., Zhao, Y. (2022). Strongsort: Make deepsort great again. In *arXiv*
Fabbri, M., Lanzi, F., Calderara, S., Palazzi, A., Vezzani, R., Cucchiara, R. (2018). Learning to detect and track visible and occluded body joints in a virtual world. In *European conference on computer vision*.
Fabbri, M., Brasó, G., Maugeri, G., Cetintas, O., Gasparini, R., Ošep, A., Calderara, S., Leal-Taixé, L., Cucchiara, R. (2021). Motsynth: How can synthetic data help pedestrian detection and tracking? In *IEEE/CVF international conference on computer vision*.
Ferryman, J. (2009). An overview of the pets2009 challenge. In *Proceedings of international workshop on pets*.
Fleuret, F., Berclaz, J., Lengagne, R., & Fua, P. (2008). Multicamera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30*(2), 267.
Gaidon, A., Wang, Q., Cabon, Y., Vig, E. (2016). Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
Gan, Y., Han, R., Yin, L., Feng, W., Wang, S. (2021). Self-supervised multi-view multi-human association and tracking. In *ACM multimedia*.
Gantmakher, F. R. (1959). *The theory of matrices*. American Mathematical Society.
Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J. (2021). Yolox: Exceeding yolo series in 2021. In *arXiv*
Han, R., Zhang, Y., Feng, W., Gong, C., Zhang, X., Zhao, J., Wan, L., Wang, S. (2019). Multiple human association between top and horizontal views by matching subjects' spatial distributions. In *arXiv*
Han, R., Feng, W., Zhao, J., Niu, Z., Zhang, Y., Wan, L., Wang, S. (2020a). Complementary-view multiple human tracking. In *AAAI conference on artificial intelligence*.
Han, R., Zhao, J., Feng, W., Gan, Y., Wan, L., Wang, S. (2020b). Complementary-view co-interest person detection. In *ACM international conference on multimedia*.
Han, R., Feng, W., Zhang, Y., Zhao, J., & Wang, S. (2022). Multiple human association and tracking from egocentric and complementary top views. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(9), 5225–5242.
Han, R., Wang, Y., Yan, H., Feng, W., & Wang, S. (2022). Multi-view multi-human association with deep assignment network. *IEEE Transactions on Image Processing, 31*, 1830–1840.
He, L., Liao, X., Liu, W., Liu, X., Cheng, P., Mei, T. (2020). Fastreid: A pytorch toolbox for general instance re-identification. In *arXiv*.
Kuo, C.H, Huang, C., Nevatia, R. (2010). Inter-camera association of multi-target tracks by on-line learned appearance affinity models. In *European conference on computer vision*.
Leal-Taixé, L., Milan, A., Reid, I., Roth, S., Schindler, K. (2015). Motchallenge 2015: Towards a benchmark for multi-target tracking. In *arXiv*.
Lealtaixe, L., Cantonferrer, C., Schindler, K. (2016). Learning by tracking: Siamese CNN for robust target association. In *IEEE conference on computer vision and pattern recognition*.
Liang, G., Lan, X., Zheng, K., Wang, S., Zheng, N. (2018). Cross-view person identification by matching human poses estimated with confidence on each body joint. In *AAAI conference on artificial intelligence*.
Liang, G., Lan, X., Chen, X., Zheng, K., Wang, S., & Zheng, N. (2019). Cross-view person identification based on confidence-weighted human pose matching. *IEEE Transactions on Image Processing, 28*(8), 3821–3835.
Lin, Y., Ezzeldeen, K., Zhou, Y., Fan, X., Yu, H., Qian, H., Wang, S. (2015). Co-interest person detection from multiple wearable camera videos. In *IEEE international conference on computer vision*.
Luiten, J., Osep, A., Dendorfer, P., Torr, P., Leibe, B. (2020). HOTA: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, pp. 1–31.
Ma, F., Shou, M.Z., Zhu, L., Fan, H., Xu, Y., Yang, Y., Yan, Z. (2022). Unified transformer tracker for object tracking. In *IEEE conference on computer vision and pattern recognition*.
Meinhardt, T., Kirillov, A., Leal-Taixe, L., Feichtenhofer, C. (2022). Trackformer: Multi-object tracking with transformers. In *IEEE conference on computer vision and pattern recognition*.
Ng, A. Y, Jordan, M. I., Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Conference on neural information processing systems*.
Redmon, J., Divvala, S. K., Girshick, R. B., Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *IEEE conference on computer vision and pattern recognition*.
Ristani, E., Tomasi, C. (2018). Features for multi-target multi-camera tracking and re-identification. In *IEEE conference on computer vision and pattern recognition*.
Ristani, E., Solera, F., Zou, R.S., Cucchiara, R., Tomasi, C. (2016). Performance measures and a data set for multi-target, multi-camera tracking. In *IEEE conference on computer vision and pattern recognition*.
Sun, X., Zheng, L. (2020). Dissecting person re-identification from the viewpoint of viewpoint. In *IEEE conference on computer vision and pattern recognition*.
Wojke, N., Bewley, A., Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. In *IEEE international conference on image processing*.
Wu, J., Cao, J., Song, L., Wang, Y., Yang, M., Yuan, J. (2021). Track to detect and segment: An online multi-object tracker. In *IEEE conference on computer vision and pattern recognition*.
Xiang, Y., Alahi, A., Savarese, S. (2015). Learning to track: Online multi-object tracking by decision making. In *IEEE international conference on computer vision*.
Xu, Y., Liu, X., Liu, Y., Zhu, S. (2016). Multi-view people tracking via hierarchical trajectory composition. In *IEEE conference on computer vision and pattern recognition*.
Xu, Y., Liu, X., Qin, L., Zhu, S. (2017). Cross-view people tracking by scene-centered spatio-temporal parsing. In *AAAI conference on artificial intelligence*.
Yang, B., & Nevatia, R. (2012a). Multi-target tracking by online learning of non-linear motion patterns and robust appearance models. In *IEEE conference on computer vision and pattern recognition*.
Yang, B., & Nevatia, R. (2012b). An online learned crf model for multi-target tracking. In *IEEE conference on computer vision and pattern recognition*.
Zamir, A. R, Dehghan, A., Shah, M. (2012). Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs. In *European conference on computer vision*.
Zelnik-Manor, L., Perona, P. (2004). Self-tuning spectral clustering. In *Conference on neural information processing systems*.
Zhang, S., Staudt, E., Faltemier, T., Roy-Chowdhury, A. K. (2015). A camera network tracking (CamNeT) dataset and performance baseline. In *Winter conference on applications of computer vision*.
Zhang, Y., Wang, C., Wang, X., Zeng, W., & Liu, W. (2021). Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision, 129*, 3069–3087.
Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X. (2022). Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*.
Zhao, J., Han, R., Gan, Y., Wan, L., Feng, W., Wang, S. (2020). Human identification and interaction detection in cross-view multi-person videos with wearable cameras. In *ACM international conference on multimedia*.

Zheng, K., Lin, Y., Zhou, Y., Salvi, D., Fan, X., Guo, D., Meng, Z., Wang, S. (2014). Video-based action detection using multiple wearable cameras. In *European conference on computer vision workshop*.

Zheng, K., Fan, X., Lin, Y., Guo, H., Wang, S. (2017). Learning view-invariant features for person identification in temporally synchronized videos taken by wearable cameras. In *IEEE international conference on computer vision*.

Zhou, X., Koltun, V., Krähenbühl, P. (2020). Tracking objects as points. In *European conference on computer vision*.

Zhou, X., Yin, T., Koltun, V., Krähenbühl, P. (2022). Global tracking transformers. In *IEEE conference on computer vision and pattern recognition*.

Zhu, J., Yang, H., Liu, N., Kim, M., Zhang, W., Yang, M. (2018). Online multi-object tracking with dual matching attention networks. In *European conference on computer vision*.

Zhu, P., Wen, L., Du, D., Bian, X., Fan, H., Hu, Q., & Ling, H. (2022). Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(11), 7380–7399.