Self-supervised Multi-view Multi-Human Association and Tracking

Yiyang Gan*
College of Intelligence and Computing, Tianjin University
realgump@tju.edu.cn

Ruize Han*†
College of Intelligence and Computing, Tianjin University
han_ruize@tju.edu.cn

Liqiang Yin
College of Intelligence and Computing, Tianjin University
yinliqiang@tju.edu.cn

Wei Feng
College of Intelligence and Computing, Tianjin University
wfeng@tju.edu.cn

Song Wang†
University of South Carolina
Columbia, USA
songwang@cec.sc.edu

ABSTRACT

Multi-view Multi-human association and tracking (MvMHAT) aims to track a group of people over time in each view, as well as to identify the same person across different views at the same time. This is a relatively new problem but is very important for multi-person scene video surveillance. Different from previous multiple object tracking (MOT) and multi-target multi-camera tracking (MTMCT) tasks, which only consider the over-time human association, MvMHAT requires to jointly achieve both cross-view and over-time data association. In this paper, we model this problem with a self-supervised learning framework and leverage an end-to-end network to tackle it. Specifically, we propose a spatial-temporal association network with two designed self-supervised learning losses, including a symmetric-similarity loss and a transitive-similarity loss, at each time to associate the multiple humans over time and across views. Besides, to promote the research on MvMHAT, we build a new large-scale benchmark for the training and testing of different algorithms. Extensive experiments on the proposed benchmark verify the effectiveness of our method. We have released the benchmark and code to the public.

CCS CONCEPTS

• **Computing methodologies** → **Tracking**; **Object identification**; **Matching**; **Unsupervised learning**.

KEYWORDS

multiple human association and tracking; multi-view cameras; self-supervised learning

[*] Equal contribution.
[†] Co-corresponding authors.

MM ’21, October 20–24, 2021, Virtual Event, China
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475177

ACM Reference Format:
Yiyang Gan, Ruize Han, Liqiang Yin, Wei Feng, and Song Wang. 2021. Self-supervised Multi-view Multi-Human Association and Tracking. In *Proceedings of the 29th ACM International Conference on Multimedia (MM ’21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475177

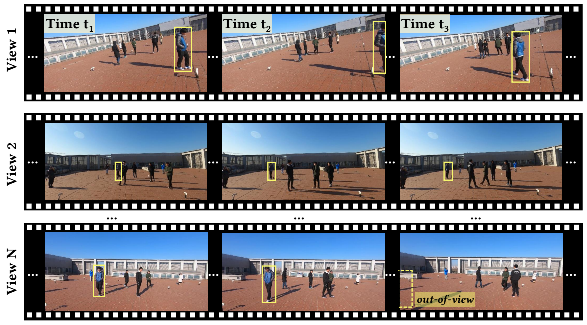


Figure 1: An illustration of the proposed MvMHAT problem.

1 INTRODUCTION

Multiple object tracking (MOT), especially multiple human tracking, is a fundamental and important task in computer vision and multimedia analysis [16, 38, 44, 45]. As an extension of MOT, multi-view multi-human association and tracking (MvMHAT) aims to continuously track a group of people in each view while simultaneously identifying the same persons across multiple views at each time [18]. With MvMHAT, we can not only record the temporal trajectories of the involved humans (referred to as subjects in this paper), but also roundly observe each subject’s detail, e.g., the human pose and behavior, from different views, which makes MvMHAT to facilitate many potential real-world applications. A typical example is video surveillance – imagining a scenario with multiple installed or wearable cameras from different views, we can employ MvMHAT to associate and analyze the collected videos for collaborative human activity recognition and important/abnormal person detection. So far, MvMHAT is a relatively new task with a handful of studies [15, 18, 54]. Among them, most works mainly study the over-time human tracking but not explicitly the cross-view human association [15, 54]. A couple of recent works [18, 19] study the MvMHAT with two complementary views for certain specific application scenarios. In this paper, as shown in Fig. 1, we are interested in MvMHAT in a more general setting where (arbitrary) multiple cameras (without calibration) are used to cover a multi-person scene from different views.

Compared to MOT, MvMHAT is a more challenging problem since we need to associate all the subjects appearing in different views during the tracking. The association usually suffers from the unknown and large view differences, illumination differences, (Views 1 and View 2 in Fig. 1), and the inconsistency of involved subjects (time t_1 and time t_3 in Views 1), etc. Moreover, with uncalibrated multi-view cameras, the human motion feature, a core cue for human matching in tracking, is usually inconsistent in different views and may not be effective when used for measuring the subject similarity for the cross-view association. This way, the subject appearance representation becomes particularly important. Considering this, in this paper, we first propose a unified framework for arbitrary-number multi-view MvMHAT. Besides, as we know, most previous works for MOT learn an appearance model from abundant labeled data for the over-time and cross-view appearance measurement. In MvMHAT, we actually have various appearance information of each subject along time and across views. The same person appearing in two views or points of time should present symmetric similarity, and such similarity among multiple views or at different points of time should be transitively consistent. This observation inspires us to adopt a self-supervised method to utilize the spatial-temporal consistency.

In this paper, we propose a self-supervised learning network to solve the MHAT problem. Specifically, given several videos capturing a group of people from different views, we first sample several frames from different views and time and apply the convolutional neural networks (CNN) to learn the embedding features of each subject. Then, we propose a spatial-temporal association network to model the over-time temporal association and the cross-view spatial association, which generates a matching matrix and can be self-supervised by a couple of symmetric-similarity (SSIM) and transitive-similarity (TSIM) losses as pretext task. In the inference stage, we leverage a new joint tracking and association scheme to solve the MvMHAT task. Moreover, the current research on MvMHAT is restricted by the lack of an appropriate public dataset that can be accessed and used to train and evaluate the deep network based algorithms. In this paper, we build a new large-scale benchmark based on several public datasets and self-collected data for the training and testing of the MvMHAT algorithm. The main contributions of this paper are:

1) We propose a self-supervised learning framework for MvMHAT. To the best of our knowledge, this is the first work to model such a problem in a self-supervised framework.
2) We propose the pairwise symmetric-similarity and triplewise transitive-similarity pretext tasks, which can be modeled as differentiable loss functions, to learn the representations for establishing the multi-view and over-time subject association and tracking.
3) We build a new benchmark for training and testing MvMHAT. Extensive experiments on the proposed datasets verify the rationality of our problem definition, the usefulness of the proposed benchmark, and the effectiveness of our method. We have released the benchmark to the public at https://github.com/realgump/MvMHAT.

2 RELATED WORK

MOT is a classical problem and has many applications in video processing and analysis. The most famous framework for MOT is the tracking-by-detection scheme, in which an object detector is first applied, and the detected subjects are then associated across frames to achieve multiple object tracking.. In this scheme, the most important issue is data association, which is mostly based on appearance similarity and motion consistency. The motion features can follow linear or nonlinear motion models. The linear model assumes the target to have a linear movement with constant velocity for a period of time [10, 41, 51], which is used in most existing trackers. The nonlinear one, to some extent, can better capture complex movements and provide a more accurate motion prediction [56, 57]. Many previous works on MOT try to develop more powerful appearance features for object association, from the hand-crafted appearance features such as color histograms [10, 58], to the recent deep network based appearance features [8, 52, 55]. This way, a key issue for such tracking-by-detection methods lies in the learning of human appearance features. More recent works also try to achieve object detection and tracking simultaneously using an end-to-end framework [3, 62]. For a more comprehensive review on MOT, we refer readers to several excellent surveys on tracking [9, 42]. Note that, the problem in this paper is extended from the MOT but not focused on the study of general MOT.

MTMCT (multi-target multi-camera tracking) is an extension of MOT, which aims to track and re-identify the targets (mainly for humans) in a large field, e.g., a campus, using many cameras installed at many sites with little or no field of view overlap. Several works [6, 7, 17, 39] focus on the inter-camera tracklet association by assuming that the within-camera tracklets in each camera are priorly given or obtained by existing algorithms. This setting is not practical in a real-world application. Several other works aim to address a more realistic problem by solving both intra- and inter-camera tracking jointly [37, 40, 41, 43]. The main thought for solving such a problem is to learn more discriminative appearance features [41] or design a more exquisite optimization model [43]. Differently, in this paper we are interested in handling a different multi-human association and tracking problem as discussed below.

MvMHAT and MTMCT both stem from the MOT task. However, they differ in two main perspectives: 1) They have different problem definitions. Besides temporal tracking, MTMCT also aims to handle the human re-identification, which is a ranking problem. Differently, MvMHAT focuses on the multi-human matching, which is a classification problem. 2) They use different camera settings. MTMCT uses multiple cameras distributed at different sites in a large area with no field-of-view (FOV) overlap. Differently, MvMHAT uses multi-view all-around cameras with overlapping FOVs covering the same scene. Several early works [1, 14, 15, 25, 28, 31] have studied a similar problem of tracking multiple humans using multiple FOV-overlapped cameras, in which the subjects commonly appear in different views at the same time. Recently, a series of works by Xu et al. [35, 53, 54] propose to track multiple people in a scene, e.g., a garden, using several cameras and collect new datasets for this research. This series of works extract various human features for tracking, including the varied poses and human actions, etc. However, the above works mainly focus on the over-time human tracking performance but not assessing the cross-view human association results. More recently, a series of works by Han et al. [18, 19] propose to jointly solve the human association and tracking problem using two complementary views, which, however, is a specific setting used in several application scenarios. Differently, in this paper, we focus on a more general setting where (arbitrary) multiple cameras observe a scene from different views. Also related to our work is a study on multi-view multi-object association (matching) [13, 20, 21, 60, 61] by exploring the matching cues, including human appearance [13, 60], spatial relation [20, 21] and motions [61], all of which only focus on cross-view association but not involving the over-time tracking.

Self-supervised Learning. Unsupervised learning has been widely used in many vision and multimedia computing tasks, including the image-based representation learning [12, 59] and video-level temporal coherence [30, 46], as well as person re-identification (re-id) [33, 34, 49] which is similar to our problem that learns the appearance similarity. The global retrieval and matching based re-id features can not be directly used for MOT, which is a local and constrained association problem, as discussed in [26]. However, there are very few works on studying the unsupervised learning based MOT, especially for the multi-human tracking [24, 27], not to mention the MvMHAT. As a special kind of unsupervised learning, self-supervised learning aims to construct the pretext tasks, commonly obtained by the acknowledged prior or self-constraint, to learn the network from unlabeled data. In this paper, we unveil the power of self-supervised learning for data association in MvMHAT.

3 OUR APPROACH

3.1 Problem Formulation

Given S synchronized videos taken from different views, we aim to address the Multi-view Multi-Human Association and Tracking (MvMHAT) task, which collaboratively tracks all the subjects in all the videos as well as identifying all the same persons appearing in different views. Specifically, we assume that the subjects have been detected in each frame in advance. The subjects are represented as bounding boxes in each view v at each time t. For each person in each view, MvMHAT aims to connect the subjects over time to form the single-view trajectory. Besides, MvMHAT also identifies the trajectories of the same person across all the views.

In this work, we formulate the above collaborative tracking, i.e., MvMHAT, as a spatial-temporal subject association problem. On one hand, the temporal (over-time) association can be regarded as a single-view multiple object tracking (MOT) problem. Similar to most MOT approaches, the goal is to solve the association matrix between tracklets $\mathcal{T}_{t-1}$ until frame $t-1$ in view v, and all the detections $\mathcal{B}^v_t = \{B_i | i = 1, 2, ..., N^v_t\}$ on frame t. Thus the association matrix is represented as $\mathbf{X}^v_t \in \mathbb{R}^{M_{t-1} \times N^v_t}$, where M_{t-1} and N^v_t denote the number of trajectories $\mathcal{T}_{t-1}$ and subjects $\mathcal{B}^v_t$, respectively. On the other hand, the spatial (cross-view) association is a multi-view subject matching problem. At each time t, we establish the association relation between different views. Taking a pair of views v and u for example, the cross-view subject association between $\mathcal{B}^v_t$ and $\mathcal{B}^u_t$ can be represented as a matching matrix $\mathbf{X}^{v,u} \in \mathbb{R}^{N^v_t \times N^u_t}$, where N^u_t denotes the number of subjects in $\mathcal{B}^u_t$.

As shown in Fig. 2, the multi-view video sequences provide the all-around and time-varying appearance of the subjects in the scene. The same person appearing in pairwise views or pairwise frames presents symmetric-similarity. They also show the cycle-consistency among different views and time. This inspires us to unveil self-supervised power for establishing the over-time and cross-view subject similarity. In the following, we adopt a self-supervised learning network for spatial-temporal subject association.

3.2 Spatial-Temporal Association Network

The spatial-temporal association network takes the video sequence without annotation as input, which learns the subject similarity used for association in a self-supervised manner. Specifically, as shown in Fig. 2, given a video frame at time t from the v-th-view video, we first apply a human detector to obtain all the subjects $\mathcal{B}^v_t$ in this frame. With the detected subjects, we apply the feature extraction network, denoted as Φ, to get the feature representation for all subjects $\mathbf{E}^v_t = \Phi(\mathcal{B}^v_t)$, by which we get $\mathbf{E}^v_t \in \mathcal{R}^{N^v_t \times D}$, where N^v_t denotes the number of subjects in view v at time t, and D denotes the dimension of feature for each subject.

With the extracted features on each frame, we can then define the subject similarity and association across frames and over time. **Spatial association.** Given a pair of frames at the same point of time t but from different views v and u, respectively, the subject similarity matrix between these two frames can be calculated by

$$\mathbf{S}^{v,u}_t = \mathbf{E}^v_t \cdot (\mathbf{E}^u_t)^{\mathrm{T}} \in \mathbb{R}^{N^v_t \times N^u_t}, \tag{1}$$

whose value at i-th row and j-th column, i.e., $\mathbf{S}^{v,u}_t(i,j)$ represents the similarity between i-th subject in $\mathcal{B}^v_t$ and j-th subject in $\mathcal{B}^u_t$. **Temporal association.** Similarly, given a pair of frames from the same view v but at different points of time t and s, respectively, the subject similarity matrix can be calculated by

$$\mathbf{S}^v_{t,s} = \mathbf{E}^v_t \cdot (\mathbf{E}^v_s)^{\mathrm{T}} \in \mathbb{R}^{N^v_t \times N^v_s}. \tag{2}$$

We then compute the matching matrix $\mathbf{X} \in [0, 1]$ based on the above similarity matrix. For clarity, we simplify both the cross-view matrix in Eq. (1) and the over-time similarity matrix in Eq. (2) as $\mathbf{S}$. We use a temperature-adaptive *softmax* operation f [23] to compute the matching matrix as

$$\mathbf{X}(r,c) = f_{r,c}(\mathbf{S}) = \frac{\exp(\tau \mathbf{S}(r,c))}{\sum_{c'=1}^{C} \exp(\tau \mathbf{S}(r,c'))}, \tag{3}$$

where r, c denote the indices of row and column in $\mathbf{S}$, respectively, and C is the number of columns for $\mathbf{S}$. Basically, we apply the *softmax* operation on each row of the matrix $\mathbf{S}$ and get $\mathbf{X}$ with same size as $\mathbf{S}$ but taking values in $[0, 1]$, as shown in Fig. 2. In Eq. (3), we use the *softmax* with an adjustable value τ as the adaptive temperature

$$\tau = \frac{1}{\epsilon} \log[\frac{\delta(C-1)+1}{1-\delta}], \tag{4}$$

which controls the soften ability of the function. ϵ and δ are two preset parameters.

So far, we get the predicted matching matrix $\mathbf{X}$. If we have the annotated data with a human identification label, the network can be trained with the supervision of ground-truth $\mathbf{X}$. In this paper, we aim to explore the symmetric-supervised pretext for learning the network without manual-annotated ground-truth labels.

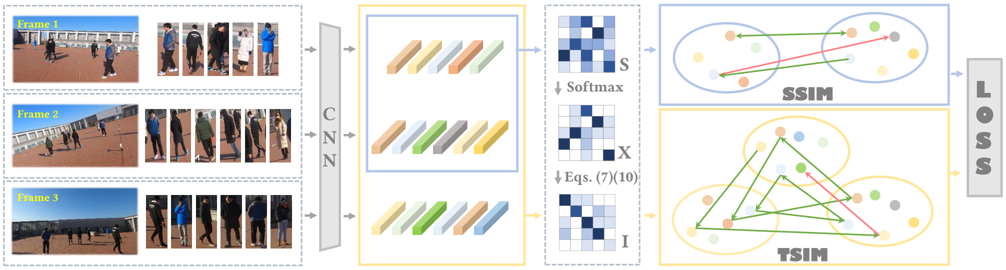
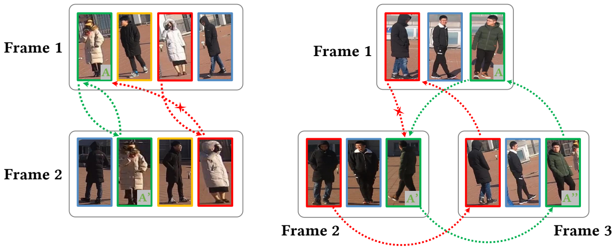


Figure 2: Overall framework of the proposed method.

3.3 Self-supervised Learning Loss

We consider two pretext tasks to construct the self-supervised loss. First, the same person appearing in a pair of frames from different views or different points of time should be consistent. As shown in Fig. 3(a), for a subject #A in Frame #1, if the subject #A' in another Frame #2 has the highest similarity with #A among all subjects; the subject #A should also be the most similar subject with #A' among all subjects in Frame #1, i.e., the self-similarity property. Second, as shown in Fig. 3(b), the subject has cycle consistency among multiple views and different points of time. If subject #A in Frame #1 is identity-consistent with #A' in Frame #2 and the subject #A'' in Frame #3, then the subjects #A' and #A'' are expected to be the same subject, i.e., the transitive-similarity property.



(a) (b)

Figure 3: An illustration of the rationale of SSIM and TSIM.

Symmetric-Similarity (SSIM): Given the similarity matrix between the subjects within two sets $\mathcal{I}$ and $\mathcal{J}$ (including the cross-view and over-time cases) as defined in Eqs. (1) and (2), which we denote as $\mathrm{S}_{ij} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|}$. We apply the *softmax* operation on the similarity matrix S to get the matching matrix defined in Eq. (3) :

$$\mathbf{X}_{ij} = f(\mathrm{S}_{ij}) \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|}. \tag{5}$$

The matching matrix $\mathbf{X}_{ij}$ can be regarded as a mapping (matching relation) from $\mathcal{I}$ to $\mathcal{J}$, i.e., $\mathcal{X}_{ij} : \mathcal{I} \mapsto \mathcal{J}$. Specifically, the row sum in X is equal to 1, and we can find the maximum in each row of $\mathbf{X}$ to seek the matched subject for each one in $\mathcal{I}$. Similarly, we get the mapping from $\mathcal{J}$ to $\mathcal{I}$ as

$$\mathbf{X}_{ji} = f(\mathrm{S}_{ij}^{\mathrm{T}}) \in \mathbb{R}^{|\mathcal{J}|\times|\mathcal{I}|}. \tag{6}$$

As shown in Fig. 2, we calculate the *symmetric-similarity matrix*

$$\mathbf{I}_{\mathrm{S}} = \mathbf{X}_{ij} \cdot \mathbf{X}_{ji} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{I}|}, \tag{7}$$

where $\mathbf{I}_{\mathrm{S}}$ can be regarded as the mapping: $\mathcal{I} \mapsto \mathcal{J} \mapsto \mathcal{I}$. Ideally, if the subjects in $\mathcal{I}$ and $\mathcal{J}$ are same, the result $\mathbf{I}_{\mathrm{S}}$ should be an identity matrix. This way, we can calculate the loss between the predicted $\mathbf{I}_{\mathrm{S}}$ and the identity matrix. However, this condition is not always satisfied due to possible occluded or out-of-view subjects over time and the field-of-view difference across views, resulting in all-zero rows in $\mathbf{X}_{ij}$ or $\mathbf{X}_{ji}$. Therefore, we can not always compel the result $\mathbf{I}_{\mathrm{S}}$ to approximate the identity matrix and have to apply more deliberate supervision on it, which we will discuss later.

Transitive-Similarity (TSIM): Besides the pairwise symmetric similarity, we also consider the triplewise transitive similarity. Given the similarity matrix S_{ij} between two sets $\mathcal{I}$ and $\mathcal{J}$, and S_{jk} between $\mathcal{J}$ and $\mathcal{K}$, we also consider the consistent similarity among this triplet, i.e., $\mathcal{I}$, $\mathcal{J}$ and $\mathcal{K}$. We first compute the third-order similarity matrix as

$$\mathring{\mathrm{S}}_{ik} = \mathrm{S}_{ij} \cdot \mathrm{S}_{jk}, \; (i \neq j \neq k), \tag{8}$$

where $\mathring{\mathrm{S}}_{ik} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{K}|}$ actually represents the similarity between the subjects in $\mathcal{I}$ and $\mathcal{K}$, through the set of $\mathcal{J}$. We then compute the matching matrix as

$$\mathring{\mathbf{X}}_{ik} = f(\mathring{\mathrm{S}}_{ik}) \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{K}|}, \quad \mathring{\mathbf{X}}_{ki} = f(\mathring{\mathrm{S}}_{ik}^{\mathrm{T}}) \in \mathbb{R}^{|\mathcal{K}|\times|\mathcal{I}|}, \tag{9}$$

where $\mathring{\mathrm{S}}_{ik}^{\mathrm{T}}$ denotes the transposition of $\mathring{\mathrm{S}}_{ik}$. The result $\mathring{\mathbf{X}}_{ik}$ represents the mapping $\mathcal{I} \mapsto \mathcal{J} \mapsto \mathcal{K}$. In contrast, $\mathring{\mathbf{X}}_{ki}$ represents the mapping along $\mathcal{K} \mapsto \mathcal{J} \mapsto \mathcal{I}$.

Therefore, we calculate the *transitive-similarity matrix*

$$\mathbf{I}_{\mathrm{T}} = \mathring{\mathbf{X}}_{ik} \cdot \mathring{\mathbf{X}}_{ki} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{I}|}, \tag{10}$$

where $\mathbf{I}_{\mathrm{T}}$ can be regarded as the mapping: $\mathcal{I} \mapsto \mathcal{J} \mapsto \mathcal{K} \mapsto \mathcal{J} \mapsto \mathcal{I}$. Similar to the matrix $\mathbf{I}_{\mathrm{S}}$ defined in Eq. (7), we need to apply an appropriate supervision on $\mathbf{I}_{\mathrm{T}}$.

Self-supervised loss function. As discussed above, the matrix $\mathbf{I}_{\mathrm{S}}$ and $\mathbf{I}_{\mathrm{T}}$, which we uniformly denote as $\mathbf{I}$ provisionally, have the property that their diagonal elements are 1 or 0, while other elements are all 0 in the ideal case. In other words, the diagonal elements of $\mathbf{I}$ should be equal to or greater than the others. With this constraint, we apply the auxiliary loss function,

$$\mathrm{L}(\mathbf{I}) = \sum_{r=1}^{|\mathcal{I}|} \mathrm{relu}(\max_{c\neq r} \mathbf{I}(r,c) - \mathbf{I}(r,r) + m), \tag{11}$$

where r, c denote the indices of row and column in $\mathbf{I}$, respectively. Specifically, for each row r, if the non-diagonal elements $\mathbf{I}(r,c)$ with $c \neq r$ are greater than the corresponding diagonal elements $\mathbf{I}(r,r)$, the loss will increase [47]. We only use the maximum non-diagonal element $\max_{c\neq r} \mathbf{I}(r,c)$ in each row to punish the hardest negative sample. The margin $m \geq 0$ is a pre-set parameter, which controls the punishment scope for the gap between $\mathbf{I}(r,c)$ and $\mathbf{I}(r,r)$. In other word, the loss will take effect iff $\mathbf{I}(r,r) - \max_{c\neq r} \mathbf{I}(r,c) \leq m$. This setting expects the diagonal element to be greater than the other elements by a margin m. This way, we define the loss

$$\mathcal{L}_{\mathrm{SSIM}} = \mathrm{L}(\mathbf{I}_{\mathrm{S}}) + \mathrm{L}(\mathbf{I}_{\mathrm{S}}^{\mathrm{T}}), \tag{12}$$

where $L(\mathbf{I}_S)$ and $L(\mathbf{I}_S^T)$ compels $\mathbf{I}_S$ to satisfy the above constraints for all rows and columns, respectively. Similarly, we define the TSIM loss

$$\mathcal{L}_{\mathrm{TSIM}} = L(\mathbf{I}_T) + L(\mathbf{I}_T^T). \tag{13}$$

The total loss is then calculated as $\mathcal{L} = \mathcal{L}_{\mathrm{SSIM}} + \mathcal{L}_{\mathrm{TSIM}}$.

Discussion. Up to now, we have present the pairwise symmetric similarity, the triplewise transitive similarity constraints, and the corresponding loss functions. Actually, there are higher-order conditions involving more views or time. In this section, we discuss the generalization of the proposed self-supervised loss.

We denote the universal set of the indices, e.g., i, j in Eq. (5), for subject group on each frame as $\mathcal{F}$, i.e., $i, j \in \mathcal{F}$. In order to associate the subjects between each pair of (over-time or cross-view) frames, we denote the subject association mapping between arbitrary pair (i, j) as $\mathcal{M} = \{\varphi_{i,j} | \forall i, j \in \mathcal{F}\}$. We first assume all the frames share the same set of subjects. Thus $\Phi_{i,j}$ is a bijection.

Given any two elements, i.e., $\forall f_1, f_n \in \mathcal{F}$, the mapping $\varphi_{f_1 f_n}$ from f_1 to f_n can be decomposed as the combination of any number of mappings.

$$\varphi_{f_1 f_n} = \varphi_{f_{n-1} f_n} \circ \varphi_{f_{n-2} f_{n-1}} \circ \cdots \circ \varphi_{f_1 f_2}, \tag{14}$$

where $f_1, f_2, ..., f_n \in \mathcal{F}$, and $\circ$ denotes the mapping composition operation. Next, the arbitrary-order combined mapping can be derived by the proposed pairwise and triplewise mapping. Specifically, Eq. (14) is equivalent to

$$\varphi_{ik} = \varphi_{jk} \circ \varphi_{ij}, \text{ for } \forall i, j, k \in \mathcal{F}, \tag{15}$$

From Eq. (15), we discuss the situations for three cases: (i) $i = k = j$; (ii) $i = k \neq j$; (iii) $i \neq k \neq j$, from which we get the variations of Eq. (15) as

$$\text{(i) Self-Identity} \qquad \varphi_{ii} = id \tag{16}$$

$$\text{(ii) Symmetric-Identity} \qquad \varphi_{ji} \circ \varphi_{ij} = id \tag{17}$$

$$\text{(iii) Transitive-Identity} \qquad \varphi_{ki} \circ \varphi_{jk} \circ \varphi_{ij} = id \tag{18}$$

where id is the identity mapping. The self-Identity condition naturally holds in our problem because the same person in the same frame has the same feature vector. This way, we only consider the conditions in Eq. (17) and Eq. (18), for which we use doubly stochastic matrix $\mathbf{X}$ to represent the association map φ. We get

$$\text{(Symmetric-Similarity)} \qquad \mathbf{X}_{ij} \cdot \mathbf{X}_{ji} = \mathbf{I} \tag{19}$$

$$\text{(Transitive-Similarity)} \qquad \mathbf{X}_{ij} \cdot \mathbf{X}_{jk} \cdot \mathbf{X}_{ki} = \mathbf{I} \tag{20}$$

where $\mathbf{I}$ is the identity matrix. In our condition, the numbers of persons in different frames can be unequal. Thus, we only assume $\mathbf{X}$ is the row stochastic matrix (sum of each row is 1) and use the margin in Eq. (11) to relax the identity matrix $\mathbf{I}$. As discussed above, the proposed method with symmetric-similarity and transitive-similarity self-supervision, can be regarded as the self-constraints for the matching-relation mapping with arbitrary order.

3.4 The New Association and Tracking Scheme

With the above spatial-temporal association network in Section 3.2 and the proposed self-supervision loss in Section 3.3, our method can be trained with the videos without tracking and association labels. In the inference stage, we propose a new scheme to jointly address the association and tracking tasks. Different from the training stage, after computing matching matrices of spatial and temporal association, we then use Hungarian Algotithm [29] to get permutation matrix $\mathbf{P} \in \{0, 1\}$. The proposed MvMHAT scheme can be summarized in Algorithm 1. Specifically for the human ID assignment strategy, we can explain it by an example. In 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 to P and initialize a new tracklet. At t_2, we mark the unmatched tracklet to be 'sleep' in view v_1. Here the tracklet of subject P in v_1 interrupts but the multi-view tracklet of P maintains because it still appears in other views. At t_3, we use the multi-view subject association results to help match P to the slept tracklet in view v_1. This is better than the traditional MOT, which has difficulty to continuously track P if it disappears for a long time – P is usually assigned with a new ID when it re-appears. However, as a limitation of our method, incorrect cross-view association results at t_3 will cause wrong tracking results.

Algorithm 1: MvMHAT framework:

```
Input: $\mathcal{V} = \{\mathcal{V}_i | i = 1, ..., S\}$: a group of temporal-consecutive
       videos (T frames for each) captured from different views.
Output: Tracked subject bounding boxes with associated ID.
for $t = 1 : T$ do
    Detect the subjects in frame $t$ for each view:
      $\mathcal{B}_t^v = \{B_{t_j}^v | j = 1, 2, ..., N_t^v\} (v = 1, 2, ..., S)$.
    Generate spatial permutation matrixs
      $\mathbf{P}_t^{v,u}, (v = 1, 2, ..., S; u = 1, 2, ..., S; v \neq u)$, and temporal
      permutation matrixs $\mathbf{P}_{t,t-1}^v, (v = 1, 2, ..., S)$.
    for $v = 1 : S$ do
        for $p = 1 : N_t^v$ do
            if $\exists r, s.t.\ \mathbf{P}_{t,t-1}^v(p, r) = 1$ then
                Associate tracklet $\mathcal{T}_r$ with $\mathcal{B}_{t_p}^v$.
            else if $\exists (u, q), s.t.\ u < v \wedge \mathbf{P}_t^{v,u}(p, q) = 1$ then
                Associate tracklet $\mathcal{T}_q$ with $\mathcal{B}_{t_p}^v$.
            else
                Initialize new tracklet $\mathcal{T}_p$ with $\mathcal{B}_{t_p}^v$.
return Bounding boxes with ID numbers
```

Implementation details. In the training stage, we take the frames across all views from two different points of time as a group of inputs of the network. We traverse all the frames from different views along the whole video. We use annotated subject detections in training and use results of Detectron [50] in inference. ResNet-50 [22] is used as the backbone network in all experiments, which has outputs of 1,000-dimensional features. In Eq. (4), we set $\epsilon = 0.1$ and $\delta = 0.5$. In Eq. (11), we set $m = 0.5$. In the inference stage, we also apply the Kalman filtering for over-time subject assocition following the MOT algorithm DeepSort [5]. We use the Pytorch backend for implementing the proposed network and run it on a computer with RTX 2080Ti GPU. Our network is trained on 8,700 groups of frames for less than 10 epochs with the initial learning rate 10^{-5}, and the inference speed is 30+ FPS.

4 PROPOSED MVMHAT BENCHMARK

Dataset Collection. To the best of our knowledge, previous MOT datasets with multiple views covering an overlapped region are relatively small and only used for algorithm testing. To comprehensively train and test the proposed framework, we build a new large-scale video dataset – MvMHAT benchmark, for the multi-view multi-human association and tracking task. To reduce the cost of data collection and annotation while maintaining the usefulness and credibility of the proposed dataset, part of videos and corresponding annotations in MvMHAT was drawn from two available datasets of Campus [53] and EPFL [15]. Besides, we have also collected 12 video groups containing 46 sequences, with each group containing three to four views. To enrich the diversity of the collected data, different from previous videos captured by fixed cameras, these videos are collected with four wearable cameras, i.e., GoPro, by covering an overlapped area present with multiple people from significantly different directions, e.g., near 90-degree view-angle difference. We then manually synchronize them and annotate the bounding box and the ID for each subject on all 30, 900 frames.

Table 1: Data source and statistics for the MvMHAT dataset.

Source	Campus	EPFL	Self Collected	Total
# Group	6	8	12	26
# Sequence	22	30	46	98
# View	3-4	3-4	3-4	3-4
Avg. Length	1,500	900	672	928
Avg. # Subject	14	8	10	10
Full Length	33,000	27,000	30,900	90,900

Dataset statistics and splitting. As shown in Table 1, in total, our dataset for MvMHAT contains 26 multi-view video groups with 98 (single-view) sequences. Each video group contains several temporal-synchronized videos with multiple views, e.g., 3 – 4 views. The average length of each sequence is 928 frames, and in average ten subjects appear in each video. The dataset, in total, contains over 90 thousand frames. We further split the dataset into training and testing datasets, each of which contains 13 video groups, and the ratio of frames between the training and the testing datasets is about 2:1. To guarantee the diversity of the testing data, the testing videos contain the videos from Campus, EPFL, and Self collected.

Evaluation metrics. *MHT metrics*: We apply the commonly used MOT metrics for the single-view tracking performance evaluation as in MOT Challenge [32], including multiple object tracking precision (MOTP) and multiple object tracking accuracy (MOTA) proposed in [4]. A key task of the MvMHAT task is to associate and track the same subject over time. We are more concerned about the ID related metrics [40] – ID precision (IDP), ID recall (IDR), and ID F_1 measure (IDF$_1$) in evaluation.

MHA metrics: We further apply the metrics for cross-view association evaluation, i.e., AIDP, AIDR, and AIDF$_1$, by expanding the cross-view association metrics in [18, 19]. Specifically, AIDP and AIDR denote the multi-view subject association precision and recall, respectively. Given the subject IDs in all views, we take two views each time and compute the pairwise subject matching performance as in [18, 19], whose average on all view pairs is used as a multi-view association metric. Based on AIDP and AIDR, the association F_1 score is computed as $\text{AIDF}_1 = \frac{2\times\text{AIDP}\times\text{AIDR}}{\text{AIDP}+\text{AIDR}}$. Following [4, 18], we apply multi-view multi-human association accuracy $\text{MHAA} = 1 - \frac{\sum_t(\text{MS}_t+\text{FP}_t+2\text{MM}_t)}{\sum_t N_t}$, where MS_t, FP_t, MM_t are the false negative, false positives, and mismatched pairs, respectively, and N_t is the total number of subjects in all views, at frame t.

Overall metrics: For the overall performance evaluation of MvMHAT problem, we take a simple average and get the MvMHAT F_1 *score*, i.e., MHAT.F$_1$ = mean(IDF$_1$, AIDF$_1$) and MvMHAT *accuracy score*, i.e., MHAT.Acc = mean(MOTA, MHAA).

5 EXPERIMENTAL RESULTS

5.1 Comparison Results

Baseline methods. As discussed above, we actually did not find existing methods that can directly handle our MvMHAT problem for comparison. Therefore, we try to include sufficient related approaches with necessary modifications for comparison.

- We first select three state-of-the-art MOT methods for single-view videos, including CenterTrack [62], Tacktor++ [3] and TraDeS [48] for comparison.
- We also include a multi-target multi-camera tracking (MTMCT) method DeepCC [41] for comparison. Note that, DeepCC is used to handle the human tracking and re-identification (re-id) using multiple cameras covering different areas. We modify it to handle the proposed MvMHAT following the deep re-id features and BIP based correlation clustering method used in DeepCC.
- Besides, we also take the approach CVMHT in [19] as a comparison method, which takes pairwise-view videos as input and cannot directly handle our problem with multiple (≥ 2) views. We divide the video groups in our dataset into pairwise video pairs and evaluate CVMHT on each pair, respectively.

For a fair comparison, we use the same human detections provided by a commonly used detector [50], for all the comparison methods and the proposed method. We also reserve the results provided by Tacktor++ and TraDeS using the private detector, which will be shown in the following Section 5.1. We clarify that all the networks in the comparison methods are the public version trained on the original training dataset. To be relatively fair, we do not retrain them because all these methods need supervision of labeled data, which is not used in our self-supervised method.

Table 2 shows the comparative results of our methods with the baseline methods. For the single-view MOT methods, i.e., Tracktor++, CenterTrack and TraDes, we first evaluate the over-time tracking performance using the standard MOT metrics. We can see that the state-of-the-art approach TraDes with the private detector provide the best MOTA score among all competitors. Note that, MOTA and MOTP mainly focus on the object detection accuracy and precision during tracking [36]. On the contrary, the ID-based metrics, i.e., IDP, IDR and IDF$_1$ evaluate the ID association and consistency over time. This paper is more concerned about the latter. We can see that the proposed method outperforms all the above methods in IDF$_1$ score. We then show the cross-view association results in the middle of Table 2. We know that the single-view MOT methods only handle the tracking in each video but not including the cross-view association. For comparison, we additionally help them by providing the ground-truth IDs for the subjects across different views when they appear in each video for the first time. The over-time tracking on each video can propagate the IDs to subsequent frames, which we can use to associate the subjects across views and over time. From the first five rows, we can see that the cross-view association performances provided by the single-view MOT methods are poor even with the above help. The main reason

Table 2: Comparative results of different methods on the proposed MvMHAT benchmark.

Method	Over-Time Tracking					Cross-View Association				Overall	
	IDP	IDR	IDF$_1$	MOTP	MOTA	AIDP	AIDR	AIDF$_1$	MHAA	MHAT.Acc	MHAT.F1
Tracktor++ [2]	54.2	40.1	46.1	79.4	66.5	34.3	14.0	20.5	37.1	51.8	33.3
CenterTrack [62]	44.3	33.5	38.1	79.2	63.5	29.7	9.1	13.9	34.1	48.8	26.0
TraDeS [48]	46.7	43.2	44.9	77.5	69.5	32.4	14.0	19.6	36.0	52.8	32.2
CenterTrack [62] (Private)	43.8	33.7	38.1	79.2	63.1	31.9	9.3	14.4	34.3	48.7	26.2
TraDeS [48] (Private)	53.9	50.5	52.1	77.4	70.8	38.7	19.7	26.1	38.5	54.6	39.1
DeepCC [41]	40.8	24.6	30.7	82.2	47.5	11.2	2.9	4.6	30.4	39.0	17.6
CVMHT [19]	51.1	36.2	42.4	82.3	54.1	32.8	26.4	29.2	41.7	47.9	35.8
Ours	53.0	51.9	52.4	79.2	64.7	53.0	46.4	49.5	51.7	58.2	51.0

Table 3: Ablation study of different variations of our method.

Method	Over-Time Tracking					Cross-View Association				Overall	
	IDP	IDR	IDF$_1$	MOTP	MOTA	AIDP	AIDR	AIDF$_1$	MHAA	MHAT.Acc	MHAT.F1
w/o Training	34.2	34.6	34.4	78.3	57.2	23.3	15.1	18.3	27.7	42.5	26.4
w/o $\mathcal{L}_{\text{SSIM}}$	52.7	51.7	52.2	79.2	64.5	56.4	41.7	48.0	49.2	56.9	50.1
w/o $\mathcal{L}_{\text{TSIM}}$	49.4	48.6	49.0	79.2	65.3	54.7	36.1	43.5	46.2	55.8	46.2
w/o Association	63.2	61.3	62.2	79.4	67.7	22.3	5.1	8.3	30.1	48.9	35.2
w/o Tracking	59.1	42.0	49.1	79.8	42.5	55.8	44.4	49.4	47.6	45.0	49.2
w/o Relax	48.1	47.4	47.8	79.1	64.1	43.9	24.9	31.8	39.2	51.6	39.8
w/o Margin	31.2	29.6	30.4	79.0	61.0	24.2	11.0	15.2	29.0	45.0	22.8
w/o Temperature	38.0	38.1	38.1	78.6	59.7	16.6	9.6	12.2	26.6	43.2	25.1
Ours	53.0	51.9	52.4	79.2	64.7	53.0	46.4	49.5	51.7	58.2	51.0

is that, without the cross-view re-association mechanism during tracking, the association will fail once the subject ID switch occurs.

Multi-view tracking approaches, i.e., DeepCC and CVMHT, also perform not very well particularly for the cross-view association. This is because DeepCC is mainly used for long-term trajectory re-identification but not good at synchronously associating the subjects across views simultaneously. For CVMHT [19], it emphasizes the subject spatial distribution but simplifies the appearance for subject matching, which is not very suitable in our setting. For the overall performance metrics, i.e., MHAT.Acc, MHAT.F1, our method achieves the best performance on both of them.

5.2 Ablation Study

- w/o Training: We directly use the Resnet-50 [22] pre-trained on ImageNet [11] to extract features instead of the one using the self-supervised training in our method.
- w/o $\mathcal{L}_{\text{SSIM}}$: Remove the SSIM loss $\mathcal{L}_{\text{SSIM}}$ in Eq. (12).
- w/o $\mathcal{L}_{\text{TSIM}}$: Remove the TSIM loss $\mathcal{L}_{\text{SSIM}}$ in Eq. (13).
- w/o Association: Remove the association module during tracking, and generate the association results when the object appears the first time.
- w/o Track: Remove the collaboratively tracking module, and implement the tracking on one view to obtain the subject ID in other views by cross-view association.
- w/o. Relax: Replace the relaxed margin by a strict one namely set $m = 1$ in Eq. (11).
- w/o Margin: Remove the margin namely set $m = 0$ in Eq. (11).
- w/o Temperature: Remove the temperature mechanism namely we set $\tau = 1$ in Eq. (3).

Effectiveness of self-supervised loss. As shown in Table 3, we can see that '*w/o Training*' has a poor performance, which shows the challenge of the MvMHAT task and the importance of proposed self-supervised training. Our loss functions enable the network to unsupervisedly find the potential characteristics of data. '*w/o* $\mathcal{L}_{\text{SSIM}}$' and '*w/o* $\mathcal{L}_{\text{TSIM}}$' show the effectiveness of $\mathcal{L}_{\text{TSIM}}$ loss function and $\mathcal{L}_{\text{SSIM}}$ function. $\mathcal{L}_{\text{SSIM}}$ only considers the relationship between pairwise frames, which leads to the drop of performance. Similarly, $\mathcal{L}_{\text{TSIM}}$ focuses on global information and only considering it does not perform as well as the proposed combined loss.

Evaluation of tracking & association mechanism. We can see from Table 3 that '*w/o Association*' provides very promising tracking results. It can be explained that, without considering the cross-view association, the method under '*w/o Association*' can avoid more ID switches during tracking. This can be regarded as an upper bound of our method only for tracking, which, however, naturally generates a poor association performance. From comparing the results of '*w/o Tracking*' and '*Ours*', we are surprised to find that with the aggregation of tracking, our method not only improves the performance of over-time tracking but also the cross-view association. It indicates that the temporal tracking can be used as a favor for association. Overall, the integration of collaborative tracking & association mechanism generates the best results.

Influences of setups. The results generated by '*w/o Relax*' and '*w/o Margin*' verify the explanation in Section 3.3. In '*w/o Relax*', we assume **I** in Eq. (11) to be an identity matrix. However, with frequent occlusions and out-of-view of subjects, frames from different views or different points of time hardly share the exact same people. In this case, an over-constrained **I** tends to give the unmatched people incorrect matchings. On the contrary, '*w/o Margin*' means the strongest relaxation of **I**, which seems to give a too weak penalty for incorrect matchings. Finally, in a real-world scenario, the spatial-temporal distribution of people is always ruleless, which leads to the different number of people in different views at different time. However, softmax has different soften abilities when the input size changes, which makes these output values to be affected by the number of people. This way, the *temperature* in softmax is beneficial.

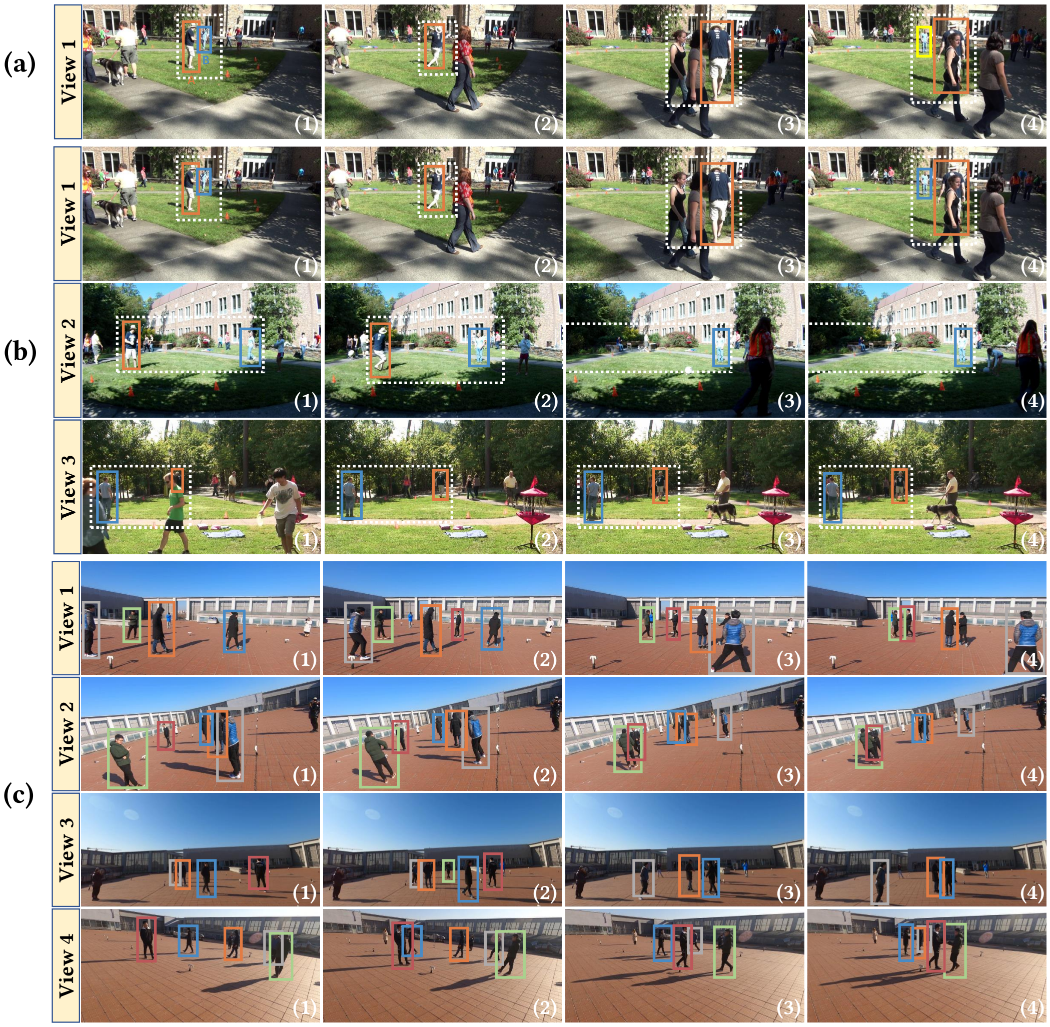


Figure 4: Samples of the qualitative results.

5.3 Qualitative Evaluation

Figure 4(a) shows a long-term tracking result of Tracktor++. We can see that two players are practicing baseball in the white dotted area in (1). However, because of the frequent moving of the player A in (2), and comings and goings of pedestrians in (3), the player B is occluded frequently. Thus, the single view MOT methods have to always try to match B with all the tracklets, which leads to lots of incorrect ID switches, e.g., in (4). As shown in Fig. 4(b), our method leverages the complementary characteristic of multiple views, and we ensure that any people can be observed in at least one view at a time. The integration of multi-view tracklets makes good use of both temporal and spatial information. The proposed MvMHAT scheme, to some extent, can help track the re-appearing detections that cannot be matched by the tracking module, while the continual tracking can help associate the subjects that cannot be matched by the association module. This way, a spatio-temporal connection is established for all observed people, which *has the potential to better handle the long-term tracking*. In Fig. 4(c), people are walking and interacting with each other in various activities. Based on the results of MvMHAT, we can capture the details of every subject from all-around perspectives. This demonstrates the *potential of the proposed MvMHAT for broad applications*, e.g., sports games and outdoor surveillance, which aim to capture both global and local details of involved people.

6 CONCLUSION

In this paper, we have studied a relatively new problem – MvMHAT, which is different from the existing MOT problem and has many applications. For fully excavating the peculiarity of MvMHAT, we model the problem as a self-supervised learning task and propose an end-to-end framework to handle it. To promote the study on this new topic, we have also built a new MvMHAT benchmark for performance evaluation. Experimental results verify the rationality of our problem formulation, the usefulness of the proposed benchmark and the effectiveness of our method.

Acknowledgement. This work was supported in part by the NSFC under Grants U1803264, 62072334, 61672376, 61671325.

REFERENCES

[1] Mustafa Ayazoglu, Binlong Li, Caglayan Dicle, Mario Sznaier, and Octavia I Camps. 2011. Dynamic subspace-based coordinated multicamera tracking. In *ICCV*.

[2] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. 2019. Tracking without bells and whistles. In *ICCV*.

[3] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixé. 2019. Tracking Without Bells and Whistles. In *ICCV*.

[4] Keni Bernardin and Rainer Stiefelhagen. 2008. Evaluating multiple object tracking performance. *EURASIP Journal on Image and Video Processing* 2008 (2008), 1–10.

[5] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. 2016. Simple online and realtime tracking. In *ICIP*.

[6] Yinghao Cai and Gerard Medioni. 2014. Exploring context information for inter-camera multiple target tracking. In *WACV*.

[7] Xiaotang Chen, Kaiqi Huang, and Tieniu Tan. 2014. Object tracking across non-overlapping views by learning inter-camera transfer models. *Pattern Recognition* 47, 3 (2014), 1126–1137.

[8] Peng Chu and Haibin Ling. 2019. Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking. In *ICCV*.

[9] Gioele Ciaparrone, Francisco Luque Sánchez, Siham Tabik, Luigi Troiano, Roberto Tagliaferri, and Francisco Herrera. 2020. Deep learning in video multi-object tracking: A survey. *Neurocomputing* 381 (2020), 61–88.

[10] Afshin Dehghan, Shayan Modiri Assari, and Mubarak Shah. 2015. GMMCP Tracker: Globally Optimal Generalized Maximum Multi Clique Problem for Multiple Object Tracking. In *CVPR*.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

[12] Carl Doersch, Abhinav Gupta, and Alexei A Efros. 2015. Unsupervised visual representation learning by context prediction. In *ICCV*.

[13] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. 2019. Fast and Robust Multi-Person 3D Pose Estimation from Multiple Views. In *CVPR*.

[14] Ran Eshel and Yael Moses. 2010. Tracking in a dense crowd using multiple cameras. *IJCV* 88, 1 (2010), 129–143.

[15] Francois Fleuret, Jerome Berclaz, Richard Lengagne, and Pascal Fua. 2008. Multi-camera people tracking with a probabilistic occupancy map. *IEEE TPAMI* 30, 2 (2008), 267.

[16] Xu Gao and Tingting Jiang. 2018. OSMO: Online Specific Models for Occlusion in Multiple Object Tracking under Surveillance Scene. In *ACM MM*.

[17] Andrew Gilbert and Richard Bowden. 2006. Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity. In *ECCV*.

[18] Ruize Han, Wei Feng, Yujun Zhang, Jiewen Zhao, and Song Wang. 2021. Multiple Human Association and Tracking from Egocentric and Complementary Top Views. *IEEE TPAMI* (2021).

[19] Ruize Han, Wei Feng, Jiewen Zhao, Zicheng Niu, Yujun Zhang, Liang Wan, and Song Wang. 2020. Complementary-View Multiple Human Tracking. In *AAAI*.

[20] Ruize Han, Yujun Zhang, Wei Feng, Chenxing Gong, Xiaoyu Zhang, Jiewen Zhao, Liang Wan, and Song Wang. 2019. Multiple Human Association between Top and Horizontal Views by Matching Subjects' Spatial Distributions. In *arXiv*.

[21] Ruize Han, Jiewen Zhao, Wei Feng, Yiyang Gan, Liang Wan, and Song Wang. 2020. Complementary-View Co-Interest Person Detection. In *ACM MM*.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*.

[23] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. In *arXiv*.

[24] Kalun Ho, Janis Keuper, and Margret Keuper. 2020. Unsupervised multiple person tracking using autoencoder-based lifted multicuts. In *arXiv*.

[25] Martin Hofmann, Daniel Wolf, and Gerhard Rigoll. 2013. Hypergraphs for joint multi-view reconstruction and multi-object tracking. In *CVPR*.

[26] Yunzhong Hou, Liang Zheng, Zhongdao Wang, and Shengjin Wang. 2019. Locality Aware Appearance Metric for Multi-Target Multi-Camera Tracking. In *arXiv*.

[27] Shyamgopal Karthik, Ameya Prabhu, and Vineet Gandhi. 2020. Simple unsupervised multi-object tracking. In *arXiv*.

[28] Saad M Khan and Mubarak Shah. 2006. A multiview approach to tracking people in crowded scenes using a planar homography constraint. In *ECCV*.

[29] Harold W. Kuhn. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly* 2, 1 (1955), 83–97.

[30] Zihang Lai and Weidi Xie. 2019. Self-supervised learning for video correspondence flow. In *BMVC*.

[31] Laura Leal-Taixe, Gerard Pons-Moll, and Bodo Rosenhahn. 2012. Branch-and-price global optimization for multi-view multi-target tracking. In *CVPR*.

[32] Laura Lealtaixé, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. 2015. MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking. In *arXiv*.

[33] Minxian Li, Xiatian Zhu, and Shaogang Gong. 2018. Unsupervised person re-identification by deep learning tracklet association. In *ECCV*.

[34] Minxian Li, Xiatian Zhu, and Shaogang Gong. 2019. Unsupervised tracklet person re-identification. *IEEE TPAMI* 42, 7 (2019), 1770–1782.

[35] Xiaobai Liu, Yuanlu Xu, Lei Zhu, and Yadong Mu. 2017. A stochastic attribute grammar for robust cross-view human tracking. *IEEE TCSVT* 28, 10 (2017), 2884–2895.

[36] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. 2020. HOTA: A Higher Order Metric for Evaluating Multi-Object Tracking. *IJCV* 129, 2 (2020), 1–31.

[37] Andrii Maksai, Xinchao Wang, Francois Fleuret, and Pascal Fua. 2017. Non-markovian globally consistent multi-object tracking. In *ICCV*.

[38] Jinlong Peng, Yueyang Gu, Yabiao Wang, Chengjie Wang, Jilin Li, and Feiyue Huang. 2020. Dense Scene Multiple Object Tracking with Box-Plane Matching. In *ACM MM*.

[39] Bryan James Prosser, Shaogang Gong, and Tao Xiang. 2008. Multi-camera Matching using Bi-Directional Cumulative Brightness Transfer Functions. In *BMVC*.

[40] Ergys Ristani, Francesco Solera, Roger S Zou, Rita Cucchiara, and Carlo Tomasi. 2016. Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. In *CVPR*.

[41] Ergys Ristani and Carlo Tomasi. 2018. Features for Multi-Target Multi-Camera Tracking and Re-Identification. In *CVPR*.

[42] Arnold WM Smeulders, Dung M Chu, Rita Cucchiara, Simone Calderara, Afshin Dehghan, and Mubarak Shah. 2013. Visual tracking: An experimental survey. *IEEE TPAMI* 36, 7 (2013), 1442–1468.

[43] Yonatan Tariku Tesfaye, Eyasu Zemene, Andrea Prati, Marcello Pelillo, and Mubarak Shah. 2019. Multi-target tracking in multiple non-overlapping cameras using fast-constrained dominant sets. *IJCV* 127, 9 (2019), 1303–1320.

[44] Gaoang Wang, Yizhou Wang, Haotian Zhang, Renshu Gu, and Jenq-Neng Hwang. 2019. Exploit the Connectivity: Multi-Object Tracking with TrackletNet. In *ACM MM*.

[45] Sibo Wang, Ruize Han, Wei Feng, and Song Wang. 2021. Multiple Human Tracking in Non-Specific Coverage with Wearable Cameras. In *ICASSP*.

[46] Xiaolong Wang, Allan Jabri, and Alexei A Efros. 2019. Learning correspondence from the cycle-consistency of time. In *CVPR*.

[47] Zhongdao Wang, Jingwei Zhang, Liang Zheng, Yixuan Liu, Yifan Sun, Yali Li, and Shengjin Wang. 2020. CycAs: Self-supervised Cycle Association for Learning Re-identifiable Descriptions. In *ECCV*.

[48] Jialian Wu, Jiale Cao, Liangchen Song, Yu Wang, Ming Yang, and Junsong Yuan. 2021. Track to Detect and Segment: An Online Multi-Object Tracker. In *CVPR*.

[49] Jinlin Wu, Yang Yang, Hao Liu, Shengcai Liao, Zhen Lei, and Stan Z Li. 2019. Unsupervised graph association for person re-identification. In *ICCV*.

[50] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

[51] Yu Xiang, Alexandre Alahi, and Silvio Savarese. 2015. Learning to track: Online multi-object tracking by decision making. In *ICCV*.

[52] Jiarui Xu, Yue Cao, Zheng Zhang, and Han Hu. 2019. Spatial-temporal relation networks for multi-object tracking. In *ICCV*.

[53] Yuanlu Xu, Xiaobai Liu, Yang Liu, and Songchun Zhu. 2016. Multi-View People Tracking via Hierarchical Trajectory Composition. In *CVPR*.

[54] Yuanlu Xu, Xiaobai Liu, Lei Qin, and Song-Chun Zhu. 2017. Cross-view people tracking by scene-centered spatio-temporal parsing. In *AAAI*.

[55] Yihong Xu, Aljosa Osep, Yutong Ban, Radu Horaud, Laura Leal-Taixé, and Xavier Alameda-Pineda. 2020. How to train your deep multi-object tracker. In *CVPR*.

[56] Bo Yang and Ram Nevatia. 2012. Multi-target tracking by online learning of non-linear motion patterns and robust appearance models. In *CVPR*.

[57] Bo Yang and Ram Nevatia. 2012. An online learned CRF model for multi-target tracking. In *CVPR*.

[58] Amir Roshan Zamir, Afshin Dehghan, and Mubarak Shah. 2012. GMCP-Tracker: Global Multi-Object Tracking Using Generalized Minimum Clique Graphs. In *ECCV*.

[59] Richard Zhang, Phillip Isola, and Alexei A Efros. 2016. Colorful image colorization. In *ECCV*.

[60] Jiewen Zhao, Ruize Han, Yiyang Gan, Liang Wan, Wei Feng, and Song Wang. 2020. Human Identification and Interaction Detection in Cross-View Multi-Person Videos with Wearable Cameras. In *ACM MM*.

[61] Kang Zheng, Xiaochuan Fan, Yuewei Lin, Hao Guo, and Song Wang. 2017. Learning View-Invariant Features for Person Identification in Temporally Synchronized Videos Taken by Wearable Cameras. In *ICCV*.

[62] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. 2020. Tracking objects as points. In *ECCV*.