Identity-Quantity Harmonic Multi-Object Tracking

Yuhang He, Xing Wei, Xiaopeng Hong, Wei Ke, *Member, IEEE*, and Yihong Gong, *Fellow, IEEE*

***Abstract*— The *data association* problem of multi-object tracking (MOT) aims to assign IDentity (ID) labels to detections and infer a complete trajectory for each target. Most existing methods assume that each detection corresponds to a unique target and thus cannot handle situations when multiple targets occur in a single detection due to detection failure in crowded scenes. To relax this strong assumption for practical applications, we formulate the MOT as a Maximizing An Identity-Quantity Posterior (MAIQP) problem on the basis of associating each detection with an identity and a quantity characteristic and then provide solutions to tackle two key problems arising. Firstly, a local target quantification module is introduced to count the number of targets within one detection. Secondly, we propose an identity-quantity harmony mechanism to reconcile the two characteristics. On this basis, we develop a novel Identity-Quantity HArmonic Tracking (IQHAT) framework that allows assigning multiple ID labels to detections containing several targets. Through extensive experimental evaluations on five benchmark datasets, we demonstrate the superiority of the proposed method.**

***Index Terms*— Multi-object tracking, maximizing an identity-quantity posterior, identity-quantity reconciliation.**

Fig. 1. Multi-object tracking in crowded scenes. Each bounding box denote a target trajectory and the number upon each bounding box denotes the trajectory's ID number. MOT remains an extremely challenging task for frequent target interactions and occlusions. The images are from the MOT20 dataset.

I. Introduction

MULTI-OBJECT Tracking (MOT) aims to locate the positions of interested targets, maintain their identities across frames and infer a complete trajectory for each target over time. It has a wide range of applications in video surveillance [1], [2], crowd behavior analysis [3], [4], pedestrian monitoring [5]–[7], etc. MOT remains a very challenging task due to severe background clutter, object occlusions, target interactions, and so forth. This is especially true for tracking targets in very crowded scenes as illustrated in Figure 1.

Recently, with the rapid progress in object detection [8]–[10], the *tracking-by-detection* paradigm [11]–[15] has become the mainstream of MOT. Taking detected bounding boxes as input, its core problem is to assign IDentity (ID) labels to different detections and infer a complete trajectory for each target over time, *i.e.*, the *data association* problem. Depending on whether solving the data association problem using a batch of images, the MOT methods can be further divided into two subcategories: *offline* and *online* methods. Offline MOT methods [16]–[22] solve the data association problem using a batch of frames (or even the entire sequence) and link detections into trajectories in a large temporal window. Thus, they are inappropriate for real-time tracking. To meet this challenge, online MOT methods [12], [15], [23], [24] perform data association frame by frame. Given existing trajectories and newly obtained detections at each time step, they match the trajectories with detections and then extend the trajectories accordingly [11], [15], [23]–[26].

These methods usually take the assumption that each detection corresponds to a unique target, and attempt to assign an "optimal" ID label to each detection by Maximizing A Posterior (MAP) [27]. However, in complex scenarios such as crowded groups, this assumption does not hold when multiple targets are covered by a single detection because of serious occlusions or complex interactions among targets. Finding the single most likely ID label in such cases leads to an identity competition among multiple targets, and results in tracking failures such as *missed trajectories* and *ID switches*. Figure 2 (a) shows an example, where two persons are contained in a single detection (*i.e.*, the blue bounding box) due to partial occlusion. Most existing MOT methods will compute an identity posterior and assign the most likely ID label (Person #2) to this detection. Thus, they will miss tracking the

Manuscript received September 14, 2021; revised January 10, 2022 and February 5, 2022; accepted February 5, 2022. Date of publication March 2, 2022; date of current version March 8, 2022. This work was supported in part by the National Key Research and Development Project of China under Grant 2020AAA0105600; in part by the National Natural Science Foundation of China under Grant U21B2048, Grant 62076195, Grant 62006182, and Grant 62006183; in part by the China Postdoctoral Science Foundation under Grant 2020M683489; and in part by the Fundamental Research Funds for the Central Universities under Grant xzy012020013. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Weiyao Lin. *(Corresponding author: Yihong Gong.)*

The authors are with the Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China (e-mail: hyh1379478@stu.xjtu.edu.cn; weixing@mail.xjtu.edu.cn; hongxiaopeng@mail.xjtu.edu.cn; wei.ke@mail.xjtu.edu.cn; ygong@mail.xjtu.edu.cn).

Digital Object Identifier 10.1109/TIP.2022.3154286

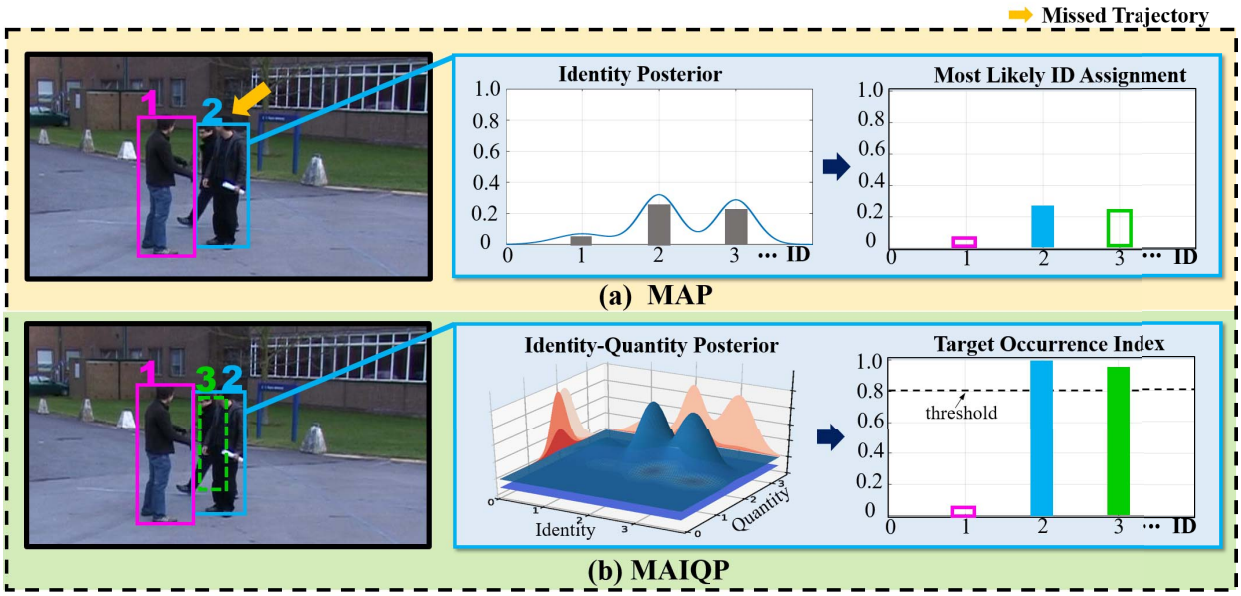


Fig. 2. Illustrations of (a) the MAP and (b) the MAIQP for multi-object tracking. The insert plate illustrates the ID assignment of the blue bounding box. The solid bounding box denotes the input detections and the dashed one denotes the detection recovered by the MAIQP method. The number upon each bounding box denotes the ID label and the yellow arrow points to missing tracking.

occluded one (Person #3) pointed by the yellow arrow. Even worse, as Persons #2 and #3 have similar ID posterior probabilities, they have to compete for the ID assignment intensely in tracking. As a result, any slight probability perturbation may lead to ID switches.

To address these problems, we associate each detection response with an identity and a quantity characteristic and formulate the MOT as a Maximizing An Identity-Quantity Posterior (MAIQP) problem. Nonetheless, there are two key problems arising with this MAIQP formulation. First, as the quantity of targets is unknown and time-varying in practice, it is critical to determine how many ID labels should be assigned to each detection response. Second, such an MOT system will be defective if the results of target identification and quantification are contradictory. For the first problem, we introduce a local target quantification module to count the targets in each detection. For the second problem, we design an identity-quantity harmony module to reconcile these two characteristics. On this basis, we propose a novel Identity-Quantity HArmonic Tracking framework (IQHAT) that allows to assign multiple ID labels to detections containing several targets. The proposed method first estimates the target identity and quantity of each detection using a target identification and a local target quantification module, respectively, and then reconciles the target identity and quantity using an identity-quantity harmony module. At the inference stage, we compute a *target occurrence index* for each target, which is a non-negative number that indicates the occurrence of a target in a detection. Then we collect the ID labels according to the target occurrence indexes and infer a complete trajectory for each target. As shown in Figure 2 (b), the MAIQP first optimizes an identity-quantity posterior of the blue bounding box, and then calculates the target occurrence indexes of different targets. Both ID labels #2 and #3 (whose occurrence indexes are larger than a threshold) are attached to the detection and the missed trajectory of Person #3 is repaired by estimating a bounding box (denoted by the dashed bounding box).

We provide comprehensive evaluations and ablation studies on five benchmark datasets, *i.e.*, 2D MOT15, MOT16, MOT17, MOT20 and HiEve. The IQHAT achieves state-of-the-art performance on these datasets and ablation studies demonstrate the superiority and efficiency of the proposed method.

In summary, the main contributions of this paper include:

- We formulate MOT as a Maximizing An Identity-Quantity Posterior (MAIQP) problem on the basis of associating each detection with a quantity and an identity characteristic.
- We propose a novel Identity-Quantity Harmonic Tracking (IQHAT) framework that allows multiple ID assignment for detections containing several targets.
- We design an identity-quantity harmony module to reconcile the target identity and quantity.

II. Related Work

The tracking-by-detection paradigm has become the mainstream of multi-object tracking, which first detects interested objects using modern object detectors [8]–[10], [28]–[30] and then associates obtained detections into trajectories by data association. Depending on whether solving the data association problem using a batch of frames or frame-by-frame, the MOT methods can be further divided into two sub-categories: offline and online methods.

A. Offline Multi-Object Tracking

Offline MOT methods solve the data association problem using a batch of frames or even the entire sequence.

A few works [14], [17]–[20], [31] are based on graph models, where each node can be either a detection or a tracklet (a short trajectory), and MOT is accomplished by linking these nodes into trajectories. The methods in [17], [18], [20], [31] build graphs based on detections, where each node denotes a detection response and edges represent affinities between detections. Detections in these graphs are linked into trajectories using hierarchical clustering [20], hybrid data association [31], or heterogeneous association fusion [18]. Tang *et al.* [17] introduce lifted edges into a multicut problem, *i.e.*, cutting the graph into several sub-graphs and each sub-graph corresponds to a trajectory, and formulate the MOT as a minimum cost Lifted Multicut Problem (LMP). Wang *et al.* [14] propose a method named TrackletNet Tracker (TNT), where each node in the graph model is a tracklet. They collect these tracklets into trajectories by minimizing a clustering cost. The method in [32] develops a Tracklet-Plane Matching (TPM) algorithm to improve the tracking performance with noise detection results, where a detection reliability evaluation scheme is introduced to identify reliable ones during association. Similarly, LinkBox [33] proposes a box-plane matching method to track targets in dense scenarios.

There are also methods [22], [34] track multiple objects by finding the most likely tracking proposals. The method in [35], namely Multiple Hypothesis Tracking (MHT), enumerates multiple tracking hypothesis and selects the most likely ones by minimizing an assignment cost. Kim *et al*. [35] propose a bilinear LSTM (bLSTM) to learn a long-term appearance model, and solve the MOT problem using the MHT algorithm. Sheng *et al*. formulate the multiple hypothesis tracking as a Maximum Weighted Independent Set (MWIS) problem, and develop an iterative hypothesis updating mechanism to generate trajectories.

Besides, Keuper *et al*. propose a Correlation Co-Clustering (CCC) [36] algorithm for tracking, which combines bottom-up motion segmentation by with multi-object tracking by clustering bounding boxes in a unified framework. The method [37] formulates MOT as a Lifted disjoint paths Tracking (LifT) problem, where lifted edges are designed to provide path connective information. Wang *et al.* [38] design a Spatial-Temporal Point Process (STPP) to mask-out noisy detections. Zhang *et al*. [39] propose a Deep Tracklet Association (DTA) method to improve long-term tracking performance, where a motion evaluation and an appearance evaluation network are designed to learn robust representation of tracklets. The method [40] solve the data association by finding a network flow by minimizing certain association costs and Braso *et al*. [41] proposes a differentiable framework based on Message Passing Network (MPN) to solve the network-flow problem. The method [42] proposes a Multiplex Labeling Graph (MLG) for tracking and constructs a long short-term memory network for MLG optimization.

B. Online Multi-Object Tracking

In online multi-object tracking [23], [24], [43]–[46], the data association problem is solved in a frame-by-frame manner. There is abundant literature on the use of bipartite matching for online MOT [15], [23], [24], [43], [47]–[49]. The method Simple Online Real-time Tracking (SORT) [47] divides existing trajectories and newly obtained detections into two disjoint sets, and then match the trajectories with the detections using the Hungarian algorithm [26]. On this foundation, Wojke *et al*. [24] further introduce deep CNN model into the tracking paradigm to extract robust appearance representation of targets. Furthermore, the method [23] integrates Appearance, Motion and Interaction information (AMIR) using a recurrent neural network, where multiple cues of targets are integrated into a unified representation across long-term temporal dependencies. To handle target interactions, they design a target-centered occupancy grid, in which the target and its neighbor targets relative locations are modeled. Feeding the occupancy grid into an LSTM network, the target relative locations are regarded as an affinity measurement for data association. The method [50] proposes a Spatial-Temporal Attention Mechanism (STAM) to handle the interactions and occlusions between targets. It learns a visibility map of the target and infer a spatial attention map. Once the target is occluded, they estimate the occlusion status using the visibility map and control the appearance feature extraction based on the spatial attention map. The Tracklet Association Tracker (TAT) [51] jointly learns feature representation and data association in a unified framework, and directly outputs data association results using target representations. Xu *et al.* [48] propose a Spatial-Temporal Relation Network (STRN) for target similarity measurement, which combines temporal and spatial information and outputs a unified feature for each target. Baisa [52] design an MOT tracker based on Gaussian Mixture Probability Hypothesis Density (GM-PHD) filter, where the filter estimates the state and cardinality of targets. Besides, Fang *et al*. [53] propose a Recurrent Auto-regressive Network (RAN) to characterize target appearance and motion dynamics over time. Yuan *et al.* [54] propose an effective self-supervised learning-based tracker for visual tracking, where a multi-cycle consistency loss is proposed for feature learning and a similarity dropout strategy is developed to suppress the low-quality training samples. Ren *et al.* [55] provide a comprehensive overview of the Neural Architecture Search (NAS) researches, where detailed and comprehensive comparisons of representative NAS methods are provided. These self-supervised and NAS-based feature learning methods can be beneficial for the target representation for multi-object tracking.

The methods [15], [56]–[58] integrates single object trackers in MOT. Chu *et al.* [15] propose an Instance Aware Tracker (IAT) to integrates single-object tracker into MOT, where each tracker corresponds to one target and the single object trackers are iteratively updates during the tracking process. Zhu *et al*. [57] integrate single object tracking with data association in a unified framework, and propose a Dual Matching Attention Networks (DMAN) to suppress noisy detections during tracking. The method [58] combines Data Association with Single Object Tracking (DASOT) with a unified convolution network, which computes correlation features for all positions and simultaneously obtain correlation heatmaps for all the targets. SiamMOT [59] proposes

a region-based MOT method, where detections are linked into trajectories using Siamese tracker.

There are MOT methods [12], [44], [45] integrate object detectors with feature extractor in a unified network. Bergmann *et al*. [12] convert a detector into a Tracktor++, which exploits the bounding box regression head of the detector to predict the position target in the next frame. The CenterTrack [45] integrates a detector into the tracking framework and locates objects as points. It associates detections between adjacent frames by predicting target position offsets. FCS-Track [60] designs a Siamese-based network to predict target motion. The FAMNet method [11] integrates feature extraction, affinity estimation and multi-dimensional assignment into a single network. Selective JDE [61] follows a teacher-student paradigm to adapt the tracking model to unseen scenes by transductive interactive self-training. Liu *et al*. design a Graph Similarity Model (GSM) [62] that integrates target features and relations for similarity measurement in a unified model. Moreover, the method [63] uses a differentiable Deep Hungarian Net (DHN) for end-to-end MOT tracking.

All these methods, however, implicitly assume that each detection corresponds to a unique target and attempt to find a single ID label for each detection. This assumption brings difficulties in tracking targets in crowd scenarios and may easily results in missed trajectories and ID switches. In contrast, the proposed method jointly optimizes the target identity and quantity of each detection and allows to assign multiple ID labels to a single detection. This novel formulation not only decreases missed trajectories especially in crowded scenes but also alleviates ID switches.

It is worth mentioning that there are a couple of research studies such as Tracking-By-Counting (TBC) [46] to incorporate the density map into the tracking process. Nonetheless, the differences between these methods and the IQHAT are distinct. First, these methods are offline methods and they solve the data association problem by following a classical network flow formulation [16]. The IQHAT, on the contrary, is an online method and solves the data association problem following the novel MAIQP formulation. Second, these assume that each detection corresponds to one target and density maps are used as a count constrain to improve detection performance. While in IQHAT, density maps are used to quantify the target number in each detection and serve for solving the data association problem. To the best of our knowledge, the IQHAT is the first work that formulates MOT as a maximizing an identity-quantity posterior problem and allows to assign multiple ID labels to detections containing several targets.

III. Overview of the Proposed Method

In this section, we first briefly introduce the background and then formulate the proposed method.

A. Background and Motivation

Given an image sequence $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^{T}$ of length T, we use $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_T\}$ to denote the detection collection of the image sequence, where $\mathcal{X}_t$ is the detection set of the *t-th* frame, and $\mathcal{X}_t^i$ denote the *i-th* detection of $\mathcal{X}_t$. Each detection is represented by a 2D bounding box (x, y, w, h), where (x, y) is the coordinate of the bounding box center, and w and h are the width and height, respectively. Let $y_t^i \in \{1, \ldots, M\}$ be the ID label of a detection $\mathcal{X}_t^i$, where M is the number of targets, $\mathcal{Y}_t$ and $\mathcal{Y} = \{\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_T\}$ be the ID label collection of $\mathcal{X}_t$ and $\mathcal{X}$, respectively. Most existing MOT methods formulate multi-object tracking as a Maximizing A Posterior (MAP) problem [27], which can be written as:

$$\mathcal{Y}^* = \underset{\mathcal{Y}}{\operatorname{argmax}}\ P(\mathcal{Y}|\mathcal{X}), \tag{1}$$

where $P(\mathcal{Y}|\mathcal{X})$ is an identity posterior given detections $\mathcal{X}$. At the inference stage, each detection is assigned with the single most likely ID label according to the identity posterior. Detections attached with the same ID label are associated into a trajectory.

As we discussed in Section I, this formulation has the difficulties to track targets in crowded scenarios. When there are interactions or occlusions between targets, a single detection may contains multiple targets. Assigning the single most likely ID label to such detections will result in missed trajectories of occluded targets and/or frequent ID switches.

B. Identity-Quantity Posterior

To tackle this problem, we associate each detection with a quantity and an identity characteristics and formulate the multi-object tracking as a Maximizing An Identity-Quantity Posterior (MAIQP) problem. Specifically, let $c_t^i \in \mathbb{R}_{\geq 0}$ be the *local quantity number* of a detection $\mathcal{X}_t^i$, which is a non-negative variable that counts the number of targets occurs in $\mathcal{X}_t^i$. We denote by $\mathcal{C}_t$ and $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_T\}$ the local quantity number set of $\mathcal{X}_t$ and $\mathcal{X}$, respectively. Given the detection set $\mathcal{X}$, the MAIQP problem can be written as:

$$\max_{\mathcal{Y},\mathcal{C}}\ P(\mathcal{Y}, \mathcal{C}|\mathcal{X}), \tag{2}$$

where $P(\mathcal{Y}, \mathcal{C}|\mathcal{X})$ is an identity-quantity posterior given $\mathcal{X}$. The MAIQP aims to jointly optimize the target identity and quantity of each detection, which requires the prediction of identity and quantity of each detection as well as the reconciliation of these two characteristics. Pursuant to this formulation, we propose a novel Identity-Quantity HArmonic Tracking (IQHAT) framework that allows to assign multiple ID labels to a single detection.

Compared with state-of-the-art methods that optimize an identity posterior and attach the most likely ID label to each detection, the IQHAT takes both target identities and quantities into account and allows to assign multiple ID labels to detections containing several targets. This solution not only decreases missed trajectories of occluded targets but also alleviates the ID competition in crowded scenarios.

IV. Identity-Quantity Harmonic Multi-Object Tracking

Figure 3 depicts the proposed IQHAT framework, an online MOT method containing four major modules: 1) the target

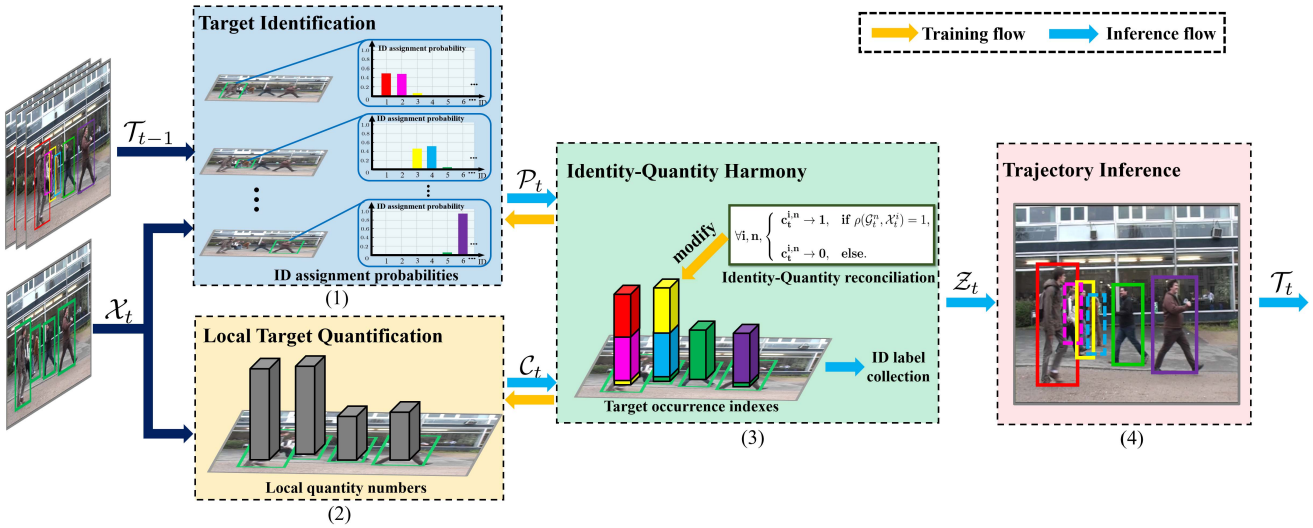


Fig. 3. Overview of the IQHAT framework. (1) Target identification module. (2) Local target quantification module. (3) Identity-quantity harmony module. (4) Trajectory inference module. The yellow and blue arrows indicate the training and inference flows, respectively.

identification module, 2) the local target quantification module, 3) the identity-quantity harmony module and 4) the trajectory inference module. At each time step t, the input to the framework is a detection set $\mathcal{X}_t$ and a collection of target trajectories $\mathcal{T}_{t-1} = \{\mathcal{T}_{t-1}^1, \mathcal{T}_{t-1}^2, \ldots, \mathcal{T}_{t-1}^{K_{t-1}}\}$ of K_{t-1} targets at time $t-1$, where $\mathcal{T}_{t-1}^n$ is the trajectory of target n. Module 1 takes $\mathcal{T}_{t-1}$ and $\mathcal{X}_t$ as input, and computes a set of ID assignment probabilities $P(y_t^i = n|\mathcal{X}_t, \mathcal{T}_{t-1}), n = 1, \ldots, M$, for each detection $\mathcal{X}_t^i$, where $P(y_t^i = n|\mathcal{X}_t, \mathcal{T}_{t-1})$ is the probability of assigning ID label n to $\mathcal{X}_t^i$. The output $\mathcal{P}_t$ is the collection of ID assignment probabilities of all the $|\mathcal{X}_t|$ detections. Module 2 estimates the number of targets in each detection $\mathcal{X}_t^i$, *i.e.*, the local quantity number c_t^i, using a local quantification algorithm, and outputs the local quantity number set $\mathcal{C}_t$ of $\mathcal{X}_t$. Taking the outputs from Modules 1 and 2, the third module computes target occurrence indexes $c_t^{i,n}, n = 1, \ldots, M$, for each detection $\mathcal{X}_t^i$ and collects an ID label set $\mathcal{Z}_t^i \subset \{1, \ldots, M\}$ for the detection according to the occurrence indexes. The output $\mathcal{Z}_t = \{\mathcal{Z}_t^i\}_{i=1}^{|\mathcal{X}_t|}$ collects the ID labels of the $|\mathcal{X}_t|$ detections. In the end, in accordance with the ID labels $\mathcal{Z}_t$, Module 4 infers an accurate and complete trajectory for each target and outputs updated trajectory set $\mathcal{T}_t$. It is worth noting that, in Module 3, if the ID assignment probability and the local target number of a detection are inconsistent, *e.g.* , it is estimated that there are two targets in a detection while only one ID is likely to be assigned according to the ID assignment probabilities, it is difficult to decide how many and which IDs should be assigned to the detection. To mitigate this discrepancy, we develop an identity-quantity reconciliation objective to jointly optimize the target identification and quantification modules. In the following, we will provide detailed descriptions of each module.

A. Target Identification

Given the existing trajectories $\mathcal{T}_{t-1}$ and newly obtained detections $\mathcal{X}_t$, this module aims to identify the targets contained in each detection and compute a probability of assigning the ID label n to a detection $\mathcal{X}_t^i$, *i.e.*, $P(y_t^i = n|\mathcal{X}_t, \mathcal{T}_{t-1})$.

The trajectory $\mathcal{T}_{t-1}^n$ of each target n is composed of a serious of tuples over a period of time:

$$\mathcal{T}_{t-1}^n = \{(k, a_k^n, b_k^n),\ k \in \tau_{t-1}^n\}, \tag{3}$$

where τ_{t-1}^n is the time index set of $\mathcal{T}_{t-1}$, a_k^n and b_k^n are the appearance feature and bounding box of the target n at time k, respectively. Taking the trajectory set $\mathcal{T}_{t-1}$ and detection set $\mathcal{X}_t$ as input, we first compute a pairwise similarity matrix $\mathbf{S} \in [0, 1]^{|\mathcal{T}_{t-1}| \times |\mathcal{X}_t|}$ between $\mathcal{T}_{t-1}$ and $\mathcal{X}_t$. Each element $\mathbf{S}(u, v)$ represents the similarity between the trajectory $\mathcal{T}_{t-1}^u$ and the detection $\mathcal{X}_t^v$:

$$\mathbf{S}(u, v) = \varphi_{app}(\mathcal{T}_{t-1}^u, \mathcal{X}_t^v) \cdot \varphi_{mot}(\mathcal{T}_{t-1}^u, \mathcal{X}_t^v), \tag{4}$$

where $\varphi_{app}(\cdot, \cdot)$ outputs a normalized cross-correlation of the trajectory-detection appearance features and $\varphi_{mot}(\cdot, \cdot)$ outputs an Intersection-over-Union (IoU) score based on target motion patterns. More specifically, the appearance similarity is calculated by a temporal weighted average of appearance feature correlations of $\mathcal{T}_{t-1}^u$ and $\mathcal{X}_t^v$:

$$\varphi_{app}(\mathcal{T}_{t-1}^u, \mathcal{X}_t^v) = \frac{\sum_{\forall k \in \tau_{t-1}^u} \left[e^{k-t} \cdot Corr(a_k^u, \phi(\mathcal{X}_t^v))\right]}{\sum_{\forall k \in \tau_{t-1}^u} e^{k-t}}, \tag{5}$$

where τ_{t-1}^u is the time index set of trajectory $\mathcal{T}_{t-1}^u$ and a_k^u is the appearance feature of the trajectory at time k. Note that as $k \leq t$, e^{k-t} is a real number in range $(0, 1]$. The motion similarity of trajectory $\mathcal{T}_{t-1}^u$ and detection $\mathcal{X}_t^v$ is computed based on IoU of bounding boxes:

$$\varphi_{mot}(\mathcal{T}_{t-1}^u, \mathcal{X}_t^v) = IoU(Pred(\mathcal{T}_{t-1}^u), \mathcal{X}_t^v), \tag{6}$$

where $Pred(\cdot)$ predicts a bounding box according to the input trajectory using target motion models, and $IoU(\cdot, \cdot)$ outputs the IoU score of the input detections.

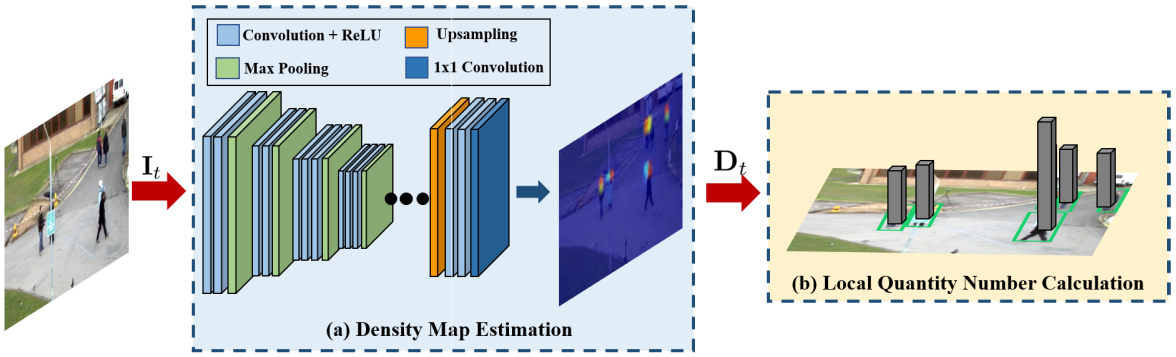


Fig. 4. Illustration of the local target quantification, where (a) a density map for the input image is estimated and (b) the local quantity number of each detection based on the density map is computed.

We then compute the ID assignment probability $P(y_t^i = n|\mathcal{X}_t, \mathcal{T}_{t-1})$ according to the similarity matrix $\mathbf{S}$:

$$P(y_t^i = n|\mathcal{X}_t, \mathcal{T}_{t-1}) = \frac{\mathbf{S}(n,i)}{||\mathbf{S}(:,i)||_1 + \varepsilon}, \tag{7}$$

where $||\cdot||_1$ outputs the L_1-norm of the input vector and $\varepsilon = 10^{-4}$ to avoid division by zero. We use $\mathcal{P}_t = \{P(y_t^i = n|\mathcal{X}_t, \mathcal{T}_{t-1}), \forall i, n\}$ to denote the ID assignment probabilities of $\mathcal{X}_t$. For computation efficiency, similar to [24], we filter out infeasible matching candidates, *i.e.*, setting $\mathbf{S}(u,v) = 0$, based on target movement consistency. In Section V-F, we conduct ablation studies to analyze the influence of using different appearance features and motion models in the trajectory-detection similarity calculation.

B. Local Target Quantification

As illustrated in Figure 4, the local target quantification module first computes a density map $\mathbf{D}_t$ for the input image frame $\mathbf{I}_t$ and then calculates the local quantity number c_t^i of each detection $\mathcal{X}_t^i$ by summing over sliced densities in the detection.

1) Density Map Estimation: Taking the *t-th* image frame $\mathbf{I}_t$ as input, the density map estimator produces a density map $\mathbf{D}_t \in \mathbb{R}_{\geq 0}^{W_D \times H_D}$ for the image, where W_D and H_D denote the width and height of the density map, respectively. Each element $\mathbf{D}_t(u,v)$ in $\mathbf{D}_t$ represents the number of targets at position (u,v). The total number of targets in the frame can be obtained by $\sum_{u=1}^{W_D}\sum_{v=1}^{H_D}\mathbf{D}_t(u,v)$. The density map estimator mainly contains a fully convolution network for backbone feature extraction and a regression header to produce density map. We take the VGG-19 [64] as the backbone of our density map estimator, where the fully-connected layers are removed. We upsample backbone features to 1/8 of the input image size by bilinear interpolation and then feed the features to a regression header, which consists of two 3×3 and a 1×1 convolutional layers, to generate density maps.

2) Local Quantity Number Calculation: To calculate the local quantity number c_t^i of each detection $\mathcal{X}_t^i$, a straightforward solution is to sum over the densities in the detection. However, as there is overlapping between different detections and a density value $\mathbf{D}_t(u,v)$ may be summed several times, the local quantity numbers of all the detections could be overestimated, *i.e.*, $\sum_{i=1}^{|\mathcal{X}_t|} c_t^i > \sum_{\forall u,v}\mathbf{D}_t(u,v)$.

To settle this problem, we first slice each density value $\mathbf{D}_t(u,v)$ to different detections and then sum over the sliced densities to calculate the local quantity number c_t^i. For a density value $\mathbf{D}_t(u,v)$, its sliced one $\mathbf{D}_t^i(u,v)$ of detection $\mathcal{X}_t^i$ is computed by:

$$\mathbf{D}_t^i(u,v) = \begin{cases} \dfrac{\mathbf{D}_t(u,v)}{\sum_{j=1}^{|\mathcal{X}_t|}\delta\big((u,v) \in \zeta(\mathcal{X}_t^j)\big)}, & \text{if } (u,v) \in \zeta(\mathcal{X}_t^i), \\ 0, & \text{else}, \end{cases} \tag{8}$$

where $\zeta(\cdot)$ outputs the point coordinate collection contained in the input detection, and $\delta(cond) = 1$ if $cond$ is true and equals zero otherwise. Then, the local quantity number c_t^i of detection $\mathcal{X}_t^i$ is obtained by summing over the sliced densities contained in detection $\mathcal{X}_t^i$:

$$c_t^i = \sum_{\forall (u,v) \in \zeta(\mathcal{X}_t^i)} \mathbf{D}_t^i(u,v). \tag{9}$$

We use $\mathcal{C}_t = \{c_t^i\}_{i=1}^{|\mathcal{X}_t|}$ to denote the local quantity number collection of the $|\mathcal{X}_t|$ detections.

C. Identity-Quantity Harmony

Taking the ID assignment probabilities $\mathcal{P}_t$ and the local quantity number set $\mathcal{C}_t$ as input, this module aims to collects an ID label set $\mathcal{Z}_t^i \subset \{1, \ldots, M\}$ for each detection $\mathcal{X}_t^i$. To mitigate the discrepancy between ID assignment probabilities and local quantity numbers, we proposes an identity-quantity reconciliation objective to jointly optimize and refine the target identification and quantification modules.

1) Identity-Quantity Reconciliation: Considering a detection containing L targets, we use I_l $(l = 1, \ldots, L)$ to denote the ID labels of these targets. The probability of assigning an ID label I_l to the detection is $P(I_l) = 1/L$. On the other hand, the probability of assigning the ID label of a non-occurred target q $(q \neq I_l)$ is $P(q) = 0$. By multiplying the ID assignment probability $P(n)$ with the target number L, we obtain a binary number $o^n \in \{0, 1\}$ of each target n:

$$o^n = P(n) \cdot L, \tag{10}$$

where if the target n appears in the detection, we have $o^n = 1$. On the contrary, we have $o^n = 0$.

The above observation suggests that the number o^n indicates the occurrence of a target n in the detection. Generalizing the ideal $P(n)$ and L to float values for practice, we calculate a target occurrence index $c_t^{i,n} \in \mathbb{R}_{\geq 0}$ by multiplying the ID assignment probability $P(y_t^i = n|\mathcal{X}_t, \mathcal{T}_{t-1})$ with the local quantity number c_t^i:

$$c_t^{i,n} = P(y_t^i = n|\mathcal{X}_t, \mathcal{T}_{t-1}) \cdot c_t^i, \tag{11}$$

where $c_t^{i,n}$ indicates the occurrence of a target n in the detection $\mathcal{X}_t^i$, *i.e.*, $c_t^{i,n} \to 1$ represents the target n is very likely to appear in $\mathcal{X}_t^i$ and $c_t^{i,n} \to 0$ for the opposite.

The target occurrence index $c_t^{i,n}$ blends the target identity and quantity characteristics into a unified one, which enables us to jointly optimize the ID assignment probabilities and the local quantity number by modifying $c_t^{i,n}$. Specifically, at the training stage, given the annotation set $\mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_T\}$ of the image sequence $\mathcal{I}$, where $\mathcal{G}_t = \{\mathcal{G}_t^n\}_{n=1}^{N_t}$ is the annotation collection of N_t targets at time t and $\mathcal{G}_t^n$ is the ground-truth bounding box of target n. For a detection $\mathcal{X}_t^i$, if a target n occurs in the detection, we should have $c_t^{i,n} = 1$. On the contrary, if the target does not appear, we expect $c_t^{i,n} = 0$. Transforming this objective into equations, we have an identity-quantity reconciliation objective:

$$\begin{aligned} L_{rec}(\mathcal{X}_t) &= \sum_{i=1}^{|\mathcal{X}_t|} \sum_{n=1}^{N_t} ||c_t^{i,n} - \rho(\mathcal{G}_t^n, \mathcal{X}_t^i)||, \\ \rho(\mathcal{G}_t^n, \mathcal{X}_t^i) &= \begin{cases} 1, & \text{if } IoU(\mathcal{G}_t^n, \mathcal{X}_t^i) \geq \theta \text{ and} \\ & \underset{m}{\text{argmax}}(IoU(\mathcal{G}_t^m, \mathcal{X}_t^i)) = n, \\ 0, & \text{else.} \end{cases} \end{aligned} \tag{12}$$

where $\rho(\mathcal{G}_t^n, \mathcal{X}_t^i)$ takes binary values of 1 or 0, which corresponds to the occurrence and non-occurrence of a target n in the detection $\mathcal{X}_t^i$, respectively, and $IoU(\cdot, \cdot)$ outputs the IoU score of input detections. θ is a IoU threshold and we empirically set $\theta = 0.5$ in line with [8]. On this basis, we can jointly reconcile the identification and quantification modules using Eq. (12). In Section V-G, we conduct ablation studies to study the effectiveness of the identity-quantity reconciliation, where the identification and quantification modules of our counterpart are trained separately. Experimental results show that the joint training of the identification and quantification modules using the proposed objective function can steadily improve the tracking performance.

2) ID Label Collection: At the inference stage, we collect an ID label set $\mathcal{Z}_t^i \subset \{1, \ldots, M\}$ for each detection $\mathcal{X}_t^i$ according to the target occurrence index $c_t^{i,n}$:

$$\mathcal{Z}_t^i = \{n \mid \ \forall n = 1, \ldots, M, \ \text{if } c_t^{i,n} \geq \alpha\}, \tag{13}$$

where α is a threshold. The influence of α is studied in Section V-E and we set $\alpha = 0.8$ according to the experiment results. By introducing the target identity and quantity to multi-object tracking and allowing multiple ID labels to a single detection, the proposed method is capable of different detection conditions: 1) when the detection $\mathcal{X}_t^i$ contains no targets (*i.e.*, a false positive detection), the ID label set $\mathcal{Z}_t^i$ is an empty one; 2) when $\mathcal{X}_t^i$ contains one target, $\mathcal{Z}_t^i$ consists a single ID label; 3) when $\mathcal{X}_t^i$ contains several targets, $\mathcal{Z}_t^i$ is composed of multiple ID labels of these targets.

D. Trajectory Inference

This module aims to infer an accurate and complete trajectory for each target. We infer target trajectories according to the status of ID label set $\mathcal{Z}_t^i$. First, if $|\mathcal{Z}_t^i| = 0$, the detection $\mathcal{X}_t^i$ corresponds to a new occurred target n or a false positive detection. According to the local quantity number c_t^i, we initialize a trajectory $\mathcal{T}_t^n$ for the new target n or neglect the detection $\mathcal{X}_t^i$ as a false positive detection:

$$\begin{cases} \mathcal{T}_t^n = \{(t, \phi_{app}(\mathcal{X}_t^i), \phi_{bbox}(\mathcal{X}_t^i))\}, & \text{if } c_t^i \geq \beta, \\ \text{neglect } \mathcal{X}_t^i, & \text{else,} \end{cases} \tag{14}$$

where β is a threshold, $\phi_{app}(\cdot)$ extracts appearance feature using a feature extractor and $\phi_{bbox}(\cdot)$ outputs the bounding box of the input detection. The influence of β is studied in Section V-E and we set $\beta = 1.1$ according to experiment results.

Second, if $\mathcal{Z}_t^i$ contains a single ID label n, we update the trajectory $\mathcal{T}_{t-1}^n$ of the target n with the detection:

$$\mathcal{T}_t^n = \mathcal{T}_{t-1}^n \cup \{(t, \phi_{app}(\mathcal{X}_t^i), \phi_{bbox}(\mathcal{X}_t^i))\}. \tag{15}$$

Thirdly, $\mathcal{Z}_t^i$ contains multiple ID labels of the targets. For each target n contained in $\mathcal{Z}_t^i$, we first estimate a bounding box b_t^n for the target and then update the target trajectory using b_t^n:

$$b_t^n = \eta(\mathcal{T}_{t-1}^n, \mathcal{X}_t^i), \tag{16}$$

$$\mathcal{T}_t^n = \mathcal{T}_{t-1}^n \cup \{(t, \phi_{app}(b_t^n), \phi_{bbox}(b_t^n))\}, \tag{17}$$

where $\eta(\cdot, \cdot)$ estimates a bounding box based on the input trajectory and detection by motion prediction. The influence of using different motion models is studied in Section V-F, based on which we adopt the Kalman filter [65] for a good trade-off between the accuracy and speed.

The entire algorithm of IQHAT is summarized in Algorithm 1. Taking detection collection $\mathcal{X}$ and image sequence $\mathcal{I}$ as input. At steps 4-8, according to ID assignment probabilities $P(y_t^i = n|\mathcal{X}_t, \mathcal{T}_{t-1})$ and local quantity numbers c_t^i, we compute a target occurrence index $c_t^{i,n}$ for each target n. During training (*i.e.*, steps 9-10), we jointly optimize the target identification module and the local target quantification module using the identity-quantity reconciliation objective. At inference stage (*i.e.*, steps 12-17), we first collect an ID label set $\mathcal{Z}_t^i$ for each detection and then infer target trajectories according to the assigned ID labels.

V. EXPERIMENT

In this section, we compare the proposed IQHAT with state-of-the-art methods on five benchmark datasets and then provide complementary experiments to analyze the efficiency and robustness of different components.

Algorithm 1 Identity-Quantity Harmonic Multi-Object Tracking

Data: Image sequence $\mathcal{I}$; Detection collection $\mathcal{X}$; Ground-truth annotation $\mathcal{G}$ for *training*;
Result: Trajectory set $\mathcal{T}$.

$\mathcal{T} = \emptyset$;
for $t = 1 : T$ **do**
  $\mathcal{T}_t = \emptyset$;
  for $\mathcal{X}_t^i \in \mathcal{X}_t$ **do**
    Compute the ID assignment probability $P(y_t^i = n|\mathcal{X}_t, \mathcal{T}_{t-1})$ of each target n using Eq. (7);
    Compute the local quantity number c_t^i using Eq. (9);
    Measure the target occurrence index $c_t^{i,n}$ of each target n using Eq. (11);
  end
  if *training* **then** // training stage
    Jointly optimize the target identification and quantification modules using Eq. (12);
  else // inference stage
    Collect ID label set $\mathcal{Z}_t = \{\mathcal{Z}_t^i\}_{i=1}^{|\mathcal{X}_t|}$ of the $|\mathcal{X}_t^i|$ detections using Eq. (13);
    for $\mathcal{Z}_t^i \in \mathcal{Z}_t$ **do**
      Initialize or update the trajectory $\mathcal{T}_t^n$ of target n using Eq. (14)-Eq. (17);
      $\mathcal{T}_t \leftarrow \mathcal{T}_t^n$;
    end
    $\mathcal{T} \leftarrow \mathcal{T}_t$;
  end
end

A. Dataset

We use five public benchmark datasets, *i.e.* , 2D MOT15, MOT16, MOT17, MOT20 and HiEve, for performance evaluation.

1) 2D MOT15 [67]: The 2D MOT15 dataset contains 11 sequences in the training set and 11 sequences in the test set, including a total number of 11,283 frames of 1,221 targets. The frame rates vary from 7 to 30 frames per second (FPS), and image resolutions range from 640×480 to 1920×1080. The dataset provides public available detection results generated by the ACF [68] detector.

2) MOT16 and MOT17 [69]: The MOT17 dataset contains 14 sequences, where 7 are used for training and the others for testing. It captures 33,741 image frames for 3,828 targets, and provides detection results from three detectors, *i.e.* , DPM [70], Faster-RCNN [8], and SDP [71]. The frame rates are in range from 14 to 30 FPS and the resolutions of most sequences are 1920×1080. The MOT16 contains the same sequences as MOT17 and provides detection results from DPM [70].

3) MOT20 [72]: The MOT20 captures 8 sequences from 3 very crowded scenarios, where 4 sequences are used for training and the others are used for testing. It contains a total number of 13,410 frames of 3,833 targets and captures more than 200 targets in each frame to evaluate tracking performance in extremely crowded scenarios. The resolution ranges from 1173×800 to 1920×1080 and the frame rates are 25 FPS. It provides public detection results using Faster-RCNN [8].

4) HiEve [73]: This dataset contains 32 video sequences collected from 9 challenging scenarios, including airport, dining hall, indoor, jail, mall, square, school, station, and street. 19 video sequences are used for training and the rest 13 are used for testing. The dataset provides public detection results using the Faster R-CNN [8] detector.

For fair comparisons, we take the public detection results as the input of our tracking framework.

B. Evaluation Metrics

We adopt two widely used metrics, *i.e.*, the MOTA↑ [74] and the IDF1↑ [75], for evaluation. Besides, we additionally report the mostly tracked (MT↑) and mostly lost (ML↓) and the number of ID switches (IDS↓) [76]. Finally, we provide the processing speed (FPS↑) to analyze the computation efficiency. Here, "↑" indicates the higher score is better and "↓" is the opposite.

- *MOTA*↑: Multi-Object Tracking Accuracy. This metric combines the number of false positives, missed targets and identity switches and measures the trajectory coverage of each target.
- *IDF1*↑: ID F1 score. This metric evaluates the ratio of correctly identified detections w.r.t. the ground-truth and measures the trajectory consistency of each target. Normalized in range [0,100%].
- *MT*↑: The ratio of mostly tracked targets (covered by trajectories by more than 80%). Ranged in [0,100%].
- *ML*↓: The ratio of mostly lost targets (covered by trajectories by less than 20%). Ranged in [0,100%].
- *IDS*↓: The total number of ID switches.
- *FPS*↑: Frame per second. Evaluate the processing speed of trackers.

C. Implementation Details

Network Architecture: The backbone of our appearance feature extractor is a ResNet50 [77] network, which is pre-trained on the CUHK03 [78] using a triplet loss [79]. We resize input images to 256×128 and take the output of the second fully-connected layer as the target appearance feature. We employ the Kalman filter [65] to model the target motions for a good trade-off between speed and accuracy. As described in Section IV-B, the density map estimator takes VGG-19 [64] as backbone and is pre-trained on the QNRF dataset [80] using a Bayesian loss [81]. **Training details.** The IQHAT tracker is trained on the training set of MOT datasets. The components of IQHAT (appearance feature extractor, motion models and density map estimator) are jointly optimized and refined using the identity-quantity reconciliation objective, where the networks are optimized using Adam optimizer [82] with an initial learning rate of 10^{-5}. The meta-parameter θ in Eq. (12) is set to $\theta = 0.5$ in accordance with [12]. **Inference details.** We set $\alpha = 0.8$ and $\beta = 1.1$ according to

TABLE I
EVALUATION RESULTS ON THE BENCHMARK DATASETS

Dataset	Method	Type	MOTA↑ (%)	IDF1↑ (%)	MT↑ (%)	ML↓ (%)	IDS ↓	FPS ↑
2D MOT15	MHT [35]	offline	32.4	45.3	16.0	43.8	435	0.7
	CCC [36]	offline	35.6	45.1	23.2	39.3	457	0.6
	TPM [32]	offline	36.2	43.6	15.4	42.6	420	0.8
	MPN [41]	offline	51.5	58.6	31.2	25.9	*375*	*6.5*
	LifT [37]	offline	*52.5*	*60.0*	*33.8*	*25.8*	730	1.5
	IAT [15]	online	38.9	44.5	16.6	31.5	720	0.3
	AMIR [23]	online	37.6	46.0	15.8	26.8	1026	1.9
	STRN [48]	online	38.1	46.6	11.5	33.4	1033	13.8
	DHN [63]	online	44.1	46.0	17.2	26.6	1347	1.6
	Tracktor++ [12]	online	46.6	47.6	18.2	27.9	1290	1.4
	Baseline	online	42.8	45.9	15.4	32.5	1047	**20.9**
	Ours	online	**48.7**	**58.4**	**29.3**	**23.4**	**567**	11.6
MOT16	MHT [35]	offline	45.8	46.1	16.2	43.2	590	0.8
	LMP [17]	offline	48.8	51.3	18.2	40.1	481	0.5
	MWIS [66]	offline	48.7	55.3	15.7	44.5	413	4.8
	TNT [14]	offline	49.2	56.1	17.3	40.3	606	0.7
	DTA [39]	offline	54.9	63.1	24.4	38.1	1088	2.5
	MPN [41]	offline	58.6	61.7	*27.3*	34.0	*354*	*6.5*
	LifT [37]	offline	*61.3*	*64.7*	27.0	*34.0*	389	0.5
	RAN [53]	online	45.9	48.8	13.2	41.9	648	0.9
	STAM [50]	online	46.0	50.0	14.6	43.6	473	0.2
	DMAN [57]	online	46.1	54.8	17.4	42.7	1616	0.5
	DASOT [58]	online	46.1	49.4	14.6	41.6	2057	9.0
	AMIR [23]	online	47.2	46.3	14.0	41.6	774	1.0
	IAT [15]	online	48.8	47.2	15.8	38.1	906	0.1
	STRN [48]	online	48.5	53.9	17.0	34.9	747	13.5
	DHN [63]	online	54.8	53.4	19.1	37.0	645	1.6
	Tracktor++ [12]	online	56.2	54.9	20.7	35.8	1068	1.6
	GSM [62]	online	57.0	58.2	22.0	34.5	475	7.6
	Baseline	online	50.7	51.8	18.2	37.8	954	**14.0**
	Ours	online	**58.6**	**62.4**	**26.1**	**36.5**	**370**	7.9
MOT17	bLSTM [21]	offline	47.5	51.9	18.2	41.7	2069	1.9
	MWIS [66]	offline	50.6	56.5	17.6	43.4	1407	2.6
	TNT [14]	offline	51.9	58.1	23.1	35.5	2288	0.7
	DTA [39]	offline	54.9	63.1	24.4	38.1	*1088*	2.5
	MPN [41]	offline	58.8	61.7	*28.8*	*33.5*	1185	*6.5*
	LifT [37]	offline	*60.5*	*65.6*	27.0	33.6	1189	0.7
	DMAN [57]	online	48.2	55.7	19.3	38.3	2194	0.5
	GM-PHD [52]	online	49.6	45.2	18.9	**33.1**	5567	1.2
	DASOT [58]	online	49.5	51.8	20.4	34.6	4142	9.1
	STRN [48]	online	50.9	56.0	18.9	33.8	2397	**13.8**
	FAMNet [11]	online	52.0	48.7	19.1	33.4	3072	0.6
	DHN [63]	online	53.7	53.8	19.4	36.6	1947	4.9
	Tracktor++ [12]	online	56.3	55.1	21.1	35.3	1987	1.5
	GSM [62]	online	56.4	57.8	22.2	34.5	1485	8.7
	Baseline	online	51.9	53.1	19.2	37.5	2638	12.9
	Ours	online	**58.4**	**61.8**	**24.1**	35.2	**1262**	8.1
MOT20	MLG [42]	offline	48.9	54.6	30.9	*22.1*	2187	–
	TBC [46]	offline	54.5	50.1	33.4	19.7	2449	5.9
	MPN [41]	offline	*57.6*	*59.1*	*38.2*	22.5	*1210*	6.5
	SORT [47]	online	42.7	45.1	16.7	26.2	4470	**57.3**
	Tracktor++ [12]	online	52.6	52.7	29.4	26.7	4374	1.2
	Baseline	online	44.3	46.9	20.8	33.2	6486	10.5
	Ours	online	**57.1**	**57.7**	**40.8**	**20.0**	**1875**	7.4

the experiment results in Section V-E. Based on the ablation study in Section V-F, the Kalman filter [65] is used to model target motions and recover the missed trajectories of occluded targets.

D. Quantitative Comparisons With State-of-the-Art Methods

In Table I, we compare the IQHAT method with baseline and state-of-the-art methods on the benchmark datasets mentioned in Section V-A. To investigate the efficiency of the MAIQP, we employ a typical MOT method, i.e., DeepSort [24], as the *Baseline*, which tracks targets using both appearance and motion cues and assigns a unique ID label to each detection using the Hungarian algorithm [26]. For fair comparison, the *Ours* uses the same appearance and motion models as the ones of *Baseline*, while solves the data association problem following the MAIQP formulation and allows to assign multiple ID labels to detections containing plural targets. From Table I, we make the following important observations.

1) Comparisons With Online Methods: The proposed method achieves state-of-the-art performance on the five benchmark datasets and significantly outperforms the other online MOT trackers by a large margin (at least 1.6 and 3.2 advances on MOTA and IDF1, respectively). Particularly, compared to the representative Tracktor++ method [12] that assigns a single ID label to each detection, the proposed method dramatically decreases the IDS number (567 vs. 1290, 370 vs. 1068, 1262 vs. 1987, 1875 vs. 4374 on 2D MOT15, MOT16, MOT17 and MOT20, respectively) by allowing to

Fig. 5. Qualitative tracking examples of our proposed method in crowded scenarios. From top to bottom: PETS09-S2L2, MOT17-03, MOT20-04 and MOT20-06. The MOT20-04 and MOT20-06 are very crowded scenarios with about 200 targets per image.

assign plural ID labels to each detection and achieves more than 5.0 improvement on the IDF1 score. Especially, on the crowded MOT20 dataset, the performance of Tracktor++ dramatically decrease due to the frequent occlusions and interactions between targets. By allowing to assign multiple ID labels to detections and infer an accurate a complete trajectory for each target, the proposed method achieves the best performance in all metrics. This is a strong evidence that the IQHAT is very effective for handling target occlusion and interaction in crowded scenarios. We provide qualitative tracking results in very crowded scenes in Figure 5.

2) Comparisons With Offline Methods: Though the proposed method is an online method and solve the data association problem between consecutive frames, it achieves highly competitive results with the offline ones. The main reason is that, the IQHAT takes both the identity and quantity into account during tracking and allows to assign multiple IDs to detections. As we discussed in Section I, this not only decreases missed trajectories of targets but also alleviates ID competition in tracking. For example, the state-of-the-art offline MPN [41] method formulates MOT as a network flow problem, which aims to find a global optimal linkage for each target. However, when multiple targets occurred in a single detection, only one linkage between the detection and other detections can be established, this will lead to missed trajectories of the occluded targets and ID switches during tracking. Consequently, we can see the proposed method achieves highly competitive results with MPN and even outperforms the MPN tracker on the crowded MOT20 dataset in terms of MT and ML.

3) Comparisons With Baseline: Compared to the *Baseline* method that assigns a single ID label to each detection, the proposed method dramatically decreases the IDS number by allowing multiple ID labels assigning to a single detection, and significantly improves the tracking performance by a large margin (at least 5.9, 8.7, 4.9 and 2.3 improvements on MOTA, IDF1, MT and ML, respectively). In the crowded MOT20 dataset, we can see that the proposed method improves the tracking performance by a large margin, where the MOTA and IDF1 achieve about 13 points improvement and the IDS number is reduced by three quarters. To get more insight of the differences between *Baseline* and *Ours*, we show tracking examples of them in Figure 6. We can see that 1) there are missed trajectories and frequent ID switches in *Baseline*

Fig. 6. Tracking examples of the *Baseline* and *Ours*. The number upon each bounding box denotes ID label. The yellow and red arrows point to missed trajectories and ID switches, respectively.

TABLE II
EVALUATION RESULTS ON THE HIEVE DATASET

Method	Type	MOTA↑ (%)	IDF1↑ (%)	MT↑ (%)	ML↓ (%)	IDS ↓
TPM [32]	offline	33.6	37.7	10.7	31.2	4287
STPP [38]	offline	37.5	40.2	20.4	29.8	4536
MPN [41]	offline	47.9	**53.3**	**33.6**	**23.8**	**1243**
LinkBox [33]	offline	**51.4**	47.2	29.3	29.1	1725
DeepSort [24]	online	27.1	28.6	8.5	41.4	2220
CenterTrack [45]	online	31.1	41.8	8.6	27.9	2767
GMPHD-ReId [52]	online	31.3	37.7	**36.0**	**24.3**	4392
FCS-Track [60]	online	47.8	49.8	25.3	30.2	1658
Selective JDE [61]	online	50.6	56.8	25.1	30.3	1719
SiamMOT [59]	online	53.2	51.7	26.7	27.5	1308
Ours	online	**53.4**	**57.4**	28.7	28.7	**1011**

(pointed by the yellow and red arrows, respectively) due to the singular ID assignment. 2) By allowing multiple ID labels assigning to a single detection and repairing missed trajectories of occluded targets, the tracking failures in *Baseline* are corrected by the *Ours*. In scene 1, the *Ours* assigns ID labels #3 and #4 to the red detection in frame 55. In scene 2, it assigns ID labels #2 and #3 to the deep green detection in frame 175, and ID labels #1 and #2 to the pink detection in frame 193. Missed trajectories of occluded targets are repaired by bounding box repairing (the dashed ones).

4) Runtime Performance: We provide the processing speed of our tracker in Table I, where the speed is tested on a desktop with an Intel I7 CPU and an NVIDIA 2080 Ti GPU. For fair comparisons, we follow the comparison settings in other methods such as [12], [41], [48], [62] and take the whole system (all the four modules) into account for speed evaluation. We can see that, 1) as our proposed method requires neither complicated occlusion reasoning nor very deep target features, it achieves a competitive running speed of about 8 FPS and ranked #3 against the other representative trackers. 2) Compared with the *Baseline* method, the proposed method is slightly slower than *Baseline* (taking about 0.05s additional per image) while significantly improves the tracking accuracy by a large margin (at most 13.5 and 12.8 improvements on MOTA and IDF1, respectively).

We provide experimental results on the HiEve dataset in Table II. The evaluation results are provided by the online evaluation system.[1] We can see that the proposed method achieves the top performance on the challenging HiEve dataset in terms of the MOTA and IDF1 metrics. Specifically, compared with the competitive SiamMOT method, the proposed method outperforms SiamMOT on MOTA (0.2%) and

[1] http://humanineventsorg/

TABLE III
INFLUENCE OF α AND β

α	0.4	0.6	0.8	1.0	1.2	1.4	1.6
MOTA(%)↑	70.6	72.4	**74.3**	73.7	71.3	70.8	67.5
IDF1(%)↑	71.0	73.2	**75.8**	74.8	73.1	71.8	68.1
β	0.8	0.9	1.0	1.1	1.2	1.3	1.4
MOTA(%)↑	74.2	74.2	74.3	**74.4**	74.0	71.4	68.5
IDF1(%)↑	75.6	75.6	75.8	**76.0**	75.2	72.7	66.8

TABLE IV
INFLUENCE OF USING DIFFERENT ATTRIBUTE MODELS

Methods	MOTA↑ (%)	IDF1↑ (%)	IDS↓	FPS↑
Baseline	54.3	56.4	375	**13.2**
Ours-A_{tri}	63.9	63.3	69	7.9
Ours-M_{reg}	**64.8**	63.5	56	5.4
Ours-D_{lsc}	63.1	62.5	80	8.1
Ours	64.5	**63.7**	**52**	7.9

TABLE V
EFFICIENCY OF THE IDENTITY-QUANTITY RECONCILIATION

Methods	MOTA↑ (%)	IDF1↑ (%)	MT↑(%)	IDS↓
Ours (w/o rec)	62.6	61.8	38.4	90
Ours	**64.5**	**63.7**	**42.3**	**52**

MT (2.0%), and significantly improves the IDF1 score (5.7%) and decreases the ID switch number (1308 vs. 1011), while achieving competitive results on the other metrics. This experimental result further demonstrates the efficiency and superiority of the proposed method..

Finally, we discuss two remained challenges of the proposed method: long-time detection missing and attribute confusion. First, as our tracker follows a tracking-by-detection paradigm, when the detector fails to detect a target in a long time or even all the time, the tracker may fail to generate an accurate trajectory for the target as the connection between frames are broken. Second, when some targets interact with each other and have very similar attributes, such as appearance and motion patterns, the IQHAT may fail to identify these two targets correctly after the interaction. To further improve the tracking performance, more advanced object detectors and discriminative features are desired to be involved.

E. Influence of α and β

In Table III, we conduct experiments using the validation data of the 2D MOT15 to study the influence of α and β.

1) Effect of α: We fix $\beta = 1.0$ to study the influence of α. The α in Eq. (13) affects the ID collection of each detection. The larger the α is, the fewer ID labels will be assigned. By contrary, a smaller α will lead to assigning more ID labels. According to the experiment results, we set $\alpha = 0.8$ to achieve the best performance.

2) Effect of β: We fix $\alpha = 0.8$ to study the influence of β. The β in Eq. (14) affects the initialization of trajectories. We set $\beta = 1.1$ according to the experiment results. When $\beta \leq 1.1$, the tracking performance keep steady and the best performance is achieved at $\beta = 1.1$. When $\beta > 1.1$, the performance begin to decrease. This is mainly because when β is too large, fewer trajectories are initialized and more targets are missed tracked.

F. Influence of Using Different Attribute Models

In Table IV, we conduct ablation study on MOT17 validation data to analyze the influence of using different attribute models in IQHAT. The *Baseline* outputs tracking results of the typical MAP tracker [24]. The *Ours-A_*(M$_*$ or D$_*$)* output tracking results which are obtained by substituting the (A)ppearance, (M)otion or (D)ensity models of the IQHAT with respective attribute models, respectively, while keeping all the other parts unchanged. **Appearance model** affects the ID probability computation in Eq.(7). We replace our appearance feature extractor with a modern ReID model [83], *i.e.*, *Ours-A_{tri}*. **Motion model** affects the ID probability computation and the estimation of missed trajectory in Eq.(17). We use a CNN-based regression model [84] to replace the Kalman filter in IQHAT, *i.e.*, *Ours-M_{reg}*. **Density map estimator** affects the local target quantification. We replace our density map estimator with a representative counting model [85], *i.e.*, *Ours-D_{lsc}*. We can see that, 1) following the MAIQP formulation, all the *Ours-** methods steadily and significantly outperforms the MAP method (*i.e.*, Baseline) by a large margin (at least 8.8, 6.7 and 295 improvement on MOTA, IDF1 and IDS, respectively). This demonstrates the efficiency and superiority of the novel IQHAT framework. 2) The *Ours-M_{reg}* slightly outperforms *Ours* in MOTA while suffers a lower FPS. Give consideration to both the accuracy and speed, we adopt the Kalman filter to model target motions. 3) The IQHAT is robust to different attribute models and does not lean to certain identification or quantification methods. Thus, it is a robust and general framework for multi-object tracking.

G. Efficiency of Identity-Quantity Reconciliation

Table V conducts experiment results to study the efficiency of identity-quantity reconciliation. The *Ours(w/o rec)* outputs tracking results without using the identity-quantity reconciliation, where the target identification module and the local target quantification modules are trained separately. More specifically, the ReID model is solely trained on the MOT dataset using a triplet loss [79], where we randomly sample 3 cropped images from each trajectory and then collect triplet tuples using the cropped images for training. As for the density map estimator, we first convert the MOT ground-truth bounding boxes into target point annotations, where the head center the of each bounding box is regarded as a target point. Then we train the density map estimator with the point annotations using a Bayesian loss [81]. From Table V, we can see that, compared to the *Ours(w/o rec)*, the *Ours* method suffers less IDS number and generates more accurate results, which steadily improves the MOTA and IDF1 scores (1.9 and 1.9 improvement, respectively), significantly decreases ID switches (reduces IDS number by 42%) and produce more accurate trajectories (3.9 improvement on MT).

VI. Conclusion

In this paper, we formulate multi-object tracking as a Maximizing An Identity-Quantity Posterior (MAIQP) problem and develop an Identity-Quantity HArmonic Tracking (IQHAT) framework that allows to assign multiple ID labels to a single detection. Different to existing MOT methods that track multiple targets by maximizing an identity posterior, the proposed method jointly optimizes the target identity and quantity, which not only decreases missed trajectories of occluded targets but also alleviates ID switches in crowd scenarios. Experimental evaluations and ablation studies on five benchmark datasets demonstrate the efficiency and superiority of IQHAT. The proposed IQHAT is a general framework for multi-object tracking. To further improve the tracking accuracy, more attribute information such as optical flow and semantic features can be involved in the IQHAT framework. Besides, to improve the computation efficiency, integrating different components (such as density map estimator and object detector) into a unified network can be beneficial.

References

[1] F. Yang, H. Lu, and M.-H. Yang, "Robust superpixel tracking," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1639–1651, Apr. 2014.

[2] J. Fan, X. Shen, and Y. Wu, "What are we tracking: A unified approach of tracking and recognition," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 549–560, Feb. 2013.

[3] J. Ferryman and A. Shahrokni, "PETS2009: Dataset and challenge," in *Proc. IEEE Int. Workshop Perform. Eval. Tracking Surveill.*, Dec. 2009, pp. 1–6.

[4] Y. He, Z. Ma, X. Wei, X. Hong, W. Ke, and Y. Gong, "Error-aware density isomorphism reconstruction for unsupervised cross-domain crowd counting," in *Proc. AAAI*, 2021, pp. 1–9.

[5] S. Zhang, J. Wang, Z. Wang, Y. Gong, and Y. Liu, "Multi-target tracking by learning local-to-global trajectory models," *Pattern Recognit.*, vol. 48, no. 2, pp. 580–590, 2015.

[6] Y. He, X. Wei, X. Hong, W. Shi, and Y. Gong, "Multi-target multi-camera tracking by tracklet-to-target assignment," *IEEE Trans. Image Process.*, vol. 29, pp. 5191–5205, 2020.

[7] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1335–1344.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[9] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[11] P. Chu and H. Ling, "FAMNet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking," 2019, *arXiv:1904.04989*.

[12] P. Bergmann, T. Meinhardt, and L. Leal-Taixe, "Tracking without bells and whistles," 2019, *arXiv:1903.05625*.

[13] H. Jiang, J. Wang, Y. Gong, N. Rong, Z. Chai, and N. Zheng, "Online multi-target tracking with unified handling of complex scenarios," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3464–3477, Nov. 2015.

[14] G. Wang, Y. Wang, H. Zhang, R. Gu, and J.-N. Hwang, "Exploit the connectivity: Multi-object tracking with TrackletNet," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 482–490.

[15] P. Chu, H. Fan, C. C. Tan, and H. Ling, "Online multi-object tracking with instance-aware tracker and dynamic model refreshment," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 161–170.

[16] L. Zhang, Y. Li, and R. Nevatia, "Global data association for multi-object tracking using network flows," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[17] S. Tang, M. Andriluka, B. Andres, and B. Schiele, "Multiple people tracking by lifted multicut and person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3539–3548.

[18] H. Sheng, Y. Zhang, J. Chen, Z. Xiong, and J. Zhang, "Heterogeneous association graph fusion for target association in multiple object tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 11, pp. 3269–3280, Nov. 2019.

[19] A. R. Zamir, A. Dehghan, and M. Shah, "GMCP-tracker: Global multi-object tracking using generalized minimum Clique graphs," in *Proc. Eur. Conf. Comput. Vis.* New York, NY, USA: Springer, 2012, pp. 343–356.

[20] L. Ma, S. Tang, M. J. Black, and L. Van Gool, "Customized multi-person tracker," in *Proc. Asian Conf. Comput. Vis.* New York, NY, USA: Springer, 2018, pp. 612–628.

[21] C. Kim, F. Li, and J. M. Rehg, "Multi-object tracking with neural gating using bilinear LSTM," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 200–215.

[22] M. Han, W. Xu, H. Tao, and Y. Gong, "An algorithm for multiple object trajectory tracking," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun./Jul. 2004, pp. 1–8.

[23] A. Sadeghian, A. Alahi, and S. Savarese, "Tracking the untrackable: Learning to track multiple cues with long-term dependencies," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 300–311.

[24] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3645–3649.

[25] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Online multiperson tracking-by-detection from a single, uncalibrated camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1820–1833, Sep. 2011.

[26] J. Munkres, "Algorithms for the assignment and transportation problems," *J. Soc. Ind. Appl. Math.*, vol. 5, no. 1, pp. 32–38, 1957.

[27] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, and T.-K. Kim, "Multiple object tracking: A literature review," 2014, *arXiv:1409.7618*.

[28] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.

[29] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[30] Q. Zhao *et al.*, "M2Det: A single-shot object detector based on multi-level feature pyramid network," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Jan. 2019, pp. 9259–9266.

[31] M. Yang, Y. Wu, and Y. Jia, "A hybrid data association framework for robust online multi-object tracking," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5667–5679, Dec. 2017.

[32] J. Peng *et al.*, "TPM: Multiple object tracking with tracklet-plane matching," *Pattern Recognit.*, vol. 107, Nov. 2020, Art. no. 107480.

[33] J. Peng, Y. Gu, Y. Wang, C. Wang, J. Li, and F. Huang, "Dense scene multiple object tracking with box-plane matching," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 4615–4619.

[34] I. J. Cox and S. L. Hingorani, "An efficient implementation of Reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 2, pp. 138–150, Feb. 1996.

[35] C. Kim, F. Li, A. Ciptadi, and J. M. Rehg, "Multiple hypothesis tracking revisited," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4696–4704.

[36] M. Keuper, S. Tang, B. Andres, T. Brox, and B. Schiele, "Motion segmentation & multiple object tracking by correlation co-clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 140–153, Jan. 2020.

[37] A. Hornakova, R. Henschel, B. Rosenhahn, and P. Swoboda, "Lifted disjoint paths with application in multiple object tracking," 2020, *arXiv:2006.14550*.

[38] T. Wang *et al.*, "Spatio-temporal point process for multiple object tracking," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 8, 2020, doi: 10.1109/TNNLS.2020.2997006.

[39] Y. Zhang *et al.*, "Long-term tracking with deep tracklet association," *IEEE Trans. Image Process.*, vol. 29, pp. 6694–6706, 2020.

[40] S. Schulter, P. Vernaza, W. Choi, and M. Chandraker, "Deep network flow for multi-object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6951–6960.

[41] G. Braso and L. Leal-Taixe, "Learning a neural solver for multiple object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6247–6257.

[42] Y. Zhang, H. Sheng, Y. Wu, S. Wang, W. Ke, and Z. Xiong, "Multiplex labeling graph for near-online tracking in crowded scenes," *IEEE Internet Things J.*, vol. 7, no. 9, pp. 7892–7902, Sep. 2020.

[43] S. Sun, N. Akhtar, H. Song, A. S. Mian, and M. Shah, "Deep affinity network for multiple object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 104–119, Jan. 2021.

[44] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "FairMOT: On the fairness of detection and re-identification in multiple object tracking," 2020, *arXiv:2004.01888*.

[45] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," in *Proc. Eur. Conf. Comput. Vis.* New York, NY, USA: Springer, 2020, pp. 474–490.

[46] W. Ren, X. Wang, J. Tian, Y. Tang, and A. B. Chan, "Tracking-by-counting: Using network flows on crowd density maps for tracking multiple targets," *IEEE Trans. Image Process.*, vol. 30, pp. 1439–1452, 2021.

[47] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3464–3468.

[48] J. Xu, Y. Cao, Z. Zhang, and H. Hu, "Spatial-temporal relation networks for multi-object tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3988–3998.

[49] Y. Xiang, A. Alahi, and S. Savarese, "Learning to track: Online multi-object tracking by decision making," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4705–4713.

[50] Q. Chu, W. Ouyang, H. Li, X. Wang, B. Liu, and N. Yu, "Online multi-object tracking using CNN-based single object tracker with spatial-temporal attention mechanism," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4836–4845.

[51] H. Shen, L. Huang, C. Huang, and W. Xu, "Tracklet association tracker: An end-to-end learning-based association approach for multi-object tracking," 2018, *arXiv:1808.01562*.

[52] N. L. Baisa, "Online multi-object visual tracking using a GM-PHD filter with deep appearance learning," in *Proc. 22th Int. Conf. Inf. Fusion (FUSION)*, Jul. 2019, pp. 1–8.

[53] K. Fang, Y. Xiang, X. Li, and S. Savarese, "Recurrent autoregressive networks for online multi-object tracking," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 466–475.

[54] D. Yuan, X. Chang, P.-Y. Huang, Q. Liu, and Z. He, "Self-supervised deep correlation tracking," *IEEE Trans. Image Process.*, vol. 30, pp. 976–985, 2021.

[55] P. Ren *et al.*, "A comprehensive survey of neural architecture search: Challenges and solutions," *ACM Comput. Surveys*, vol. 54, no. 4, pp. 1–34, May 2022.

[56] J. Yin, W. Wang, Q. Meng, R. Yang, and J. Shen, "A unified object motion and affinity model for online multi-object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6768–6777.

[57] J. Zhu, H. Yang, N. Liu, M. Kim, W. Zhang, and M.-H. Yang, "Online multi-object tracking with dual matching attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 366–382.

[58] Q. Chu, W. Ouyang, B. Liu, F. Zhu, and N. Yu, "DASOT: A unified framework integrating data association and single object tracking for online multi-object tracking," in *Proc. AAAI*, 2020, pp. 10672–10679.

[59] B. Shuai, A. Berneshawi, X. Li, D. Modolo, and J. Tighe, "SiamMOT: Siamese multi-object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12372–12382.

[60] B. Shuai *et al.*, "Application of multi-object tracking with Siamese track-RCNN to the human in events dataset," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 4625–4629.

[61] A. Wu, C. Lin, B. Chen, W. Huang, Z. Huang, and W.-S. Zheng, "Transductive multi-object tracking in complex events by interactive self-training," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 4620–4624.

[62] Q. Liu, Q. Chu, B. Liu, and N. Yu, "GSM: Graph similarity model for multi-object tracking," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 530–536.

[63] Y. Xu, A. Sep, Y. Ban, R. Horaud, L. Leal-Taixe, and X. Alameda-Pineda, "How to train your deep multi-object tracker," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6787–6796.

[64] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[65] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME, D, J. Basic Eng.*, vol. 82, no. 1, pp. 35–45, 1960.

[66] H. Sheng, J. Chen, Y. Zhang, W. Ke, Z. Xiong, and J. Yu, "Iterative multiple hypothesis tracking with tracklet-level association," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 12, pp. 3660–3672, Dec. 2019.

[67] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "MOTChallenge 2015: Towards a benchmark for multi-target tracking," 2015, *arXiv:1504.01942*.

[68] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.

[69] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," 2016, *arXiv:1603.00831*.

[70] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[71] F. Yang, W. Choi, and Y. Lin, "Exploit all the layers: Fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2129–2137.

[72] P. Dendorfer *et al.*, "MOT20: A benchmark for multi object tracking in crowded scenes," 2020, *arXiv:2003.09003*.

[73] W. Lin *et al.*, "Human in events: A large-scale benchmark for human-centric video analysis in complex events," 2020, *arXiv:2005.04490*.

[74] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *EURASIP J. Image Video Process.*, vol. 2008, pp. 1–10, Dec. 2008.

[75] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Eur. Conf. Comput. Vis.* New York, NY, USA: Springer, 2016, pp. 17–35.

[76] B. Wu and R. Nevatia, "Tracking of multiple, partially occluded humans based on static body part detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2006, pp. 951–958.

[77] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[78] W. Li and X. Wang, "Locally aligned feature transforms across views," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3594–3601.

[79] H. Luo *et al.*, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2597–2609, Oct. 2020.

[80] H. Idrees *et al.*, "Composition loss for counting, density map estimation and localization in dense crowds," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 532–546.

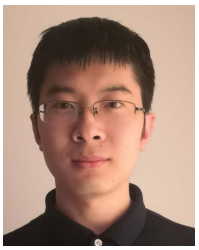
[81] Z. Ma, X. Wei, X. Hong, and Y. Gong, "Bayesian loss for crowd count estimation with point supervision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6142–6151.

[82] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[83] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*.

[84] D. Held, S. Thrun, and S. Savarese, "Learning to track at 100 FPS with deep regression networks," in *Proc. Eur. Conf. Comput. Vis.* New York, NY, USA: Springer, 2016, pp. 749–765.

[85] D. B. Sam, S. V. Peri, M. N. Sundararaman, A. Kamath, and V. B. Radhakrishnan, "Locate, size and count: Accurately resolving people in dense crowds via detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2739–2751, Aug. 2021.

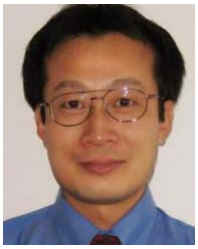


Yuhang He received the B.S. degree in control science and engineering from Xi'an Jiaotong University, Shaanxi, China, in 2016, where he is currently pursuing the Ph.D. degree with the College of Artificial Intelligence (CAI). His current research interests include image classification, object tracking, and multi-modal learning.

Xing Wei received the B.E. degree in automation and the Ph.D. degree in control science and engineering from Xi'an Jiaotong University, Shaanxi, China, in 2013 and 2019, respectively. His research interests include computer vision, pattern recognition, and machine learning, specifically in the areas of visual scene analysis, image retrieval, object recognition, and low-level vision.

Xiaopeng Hong received the Ph.D. degree in computer application and technology from the Harbin Institute of Technology (HIT), China, in 2010. He is a Professor with the HIT. He had been a Professor with Xi'an Jiaotong University, China, and an Adjunct Professor with the University of Oulu, Finland. He has authored over 50 articles in top-tier publications and conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, CVPR, ICCV, and AAAI. He has served as an/a Area Chair/Senior Program Committee Member for ACM MM, AAAI, IJCAI, and ICME; a Co-Organizer for six international workshops in conjunction with IEEE CVPR, ACM MM, and IEEE FG; and a Co-Lecturer for two tutorials in conjunction with ACM MM21 and IJCB21. His studies about subtle facial movement analysis have been reported by the international media like *MIT Technology Review* and been awarded the 2020 IEEE Finland Section Best Student Conference Paper. He has served as a Guest Editor for peer-reviewed journals like *Pattern Recognition Letters* and *Signal, Image and Video Processing*.

Wei Ke (Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Chinese Academy of Sciences, Beijing, China, in 2018. He was a Postdoctoral Researcher with the Robotics Institute, Carnegie Mellon University, until 2020. In 2016, he visited the Center for Machine Vision and Signal Analysis, University of Oulu, as a joint Ph.D. student, supported by the China Scholarship Council (CSC). He is currently an Associate Professor with Xi'an Jiaotong University. He has published about 20 papers in refereed conferences and journals including IEEE CVPR and ECCV. His research interests include computer vision and deep learning. He is the winner of the President Award of Chinese Academy of Sciences in 2017.



Yihong Gong (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from The University of Tokyo, Japan, in 1987, 1989, and 1992, respectively. In 1992, he joined Nanyang Technological University, Singapore, as an Assistant Professor with the School of Electrical and Electronic Engineering. From 1996 to 1998, he was a Project Scientist with the Robotics Institute, Carnegie Mellon University, USA. Since 1999, he has been with the Silicon Valley Branch, NEC Labs America, as a Group Leader, the Department Head, and the Branch Manager. In 2012, he joined Xi'an Jiaotong University, China, as a Distinguished Professor. His research interests include image and video analysis, multimedia database systems, and machine learning.