UTM: A Unified Multiple Object Tracking Model with Identity-Aware Feature Enhancement

Sisi You\textsuperscript{1} Hantao Yao\textsuperscript{2} Bing-kun Bao\textsuperscript{1} Changsheng Xu\textsuperscript{2,3}\textsuperscript{*}

\textsuperscript{1}Nanjing University of Posts and Telecommunications
\textsuperscript{2}State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences (CASIA)
\textsuperscript{3}University of Chinese Academy of Sciences

ssyou@njupt.edu.cn, hantao.yao@nlpr.ia.ac.cn, bingkunbao@njupt.edu.cn, csxu@nlpr.ia.ac.cn

Abstract

Recently, Multiple Object Tracking has achieved great success, which consists of object detection, feature embedding, and identity association. Existing methods apply the three-step or two-step paradigm to generate robust trajectories, where identity association is independent of other components. However, the independent identity association results in the identity-aware knowledge contained in the tracklet not be used to boost the detection and embedding modules. To overcome the limitations of existing methods, we introduce a novel Unified Tracking Model (UTM) to bridge those three components for generating a positive feedback loop with mutual benefits. The key insight of UTM is the Identity-Aware Feature Enhancement (IAFE), which is applied to bridge and benefit these three components by utilizing the identity-aware knowledge to boost detection and embedding. Formally, IAFE contains the Identity-Aware Boosting Attention (IABA) and the Identity-Aware Erasing Attention (IAEA), where IABA enhances the consistent regions between the current frame feature and identity-aware knowledge, and IAEA suppresses the distracted regions in the current frame feature. With better detections and embeddings, higher-quality tracklets can also be generated. Extensive experiments of public and private detections on three benchmarks demonstrate the robustness of UTM.

1. Introduction

Multiple Object Tracking (MOT) aims at locating and identifying all of the moving objects in the video, which has broad application prospects in visual surveillance, human-computer interaction, virtual reality, and unmanned vehicles. With the rapid development of object detection [12, 34, 35, 66], Tracking-By-Detection has became a favorite paradigm in MOT. Recently, a number of Tracking-By-Detection approaches have been proposed, which can be divided into two paradigms: Separate Detection and Embedding (SDE) [1, 2, 4, 5, 8, 10, 16, 25, 38, 43, 52, 54], and Joint Detection and Embedding (JDE) [15, 29, 45, 50, 64].

As illustrated in Figure 1(a), SDE can be divided into three independent components: object detection, feature embedding, and identity association\textsuperscript{1}. Candidate bounding boxes (bboxes) are obtained by standard detectors in each frame first, then identity embedding of each bbox is extracted by the re-identification algorithms, finally linked across frames through identity association to generate tra-

\textsuperscript{*}indicates corresponding author: Changsheng Xu.

Figure 1. Comparison of different MOT frameworks in existing methods. (a) Separate Detection and Embedding (SDE) [2]. (b) Joint Detection and Embedding (JDE) [50]. (c) The proposed Unified Tracking Model (UTM) that constructed with Identity-Aware Feature Enhancement (IAFE) module.
jectories. Therefore, SDE mainly exploits refining detection [2, 16], enhancing identity embedding [1, 25, 38, 52], or designing robust association algorithms [5, 8, 54] to improve tracking performance. With the maturity of multi-task learning, some approaches [45, 50, 64] propose JDE framework to reduce the computation cost. Different from SDE, JDE applies the one-shot tracker to generate object detections and corresponding visual embeddings simultaneously, thus treating MOT as two independent components, shown in Figure 1(b). Since the identity association in the above two paradigms is independent of object detection and feature embedding, the significant clues contained in the tracklet cannot be applied to enhance the detection and embedding modules.

To address the above problem, a feasible idea is introducing an auxiliary module to associate and propagate identity-aware knowledge between the identity association and the other two components. Therefore, we design a novel Identity-Aware Feature Enhancement (IAFE) module to achieve information interaction between different components, shown in Figure 1(c). In detail, IAFE propagates the identity-aware knowledge generated by identity association module to enhance the backbone feature of object detection. Meanwhile, the enhanced backbone feature can be utilized by the feature embedding module to generate discriminative embeddings. With more accurate detections and robust embeddings, the identity association module can produce higher-quality tracklets. Therefore, object detection, feature embedding, and identity association are involved with each other, thus forming a positive feedback loop with mutual benefits.

As shown in Figure 2, the proposed Unified Tracking Model (UTM) is composed of IAFE, detection branch, embedding branch, identity association branch, and memory aggregation module. The detection and embedding branches are applied to locate and identify each object of the current frame, and the identity association branch applies graph matching to associate the candidate bboxes with history tracklets. To achieve information interaction, IAFE is proposed to bridge and benefit these three branches, which utilizes the identity-aware knowledge to boost the detection and embedding. Specifically, IAFE consists of two modules: Identity-Aware Boosting Attention (IABA) and Identity-Aware Erasing Attention (IAEA), where IABA boosts the consistent regions between the current frame feature and identity-aware knowledge, and IAEA erases the distracted regions in the current frame feature. With the designed IAFE, UTM constructs a positive feedback loop among all the three branches to improve the performance of MOT. Furthermore, we introduce a memory aggregation module to capture identity-aware knowledge through adaptively selecting features of history frames, which can alleviate the effect of identity switches.

The main contributions of the proposed method can be summarized as follows:

- We design an Identity-Aware Feature Enhancement (IAFE) module to bridge and benefit different components in MOT. Specifically, it utilizes the identity-aware knowledge to boost backbone feature, which is further used to enhance the detection and embedding.
- With the proposed IAFE, we construct a Unified Tracking Model (UTM) to form a positive feedback loop with mutual benefits.
- The evaluation of public and private detections on three benchmarks verifies the effectiveness and generalization ability of the proposed UTM.

2. Related Work

MOT has been received more and more attention from industry and academia in the past years. We review the most relevant work of MOT, i.e., Separate Detection and Embedding, Joint Detection and Embedding.

2.1. Separate Detection and Embedding

Separate Detection and Embedding (SDE) is the most popular framework for MOT, which consists of three independent components: object detection, feature embedding, and identity association. Existing SDE methods exploit refining detection [2, 5, 15, 16, 59] or the provided public detections [11–13, 35] for MOT. Then, the appearance embedding model [1, 2, 25, 38, 52, 62] is used to extract discriminative feature for each candidate bbox, e.g., Bae et al. [1] boost the representation learning by training the visual model on person re-identification datasets. Besides the appearance model, the motion model [10, 15, 39, 52, 58] is also applied to describe the motion information of each object, e.g., the Kalman filter [22] is widely used in MOT. Finally, the matching algorithm is used to solve the problem of identity association, e.g., Hungarian algorithm [24], multi-cut [19, 20], min-cost max-flow network [61], and conditional random field [55]. Except for the traditional matching algorithms, more and more methods [5, 8, 16, 36, 49, 51, 54] apply deep learning for identity association. Xu et al. [54] propose a Deep Hungarian Network module to substitute the Hungarian algorithm. Some approaches [5, 8, 16, 49, 51] formulate identity association as a graph optimization problem. The work most related to ours is MPNTrack [5], which treats each object as a graph node and applies the edge feature for classification. The major difference is that we construct the detection graph and tracklet graph for identity association according to high-order context information, which brings the better performance of our model.

2.2. Joint Detection and Embedding

To construct a real-time tracker, Joint Detection and Embedding (JDE) methods have begun to attract more atten-
Identity-Aware Feature Enhancement

Backbone $F_t$

Detection branch

$F_t^\ast$

Detection Graph

Tracklet Graph

Identity ... $T_{t-1}$

$E_t$

Memory Aggregation Linear Layer

ReLU

Memory bank

... to enhance detection and embedding modules in the multi-step methods. To address these problems, we introduce the identity-aware feature enhancement module to form a unified tracking model, which can bridge and benefit the above three components to form a positive feedback loop.

3. Methodology

As shown in Figure 2, UTM is composed of the Identity-Aware Feature Enhancement (IAFE) module, detection branch, embedding branch, identity association branch, and memory aggregation module. Given the $t$-th frame, we first apply the backbone network to obtain the backbone feature $F_t$. Then, IAFE takes the efficient tracklet features as the identity-aware knowledge to enhance the backbone feature of each tracked object, where tracklet feature set and enhanced backbone feature are denoted as $E_{t-1}$ and $\tilde{F}_t$, respectively. Next, the detection and embedding branches utilize $F_t$ to obtain candidate bbox set $B_t$ and object embedding set $\tilde{F}_t$. Then, the identity association branch utilizes the high-order contextual information to associate the candidate bboxes and history tracklets. Finally, the memory aggregation module is used to update the tracklet feature that is applied to enhance the backbone feature with IAEA in the next frame. In the following, we will describe the detail of each module.

3.1. Identity-Aware Feature Enhancement

To construct a Unified Tracking Model, we propose an Identity-Aware Feature Enhancement (IAFE) module to bridge the detection, embedding, and identity association branches to generate a positive feedback loop. Specifically, IAFE leverages the efficient tracklet features to boost the detection and embedding branches, thereby generating high-quality tracklets with the identity association branch. As shown in Figure 3, IAFE consists of the Identity-Aware Boosting Attention (IABA) and Identity-Aware Erasing Attention (IAEA), where IABA is utilized to boost the features of consistent regions between history tracklets and current frame, and IAEA is applied to suppress the features of dis-
Figure 3. Illustration of the proposed IAFE module. BBL indicates the block binarization layer, \( R \) and \( M \) represent the correlation feature and binary mask, respectively.

truncated regions in the current frame.

Following Tracker [2], we combine the corresponding public detections and tracked objects in the previous frame as the candidate proposals \( P_t \) of the current frame \( I_t \). After that, IAFE applies the tracklet features \( E_{t-1} \) to enhance the backbone features of candidate proposals, where \( E_{t-1} = \{ E_{t-1,1}, \ldots, E_{t-1,m} \} \), \( m \) is the number of history tracklets. For example, given the candidate proposal \( P_t,*, \in P_t \) along with the backbone feature \( F_{t,*} \), we can leverage \( E_{t-1} \) to generate the enhanced feature \( \tilde{F}_{t,*} \) as follows:

\[
\tilde{F}_{t,*} = F_{t,*} \oplus [f(F_{t,*}, E_{t-1}) \ominus g(F_{t,*}, E_{t-1})],
\]

where \( \oplus \) and \( \ominus \) indicate element-wise addition and subtraction. \( f(F_{t,*}, E_{t-1}) \) and \( g(F_{t,*}, E_{t-1}) \) represent IABA and IAEA, respectively. In the following, we give a detailed description about the boosting attention module \( f(\cdot) \) and erasing attention module \( g(\cdot) \).

The boosting attention module \( f(\cdot) \) leverages \( E_{t-1} \) to enhance the consistent feature between \( E_{t-1} \) and \( F_{t,*} \). Given the proposal feature \( F_{t,*} \in \mathbb{R}^{C \times H \times W} \) and \( E_{t-1} \), \( f(F_{t,*}, E_{t-1}) \) can be formulated as follows:

\[
f(F_{t,*}^i, E_{t-1}) = \sum_{k=1}^{m} \lambda_{*,k} \sum_{j \neq i} h(F_{t,*}^i, E_{t-1,k}^j) \rho(E_{t-1,k}),
\]

where \( E_{t-1,k} \in E_{t-1} \) is the tracklet feature of the \( k \)-th tracklet \( T_{t-1,k} \), and \( F_{t,*}^i, E_{t-1,k}^j \in \mathbb{R}^{C} \) are the features sampled from \( F_{t,*} \) and \( E_{t-1,k} \). \( h(x_i, x_j) = e^{\psi(x_i)^T \varphi(x_j)} \), where \( \psi, \varphi, \rho \) are convolution layers. \( \lambda_{*,k} \) is an indicator and is defined as follows:

\[
\lambda_{*,k} = \begin{cases} 
1 & \text{IoU}(P_{t,*}, B_{t-1,k}) > \lambda_{iou}, \\
0 & \text{otherwise},
\end{cases}
\]

where \( \text{IoU}(P_{t,*}, B_{t-1,k}) \) is the geometric similarity between the candidate proposal \( P_{t,*} \) and the last box \( B_{t-1,k} \) of \( T_{t-1,k} \), and \( \lambda_{iou} \) is the geometric threshold.

To eliminate the background information, we apply the tracklet features to erase distracted regions in the backbone feature \( F_{t,*} \), which can be formulated as follows:

\[
g(F_{t,*}, E_{t-1}) = \sum_{k=1}^{m} \lambda_{*,k} F_{t,*} \ominus R(F_{t,*}, \phi(E_{t-1,k}))
\]

\[
\ominus M(F_{t,*}, \phi(E_{t-1,k})),
\]

where \( \phi \) indicates the average pooling layer, \( \odot \) is element-wise product operation, \( R(x_i, x_j) \) and \( M(x_i, x_j) \) represent the correlation feature and binary mask, respectively. The correlation feature denotes the dot-produce similarity between tracklet feature and all the local features in \( F_{t,*} \).

\[
\mathcal{R}(F_{t,*}[i,j], \phi(E_{t-1,k})) = (F_{t,*}[i,j])^T \phi(E_{t-1,k}),
\]

where \( F_{t,*}[i,j] \) denotes the local feature of spatial location \( (i, j) \) of \( F_{t,*} \).

After obtaining the correlation feature, we apply \( \mathcal{R}(\cdot) \) to generate the binary mask \( \mathcal{M}(\cdot) \) between backbone feature \( F_{t,*} \) and tracklet feature \( E_{t-1,k} \). Inspired by [21], we apply a block binarization layer to generate the mask for erasing a contiguous region in \( F_{t,*} \). We exploit a sliding block with the size of \( 3 \times 3 \) and strides \( = 1 \) to search the most highlighted continuous area in the correlation feature, and define the correlation value of each block as the sum of the values in the block. Then, we select the candidate block with the highest correlation value as the block to be reversed, i.e., the binary mask of correlation feature is obtained by setting the values of the selected block to 0 and others to 1. Next, we apply a softmax layer to generate a gate map for end-to-end training, where gate map is obtained by the element-wise produce operation between softmax-based correlation feature and binary mask. Finally, the erasing attention can be obtained based on the gate map and binary mask.

### 3.2. Unified Tracking Model

The proposed IAFE in Sec. 3.1 utilizes the identity-aware knowledge to achieve information interaction between different components in MOT, thus constructs a Unified Tracking Model (UTM) to form a positive feedback loop. Except for IAFE, UTM also contains the detection, embedding, identity association branches, and memory aggregation module, which will be described in the following.

**Detection Branch.** Inspired by Tracktor [2], we adopt Faster R-CNN [35] as the detection branch, where the regression head and classification head are utilized to refine boxes and infer classes of objects inside boxes. Different from the candidate proposals generated by RPN in Faster R-CNN, we combine the public detections in the current frame and the tracked objects in the previous frame as the candidate proposals. Then the regression head is exploited to predict candidate boxes \( B_t = \{ B_{t,1}, \ldots, B_{t,n} \} \) based on the enhanced backbone feature \( \tilde{F}_t \) which obtained by IAFE, and the classification head gives the confidence score for
each bbox. Note that \( B_{t,i,j} \) indicates a bbox in \( B_t \) and \( n \) is the number of candidate bboxes. It is worth noting that we merely apply the regression head to refine the public detections and tracked objects, while not generate new candidate bboxes. Simultaneously, the regression head is trained with L1 loss on displacements, and the classification head is learned with a cross-entropy loss.

**Embedding Branch.** After obtaining the candidate bboxes \( B_t \), the embedding branch targets to generate their discriminative embeddings. Given the enhanced backbone feature \( F_t \) and a bbox \( B_{t,i} \), the corresponding embedding \( \hat{F}_{t,i} \) can be formulated as follows:

\[
\hat{F}_{t,i} = \mathcal{E}(\hat{F}_t, B_{t,i}),
\]

where \( \mathcal{E} \) represents convolution layer on top of the ROI-Align layer. Then, the embeddings of \( B_t \) can be denoted as \( \hat{F}_t = \{\hat{F}_{t,1}, \ldots, \hat{F}_{t,i}, \ldots, \hat{F}_{t,n}\} \). Furthermore, the embedding branch is trained with the combination of cross-entropy loss and triplet loss to learn the discriminative identity embeddings.

**Identity Association Branch.** In this paper, we formulate identity association as a graph matching problem between the candidate bboxes \( B_t \) and history tracklets \( T_{t-1} \), where \( B_t = \{B_{t,1}, \ldots, B_{t,n}\} \), \( T_{t-1} = \{T_{t-1,1}, \ldots, T_{t-1,m}\} \), and \( n \) and \( m \) represent the number of candidate bboxes and history tracklets, respectively. We first construct the detection graph and tracklet graph to describe the relationships of different objects in the candidate bboxes and history tracklets, respectively. Then, the cross-graph message passing between the detection graph and tracklet graph is applied to enhance the node features. Finally, the graph matching layer applies high-order context information to associate the detection graph with the tracklet graph.

The detection graph is defined as \( G_D = (V_D, E_D) \), \( V_D = \{B_{t,i}, \hat{F}_{t,i}\} (i \in [1, n]) \) represents the node set, where \( \hat{F}_{t,i} \) is the embedding of the i-th bbox \( B_{t,i} \) for the t-th frame. \( E_D = \{[\hat{F}_{t,i}, \hat{F}_{t,i}]\} \) indicates the edge set, where \( [\cdot] \) indicates the concatenation operation. Similarly, the tracklet graph is defined as \( G_T = (V_T, E_T) \), \( V_T = \{B_{t-1,j}, \mathbf{E}_{t-1,j}\} \), \( E_T = \{[\mathbf{E}_{t-1,j}, \mathbf{E}_{t-1,j}]\} (j \in [1, m]) \), where \( B_{t-1,j} \) and \( \mathbf{E}_{t-1,j} \) indicate the last bbox and tracklet feature of the j-th history tracklet \( T_{t-1,j} \).

To model the feature interaction between the detection graph \( G_D \) and the tracklet graph \( G_T \), we adopt the cross-graph message passing to propagate the information across these two graphs. Let \( \mathbf{F}_{t,i}^{(0)} = \hat{F}_{t,i} \) and \( \mathbf{E}_{t-1,j}^{(0)} = \mathbf{E}_{t-1,j} \) be the initial feature of each node in \( V_D \) and \( V_T \). We analyze the effect of three node aggregation rules.

\[
\begin{align*}
\text{(Type 1)} & \quad \mathbf{F}_{t,i}^{(t+1)} = \text{agg}^{(1)}(\mathbf{F}_{t,i}^{(0)}), \\
\text{(Type 2)} & \quad \mathbf{F}_{t,i}^{(t+1)} = \mathcal{N}_c(\mathbf{F}_{t,i}^{(t)} + \frac{1}{m} \sum_{j=1}^{m} \mathbf{E}_{t-1,j}^{(t)}), \\
\text{(Type 3)} & \quad \mathbf{F}_{t,i}^{(t+1)} = \mathcal{N}_c(\mathbf{F}_{t,i}^{(t)} + \sum_{j=1}^{m} w^{(t)} \mathbf{E}_{t-1,j}^{(t)}),
\end{align*}
\]

where \( \mathbf{F}_{t,i}^{(l)} \) and \( \mathbf{E}_{t-1,j}^{(l)} \) are the features of the l-th propagation, \( w^{(t)} = \cos(\mathbf{F}_{t,i}^{(l)}, \mathbf{E}_{t-1,j}^{(l)}) + \text{IoU}(B_{t,i}, B_{t-1,j}) \) is the combination of cosine similarity and geometric similarity obtained by the Intersection over Union (IoU) of two bboxes, and \( \mathcal{N}_c \) represents learnable function, e.g., MLP. The final feature of the node in \( V_D \) can be denoted as \( \mathbf{F}_{t,i} = \mathbf{F}_{t,i}^{(L)} \), where \( L \) is the total steps of the message passing. Meanwhile, the similar operation is also applied to the tracklet graph \( G_T \).

After that, we utilize the first-order node-to-node similarity and the second-order edge-to-edge similarity to compute the affinity matrix \( M \), where both of the similarities are described with cosine similarity. Furthermore, two nodes between the detection graph and tracklet graph are matched if the corresponding similarity in \( M \) is higher than an affinity threshold \( \gamma \). Finally, the optimal matching \( Y^* \) can be obtained with:

\[
Y^* = \arg\max_Y Y^T MY,
\]

where \( Y \in \{0, 1\}^{n \times m} \) is a permutation matrix that denotes the matching result between the detection and tracklet graphs. The more details of the optimization of \( Y^* \) can be referred to [60]. Moreover, the identity association branch is trained with weighted binary cross-entropy loss.

**Memory Aggregation.** The identity-aware feature enhancement module is described in Sec. 3.1, which applies the efficient tracklet features to enhance the detection and embedding branches. Although storing identity embeddings formed in the previous frames is an intuitive and accessible way to obtain tracklet feature, the identity embeddings may noisy due to occlusion and identity switches. To enhance the robustness of tracklet feature, we introduce a memory bank to aggregate effective tracklet feature for each object, so that all of the object detection, feature embedding, and identity association can be further boosted. Furthermore, the memory bank groups cached memories in different memory units for different objects, each memory unit can be denoted as \( \mathbf{F}_{t-1,j} = \{\mathbf{F}_{t-1,j}, \ldots, \mathbf{F}_{t-1,j}\} \), where \( \mathbf{F}_{t-1,j}^{m} \) and \( \mathbf{F}_{t-1,j} \) denote the memory and identity embedding of the j-th object in the (t-1)-th frame, and \( \eta \) is the memory length. Specifically, we adopt a learnable memory aggregation module to adaptively select valid identity embeddings, which consists of a linear layer with a ReLU activation function. Each tracklet feature \( \mathbf{E}_{t-1,j} \) can be updated as follows:

\[
\mathbf{E}_{t-1,j} = \theta(\mathbf{F}_{t-1,j}^{m}, \mathbf{E}_{t-1,j}),
\]

where \( \theta \) represents the memory aggregation operation.

**Optimization.** To achieve faster convergence, we adopt a multi-step optimization strategy. Firstly, we jointly train the detection branch, embedding branch, and memory aggregation module. Specifically, the detection branch is
constrained with L1 loss and cross-entropy loss, the embedding branch and memory aggregation module are constrained with two discriminative identity losses that consists of cross-entropy loss and triplet loss. Secondly, we utilize the output of detection branch and embedding branch to train the identity association branch while the detection and embedding branches are fixed. Finally, we jointly fine-tune all the components in the full unified model.

4. Experiments

Datasets and evaluation metrics. We conduct all the experiments on three MOT benchmarks, e.g., MOT16 [31], MOT17 [31], and MOT20 [9]. Following the CLEAR MOT Metrics [3], IDF1 Score [37], and HOTA [30], we apply some basic items for quantitative evaluation, e.g., Multiple Object Tracking Accuracy (MOTA $\uparrow$), Higher Order Tracking Accuracy (HOTA $\uparrow$), Mostly Lost (ML $\downarrow$), Mostly Tracked (MT $\uparrow$), False Positives (FP $\downarrow$), False Negatives (FN $\downarrow$), and Identity Switches (IDS $\downarrow$).

Implementation details. The proposed method is implemented by Pytorch with RTX 3090. We adopt Faster R-CNN [35] with Feature Pyramid Network (FPN) [27] as the detection branch. For public detection, we pre-train the backbone of ResNet101 [18] on COCO dataset [28]. For private detection, we pre-train the detection branch on CrowdHuman [40] and MOT training datasets. Simultaneously, the embedding branch is pre-trained on Market1501 [65] and CUHK03 [26] datasets. Then we refine the whole model with the MOT training datasets. The initial learning rate is set to 0.002 with a decay factor 0.5 at every 3 epochs up to 30 epochs. Adam optimizer [23] is used with a mini-batch size of 2. We set the number of message passing steps $L = 3$, geometric threshold $\lambda_{gea} = 0.7$, affinity threshold $\gamma = 0.6$, and the memory length $\eta = 30$.

4.1. Benchmark Evaluation

We compare the proposed Unified Tracking Model (UTM) with several methods on three benchmarks, e.g., MOT16, MOT17, and MOT20. The benchmark evaluation can be divided into public detection and private detection.

Public Detection. We compare the proposed method with traditional tracking-by-detection methods that apply Tracktor for refining detections on the public detection setting for a fair comparison, and the comparison with other refined detectors are provided in supplementary material. Furthermore, the compared methods can be categorized into online and offline tracking methods. As shown in Table 1, the proposed method achieves better performance than existing online methods on most of the evaluation metrics. Moreover, we compare UTM with two multi-step frameworks to demonstrate its superiority. Compared with the Separate Detection and Embedding (SDE) method Tracktor [2], UTM obtains a higher HOTA, e.g., 8.5%, 7.7%, and 11.2% improvements on MOT16, MOT17, and MOT20. Meanwhile, we compare UTM with the Joint Detection and Embedding (JDE) method TADAM [15], the proposed UTM obtains a higher IDF1, e.g., 7.6%, 6.4%, and 14.3% improvements on MOT16, MOT17, and MOT20. Among existing offline methods, the most related work to ours is GMTsI [16], which utilizes a similar graph matching method for identity association. The major difference and novelty is that UTM leverages the identity-aware knowledge to enhance the object detection and feature embedding modules. Compared with GMTsI [16], UTM achieves 4.5% improvement of the MOTA metric on MOT17 dataset. We attribute the performance improvement to the proposed UTM generates a positive feedback loop with identity-aware feature enhancement module.

Private Detection. To further verify the effectiveness of the proposed UTM, we compare UTM with some algorithms on private detection setting and summarize the re-

<table>
<thead>
<tr>
<th>Methods</th>
<th>Refined MOTA</th>
<th>HOTA</th>
<th>IDF1</th>
<th>FP</th>
<th>FN</th>
<th>IDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOT16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPNT(O) [5]</td>
<td>Tracktor</td>
<td>58.6</td>
<td>48.9</td>
<td>61.7</td>
<td>4.949</td>
<td>70,252</td>
</tr>
<tr>
<td>LPC(O) [8]</td>
<td>Tracktor</td>
<td>58.8</td>
<td>51.7</td>
<td>67.6</td>
<td>6.167</td>
<td>68,432</td>
</tr>
<tr>
<td>GMTsI(O) [16]</td>
<td>Tracktor</td>
<td>61.1</td>
<td>51.2</td>
<td>66.6</td>
<td>3.891</td>
<td>66,550</td>
</tr>
<tr>
<td>DeepMOT [54]</td>
<td>Tracktor</td>
<td>54.8</td>
<td>42.2</td>
<td>53.4</td>
<td>2.955</td>
<td>78,765</td>
</tr>
<tr>
<td>GMT [16]</td>
<td>Tracktor</td>
<td>55.9</td>
<td>48.9</td>
<td>63.9</td>
<td>2371</td>
<td>77,545</td>
</tr>
<tr>
<td>Tracktor [2]</td>
<td>Tracktor</td>
<td>56.2</td>
<td>44.6</td>
<td>54.9</td>
<td>2.394</td>
<td>76,844</td>
</tr>
<tr>
<td>ArTIST [39]</td>
<td>Tracktor</td>
<td>56.6</td>
<td>57.8</td>
<td>53.2</td>
<td>75,031</td>
<td>519</td>
</tr>
<tr>
<td>LiTISI [19]</td>
<td>Tracktor</td>
<td>57.5</td>
<td>49.6</td>
<td>64.1</td>
<td>4.249</td>
<td>72,868</td>
</tr>
<tr>
<td>TMOH* [41]</td>
<td>Tracktor</td>
<td>63.2</td>
<td>50.7</td>
<td>63.5</td>
<td>3.122</td>
<td>63,376</td>
</tr>
<tr>
<td>UTM</td>
<td>Tracktor</td>
<td>63.8</td>
<td>53.1</td>
<td>67.1</td>
<td>8.328</td>
<td>57,269</td>
</tr>
<tr>
<td>MOT17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LiTISI(O) [19]</td>
<td>Tracktor</td>
<td>58.2</td>
<td>50.7</td>
<td>65.2</td>
<td>16,850</td>
<td>217,944</td>
</tr>
<tr>
<td>MPNT(O) [5]</td>
<td>Tracktor</td>
<td>58.8</td>
<td>49.0</td>
<td>61.7</td>
<td>17,413</td>
<td>213,594</td>
</tr>
<tr>
<td>LPC(O) [8]</td>
<td>Tracktor</td>
<td>59.0</td>
<td>51.7</td>
<td>66.8</td>
<td>23,103</td>
<td>206,947</td>
</tr>
<tr>
<td>GMTsI(O) [16]</td>
<td>Tracktor</td>
<td>59.0</td>
<td>51.1</td>
<td>65.9</td>
<td>20,395</td>
<td>209,553</td>
</tr>
<tr>
<td>GMT [16]</td>
<td>Tracktor</td>
<td>56.2</td>
<td>49.1</td>
<td>63.8</td>
<td>8,719</td>
<td>236,541</td>
</tr>
<tr>
<td>Tracktor [2]</td>
<td>Tracktor</td>
<td>56.3</td>
<td>44.8</td>
<td>55.1</td>
<td>8,866</td>
<td>235,449</td>
</tr>
<tr>
<td>ArTIST [39]</td>
<td>Tracktor</td>
<td>56.7</td>
<td>57.5</td>
<td>12,353</td>
<td>230,437</td>
<td>1,756</td>
</tr>
<tr>
<td>TADAM [15]</td>
<td>Tracktor</td>
<td>59.7</td>
<td>58.7</td>
<td>6766</td>
<td>216,029</td>
<td>1,930</td>
</tr>
<tr>
<td>TMOH* [41]</td>
<td>Tracktor</td>
<td>62.1</td>
<td>50.4</td>
<td>62.8</td>
<td>10,951</td>
<td>201,195</td>
</tr>
<tr>
<td>UTM</td>
<td>Tracktor</td>
<td>63.5</td>
<td>52.5</td>
<td>65.1</td>
<td>33,683</td>
<td>170,352</td>
</tr>
<tr>
<td>MOT20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LPC(O) [8]</td>
<td>Tracktor</td>
<td>56.3</td>
<td>49.0</td>
<td>62.5</td>
<td>11,726</td>
<td>213,056</td>
</tr>
<tr>
<td>MPNT(O) [5]</td>
<td>Tracktor</td>
<td>57.6</td>
<td>46.8</td>
<td>59.1</td>
<td>16,953</td>
<td>210,384</td>
</tr>
<tr>
<td>Tracktor [2]</td>
<td>Tracktor</td>
<td>52.6</td>
<td>42.1</td>
<td>52.7</td>
<td>6,930</td>
<td>236,680</td>
</tr>
<tr>
<td>ArTIST [39]</td>
<td>Tracktor</td>
<td>53.6</td>
<td>51.0</td>
<td>7.765</td>
<td>230,576</td>
<td>1,531</td>
</tr>
<tr>
<td>TADAM [15]</td>
<td>Tracktor</td>
<td>56.6</td>
<td>51.6</td>
<td>38,407</td>
<td>182,520</td>
<td>2,690</td>
</tr>
<tr>
<td>TMOH* [41]</td>
<td>Tracktor</td>
<td>60.1</td>
<td>48.9</td>
<td>61.2</td>
<td>38,043</td>
<td>165,899</td>
</tr>
<tr>
<td>UTM</td>
<td>Tracktor</td>
<td>64.4</td>
<td>53.3</td>
<td>68.5</td>
<td>82,726</td>
<td>98,974</td>
</tr>
</tbody>
</table>

Table 1. Comparison with modern methods on MOT16, MOT17, and MOT20 benchmarks with the provided public detections. Best results are marked in BLOD. “O” and * indicate the offline methods and post processing methods.
To verify the benefits of the proposed Unified Tracking Model (UTM), we conduct several comparisons between UTM and existing multi-step paradigms, e.g., SDE and JDE. To make a fair comparison, we take the SDE method Tracktor [2] as the baseline for these three frameworks. As shown in Table 3, UTM (c) obtains a significant improvement compared with SDE (a) and JDE (b), e.g., 2.0% and 1.9% improvements on MOTA. We attribute the performance gain to UTM which can generate a positive feedback loop with mutual benefits. To compare the effectiveness of different methods on occlusion, we analyze the ratio of successfully tracked objects with respect to their occlusion ratio. The occlusion ratio is defined as the ratio between the occluded area divided by its bbox area, and the higher object occlusion ratio denotes the heavier occlusion. From Figure 4, it can be observed that the proposed UTM performs sufficiently well on most settings. Specifically, the proposed method obtains a higher performance for the severely occluded object, e.g., the object is occluded than 50%. The reason is that UTM can leverage identity-aware knowledge to enhance the object detection and feature embedding modules.

4.2. Ablation Study

To prove the effectiveness of the proposed components in UTM, we conduct some ablation studies on the MOT16 validation dataset following [5].

<table>
<thead>
<tr>
<th>Methods</th>
<th>Detector</th>
<th>MOTA</th>
<th>HOTA</th>
<th>IDF1</th>
<th>FP</th>
<th>FN</th>
<th>IDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOT16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FairMOT [64]</td>
<td>CenterNet</td>
<td>73.7</td>
<td>59.3</td>
<td>72.3</td>
<td>27,507</td>
<td>117,477</td>
<td>3,303</td>
</tr>
<tr>
<td>GRTU [48]</td>
<td>CenterNet</td>
<td>73.8</td>
<td>55.8</td>
<td>68.9</td>
<td>28,998</td>
<td>115,104</td>
<td>3,699</td>
</tr>
<tr>
<td>GRTU [48]</td>
<td>CenterNet</td>
<td>74.9</td>
<td>62.0</td>
<td>75.0</td>
<td>32,007</td>
<td>107,616</td>
<td>1,812</td>
</tr>
<tr>
<td>TLR [47]</td>
<td>CenterNet</td>
<td>76.5</td>
<td>60.7</td>
<td>73.6</td>
<td>29,808</td>
<td>99,510</td>
<td>3,369</td>
</tr>
<tr>
<td>MAA [42]</td>
<td>FRCNN</td>
<td>79.4</td>
<td>62.0</td>
<td>75.9</td>
<td>37,320</td>
<td>77,661</td>
<td>1,452</td>
</tr>
<tr>
<td>ByteTrack [63]</td>
<td>YOLOX</td>
<td>80.3</td>
<td>63.1</td>
<td>77.3</td>
<td>25,491</td>
<td>83,721</td>
<td>2,196</td>
</tr>
<tr>
<td>UTM</td>
<td>FRCNN</td>
<td>81.8</td>
<td>64.0</td>
<td>78.7</td>
<td>25,077</td>
<td>76,298</td>
<td>1,431</td>
</tr>
<tr>
<td>MOT17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FairMOT [64]</td>
<td>CenterNet</td>
<td>73.7</td>
<td>59.3</td>
<td>72.3</td>
<td>27,507</td>
<td>117,477</td>
<td>3,303</td>
</tr>
<tr>
<td>PermaTrack [44]</td>
<td>CenterNet</td>
<td>73.8</td>
<td>55.8</td>
<td>68.9</td>
<td>28,998</td>
<td>115,104</td>
<td>3,699</td>
</tr>
<tr>
<td>GRTU [48]</td>
<td>CenterNet</td>
<td>74.9</td>
<td>62.0</td>
<td>75.0</td>
<td>32,007</td>
<td>107,616</td>
<td>1,812</td>
</tr>
<tr>
<td>TLR [47]</td>
<td>CenterNet</td>
<td>76.5</td>
<td>60.7</td>
<td>73.6</td>
<td>29,808</td>
<td>99,510</td>
<td>3,369</td>
</tr>
<tr>
<td>MAA [42]</td>
<td>FRCNN</td>
<td>79.4</td>
<td>62.0</td>
<td>75.9</td>
<td>37,320</td>
<td>77,661</td>
<td>1,452</td>
</tr>
<tr>
<td>ByteTrack [63]</td>
<td>YOLOX</td>
<td>80.3</td>
<td>63.1</td>
<td>77.3</td>
<td>25,491</td>
<td>83,721</td>
<td>2,196</td>
</tr>
<tr>
<td>UTM</td>
<td>FRCNN</td>
<td>81.8</td>
<td>64.0</td>
<td>78.7</td>
<td>25,077</td>
<td>76,298</td>
<td>1,431</td>
</tr>
<tr>
<td>MOT20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FairMOT [64]</td>
<td>CenterNet</td>
<td>61.1</td>
<td>54.9</td>
<td>67.3</td>
<td>103,440</td>
<td>88,901</td>
<td>5,243</td>
</tr>
<tr>
<td>MAA [42]</td>
<td>FRCNN</td>
<td>73.9</td>
<td>57.3</td>
<td>71.2</td>
<td>24,942</td>
<td>108,744</td>
<td>1,331</td>
</tr>
<tr>
<td>ReMOT [56]</td>
<td>CenterNet</td>
<td>77.4</td>
<td>61.2</td>
<td>73.1</td>
<td>28,351</td>
<td>86,659</td>
<td>1,789</td>
</tr>
<tr>
<td>ByteTrack [63]</td>
<td>YOLOX</td>
<td>77.8</td>
<td>61.3</td>
<td>75.2</td>
<td>26,249</td>
<td>87,594</td>
<td>1,223</td>
</tr>
<tr>
<td>UTM</td>
<td>FRCNN</td>
<td>78.2</td>
<td>62.5</td>
<td>76.9</td>
<td>29,964</td>
<td>81,516</td>
<td>1,228</td>
</tr>
</tbody>
</table>

Table 2. Comparison on private detection setting of MOTChallenge benchmarks. The best results are marked in bold.

lative results in Table 2. The compared methods based on three different detectors, e.g., Faster R-CNN [35], CenterNet [66], and YOLOX [14]. As shown in Table 2, UTM achieves the better performance than existing methods on most of the evaluation metrics. In terms of the most important evaluation metric MOTA, the proposed method obtains an obvious improvement upon the current state-of-the-art performance, e.g., 4.5% improvement on MOT16 dataset. Among the compared method, the most related work to ours is FairMOT [64], which designs a joint detection and embedding network based on CenterNet. The major difference and novelty of ours is that we generate a positive feedback loop with identity-aware feature enhancement module in UTM. Compared with FairMOT, the proposed method obtains an obvious improvement, e.g., 16.4% improvements on MOT of MOT20 dataset. Compared with MAA [42], which applies the similar detector Faster R-CNN, UTM achieves a noticeable improvements on MOT17 and MOT20 datasets. We attribute the performance improvement to the proposed UTM leverages identity-aware knowledge to enhance the object detection and feature embedding modules.

Figure 4. Illustration of the relation between the tracked objects number and occlusion ratio. Wider gray bars show the occurrence of ground truth object bboxes in each occlusion level interval, while narrower colored bars illustrate the percentage of objects tracked for their respective method. Note that occurrences and tracked percentages are not drawn in the same unit.
using memory aggregation module, we simply utilize average pooling layer to generate the tracklet feature as the input of IAFE. Without the memory aggregation module, we observe the decrease on MOTA, IDF1, and HOTA. This indicates that the discriminative memory aggregation exactly obtain more robust identity-aware knowledge to benefit IAFE, so that further boost the detection and embedding.

**Effect of graph matching:** To verify the effectiveness of the graph matching layer, we compare it with the Hungarian algorithm [24]. As shown in Table 5, compared with the Hungarian algorithm, graph matching (GM) obtains the improvement of 5.1% in term of IDF1. The reason is that the Hungarian algorithm ignores the second-order edge-to-edge similarity that can model the group activity to generate more reliable tracklets. Furthermore, the obvious improvement on IDF1 demonstrates the robust association of the graph matching with high-order context information.

**Effect of node aggregation rules:** We further analyze the aggregation rules used for feature interaction. As shown in Table 6, the node aggregation rule “Type 3” obtains the best performance. The reason is that the network can focus on the node with a high affinity score between the detection graph and tracklet graph.

**Effect of message passing steps $L$:** As shown in Figure 5(a), we observe a clear upward tendency for both IDF1 and HOTA from 0 to 3 steps, and then they tend to be flat after $L = 3$. Hence, we use $L = 3$ in this work.

**Effect of geometric threshold $\lambda_1$:** We summarize the results for the effect of the geometric threshold $\lambda_1$ in Figure 5(b), it can be observed that IDF1 and HOTA metrics increase significantly from 0.1 to 0.7, and then decrease after $\lambda_1 = 0.7$. Thus, we set $\lambda_1 = 0.7$ in IAFE.

**Effect of affinity threshold $\gamma$:** We also analyze the effect of the affinity threshold $\gamma$ in graph matching and the related results are summarized in Figure 5(c), it can be observed that HOTA and IDF1 metrics first increase for $\gamma \in [0.1, 0.6]$ and then slowly decrease for $\gamma > 0.6$. Thus, the proposed method works best when $\gamma = 0.6$.

**Effect of memory length $\eta$:** We conduct an experiment to show the effect of memory length. As illustrated in Figure 5(d), the HOTA and IDF1 metrics reach the best performance for $\eta = 30$.

### 5. Conclusion

In this work, we propose a Unified Tracking Model (UTM) to generate a positive feedback loop with multi benefits, which introduces the Identity-Aware Feature Enhancement (IAFE) module to bridge and benefit object detection, feature embedding, and identity association. IAFE leverages the identity-aware knowledge to enhance the detection and embedding modules, thereby generating reliable tracklets by identity association. Specifically, IAFE consists of identity-aware boosting attention and identity-aware erasing attention, where the boosting attention and erasing attention are utilized to enhance and suppress regions of the current frame feature. The proposed method achieves the best performances on three benchmarks, which illustrates the effectiveness of UTM. In the future, we will optimize the embedding branch to reduce the influence caused by the small batch size for the end-to-end training.

**Acknowledgements.** This work was supported by National Key Research and Development Project (2020AAA0106200), National Natural Science Foundation of China (61936005, 62036012, 61721004, U21B2044), Beijing Natural Science Foundation (L201001.422039), and Natural Science Research Start-up Foundation of Recruiting Talents of Nanjing University of Posts and Telecommunications (NY222116).
References


[16] Jiawei He, Zehao Huang, Naiyan Wang, and Zhaoxiang Zhang. Learnable graph matching: Incorporating graph partitioning with deep feature learning for multiple object tracking. In *CVPR*, pages 5299–5309, 2021. 1, 2, 6


[52] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In ICIP, pages 3645–3649, 2017. 1, 2
[61] Li Zhang, Yuan Li, and Ramakant Nevatia. Global data association for multi-object tracking using network flows. In CVPR, pages 1–8, 2008. 2