Unsupervised Embedding and Association Network for Multi-Object Tracking

Yu-Lei Li
School of Informatics, Xiamen University
yuleili2008@gmail.com

Abstract

How to generate robust trajectories of multiple objects without using any manual identity annotation? Recently, identity embedding features from Re-ID models are adopted to associate targets into trajectories. However, most previous methods equipped with embedding features heavily rely on manual identity annotations, which bring a high cost for the multi-object tracking (MOT) task. To address the above problem, we present an unsupervised embedding and association network (UEANet) for learning discriminative embedding features with pseudo identity labels. Specifically, we firstly generate the pseudo identity labels by adopting a Kalman filter tracker to associate multiple targets into trajectories and assign a unique identity label to each trajectory. Secondly, we train the transformer-based identity embedding branch and MLP-based data association branch of UEANet with these pseudo labels, and UEANet extracts branch-dependent features for the unsupervised MOT task. Experimental results show that UEANet confirms the outstanding ability to suppress IDS and achieves comparable performance compared with state-of-the-art methods on three MOT datasets.

1 Introduction

Multi-object tracking (MOT), which aims at estimating reliable trajectories of multiple targets in a video sequence, has a wide range of applications, for example, modern autonomous driving [Milan *et al.*, 2016] and intelligent robot vision analysis [Dendorfer *et al.*, 2020]. Recent MOT methods mainly adopt the tracking-by-detection paradigm, i.e., they continuously detect multiple targets with detectors and then associate those detection predictions into trajectories with different data association algorithms [Zhang *et al.*, 2021b; Wu *et al.*, 2021].

Despite different association algorithms, some MOT methods [Zou, 2020; Wu *et al.*, 2021; Zhang *et al.*, 2021b; Wang *et al.*, 2021b; Guo *et al.*, 2021; Sun *et al.*, 2021; Xu *et al.*, 2020] achieve the impressive performance by addressing the online MOT task with the following two steps: (a) detection and identity embedding (or motion predicting), and (b) embedding feature/motion prediction-based data association. The above two-step tracking procedure inspires three feasible ways to improve tracking performance. For example, some methods [Zhang *et al.*, 2021b; Wang *et al.*, 2021b] adopt the anchor-free detector [Zhou *et al.*, 2019] to discover occluded targets. Some other methods [Zou, 2020; Wu *et al.*, 2021; Zhang *et al.*, 2021b; Wang *et al.*, 2021b; Guo *et al.*, 2021] use an extra identity embedding (or motion predicting) branch for simultaneously generating detection predictions and embedding features (or motion predictions) in their unified networks with the manual annotations of targets. In terms of data association, some deep data association algorithms [Sun *et al.*, 2021; Xu *et al.*, 2020] bring lower IDS and higher Multi-Object Tracking Accuracy (MOTA [Bernardin and Stiefelhagen, 2008]) than hand-crafted constraints.

Figure 1: Visualization of our tracking results on MOT2017 (top-2 rows) and MOT2020 (bottom-2 rows). Each row shows the sampled frames with a 20-frame interval. Different colors represent different identities, and targets with the same identity are assigned to the same trajectory.

However, those previous methods mainly focus on the hand-crafted association algorithms, which suppresses the ability to effectively exploit discriminative embedding features to generate robust trajectories. Moreover, learning embedding features with the manual identity annotations brings a high economic cost for supervised MOT methods.

In this paper, we present a novel unsupervised transformer-based embedding and association network (UEANet, as shown in Figure 2). To address the problems of high-cost manual annotations and full use of embedding features, UEANet adopts pseudo identity labels to generate discriminative embedding features with a transformer-based discriminative feature enhancer (DFE) module, and it performs an end-to-end data association process in an unsupervised way.

We summarize the main contributions as follows:

- We present a novel unsupervised embedding and association network (UEANet) to learn discriminative embedding features and perform an end-to-end data association process for the unsupervised MOT task.
- We train the transformer-based identity embedding branch and MLP-based association branch of UEANet with pseudo identity labels, which go beyond the limitation of high-cost manual identity annotations used in the supervised MOT task.
- UEANet achieves comparable performance compared with previous advanced methods on three MOT datasets, and it ranks second in terms of MOTA.

2 Related Work

Recent MOT methods. Tracking by detection is the most popular tracking paradigm, as mostly adopted in recent MOT methods [Zou, 2020; Wu *et al.*, 2021; Zhang *et al.*, 2021b; Wang *et al.*, 2021b; Guo *et al.*, 2021; Sun *et al.*, 2021; Xu *et al.*, 2020]. These methods focus on reliable detection predictions and embedding features (or motion predictions) for subsequently associating targets into trajectories with different data association algorithms. For example, TraDes [Wu *et al.*, 2021] extends the advanced detector [Zhou *et al.*, 2019] with the extra motion prediction branch and uses the motion predictions to perform data association based on the bounding-box IoU constraint. Some other methods [Sun *et al.*, 2021; Xu *et al.*, 2020] adopt fully-connected (FC) network-based association algorithms, which further improve their tracking performance. For example, SST [Sun *et al.*, 2021] employs the deep affinity network to use the pairs of object appearance features between different frames for generating trajectories in the end-to-end way, which highly improves the performance on IDS. However, the hand-craft constraints or simple deep association network need discriminative embedding features (or motion predictions) learned in the supervised way. The manual identity annotations are essential in previous supervised methods. In contrast, recent unsupervised method [Liu *et al.*, 2022] trains the re-ID branch without any manual identity label in the unsupervised way. Moreover, it is adversely affected by confusing embedding features caused by unsupervised training.

We follow the unsupervised learning fashion and present a transformer-based embedding and association network to adopt the pseudo identity labels instead of manual annotations for the unsupervised MOT task.

Our unsupervised method. In this paper, we try to generate discriminative embedding features with the proposed discriminative feature enhancer module, and suppress the competition between detection and identity embedding branches. Moreover, the MLP-based data association branch takes the embedding features of candidate trajectories as input and generates final trajectories in an end-to-end way. Note that we train the identity embedding and data association branches without using any manual identity label. The proposed unsupervised embedding and association network (UEANet) breaks through the limitation of high-cost manual annotations, and achieves comparable performance against previous advanced methods.

3 Proposed Method

As the overview of UEANet shown in Figure 2, Our method consists of the detection and identity embedding branches integrated with the transformer-based discriminative feature enhancer (DFE) module, and the subsequent MLP-based data association branch. We train the identity embedding and data association branches with the pseudo identity labels generated by the Kalman filter tracker of ByteTrack [Zhang *et al.*, 2021a]. Next, we introduce the motivation of UEANet to extract discriminative embedding features with the DFE module in the unsupervised way.

3.1 Motivation of Our Unsupervised Method

Due to the high economic cost of manual identity annotations in the supervised methods, we train the proposed UEANet with pseudo identity labels. As the pseudo identity labels are obtained with relatively poor detection predictions, they lack the interfering identity labels of the near-completely occluded targets. However, the manual detection labels and the pseudo identity labels are not one-to-one correspondences, and bring the competition of network optimization between the detection and identity embedding branches, resulting in confusing embedding features for data association. Unlike the previous unsupervised method [Liu *et al.*, 2022] suffering from the confusing embedding features $\mathbf{e}_t^c$ (formulated as Eq.(1)), we develop the DFE module to concurrently enhance the embedding and detection features for generating reliable detection predictions $\mathbf{d}_t$ and discriminative embedding features $\mathbf{e}_t$, which are formulated as

$$f^P : F_t \longrightarrow \mathbf{d}_t, \mathbf{e}_t^c, \tag{1}$$

$$f^U : F_t \longrightarrow \mathbf{d}_t, \mathbf{e}_t, \tag{2}$$

where f^P and f^U denote the previous and the proposed unsupervised detection and embedding branches. F_t denotes the current-frame image. Then, we perform the end-to-end data association, mathematically defined as

$$g_{e2e} : \mathbf{d}_t, \mathbf{e}_t, \mathbf{d}_{t-1}, \mathbf{e}_{t-1} \longrightarrow \mathbf{A}, \tag{3}$$

where g_{e2e} denotes the proposed end-to-end unsupervised association branch that calculates the association matrix $\mathbf{A}$ between the detection predictions from the frames t and $t-1$.

The final trajectories are commonly obtained with the association matrix $\mathbf{A}$, which is learned with pairs of embedding features corresponding to candidate trajectories between the current frame and the associated trajectories, like SST [Sun *et al.*, 2021]. In addition, we jointly train the identity embedding and data association branches to perform MOT in the end-to-end unsupervised way.

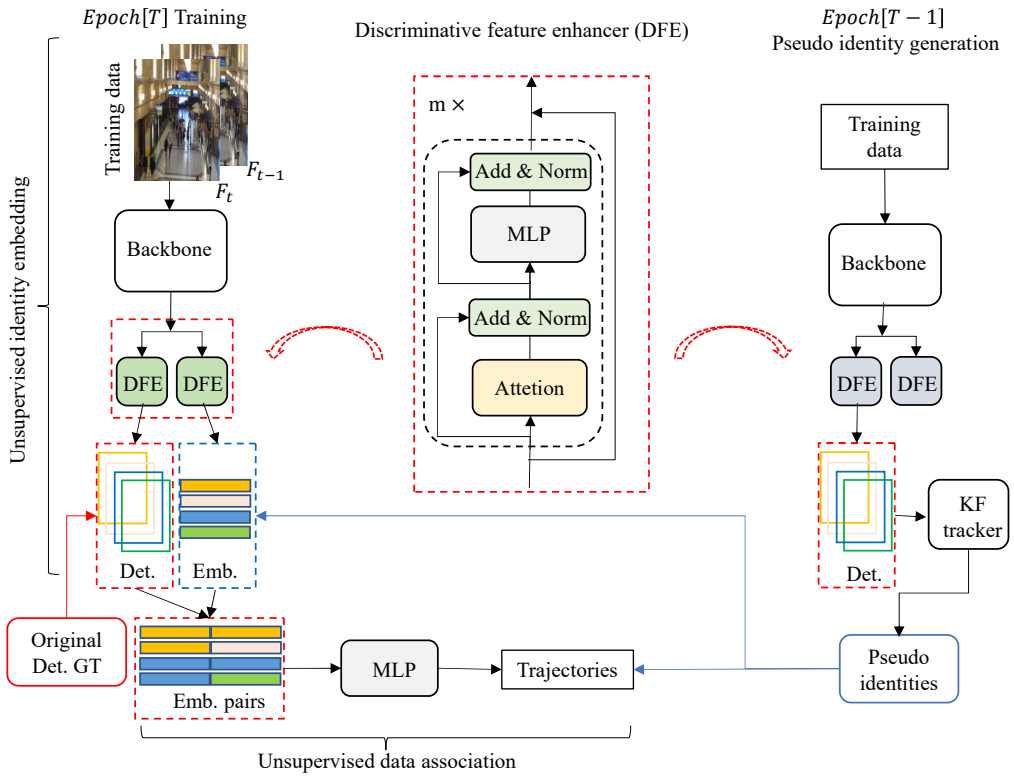

Figure 2: Overview of the proposed UEANet. We extend a detection network by integrating an identity embedding branch to generate discriminative embedding features based on pseudo identities. The subsequent MLP-based data association branch evaluates the association confidence scores of candidate trajectories between the current frame t and the past frame $t-1$. MLP and Attention respectively denote the MLP and attention components used in the discriminative feature enhancer (DFE) module. Emb. pairs represent pairs of embedding features for candidate trajectories. Original Det. GT denote the original manual labels for object detection. Therefore, we finally achieve the trajectories of t with the unsupervised end-to-end transformer-based embedding and association network.

3.2 Unsupervised Identity Embedding

We employ the proposed transformer-based discriminative feature enhancer (DFE) in the detection and identity embedding branches of UEANet, as shown in Figure 2. The DFE module aims to suppress the branch competition and generate reliable detection predictions and discriminative embedding features. It contains two components: attention and MLP.

Attention variants. The attention component of DFE is selected from multiple variants, such as self-attention [Vaswani *et al.*, 2017], SGE-attention [Li *et al.*, 2019], shuffle-attention [Zhang and Yang, 2021], and SPSelf-attention [Liu *et al.*, 2021]. It aims to suppress the branch competition by generating branch-dependent features for the detection and identity embedding branches. For example, SGE-attention [Li *et al.*, 2019] improves the learning of different semantic sub-features of each branch, and intentionally enhances the respective spatial distribution of features for the two branches. The above processes can be formulated as

$$\mathbf{x}_t^c = MLPMixer[SGE - Attention(\mathbf{x}_t)], \quad (4)$$

$$\mathbf{x}_t^p = MLPMixer[SGE - Attention(\mathbf{x}_t)], \quad (5)$$

where $\mathbf{x}_t$ is the input features of the current frame t. $\mathbf{x}_t^c$ and $\mathbf{x}_t^p$ are the detection-dependent features and identity-dependent features for the detection and identity embedding processes, respectively. $MLPMixer$ denotes the MLP component.

MLP variants. Some MLP variants, such as FFN [Vaswani *et al.*, 2017], ResMLP [Touvron *et al.*, 2021], RepMLP [Ding *et al.*, 2021] and MLPMixer [Tolstikhin *et al.*, 2021], are taken into consideration for the MLP component of DFE. For example, MLPMixer [Tolstikhin *et al.*, 2021] is to clearly separate the per-location operations and cross-location operations in multi-layer perceptrons for feature extraction, which enhances the fitting ability of the attention component of DFE, as shown in Eq.(4)&(5). With the cooperation of the two components, the proposed DFE module effectively extracts branch-dependent features to generate reliable detection predictions and discriminative embedding features for the subsequent data association process.

3.3 Unsupervised Data Association

The MLP-based data association branch is integrated into the detection and identity embedding network to replace the hand-craft association algorithms. It is jointly optimized with the detection and identity embedding branches in the unsupervised way. Note that we employ the same MLP component as the one in the DFE module.

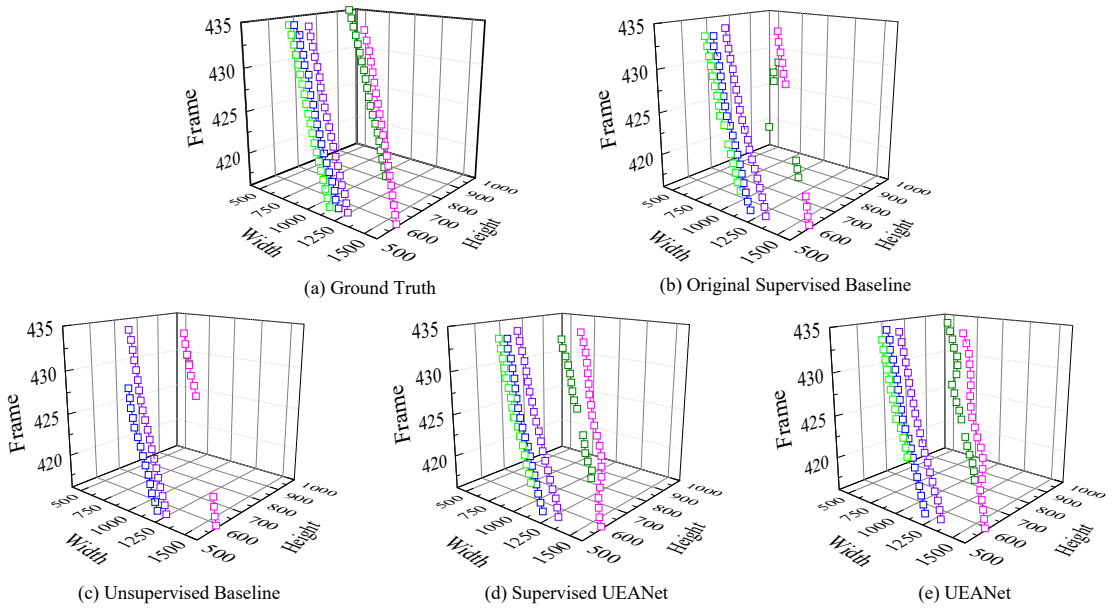

Figure 3: Qualitative comparisons of the predicted trajectories obtained by the UEANet variants in the **first** and **sixth** rows of Table 1. These trajectories are selected from Frame 416 to Frame 435 in the No. 0013 video sequence of MOT2017. The three axes (Height, Width, Frame) represent the two-dimensional spatial positions and one-dimensional temporal position of targets in the predicted trajectories. The identity of each trajectory is marked by its color.

As shown in Figure 2, each detection prediction has the corresponding embedding feature. We build embedding feature pairs to mark each candidate trajectory between detection predictions in the past frame $t-1$ and the current frame t, similar to SST [Sun *et al.*, 2021]. Moreover, for each detection prediction of t, we only select top-n embedding feature pairs to reduce computational cost. The MLP-based data association branch outputs probabilities of candidate trajectories included in the ground truth, i.e., association confidence scores c_t, which are formulated as

$$\mathbf{c}_t = MLPMixer[(\mathbf{e}_t, \mathbf{e}_{t-1})], \tag{6}$$

where $(\mathbf{e}_t, \mathbf{e}_{t-1})$ are pairs of embedding features for candidate trajectories. We keep the trajectories whose association confidence scores are larger than a confidence threshold β.

Compared with the hand-craft association algorithm in [Zhang *et al.*, 2021b] and the FC-based deep association network [Sun *et al.*, 2021], our MLP-based association branch performs the unsupervised end-to-end data association process and achieves robust trajectories (see Figure 3 and Table 2).

3.4 Training Steps and Loss

To adopt unsupervised training of the proposed UEANet, before the epoch T, we firstly generate detection predictions with the detection branch equipped with the network parameters of the past epoch $T-1$. Then, we construct trajectories with the Kalman filter tracker of ByteTrack [Zhang *et al.*, 2021a] for generating the pseudo identity labels. At the epoch T, we adopt the pseudo identity labels $\mathbf{p}_t^G$ to train the identity embedding and association branches while training the detection branch with the original manual detection labels. Moreover, embedding features $\mathbf{e}_t$ are mapped to the corresponding identities $\mathbf{p}_t$ with a FC layer and a softmax function. We split the long trajectories into tracklets to generate the data association labels $\mathbf{c}_t^G$. Therefore, we add the extra identity embedding loss L_{ie} and data association loss L_{da} to the original detection loss L_{det} utilized in the detector [Zhou *et al.*, 2019] for training the whole tracking network with Eq.7.

$$L(t) = L_{det}(\mathbf{d}_t^G, \mathbf{d}_t) + L_{ie}(\mathbf{p}_t^G, \mathbf{p}_t) + L_{da}(\mathbf{c}_t^G, \mathbf{c}_t), \tag{7}$$

where L_{ie} and L_{da} are the common K-class and 2-class cross-entropy losses. K is the number of trajectories. For L_{da}, we set the association label to 1 when the candidate trajectory belongs to the pseudo trajectories; otherwise, we set it to 0.

4 Experiments

4.1 Implementation Details

We evaluate the proposed UEANet on the MOT2016 [Milan *et al.*, 2016], MOT2017 [Milan *et al.*, 2016] and MOT2020 [Dendorfer *et al.*, 2020] datasets. For ablation studies in Section 4.2, we follow the previous methods [Zhang *et al.*, 2021b; Wang *et al.*, 2021b; Wu *et al.*, 2021] to use the first half of each video sequence of the MOT2017 training set for training while using the second half for validation. We follow the unsupervised tracking fashion to train the backbone [Zhang *et al.*, 2021b], and the detection, identity embedding and data association branches of UEANet for 30 epochs with a learning rate of 1×10^{-4} and a mini-batch size of 24 on 4 RTX2080 Ti GPUs (using 2 RTX2080 Ti GPUs for ablation

Attention	MLP	MOTA↑	IDF1↑	IDS↓
None*	None*	69.1	72.9	299
None	None	68.4	69.9	751
None	FFN	69.6	74.4	271
None	RepMLP	70.2	75.1	301
None	ResMLP	70.1	74.1	373
None	MLPMixer	70.2	74.5	372
Self-Attention	None	69.5	74.2	320
SGE-Attention	None	69.4	73.7	374
Shuffle-Attention	None	70.4	75.6	327
SPSelf-Attention	None	69.7	74.6	309
SGE-Attention	FFN	69.9	74.5	314
SGE-Attention	RepMLP	69.7	74.8	**265**
SGE-Attention	ResMLP	69.2	73.5	387
Self-Attention	MLPMixer	69.7	75.0	350
Shuffle-Attention	MLPMixer	70.5	75.1	327
SPSelf-Attention	MLPMixer	70.3	75.0	331
SGE-Attention*	MLPMixer*	70.5	75.4	298
SGE-Attention	MLPMixer	**70.7**	**75.7**	295

Table 1: Ablation studies of the attention and MLP variants on the MOT2017 validation set. Attention denotes the attention component used in the discriminative feature enhancer (DFE) module. MLP denotes the MLP component used in the DFE module and the MLP-based association branch. * means supervised training. The best result is marked in bold.

studies). We set m to 1 for both detection and identity embedding branches, and experimentally set n to 3. The association confidence threshold β is 0.4 during inference.

Following the prior methods [Zou, 2020; Wu *et al.*, 2021; Zhang *et al.*, 2021b], we use the common evaluation metrics: HOTA, IDF1, and the CLEAR metrics [Bernardin and Stiefelhagen, 2008], etc..

4.2 Ablation Studies

Supervised vs. unsupervised tracking. We evaluate the baseline [Zhang *et al.*, 2021b] and the proposed UEANet based on both supervised and unsupervised training. The results are shown in the first and sixth rows of Table 1. UEANet achieves comparable performance on MOTA, IDF1 and IDS compared with the supervised UEANet, while the supervised baseline outdistances the unsupervised one. These results indicate that UEANet effectively suppresses the adverse influence of the branch competition caused by the pseudo labels and extracts discriminative branch-dependent features by the DFE module for the unsupervised MOT task.

Attention variants. To validate the contribution of the attention component, we adopt the different attention variants in the DFE module and observe their influences on performance. The comparison results are shown in the third, fifth and sixth rows of Table 1. Compared to the unsupervised baseline, the UEANet variants only equipped with the attention component, achieve the different and better tracking results on MOTA, IDF1 and IDS. Moreover, the UEANet variants equipped with both attention and MLPMixer components, further improve the performance on MOTA. For example, UEANet (SGE-Attention + MLPMixer) achieves the improvement of 1.0%/0.7% on MOTA/IDF1 compared with UEANet (Self-Attention + MLPMixer). These observations confirm the important role of the attention component in the DFE module for generating branch-dependent features.

MLP variants. As shown in Table 1, we use four kinds of MLP variants in the detection, identity embedding and data association branches to evaluate their contributions to UEANet. The UEANet variants achieve better tracking performance than the unsupervised baseline while performing the end-to-end MLP-based data association process. Moreover, different MLP variants bring different performance in terms of the cooperation between the attention and MLP components. For example, UEANet (SGE-Attention + MLPMixer) has better performance on MOTA than UEANet (SGE-Attention + FFN), which verifies the different but effective fitting abilities of MLP variants to generate discriminative embedding features and reliable trajectories.

Visualization of trajectories. As shown in Figure 3, we present the visualization of the predicted trajectories obtained by the UEANet variants. The False Negatives (FN [Bernardin and Stiefelhagen, 2008]) of targets are displayed in the figure. we visually evaluate the performance of these variants according to the completeness of trajectories. The proposed UEANet obtains the best trajectories against other variants, which visually proves the superiority of UEANet to suppress the adverse influence of the pseudo identity labels and generate robust trajectories for the unsupervised MOT task.

4.3 Experiments on the MOT Test Sets

MOT2016 and MOT2017. For MOT2016 and MOT2017 [Milan *et al.*, 2016], the official MOTChallenge evaluation server obtains the tracking results for the proposed UEANet. As shown in the first and second rows of Table 2, UEANet ranks top-1 or top-2 among all the published methods. Compared with the second-performance method [Wang *et al.*, 2021b], it achieves an improvement of 1.3%/3.3%/2.2% on MOTA/IDF1/HOTA for MOT2016. Note that UEANet achieves the desirable performance with less computational cost than the top-1 method, ByteTrack [Zhang *et al.*, 2021a] (4 RTX2080 Ti GPUs vs. 8 V100 GPUs) and meanwhile decreases IDS by 30.2% from 2196 to 1533 on MOT2017. Moreover, it outperforms the unsupervised method, UTracK [Liu *et al.*, 2022] on MOTA, IDF1 and IDS. These results indicate that UEANet is a strong end-to-end baseline for the unsupervised tracking task. It exploits the transformer-based DFE module to learn discriminative embedding features and generate reliable trajectories based on the pseudo identity labels.

MOT2020. The MOT2020 [Dendorfer *et al.*, 2020] dataset includes much more persons per frame and heavier occlusion cases than the MOT2017 dataset. As shown in the third row of Table 2, the proposed UEANet also ranks top-2 among all the published methods. However, it only achieves the relatively lower tracking performance (on MOTA, HOTA and IDS) on MOT2020 compared to MOT2017. The reason is most likely the heavier occlusion cases of MOT2020, which have the adverse influence on the pseudo identity labels and

Dataset	Method	Pub.&Year	MOTA↑	IDF1↑	HOTA↑	MT↑	ML↓	IDS↓	FPS↑
MOT16	CTrackerV1 [Peng *et al.*, 2020]	ECCV'20	67.6	57.2	49.2	32.9	23.1	1897	6.8
	FairMOT [Zhang *et al.*, 2021b]	IJCV'21	74.9	*72.8*	59.8	44.7	*15.9*	*1074*	**25.9**
	CSTrack [Zou, 2020]	arXiv'20	75.6	73.3	*59.8*	42.3	16.5	1121	15.8
	FUFET [Shan *et al.*, 2020]	arXiv'20	*76.5*	68.6	58.3	**52.8**	**12.3**	1026	6.6
	CorrTracker [Wang *et al.*, 2021b]	CVPR'21	76.6	74.3	61.0	47.8	21.7	979	15.9
	GSDT_V2 [Wang *et al.*, 2021a]	ICRA'21	73.2	66.5	55.2	41.7	17.5	3891	1.6
	TraDeS [Wu *et al.*, 2021]	CVPR'21	70.1	64.7	53.2	37.3	20.0	1144	*22.3*
	UTrack† [Liu *et al.*, 2022]	Neuro'22	74.2	71.1	-	*44.8*	14.0	1324	*24.8*
	UEANet(Ours)†	This paper	**77.9**	**77.6**	**63.2**	43.5	17.3	**491**	25.1
MOT17	CTrackerV1 [Peng *et al.*, 2020]	ECCV'20	66.6	57.4	49.0	32.2	24.2	5529	6.8
	FairMOT [Zhang *et al.*, 2021b]	IJCV'21	73.7	72.3	59.3	43.2	17.3	3303	25.9
	CSTrack [Zou, 2020]	arXiv'20	74.9	72.6	59.3	41.5	17.5	3567	15.8
	FUFET [Shan *et al.*, 2020]	arXiv'20	76.2	*73.6*	57.9	51.1	13.6	3237	6.8
	TransCenter [Xu *et al.*, 2021]	arXiv'21	73.2	62.2	54.5	41.1	19.0	*2964*	1.0
	CorrTracker [Wang *et al.*, 2021b]	CVPR'21	*76.5*	*73.6*	*60.7*	*47.6*	**12.7**	3369	15.6
	GSDT_V2 [Wang *et al.*, 2021a]	ICRA'21	73.2	66.5	55.2	41.7	19.0	3891	4.9
	TraDeS [Wu *et al.*, 2021]	CVPR'21	69.1	63.9	52.7	36.4	21.5	3555	22.3
	ByteTrack [Zhang *et al.*, 2021a]	arXiv'21	**80.3**	**77.3**	**63.1**	**53.2**	*14.5*	2196	**29.6**
	UTrack† [Liu *et al.*, 2022]	Neuro'22	73.5	70.2	-	43.3	15.2	4110	*25.4*
	UEANet(Ours)†	This paper	77.2	77.0	62.7	41.7	19.0	**1533**	25.1
MOT20	CSTrack [Zou, 2020]	arXiv'20	66.6	68.6	54.0	50.4	15.5	*3196*	0.2
	TransCenter [Xu *et al.*, 2021]	arXiv'21	61.9	50.4	44.3	49.4	15.5	4653	1.0
	TransTrack [Sun *et al.*, 2020]	arXiv'20	64.5	59.2	48.5	49.1	13.6	3565	14.9
	FairMOT [Zhang *et al.*, 2021b]	IJCV'21	61.8	67.3	*54.6*	68.8	**7.6**	5243	*13.2*
	CorrTracker [Wang *et al.*, 2021b]	CVPR'21	65.2	*69.1*	-	*66.4*	8.9	5183	8.5
	GSDT_V2 [Wang *et al.*, 2021a]	ICRA'21	*67.1*	67.5	53.6	53.1	13.2	3230	1.5
	ByteTrack [Zhang *et al.*, 2021a]	arXiv'21	**77.8**	75.2	**61.3**	**69.2**	*9.5*	**1223**	**17.5**
	UTrack† [Liu *et al.*, 2022]	Neuro'22	68.5	69.4	-	57.9	12.2	2147	12.4
	UEANet(Ours)†	This paper	73.0	**75.6**	58.6	55.0	13.9	1423	12.8

Table 2: Comparisons of the proposed UEANet with state-of-the-art methods on the three MOT test sets. 'Pub.&Year' denotes the article publishers and publication years of previous methods. '↑'('↓') means that the higher (lower) result is the better. '†' denotes an unsupervised method. The top-3 best results for each metric are marked in **bold**, underlined and *italics*, respectively.

the unsupervised training of UEANet. Note that UEANet surpasses UTracK [Liu *et al.*, 2022] and achieves the excellent IDF1 on all three datasets, which directly demonstrates the strong ability to extract the discriminative embedding features for identifying similar targets. The results further indicate that UEANet effectively suppresses the branch competition caused by the pseudo identity labels. It effectively extracts the branch-dependent features with the DFE module for the unsupervised MOT task, even in crowded scenarios.

Visualization of test results. We show the tracking results obtained on the MOT2017 and MOT2020 test sets in Figure 1. These tracking results confirm the outstanding ability of UEANet to deal with the unsupervised MOT task.

Failure examples. As the predicted trajectories shown in Figure 3, the FN of the UEANet variants are visually displayed in the No. 0013 sequence of the MOT2017 validation set. The proposed UEANet achieves much better trajectories than the unsupervised baseline. However, there are still many FN compared with the ground-truth trajectories, which means there is still room for improvement in our method.

5 Conclusion

In this paper, we present a novel tracking network integrated with a transformer-based identity embedding branch and a MLP-based data association branch to perform the unsupervised MOT task with pseudo identity labels instead of high-cost manual identity annotations. To alleviate the adverse influence of these pseudo labels, we develop a discriminative feature enhancer (DFE) module to suppress the branch competition and extract branch-dependent features for the unsupervised identity embedding and end-to-end data association processes. We evaluate many UEANet variants in the experiments and make the proposed UEANet a strong baseline for the subsequent researches. Extensive evaluations demonstrate the superiority of our unsupervised method on the three MOT datasets.

References

[Bernardin and Stiefelhagen, 2008] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008.

[Dendorfer *et al.*, 2020] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020.

[Ding *et al.*, 2021] Xiaohan Ding, Chunlong Xia, Xiangyu Zhang, Xiaojie Chu, Jungong Han, and Guiguang Ding. Repmlp: re-parameterizing convolutions into fully-connected layers for image recognition. *arXiv preprint arXiv:2105.01883*, 2021.

[Guo *et al.*, 2021] Song Guo, Jingya Wang, Xinchao Wang, and Dacheng Tao. Online multiple object tracking with cross-task synergy. In *CVPR*, pages 8136–8145, 2021.

[Li *et al.*, 2019] Xiang Li, Xiaolin Hu, and Jian Yang. Spatial group-wise enhance: Improving semantic feature learning in convolutional networks. *arXiv preprint arXiv:1905.09646*, 2019.

[Liu *et al.*, 2021] Huajun Liu, Fuqiang Liu, Xinyi Fan, and Dong Huang. Polarized self-attention: towards high-quality pixel-wise regression. *arXiv preprint arXiv:2107.00782*, 2021.

[Liu *et al.*, 2022] Qiankun Liu, Dongdong Chen, Qi Chu, Lu Yuan, Bin Liu, Lei Zhang, and Nenghai Yu. Online multi-object tracking with unsupervised re-identification learning and occlusion estimation. *Neurocomputing*, 2022.

[Milan *et al.*, 2016] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016.

[Peng *et al.*, 2020] Jinlong Peng, Changan Wang, Fangbin Wan, Yang Wu, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. In *ECCV*, volume 12349, pages 145–161, 2020.

[Shan *et al.*, 2020] Chaobing Shan, Chunbo Wei, Bing Deng, Jianqiang, Huang, Xian-Sheng Hua, Xiaoliang Cheng, and Kewei Liang. Tracklets predicting based adaptive graph tracking. *arXiv preprint arXiv:2010.09015*, 2020.

[Sun *et al.*, 2020] Peize Sun, Yi Jiang, Rufeng Zhang, Enze Xie, Jinkun Cao, Xinting Hu, Tao Kong, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple-object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020.

[Sun *et al.*, 2021] ShiJie Sun, Naveed Akhtar, HuanSheng Song, Ajmal Mian, and Mubarak Shah. Deep affinity network for multiple object tracking. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):104–119, 2021.

[Tolstikhin *et al.*, 2021] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*, 2021.

[Touvron *et al.*, 2021] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, et al. Resmlp: Feedforward networks for image classification with data-efficient training. *arXiv preprint arXiv:2105.03404*, 2021.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[Wang *et al.*, 2021a] Qiang Wang, Yun Zheng, Pan Pan, and Yinghui Xu. Joint object detection and multi-object tracking with graph neural networks. In *ICRA*, pages 1–1, 2021.

[Wang *et al.*, 2021b] Qiang Wang, Yun Zheng, Pan Pan, and Yinghui Xu. Multiple object tracking with correlation learning. In *CVPR*, pages 3876–3886, 2021.

[Wu *et al.*, 2021] Jialian Wu, Jiale Cao, Liangchen Song, Yu Wang, Ming Yang, and Junsong Yuan. Track to detect and segment: An online multi-object tracker. In *CVPR*, pages 12352–12361, 2021.

[Xu *et al.*, 2020] Yihong Xu, Aljosa Osep, Yutong Ban, Radu Horaud, Laura Leal-Taixé, and Xavier Alameda-Pineda. How to train your deep multi-object tracker. In *CVPR*, pages 6786–6795, 2020.

[Xu *et al.*, 2021] Yihong Xu, Yutong Ban, Guillaume Delorme, Chuang Gan, Daniela Rus, and Xavier Alameda-Pineda. Transcenter: Transformers with dense queries for multiple-object tracking. *arXiv preprint arXiv:2103.15145*, 2021.

[Zhang and Yang, 2021] Qing-Long Zhang and Yu-Bin Yang. Sa-net: Shuffle attention for deep convolutional neural networks. In *ICASSP*, pages 2235–2239, 2021.

[Zhang *et al.*, 2021a] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. *arXiv preprint arXiv:2110.06864*, 2021.

[Zhang *et al.*, 2021b] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129(11):3069–3087, 2021.

[Zhou *et al.*, 2019] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

[Zou, 2020] Chao Liang Zhipeng Zhang Yi Lu Xue Zhou Bing Li Xiyong Ye Jianxiao Zou. Rethinking the competition between detection and reid in multi-object tracking. *arXiv preprint arXiv:2010.12138*, 2020.