LEARNING CORRELATION FOR ONLINE MULTIPLE OBJECT TRACKING

Ying Wang, Chihui Zhuang, Haihui Ye, Yan Yan, Hanzi Wang⋆

Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, Xiamen, China
{yingwang7, chzhuang}@stu.xmu.edu.cn, haihui_ye1@163.com, {yanyan, hanzi.wang}@xmu.edu.cn

ABSTRACT

Existing multiple object tracking methods usually strengthen data association by discriminative identity embeddings. However, many works treat object detection and association as two individual tasks, thus gaining limited benefits. In this paper, we follow the joint detection and tracking paradigm to learn correlation for online multiple object tracking. The proposed method, named LCTrack, links the two tasks by an attention mechanism. Specifically, for robust feature representations, we introduce an identity-aware attention module to extract reliable identity embeddings and model their correlation between two consecutive frames. Furthermore, for effective correlation learning, we design a target-aware loss to train the identity embedding extraction, which is well compatible with the detection task. Therefore, LCTrack can boost the position prediction and data association by the enhanced feature representation. Experimental results on the MOTChallenge benchmarks demonstrate the effectiveness and favorable performance of the proposed LCTrack in comparison with state-of-the-art methods.

Index Terms— Multi-object tracking, correlation learning, identity-aware attention, target-aware loss

1. INTRODUCTION

Multiple object tracking is one of the fundamental and significant tasks in computer vision, which benefits extensive applications such as video analysis [1], human action recognition [2], and autonomous vehicles [3].

Most recent tracking methods follow two paradigms: the tracking-by-detection (TBD) paradigm and the joint detection and tracking (JDT) paradigm. The TBD methods [4, 5, 6, 7, 8, 9] decompose multiple object tracking into object detection and data association. For object detection, an off-the-shelf detector is applied to localize objects frame by frame. For data association, existing methods either employ another network to extract re-identification (re-ID) embeddings or compute the interaction of union. Although great progress has been made on detection and association, these TBD methods cannot optimize the network in an end-to-end manner. Thus, the association measurements with stronger discriminability still achieve undesirable performance due to the low-quality detections.

The JDT methods usually share the parameters of the backbone network [10, 11, 12, 13, 14, 15, 16] and most of them employ a re-ID branch to obtain objects' identity embedding features [17, 18, 19]. To reduce the inference time, these methods simultaneously use a single backbone network to extract identity embeddings and predict detections. Without using the additional re-ID branch, CenterTrack [10] adopts an anchor-free detector [20] to localize object centers and trains the detector to learn a 2D tracking offset. Based on the predicted tracking offset and the center distance, CenterTrack utilizes a greedy matching strategy to associate the object centers between two adjacent frames. Nevertheless, the JDT methods still suffer from two limitations. On the one hand, the context relation supports trackers to distinguish different objects. However, most detectors usually localize objects without taking full use of the correlation information between consecutive frames. On the other hand, the re-ID embedding features support association across frames. However, the learning of embedding features is difficult to be compatible with object detection, because of the different objectives between the detection loss and the re-ID loss. Therefore, the performance of these JDT trackers is inferior to the TBD methods.

To address the problems mentioned above, we propose a simple yet effective method, called as LCTrack, built upon CenterTrack, where each point on feature maps stands for a target center or background. Specifically, we introduce an identity-aware attention module to exploit the correlation in the videos. We first extract point-wise identity embedding features by an embedding network and the correlation of the identity embeddings is modeled to enhance appearance features of the current frame. The enhanced features contain the correlation information between objects and their surroundings, which enables the tracker to focus on targets while suppressing distractors when occlusions. Furthermore, we design a target-aware loss to supervise the embedding extraction, which focuses on the difference of different identities and the distance between the identity embeddings of the predicted and the ground-truth center positions. Thus, our LCTrack pro-

⋆Corresponding author: Hanzi Wang, hanzi.wang@xmu.edu.cn.

This work is supported by the National Natural Science Foundation of China under Grant Nos. U21A20514, 61872307 and 62071404, and by the Natural Science Foundation of Fujian Province under Grant No. 2020J01001.

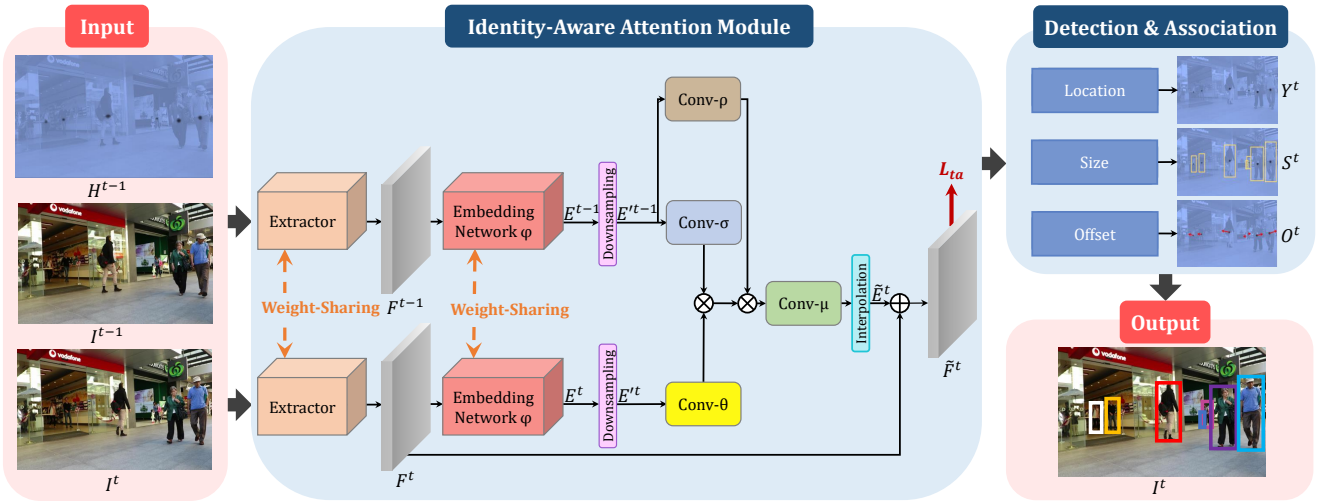

Fig. 1. Overview of our LCTrack. The appearance features of the current frame are enhanced with the identity-aware attention module to construct the correlation of identities between consecutive frames. The target-aware loss (L_{ta}) is used to supervise the identity embeddings extraction. "$\otimes$" stands for a matrix multiplication and "$\oplus$" for an element-wise sum.

vides the discriminative embeddings for the identity-aware attention module to generate robust appearance features. Finally, we predict the detections and infer the tracking offsets from the enhanced features to generate tracklets.

The main contributions of this work are summarized as follows. At first, we introduce an identity-aware attention module to model the correlation between identity embeddings in two frames for robust object detection and association. Secondly, we design the target-aware loss to provide reliable identity embeddings for learning correlation effectively, which is compatible with the detection loss. Finally, experimental results demonstrate the promising tracking performance of the proposed method against several state-of-the-art trackers on the MOTChallenge benchmarks.

2. METHODOLOGY

2.1. Overview

In this paper, we propose LCTrack, based on CenterTrack [10], to learn correlation for online multiple object tracking. The framework is shown in Figure 1.

We first take the current frame I^t, the previous frame I^{t-1}, and the heatmap H^{t-1} rendered from previous frame target centers as inputs to obtain two feature maps of the frames F^t and F^{t-1}, which are generated by a feature extractor. We introduce the identity-aware attention module to extract identity embedding features of the two frames, *i.e.*, E^{t-1} and E^t, and obtain the enhanced embeddings $\tilde{E}^t$ by calculating their correlation. Then we aggregate $\tilde{E}^t$ with the features of the current frame F^t to extract robust appearance features. The target-aware loss L_{ta} is used to help learning correlation by extracting discriminative identity embeddings. According to the enhanced appearance features, we predict targets' locations Y^t, sizes S^t, and tracking offsets O^t. Thus the final output is obtained by associating the tracking offsets.

2.2. Identity-Aware Attention Module

The correlation information supports predicting target motion between adjacent frames. To model the temporal connection of targets, we propose the identity-aware attention module to learn the correlation of identities.

Given two appearance features F^t and F^{t-1}, we use an embedding network $\varphi(\cdot)$ to extract their point-wise identity embedding features:

$$E^t = \varphi(F^t). \tag{1}$$

To achieve efficient calculation, we downsample the identity embeddings to generate E^*. Then, we obtain the identity-aware attention from the embeddings. The process can be described as follows:

$$g(E^{*t}, E^{*t-1}) = \theta(E^{*t})\sigma(E^{*t-1})\rho(E^{*t-1}), \tag{2}$$

where $\theta(\cdot)$ and $\sigma(\cdot)$ are convolution layers with 1×1 kernels to calculate the correlation of identity embeddings. $\rho(\cdot)$ is another convolution layer to generate a representation of E^{*t-1} and the embeddings containing the spatio-temporal correlation are generated. A convolution layer μ and interpolations Up are performed on these embeddings to obtain the enhanced embeddings $\tilde{E}^t$ for the output:

$$\tilde{E}^t = Up(\mu(g(E^{*t}, E^{*t-1}))). \tag{3}$$

The identity-aware attention not only captures the correlation of identities, but also takes the background information

into account. Furthermore, we aggregate the current-frame feature F^t with $\tilde{E}^t$ to generate the enhanced features $\tilde{F}^t$:

$$\tilde{F}^t = F^t \oplus \omega \tilde{E}^t, \tag{4}$$

where $\oplus$ represents the element-wise addition and ω denotes the weight to balance the attention for the output.

Note that H^{t-1} is combined with I^t to extract the appearance features F^t. Based on the point-wise embeddings, we model the correlation between adjacent frames to further enhance the feature representation in the current frame. This way can capture rich identity information for predicting the tracking displacement in two frames and allow the tracker to deal with crowded scenes. Each value in $\tilde{E}^t$ stands for the affinity between the object p_i^t at frame t and all objects at time $t-1$. A larger value indicates the objects in the corresponding center points are more likely to belong to the same identity. Conversely, a smaller value suggests that the corresponding points are prone to having a different identity or background.

2.3. Target-Aware Loss

The common re-ID losses aim at enlarging intra-identity variance while the detection losses focus on maximizing inter-identity difference and minimizing intra-identity variance. Thus, the re-ID branch is prone to deteriorate the detection quality in joint training. We design the target-aware loss to learn correlation without affecting the detection performance.

We fetch the identity embedding of a target p_i, denoted as $\tilde{E}_i$, from the enhanced embeddings $\tilde{E}$. Then, we project $\tilde{E}_i$ to a class distribution and treat all targets with the same identity in the training sequences as a class. Thus, the probability of $\tilde{E}_i$ belonging to the class k can be calculated and denoted as $\hat{\mathbf{L}}_i(k)$. The loss L_{ta} consists of two parts: one for emphasizing the intra-identity difference, the other for forcing the distance between predicted identities and ground-truth identities, which can be defined as follows:

$$L_{ta} = \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{1}{e^{\lambda_1}} \mathbf{L}_i(k) \log(\hat{\mathbf{L}}_i(k)) + \frac{d(\hat{c}_i, c_i)}{e^{\lambda_2} \gamma_i} + \lambda_1 + \lambda_2, \tag{5}$$

where N and K stand for the numbers of identities and classes, respectively. We set λ_1 and λ_2 as the learnable parameters and $\mathbf{L}_i(k)$ is the ground-truth identity label of p_i. c_i and $\hat{c}_i$ is the ground-truth and predicted positions, respectively. The distance function $d(\cdot)$ is used to calculate the predicted and the ground-truth positions, and γ_i is the diagonal distance between their bounding boxes. We only utilize the identity embeddings located at target centers for training.

The common re-ID losses usually overlook the inter-identity difference, which is one of the main focuses in object detection. Nevertheless, our L_{ta} not only forces intra-identity variance, but also emphasizes inter-identity difference when extracting embeddings. This manner provides discriminative embeddings for supporting the identity-aware attention to capture the correlation between embeddings in two frames with improved detection performance.

Table 1. Ablation study on the MOT17 validation set. "*IAA*" is the identity-aware attention module, "L_{ta}" is the target-aware loss, and "L_{ce}" is the cross-entropy loss. "Baseline+*IAA*+L_{ta}" is represented by "LCTrack". ↑ denotes that a higher value is better, ↓ denotes that a lower value is better. The best results are highlighted in **bold**.

Setup	MOTA↑	IDF1↑	FP↓	FN↓	IDS↓
Baseline	66.1	64.2	**4.5%**	28.4%	1.0%
Baseline+*IAA*	67.1	65.5	5.0%	26.8%	1.0%
Baseline+L_{ta}	67.0	66.1	5.3%	27.1%	**0.9%**
Baseline+*IAA*+L_{ce}	66.0	65.1	5.6%	27.7%	1.0%
LCTrack	**67.6**	**66.2**	4.8%	**26.1%**	**0.9%**

2.4. Tracklet Generation

We regress the center locations, sizes and predict the tracking offsets similar to CenterTrack. Based on the predicted offsets, a greedy matching algorithm is employed to associate detections and compose tracklets. Note that CenterTrack localizes and associates objects from a local perspective and only concatenates the previous frame and the current frame to construct the temporal relation. Thus, without exploiting the correlation between consecutive frames, the position prediction and association are not effective when coping with occlusion.

Different from CenterTrack, our LCTrack makes full use of the correlation between two frames via embedding learning, thus enhancing the appearance features for position prediction. According to the prior-frame features, the identity-aware attention module constructs a full correlation for all identities to enhance the features in the current frame. Meanwhile, the target-aware loss is well compatible with the detection loss and enables the tracker to effectively distinguish different identities. Based on the robust appearance features, we therefore predict the tracking offset to guide data association, which alleviates the model drift under significant occlusion.

3. EXPERIMENT

3.1. Implementation Details and Evaluation Metrics

Implementation Details. Our implementation is based on CenterTrack without Track Re-birth [10]. Specifically, we use DLA-34 [22] as the feature extractor and the resized images as the inputs. LCTrack is trained for 20 epochs, adopting a learning rate of $1.25e^{-4}$ dropped by a factor of 10 at epoch 14. The weight ω is set to 0.1 for feature aggregation empirically. Since the previous frame image and heatmap are empty in the first frame, we learn the correlation in the same images. All experiments are implemented in PyTorch and we test the inference speed with a RTX 2080 Ti GPU.

Evaluation Metrics. We evaluate our tracker on the MOT16 and MOT17 datasets [23], which are challenging and widely used for evaluating the performance of MOT methods. ***Metrics:*** The common evaluation metrics [24]: Multiple Object

Table 2. Comparison among the online state-of-the-art methods on the MOT test set under the "private detector" protocol. "D" denotes the detection time, which is usually more than 100ms. "TBD" indicates the tracking-by-detection paradigm and "JDT" indicates the joint detection and tracking paradigm. The best results in the JDT paradigm are highlighted in **bold**.

MOT16 Test Set									
Methods	Paradigm	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDS↓	Time(ms)↓
SORT [5]	TBD	59.8	53.8	25.4%	22.7%	8,698	63,245	1,423	17+D
POI [6]	TBD	66.1	65.1	34.0%	20.8%	5,061	55,914	805	101+D
TubeTK [12]	JDT	64	59.4	33.5%	19.4%	10,962	53,626	**1,117**	1,000
JDE [18]	JDT	64.4	55.8	35.4%	20.0%	-	-	1,544	45
CTracker [19]	JDT	67.6	57.2	32.9%	23.1%	**8,934**	48,305	1,897	29
LCTrack(Ours)	JDT	**69.9**	**60.4**	**40.2%**	**16.3%**	12,661	**40,599**	1,680	57
MOT17 Test Set									
DAN [21]	TBD	52.4	49.5	21.4%	30.7%	25,423	234,592	8,431	159+D
Tracktor+CTdet [11]	TBD	54.4	56.1	25.7%	29.8%	44,109	210,774	2,574	-
TubeTK [12]	JDT	63.0	58.6	31.2%	19.9%	27,060	177,483	4,137	333
CTracker [19]	JDT	66.6	57.4	32.2%	24.2%	**22,284**	160,491	5,529	29
CenterTrack [10]	JDT	67.3	**59.9**	34.9%	24.8%	23,031	158,676	**2,898**	57
LCTrack(Ours)	JDT	**69.1**	59.8	**39.0%**	**17.7%**	33,915	**135,312**	5,181	57

Tracking Accuracy (MOTA), ID F1 score (IDF1), Mostly Tracked Targets (MT), Mostly Lost Targets (ML), False Positives (FP), False Negatives (FN), and Identity Switches (IDS).

3.2. Ablation Study

For the ablation study, we split the MOT17 training set into two halves, one for training, and the other for validation, as in CenterTrack [10]. The results are reported in Table 1.

Influence of the identity-aware attention module (*IAA*). We integrate the *IAA* module into the baseline tracker without using additional supervision to learn correlation for robust appearance features. As shown in Table 1, with *IAA*, the tracking performance of our method is increased by 1.1% in terms of MOTA and by 1.3% in terms of IDF1, respectively. Compared with the baseline tracker, *IAA* reduces the number of FP, FN, and IDS by capturing the correlation between identities in consecutive frames.

Influence of the target-aware loss (L_{ta}). As reported in Table 1, we directly utilize the embedding network $\phi(\cdot)$ to extract the embeddings for associating across frames. The baseline with L_{ta} reduces IDS and brings an improvement of MOTA and IDF1 significantly. To further analyze the effectiveness of L_{ta}, we use the cross-entropy loss to supervise the embedding learning of *IAA*, as shown in the fourth row of Table 1. The cross-entropy loss deteriorates the tracking performance greatly, verifying our L_{ta} is well compatible with the detection loss to improve the overall performance.

With the identity-aware attention module and the target-aware loss, our LCTrack remarkably improves MOTA by 1.5% and IDF1 by 2.0% compared with the baseline, respectively. Although increasing FP slightly, our LCTrack achieves less FN and IDS, which verifies that LCTrack can predict target positions correctly and reduce identity changes. Experimental results demonstrate that our method improves the comprehensive tracking performance.

3.3. Benchmark Evaluations

We compare our LCTrack with several online state-of-the-art methods on the test set of the MOTChallenge benchmarks[23].

As shown in Table 2, the benchmark results of LCTrack show its favorable performance against those online private methods. In particular, our LCTrack outperforms the second-best tracker by a margin of 2.3% MOTA on MOT16 and 1.8% MOTA on MOT17, respectively, running at 14 FPS. Although IDS of LCTrack is not the best, our LCTrack achieves higher IDF1, which also implies the effectiveness of the identity embedding learning for data association. In addition, LCTrack achieves the highest MT and the least ML, which indicates our method can predict correct tracklets to cover the most ground-truth trajectories. We observe that the TBD methods, such as POI [6], obtain fewer ID switches than the JDT methods on MOT16. This is mainly because the complicated embedding learning used in these TBD methods is designed to associate objects. The experimental results show the benefit of learning correlation from identity embeddings, verifying the discriminative power of LCTrack under crowded scenes.

4. CONCLUSION

In this paper, we propose an online multiple object tracking method LCTrack to learn correlation under the joint detection and tracking paradigm. We propose the identity-aware attention module to enhance appearance features by capturing the correlation between identity embedding features. Then we design the target-aware loss to supervise the object centers for providing reliable identity embeddings and being compatible with object detection. The enhanced representations can boost position prediction and data association. Evaluations on the MOTChallenge benchmarks demonstrate that the proposed LCTrack improves the tracking performance of the baseline tracker remarkably.

5. REFERENCES

[1] K. Chumachenko, J. Raitoharju, A. Iosifidis, and Mo. Gabbouj, "Ensembling object detectors for image and video data analysis," in *ICASSP*, 2021, pp. 1515–1519.

[2] S. Wang, R. Han, W. Feng, and S. Wang, "Multiple human tracking in non-specific coverage with wearable cameras," in *ICASSP*, 2021, p. 2200–2204.

[3] S. Pang and H. Radha, "Multi-object tracking using poisson multi-bernoulli mixture filtering for autonomous vehicles," in *ICASSP*, 2021, p. 7963–7967.

[4] Z. Lu, V. Rathod, R. Votel, and J. Huang, "Retinatrack: online single stage joint detection and tracking," in *CVPR*, 2020, pp. 14668–14678.

[5] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *ICIP*, 2016, pp. 3464–3468.

[6] F. Yu, W. Li, Q. Li, Y. Liu, X. Shi, and et al, "Poi: multiple object tracking with high performance detection and appearance feature," in *ECCV Workshop*, 2016, pp. 36–42.

[7] N. Mahmoudi, S. M. Ahadi, and M. Rahmati, "Multi-target tracking using cnn-based features: cnnmtt," *MTA*, vol. 78, no. 6, pp. 7077–7096, 2019.

[8] K. Fang, Y. Xiang, X. Li, and S. Savarese, "Recurrent autoregressive networks for online multi-object tracking," in *WACV*, 2018, pp. 466–475.

[9] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *ICIP*, 2017, pp. 3645–3649.

[10] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," in *ECCV*, 2020, pp. 474–490.

[11] P. Bergmann, T. Meinhardt, and L. Leal-Taixe, "Tracking without bells and whistles," in *ICCV*, 2019, pp. 941–951.

[12] B. Pang, Y. Li, Y. Zhang, M. Li, and C. Lu, "Tubetk: adopting tubes to track multi-object in a one-step training model," in *CVPR*, 2020, pp. 6308–6318.

[13] F. Saleh, S. Aliakbarian, H. Rezatofighi, M. Salzmann, and S. Gould, "Probabilistic tracklet scoring and inpainting for multiple object tracking," in *CVPR*, 2021, pp. 14329–14339.

[14] S. Guo, J. Wang, X. Wang, and D. Tao, "Online multiple object tracking with cross-task synergy," in *CVPR*, 2021, pp. 8136–8145.

[15] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Detect to track and track to detect," in *ICCV*, 2017, pp. 3038–3046.

[16] J. Yin, W. Wang, Q. Meng, R. Yang, and J. Shen, "A unified object motion and affinity model for online multi-object tracking," in *CVPR*, 2020, pp. 6768–6777.

[17] P. Voigtlaender, M. Krause, A. Osep, J. Luiten, B. B. G. Sekar, and et al, "Mots: multi-object tracking and segmentation," in *CVPR*, 2019, pp. 7942–7951.

[18] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, "Towards real-time multi-object tracking," in *ECCV*, 2020, pp. 107–122.

[19] J. Peng, C. Wang, F. Wan, Y. Wu, Y. Wang, and et al, "Chained-tracker: chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking," in *ECCV*, 2020, pp. 145–161.

[20] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

[21] S. Sun, N. Akhtar, H. Song, A. Mian, and M. Shah, "Deep affinity network for multiple object tracking," *TPAMI*, vol. 43, no. 1, pp. 104–119, 2019.

[22] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *CVPR*, 2018, pp. 2403–2412.

[23] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "Mot16: A benchmark for multi-object tracking," *arXiv preprint arXiv:1603.00831*, 2016.

[24] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *JIVP*, pp. 1–10, 2008.