A Simple Framework for Multi-Object Tracking with Multiple Cues in DanceTrack 2022

Guangxin Han[1], Mingzhan Yang[1], Yanxin Liu[1], Shiyu Zhu[2], Yuzhuo Han[2], Xu Jia[1], and Huchuan Lu[1]

[1] Dalian University of Technology
[2] Honor Device Co. Ltd

Abstract. In this work we point out that joint utilization of multiple informative cues is the key to achieve reliable tracking on high frame rate multi-object tracking datasets like DanceTrack. A simple online tracker with limited temporal information from the perspective of joint utilization of multiple cues is proposed accordingly. Equipped with a simple detection model and a re-identification model, our method exceeds State-of-the-art methods (SOTA) by a large margin. In addition, we design two variants that learning detection and reid tasks jointly, which both achieve an excellent balance between inference speed and tracking accuracy. The proposed model, with small model size and reasonable performance, could be beneficial to practical application of multi-target tracking on edge devices.

Keywords: Multi-object tracking, Online model

1 Introduction

The current state-of-the-art MOT algorithms for person tracking are mostly based on single-stage detectors. Specifically, ByteTrack [12] additionally uses low-confidence detections to improve the recall rate of trajectories. OC-SORT [1] handles non-linear motion and frequent occlusions by introducing velocity direction consistency and last observation-based retrieval during data association, as well as generating virtual trajectories in lost period to smooth the parameters of Kalman filter.

In this paper, we emphasize that joint utilization of information from multiple cues could bring lots of benefits to tracking performance. Specifically, there are three types of cues that could be leveraged in DanceTrack [7], that is, location cues, motion cues and appearance cues. Following this idea, we design a multi-stage association strategy based on limited temporal information to ensure the simplicity of the overall framework. The proposed method is validated to be effective on challenge dataset and outperforms SOTA methods by a large margin with simple network architecture design and multi-stage association strategy. For practical applications, we also design two variants which show excellent balance between the accuracy and speed.

2 Proposed Method

In this paper, we follow the paradigm of separated detection and embedding, using independent detection and ReID models. As for association strategy, we jointly use information from multiple cues to perform data association through multi-stage matching. This method achieves the 3rd place in the 1st Multiple People Tracking in Group Dance Challenge - DanceTrack. In addition, we follow the joint detection and embedding paradigm and design two variants. In the first variant, we integrate the detection and the ReID models into a unified framework (named joint version). Based on joint version, we then choose a smaller detection model as well as optimize the network head (named light version). The two variants both show a good balance between accuracy and speed.

2.1 Final Solution

We use YOLOX [3] as out base detector, then replace the original label assignment strategy SimOTA with Task Alignment Learning (TAL) in TOOD [2]. A combination of classification scores and IoU scores is used to compute the assignment of positive and negative samples. Samples with the highest scores are selected as positive ones, while the remaining samples are set as negative ones. For each ground truth, the cost of the corresponding samples is as Eq.1.

$$t = cls^{\alpha} \times iou^{\beta} \tag{1}$$

On DanceTrack [7], the ReID models containing local features can better suppress the impact of background and other distractors under occlusion, providing more reliable embedded features. Thus, we use MGN [8] with ResNet-50 [4] as the ReID model. To ensure the convergence and generalization, we mix CUHKSYSU [9] and DanceTrack [7] for training.

About the association strategy using appearance cues, the key is how to use and update the embedding features. In the matching, we fuse the appearance cost matrix with the IoU-based motion cost matrix with a certain weight. EMA bank in FairMOT [13] is used to maintain the embedding feature of trajectories, and the features are only updated when the trajectories are matched to the detections with high confidence.

The overall association strategy is divided into 3 stages. The first stage aims to match with high score detections. Three types of cues are used here: motion cues with Kalman filter, velocity and direction cues, and appearance cues. Then the trajectories' features are updated. The second stage aims to match with low score detections. There are two differences between the second and the first stage due to the confidence of detections. We do not use velocity cues in the second stage matching and do not update embedding feature banks of corresponding trajectories. In the third stage, we recover trajectories based on the last observation i.e. location cues. After the matching procedure, we generate virtual trajectories in the lost period of recovered trajectories to smooth the Kalman filter parameters. Finally, new trajectories are initialized and old trajectories are deleted.

2.2 Joint Version

From the perspective of speed-accuracy trade-off, we design a framework for joint learning of detection and ReID tasks, which substantially improves the inference speed with a slight decrease in tracking performance. Unlike CSTrack [5] and RelationTrack [11] with complicated modules, we design a simple but efficient way to handle the competition between detection and ReID task. Specifically, we decouple the network structure of the two tasks in the middle part of the network. That is, the ReID module is branched in the middle of the backbone, takes the feature map of 8x stride from the backbone and output the ReID feature map with 32x stride in one forward operation. The embedding feature of each objects are obtained with RoIAlign w.r.t detected boxes. The overall architecture are shown in Fig.1.

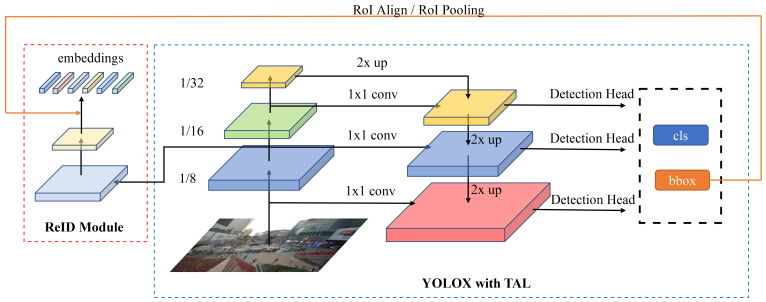

Fig. 1. Proposed method with joint detection and embedding framework

As for the loss function for ReID task, we find that OIM [10] loss eases the conflict between detection and ReID tasks and achieve better convergence under large number of Identities. OIM [10] Loss can utilize objects with and without ID information by caching the embedding feature of unlabelled objects in the circular queue, as shown in Eq.2.

$$p_i = \frac{\exp(v_i^T x/\tau)}{\sum_{j=i}^{L}\exp(v_j^T x/\tau) + \sum_{k=1}^{Q}\exp(u_k^T x/\tau)} \tag{2}$$

$$q_i = \frac{\exp(u_i^T x/\tau)}{\sum_{j=i}^{L}\exp(v_j^T x/\tau) + \sum_{k=1}^{Q}\exp(u_k^T x/\tau)} \tag{3}$$

$$L_{\mathrm{OIM}=E_x[\log p_t]} \tag{4}$$

2.3 Light Version

We refer to NanoDet-PLUS [6] and use the depth-wise separable convolution with a kernel size of 5 to replace the normal convolution in the head of YOLOX

[3]. This modification maintains the detection performance while reduces the required flops and parameters. Besides, we use YOLOX-S as the base detector.Even thought the flops and parameters are 13.3% and 9.9% of the SOTA model, we still achieve better tracking performance on DanceTrack [7].

3 Results

Table 1. Comparison of the state-of-the-art methods on DanceTrack [7] test sets. "n" means the number of detections

Method	HOTA	MOTA	IDF1	Param	GFlops
ByteTrack	47.3	89.5	52.5	99.00M	791.73
OC-SORT	55.7	92.0	54.6	99.00M	791.73
Ours-Light	56.6	90.3	59.1	**13.14M**	**78.41**
Ours-Joint	63.2	**92.4**	63.5	141.18M	987.69
Ours	**64.6**	92.3	**64.5**	99.00M + 68.67M × n	791.73 + 27.94 × n

Our best solution significantly outperforms the current SOTA algorithm OC-SORT [1] in terms of tracking performance.The huge improvement in IDF1 shows the superiority of joint utilization of multiple cues in data association, even if only short-term information is used. In our Joint Version, we significantly improve the inference speed and resource usage at the cost of slightly reduced tracking performance. Our Light Version still achieves higher tracking accuracy with only 13.3% parameters and 9.9% of GFlops of OC-SORT [1], showing a superior balance of tracking accuracy and inference speed on edge devices.

4 Limitations

The main limitation of the proposed method lies in the difficulty in handling long-time complex motions. The reliability of all cues will rapidly decrease when the trajectory is interrupted. Better performance could be expected when long-term temporal information is exploited. In addition, a more robust ReID model for occlusion should further improve tracking performance.

References

1. Cao, J., Weng, X., Khirodkar, R., Pang, J., Kitani, K.: Observation-centric sort: Rethinking sort for robust multi-object tracking. arXiv preprint arXiv:2203.14360 (2022)
2. Feng, C., Zhong, Y., Gao, Y., Scott, M.R., Huang, W.: Tood: Task-aligned one-stage object detection. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3490–3499. IEEE Computer Society (2021)
3. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
5. Liang, C., Zhang, Z., Zhou, X., Li, B., Zhu, S., Hu, W.: Rethinking the competition between detection and reid in multiobject tracking. IEEE Transactions on Image Processing **31**, 3182–3196 (2022)
6. RangiLyu: Nanodet-plus: Super fast and high accuracy lightweight anchor-free object detection model. https://github.com/RangiLyu/nanodet (2021)
7. Sun, P., Cao, J., Jiang, Y., Yuan, Z., Bai, S., Kitani, K., Luo, P.: Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20993–21002 (2022)
8. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 274–282 (2018)
9. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: End-to-end deep learning for person search. arXiv preprint arXiv:1604.01850 **2**(2), 4 (2016)
10. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3415–3424 (2017)
11. Yu, E., Li, Z., Han, S., Wang, H.: Relationtrack: Relation-aware multiple object tracking with decoupled representation. IEEE Transactions on Multimedia (2022)
12. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. arXiv preprint arXiv:2110.06864 (2021)
13. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: On the fairness of detection and re-identification in multiple object tracking. International Journal of Computer Vision **129**, 3069–3087 (2021)