The Second-place Solution for ECCV 2022 Multiple People Tracking in Group Dance Challenge

Fan Yang, Shigeyuki Odashima, Shoichi Masui, Shan Jiang
Fujitsu Research, Japan
contact: fan.yang@fujitsu.com

Abstract

This is our 2^{nd}-place solution for the ECCV 2022 Multiple People Tracking in Group Dance Challenge. Our method mainly includes two steps: online short-term tracking using our Cascaded Buffer-IoU (C-BIoU) Tracker, and, offline long-term tracking using appearance feature and hierarchical clustering. Our C-BIoU tracker adds buffers to expand the matching space of detections and tracks, which mitigates the effect of irregular motions in two aspects: one is to directly match identical but non-overlapping detections and tracks in adjacent frames, and the other is to compensate for the motion estimation bias in the matching space. In addition, to reduce the risk of overexpansion of the matching space, cascaded matching is employed: first matching alive tracks and detections with a small buffer, and then matching unmatched tracks and detections with a large buffer. After using our C-BIoU for online tracking, we applied the offline refinement introduced by ReMOTS [12].

1. Introduction

Although MOT studies have been greatly developed [2, 11, 13, 6], a new challenge has recently attracted attention: unlike conventional MOT tasks that contain objects with distinct appearances and regular motions, MOT tasks that cover animals, group dancers, and sports players, may have indistinguishable appearances and irregular motions, which could cause existing MOT methods to fail. In particular, several MOT methods [2, 11, 13, 3] that perform well on MOT17 [8], may experience a significant performance drop on the DanceTrack [10].

We presume that tracking failures are caused by two reasons: (i) The detections and tracks of identical objects do not overlap between adjacent frames (*e.g.*, due to the fast movement) and thus the tracking fails; (ii) After track initialization, unmatched tracks (*e.g.*, occluded objects) continue to update their geometric features for multiple frames, however, if their motion estimations are inaccurate (*e.g.*, due to a sudden acceleration or turning), they miss the matching opportunity when corresponding detections are available in subsequent frames. When the appearance of objects can be distinguished, appearance features could be employed to alleviate issues (i) and (ii), by matching cross-frame detections based on their appearance similarities. Nonetheless, when irregular motions are accompanied by indistinguishable appearances, most existing MOT solutions may not be able to perform a dependable tracking, so a new solution is desirable.

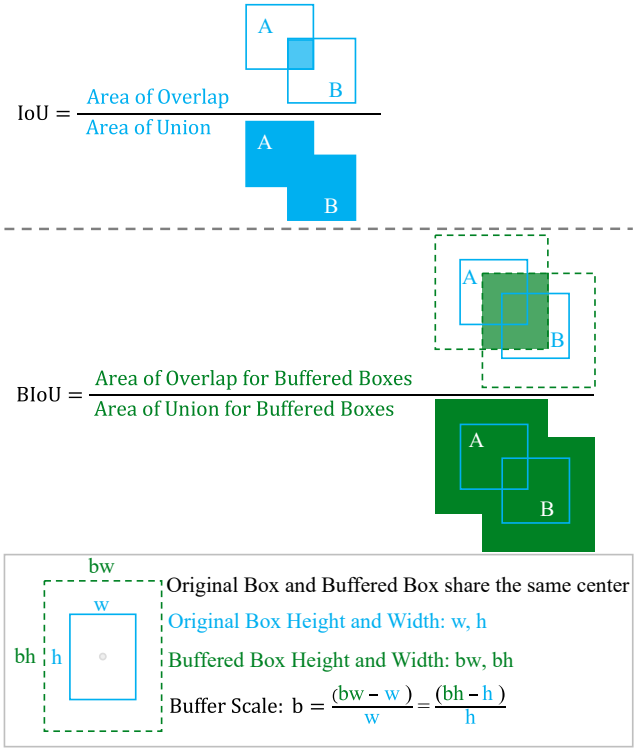

Figure 1: **Illustration of how Buffered IoU (BIoU) is calculated**. Our BIoU adds a buffer that is proportional to the original bounding box. It does not change the location center, scale ratio, and shape of the original bounding boxes but expands the original matching space.

In this study, we propose a Cascaded-Buffered Intersection over Union (C-BIoU) tracker to track multiple objects

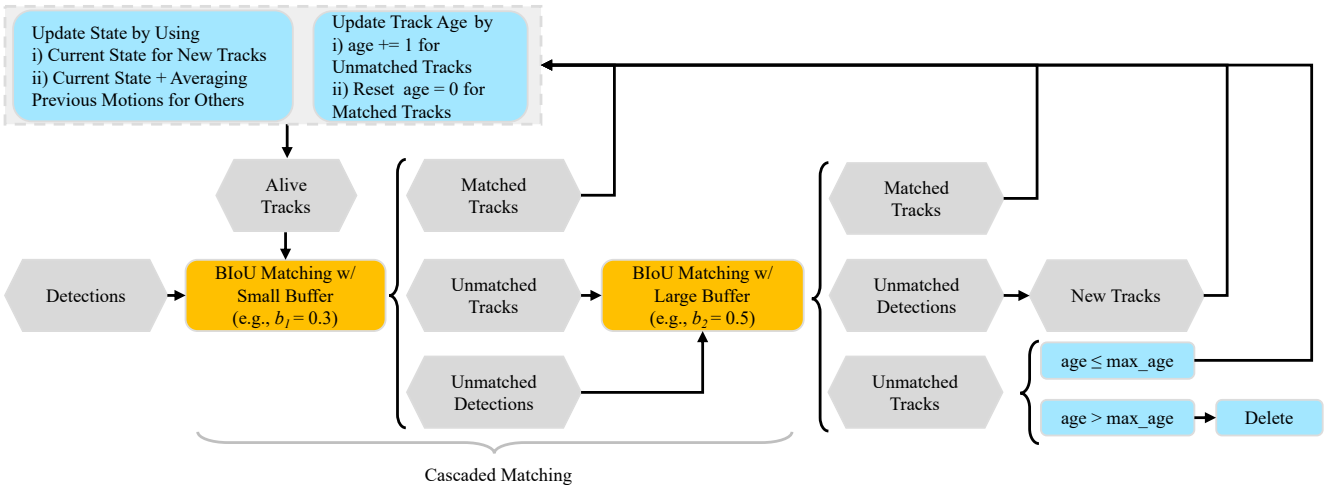

Figure 2: **Framework of Our C-BIoU Tracker.**

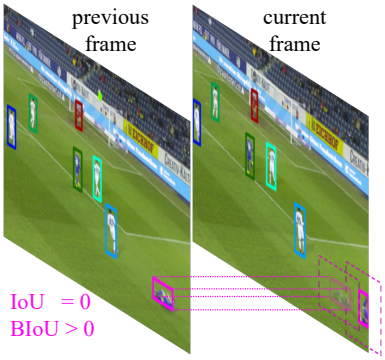

Figure 3: **An illustration of BIoU forms better cross-frame geometric consistency than IoU**. The bounding box of an identical object shares the same color. The magenta object has no overlapping detections between adjacent frames. Whether this is caused by the fast movement or incorrect motion estimation, our BIoU expands the matching space to reduce the miss matching.

that have irregular motions and indistinguishable appearances.

2. Method

2.1. Detection of C-BIoU Tracker

Our C-BIoU tracker (Fig. 2) follows the tracking-by-detection paradigm—the object detector and MOT framework are separately designed. Given a video, we applied YOLOX [4] trained by OC-SORT [3] to generate bounding boxes at each frame. We tried input sizes of $[800, 1440]$ and $[960, 1728]$. Our C-BIoU tracker then takes detected bounding boxes as inputs to produce tracking results. Such a pipeline provides great flexibility to apply our C-BIoU tracker on arbitrary detections.

2.1.1 Buffered IoU

The Buffered IoU (BIoU) is our main contribution in this work. As shown in Fig. 1, the BIoU simply adds buffers that are proportional to the original detections and tracks for calculating the IoU. Our BIoU retains the same location centers, scale ratios, and shapes of the original detections and tracks, but it expands the matching space to measure the geometric consistency. Let $\boldsymbol{o} = (x, y, w, h)$ denote an original detection and (x, y, w, h) be the top-left coordinate, width, and height of the detection, respectively. Suppose that the buffer scale is b, we have the buffered detection as $\boldsymbol{o}_b = (x - bw, y - bh, w + bw, h + bh)$. To approach our cascaded matching, we apply grid research [1] to find the best combination of two buffer scales b_1 and b_2 on the training set, and then apply them to the validation set and test set. Since we have $b_1 < b_2$, when we search for the combination of b_1 and b_2 within a certain range, the number of combinations is limited. Considering that the speed of our C-BIoU is fast, the grid search takes an acceptable time.

2.1.2 Simple Motion Estimation

Unlike most MOT methods [2, 11, 13] that apply the Kalman filter [5] for state estimation, we simply average motions of recent frames to quickly respond to unpredictable motion changes. At frame t, suppose that a track has matched detections for more than n frames, after Δ unmatched frames, its track state $\boldsymbol{s}$ can be represented as $\boldsymbol{s}^{t+\Delta} = o^t + \frac{\Delta}{n}\sum_{i=t-n+1}^{t}(\boldsymbol{o}^i - \boldsymbol{o}^{i-1})$. The matched detections between frame $t - n$ to t are used to calculate motions and the average motion is applied to update the track state. We set $2 \leq n \leq 5$ by default in our experiments. The IoU score of buffered $\boldsymbol{s}_b^{t+\Delta}$ and $\boldsymbol{o}_b^{i+\Delta}$ is used for data association at the frame $t + \Delta$. Due to the simplicity of our approach, the overall tracking speed is increased for our C-BIoU tracker.

2.1.3 Track Management

In an MOT framework, the function of track management is to decide how and when to initialize, update and termi-

Table 1: **Result details of our method in the ECCV 2022 Group Dance Challenge.** Using non maximum suppression (NMS), we merged detections generated by two input scales. The data rendered in **Bold** indicate the best results.

Tracker	Detector	HOTA↑	DetA↑	AssA↑	MOTA↑	IDF1↑
C-BIoU (Online)	YOLOX-X [800, 1440]	60.6	**81.3**	45.4	91.6	61.6
C-BIoU (Online)	Merged YOLOX-X [800, 1440] and [960, 1728]	61.7	**81.3**	47.0	92.2	65.3
C-BIoU (Online) + ReMOTS (Offline Refinement)	Merged YOLOX-X [800, 1440] and [960, 1728]	**66.6**	**81.3**	**54.7**	**92.3**	**71.4**

nate a track. We design our track management based on the mainstream solution introduced by SORT [2]. It initializes tracks from unmatched detections, applies the alive tracks to match new detections, and terminates a track when it has not been matched for a given amount of frames (*i.e.*, max_age). Two BIoUs, which respectively equip small and large buffers, are grouped into a cascaded matching. First, we match alive tracks and detections with the BIoU that has a small buffer (*i.e.*, b_1). Then, we continue to match unmatched tracks and detections with the BIoU that has a large buffer (*i.e.*, b_2). For the motion estimation, we simply average the speeds of recent frames to quickly respond to unpredictable motion changes.

2.2. Offline Refinement

Due to the occlusion, some long-term tracklets could be broken down by only referring to geometry features. To recover long-term tracklets, we employed an offline tracking with appearance features. We utilized Strong ReID [7] to obtain appearance feature. To initialize the re-id model, we utilized the DanceTrack training set. To include the tracking labels in our re-id training, we took a self-supervised learning method introduced by ReMOTS [12]. Referring to tracking ID, in each video, we construct triplets and only apply triplet loss to them.

After the training, we generated appearance features for each short-term tracklets obtained from the previous step. Within a video, we formed a distance matrix $\mathcal{D}$ between short-term tracklets as

$$\mathcal{D}_{k1,k2} = \begin{cases} inf, & if\ \Pi_{k1} \cap \Pi_{k2} \neq \varnothing \\ \frac{1}{N_{k1} N_{k2}} \sum_{i \in \Pi_{k1}} \sum_{j \in \Pi_{k2}} \left(1 - \frac{f_i^{k1} f_j^{k2}}{\|f_i^{k1}\| \|f_j^{k2}\|}\right), & otherwise \end{cases}$$

where for tracklets T_{k1} and T_{k2}, $\mathcal{D}_{k1,k2}$ is their distance; Π_{k1} and Π_{k2} are their temporal ranges; f_i^{k1} and f_j^{k2} are their appearance features at frame i and j, and N_{k1} and N_{k2} are the number of observations within the tracklets, respectively.

Based on $\mathcal{D}$, we applied hierarchical clustering to cluster short-term tracklets to long-term ones.

3. Results

3.1. Challenge Results

We obtained the 2^{nd} place in the ECCV 2022 Group Dance Challenge (Table 2), and the result details are summerized in Table 1. Since we directly applied the YOLOX trained by OC-SORT [3] to perform the detection without any fine tuning, our DetA score (*i.e.*, detection performance) is worse than other teams, which further affects our HOTA, AssA and IDF1 scores. We assume that our tracking performance could be improved if strong detections are provided.

Table 2: **The top-3 teams in the ECCV 2022 Group Dance Challenge.** The data rendered in **Bold** and Underlined indicate the best and second best results respectively.

Ranking	Team Name	HOTA↑	DetA↑	AssA↑	MOTA↑	IDF1↑
1^{st} place	mfv	**73.4**	83.7	**64.4**	92.1	**76.0**
2^{nd} place	C-BIoU (Ours)	66.6	81.3	54.7	92.3	71.4
3^{rd} place	ymzis69	64.6	82.5	50.7	92.3	64.5

Table 3: **Ablation experiments on the DanceTrack validation set [10]**. Where "C.M." and "Mo." represent the cascaded matching and motion estimation, respectively. We remove the cascaded matching and motion estimation in Fig. 2 to construct a unified framework for the IoU, GIoU [9], DIoU [14], and BIoU.

Tracker	C.M.	Mo.	HOTA↑	DetA↑	AssA↑	MOTA↑	IDF1↑
DanceTrack Validation Set [10]. Using Oracle Detections.							
IoU Tracker	✗	✗	76.6	97.5	60.2	99.2	73.6
GIoU Tracker	✗	✗	77.1	**97.6**	60.9	99.2	74.0
DIoU Tracker	✗	✗	75.1	97.0	58.2	99.2	72.9
BIoU Tracker	✗	✗	80.0	97.5	65.7	**99.3**	78.2
C-BIoU Tracker	✓	✗	80.2	97.5	65.9	**99.3**	79.3
C-BIoU Tracker	✓	✓	**81.7**	**97.6**	**68.4**	**99.3**	**80.5**

3.2. Ablation Experiments

3.2.1 Effect of Each Module in the C-BIoU Tracker

Table 3 shows the influence of each module in our C-BIoU tracker. Using the same framework, the tracker equipped with BIoU achieves a higher HOTA score than other trackers equipped with IoU, GIoU [9], or DIoU [14]. Although the GIoU and DIoU can incorporate non-overlapping boxes for geometric consistency measurement, they may not generate comparable results as our BIoU does. Integrating cascaded matching and BIoU can slightly improve the performance as compared to using BIoU alone, with a HOTA gain of 0.2. According to the results, motion estimation plays an important role in our C-BIoU tracker. Since our BIoU can compensate the matching space for incorrect motion estimation, using a simple motion estimation (*i.e.*, averaging previous motions) yields better HOTA scores than that without using motion estimation.

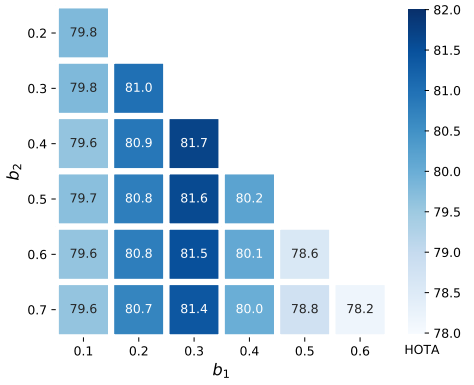

Figure 4: **Results of applying various buffer-scale combinations on the DanceTrack validation set [10].** For buffer scales b_1 and b_2, since we have $b_1 < b_2$, we only check the lower triangle of the combination matrix.

3.2.2 Effect of Buffer Scales in the C-BIoU Tracker

In our C-BIoU tracker, the buffer scales b_1 and b_2 are critical hyperparameters. Here, we perform ablation studies to investigate how buffer scales affect the tracking performance. On the DanceTrack validation set [10], we form the combination of b_1 and b_2 ranging from 0.1 to 0.7 and evaluate their tracking performance. Since we have $b_1 < b_2$, we only need to check 21 combinations. As shown in Fig. 4, the combination of $[0.3, 0.4]$ gives the maximum HOTA score. In real practice, we perform a similar approach to select the best combination on the training dataset and apply them to the test dataset.

4. Conclusion

We present a novel Cascaded-Buffered IoU (C-BIoU) tracker to track multiple objects that have indistinguishable appearances and irregular motions. The good performance of our C-BIoU tracker can be attributed to its buffered matching space, which mitigates the effect of irregular motions in two aspects: one is to directly match identical but non-overlapping detections and tracks in adjacent frames, and the other is to compensate for the motion estimation bias in the matching space.

References

[1] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

[2] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *IEEE international conference on image processing*, pages 3464–3468, 2016.

[3] Jinkun Cao, Xinshuo Weng, Rawal Khirodkar, Jiangmiao Pang, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. *arXiv preprint arXiv:2203.14360*, 2022.

[4] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[5] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960.

[6] Shuai Li, Yu Kong, and Hamid Rezatofighi. Learning of global objective for network flow in multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8855–8865, June 2022.

[7] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[8] 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.

[9] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.

[10] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 0–10, June 2022.

[11] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *IEEE international conference on image processing*, pages 3645–3649, 2017.

[12] Fan Yang, Xin Chang, Chenyu Dang, Ziqiang Zheng, Sakriani Sakti, Satoshi Nakamura, and Yang Wu. Remots: Self-supervised refining multi-object tracking and segmentation. *arXiv preprint arXiv:2007.03200*, 2020.

[13] 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.

[14] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12993–13000, 2020.