Embedding with Bounding Box Contracting for Multi-object Tracking

Like Zhang, Wenjing Kang(✉), and Guangdong Zhang

School of Information Science and Engineering, Harbin Institute of Technology, Weihai 264209, Shandong , China
kwjqq@hit.edu.cn

Abstract. The development of 5G/6G network can achieve high data transmission speed, which promotes the wide application of remote video monitoring. Multi-object tracking (MOT) aims at detecting and tracking all the objects of interesting categories in videos. Appearance and motion information of each object are significant clues utilized for finding associations between detections and tracks. Many approaches model each object appearance through bounding box region, which is vulnerable to background noise and motion deformation. In this paper, we alleviate this problem, via embedding with object bounding box contracting. We also integrate an online tracking by detection model, comprehensive use of appearance and motion information for data association. Object bounding box contracting is introduced to relieve the impact of interference and obtain high-quality re-ID embeddings. Experimental results based on the MOT17 benchmark show that the integrated tracker with bounding box contracting for embedding achieves 80.6 MOTA, 79.4 IDF1 and 64.4 HOTA.

Keywords: Multi-Object Tracking · Object Detecting · Embedding Methods

1 Introduction

Multi-object tracking (MOT), which aims to detect and estimate trajectories of multiple target objects in a video, is widely used in autonomous driving, video surveillance and more [1].

Advanced online multi-object tracking algorithms follow two main paradigms: tracking by detection and joint detection and tracking [2]. The tracker belongs to tracking by detection (TBD) paradigm, divides the detection and tracking into two independent tasks. It usually adopts off-the-shelf high-performance object detection algorithm and focuses on the research of data association algorithms. Algorithms that follow the joint detection and tracking paradigm (JDT) perform detection and tracking simultaneously. It reuses the feature extracted by the backbone, which greatly reduces the amount of computation and speeds up the inference speed of the algorithm.

Although the joint detection and tracking algorithm has a faster inference speed, its accuracy often lags behind the detection-based tracking algorithm. Especially in

A. Li et al. (Eds.): 6GN 2022, LNICST 504, pp. 65–76, 2023.
https://doi.org/10.1007/978-3-031-36011-4_6

complex scenes such as occlusion, the feature embeddings extracted by the tracking algorithm of the joint detection and tracking paradigm are insufficient to distinguish the target, resulting in target identity switching [3]. The detection-based tracking algorithm is slow, but each part of it can be independently optimized in blocks, and often has high tracking accuracy. Therefore, the tracking by detection paradigm is still the current mainstream research direction.

Algorithms that follow the tracking by detection paradigm can be divided into SORT [4] like algorithms and Deep SORT [5] like algorithms. SORT like algorithms only use motion information for data association, while Deep SORT like algorithms use target motion information and appearance information for data association. Compared with SORT like algorithms, Deep SORT like algorithms have the ability to re-identify lost targets due to the introduction of appearance information.

The Deep SORT like algorithms focus on improving the utilization of object appearance information and motion information, and have achieved high tracking performances after years of development. For the feature information modeling of the object, most tracking by detection algorithms first scale the area where the target bounding box is located to the same size, and then extract feature through an independent convolutional neural network. Although Deep SORT like algorithms have made great progress, past algorithms focused on using re-ID convolutional neural networks with stronger feature extraction capabilities [6, 7] or increasing the number of appearance feature [8, 9]. However, these algorithms neglect the interference of bounding box area itself, such as the bounding box predicted by the detection algorithm contains a large amount of background caused by motion deformation.

In this paper, we alleviate the interference of background noise through contract the bounding boxes in the re-ID embedding stage, so that the re-ID model pays more attention to the target. Through this simple operation, most of the background noise can be eliminated, and the influence of motion deformation on re-ID embedding can be avoided at the same time.

We integrate a new tracker, follows the Deep SORT like paradigm. We adopt a recent high-performance detector YOLOX [10] to obtain the detection bounding boxes and associate them using appearance and motion information. To evaluate our algorithm, experiments are performed on the MOT17 [11] test sets. Our method achieves 80.6 MOTA, 79.4 IDF1 and 64.4 HOTA.

The main contributions of our work are summarized as follows:

- We propose a simple and effective method to alleviate the influence of background noise and motion deformation during re-ID embedding, and verify its effectiveness on the Deep SORT algorithm.
- We integrate a new tracker, follows the Deep SORT like paradigm. Evaluate the tracker's performance on the MOT17 test sets.

2 Related Work

Algorithms following the tracking by detection paradigm can be divided into two separate tasks: object detection and data association. Object detection estimates the bounding boxes and data association obtains the identities.

2.1 Object Detection

Object detection is one of the hot research directions in the field of computer vision and it is the pre-task of multi-object tracking. The MOT17 dataset [11] provides detection results obtained by object detector such as DPM [12], Faster R-CNN [13] and SDP [14]. Many tracking methods based on these published detection results and focus on research on data association algorithms.

With the development of convolutional neural networks, object detection algorithms have made great progress. Existing object detection algorithms can be divided into two-stage detection algorithms and one-stage detection algorithms. The two-stage object detection algorithms need to generate candidate boxes through the region proposal network at first stage. And perform object classification and bounding box regression in the subsequent network. For example, R-CNN [15], Fast R-CNN [16] and Faster R-CNN [13]. The one-stage object detection algorithms don't need to generate candidate boxes, and directly regard the object location as a regression task. The representative algorithm has the YOLO series [17, 18], SSD [19], etc.

Compared with the one-stage object detection algorithm, two-stage object detection algorithm often has higher detection accuracy and slower speed in the past. With the continuous development of object detection algorithms, the detection accuracy of the one-stage object detection algorithm has been greatly improved. YOLOX [10] is a one-stage object detection algorithm, which is also an anchor-free detector. Benefited from its decoupled object detection head design, the detection accuracy is greatly improved. This detector is widely used in the state-of-the-art tracking algorithms.

2.2 Multi-object Tracking

Tracking by detection paradigm still dominate the field of multi-object tracking. Algorithms that follow the detection-based tracking paradigm decouple detection and data association tasks, and can flexibly use high-performance object detectors to improve tracking performance and make full use of the achievements in the field of object detection.

With the development of object detection algorithms, the performance of detectors has been continuously improved. More and more tracking by detection algorithms employ high-performance object detectors to estimate object bounding boxes for better tracking results. ByteTrack [20] adopts YOLOX [10] as the object detection algorithm, achieves high tracking performance only uses motion information for finding association. Alpha-Refine [21] corrects the object bounding box by predicting the object mask to further improve the prediction accuracy of the bounding box.

Appearance and motion feature are the main basis for data association. The motion feature of most multi-object tracking algorithms are modeled by the Kalman filter algorithm. The object appearance feature of detection-based tracking algorithms usually use independent feature extraction networks to extract feature. BoT-SORT [7] and StrongSORT [6] use stronger feature extractor BoT [22], to obtain feature vectors.

Some algorithms only use motion information for data association, such as: SORT [4] and ByteTrack [20]. Although these algorithms also achieve high-quality tracking results, they rely on the performance of object detectors too much. Besides, since the

appearance feature is not used, the tracking algorithms have no ability to re-identify the lost target leading to the ID switching problem.

3 Proposed Method

We alleviate the interference of background during the re-ID embedding stage with bounding boxes contracting. Further integrate a multi-object tracking algorithm, follow the tracking by detection paradigm. Adopt YOLOX [10] as the object detector and comprehensively using the motion and appearance information for data association.

3.1 Embedding with Bounding Box Contracting

Existing multi-object tracking algorithms neglect the interference caused by the background noise contained in the bounding boxes region output by the detector, as shown in the Fig. 1.

Fig. 1. Bounding boxes region contains background noise.

To alleviate this interference, we contract the bounding boxes horizontally during re-ID embedding. This simple but effective operation allows us to filter out a portion of the background noise in the bounding boxes. The coordinates of the object bounding boxes after contracting are calculated by the following:

$$x_i' = x_i + w_i \times perc \tag{1}$$

$$y_i' = y_i \tag{2}$$

$$w_i' = w_i(1 - 2 \times perc) \tag{3}$$

$$h_i' = h_i \tag{4}$$

where (x_i, y_i, w_i, h_i) is the i-th coordinate predicted by the object detector, (x, y) is the upper left corner coordinate, (w, h) is the width and height of the object bounding box. And (x_i', y_i', w_i', h_i') is the coordinate of the object bounding boxes after contracting. *Perc* is the contracting ratio, set to 5%.

By contracting $w_i \times perc$ width on the left and right sides of the object bounding box, effectively alleviate a large amount of background noise caused by motion deformation and other reasons in the output object bounding box of the detector. Therefore, the proportion of the human body in the bounding box is increased, the identity embedding network pays more attention to the main body of the target. The background interference and posture changes of the human body caused by motion, such as hands, feet, etc., are reduced. As shown in Fig. 2.

Fig. 2. Contracted bounding boxes for re-ID embedding.

3.2 Integrated Tracker

We utilize the advantages of existing multi-object tracking algorithms and the method proposed earlier in this paper to integrate a new multi-object tracker. Comprehensive use of object appearance and motion feature for data association, so that the algorithm has the ability to re-identify lost targets. Camera motion compensation and trajectory interpolation is employed to further improve tracking performance.

Kalman Filter. Pedestrian motion modeling using Kalman filter algorithm, propagate the track position from the previous frame to the current frame. We adopt the Kalman filter implementation form in the BoT-SORT [7] algorithm, and the variables are as follows:

$$\boldsymbol{x}_k = \left[x_c(k), y_c(k), w(k), h(k), \dot{x}_c(k), \dot{y}_c(k), \dot{w}(k), \dot{h}(k)\right]^{\mathrm{T}} \tag{5}$$

$$\boldsymbol{z}_k = \left[z_{x_c}(k), z_{y_c}(k), z_w(k), z_h(k)\right]^{\mathrm{T}} \tag{6}$$

$$\begin{aligned}\boldsymbol{Q}_k = diag((\sigma_p\hat{w}_{k-1|k-1})^2, (\sigma_p\hat{h}_{k-1|k-1})^2,\\ (\sigma_p\hat{w}_{k-1|k-1})^2, (\sigma_p\hat{h}_{k-1|k-1})^2,\\ (\sigma_y\hat{w}_{k-1|k-1})^2, (\sigma_y\hat{h}_{k-1|k-1})^2,\\ (\sigma_y\hat{w}_{k-1|k-1})^2, (\sigma_y\hat{h}_{k-1|k-1})^2)\end{aligned} \tag{7}$$

$$\begin{aligned}\boldsymbol{R}_k = diag((\sigma_m z_w(k))^2, (\sigma_m z_h(k))^2,\\ (\sigma_m z_w(k))^2, (\sigma_m z_h(k))^2)\end{aligned} \tag{8}$$

where $\boldsymbol{x}_k$ is the state vector of the k-th frame, z_k is the observation vector, $\boldsymbol{Q}_k$ is the process noise covariance and $\boldsymbol{R}_k$ is measurement noise covariance. Choose the noise factors to be $\sigma_p = 0.05$, $\sigma_y = 0.00625$ and $\sigma_m = 0.05$.

Re-ID Embedding. We utilize the method proposed in Sect. 3.1 to mitigate the influence of disturbances during re-ID embedding. Extract object appearance feature vector using OSNet [23] network. The objects area is scaled to the same size and sent to the convolutional neural network to extract feature, and finally obtain a 512-dimensional feature vector for each object.

Camera Motion Compensation. Motion feature matching heavily relies on the IoU distance between the tracklets and the detected objects bounding boxes. The camera movement will bring about a large target displacement, which will offset the trajectory position predicted by the Kalman filter. Although the camera position is kept unchanged in some scenes, factors such as vibration still cause the video picture to shake. We adopt the camera motion compensation model in BoT-SORT, to reduce ID switching and mismatches caused by camera motion.

The affine matrix $\boldsymbol{A}_{k-1}^k \in \mathbb{R}^{2\times 3}$ is solved by RANSAC, and the camera motion correction step can be performed by the following equations:

$$\boldsymbol{A}_{k-1}^k = \left[\boldsymbol{M}_{2\times 2}|\boldsymbol{T}_{2\times 1}\right] = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \tag{9}$$

$$\tilde{\boldsymbol{M}}_{k-1}^k = \begin{bmatrix} \boldsymbol{M} & 0 & 0 & 0 \\ 0 & \boldsymbol{M} & 0 & 0 \\ 0 & 0 & \boldsymbol{M} & 0 \\ 0 & 0 & 0 & \boldsymbol{M} \end{bmatrix} \tilde{\boldsymbol{T}}_{k-1}^k = \begin{bmatrix} a_{13} \\ a_{23} \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{10}$$

$$\hat{\boldsymbol{x}'}_{k|k-1} = \tilde{\boldsymbol{M}}_{k-1}^k \hat{\boldsymbol{x}}_{k|k-1} + \tilde{\boldsymbol{T}}_{k-1}^k \tag{11}$$

$$\boldsymbol{P'}_{k|k-1} = \tilde{\boldsymbol{M}}_{k-1}^k \boldsymbol{P}_{k|k-1} \tilde{\boldsymbol{M}}_{k-1}^k \tag{12}$$

where $\hat{\boldsymbol{x}}_{k|k-1}, \hat{\boldsymbol{x}'}_{k|k-1}$ is the Kalman Filter's predicted state vector of time k before and after compensation of the camera motion respectively. $\boldsymbol{P}_{k|k-1}, \boldsymbol{P'}_{k|k-1}$ is the Kalman Filter's predicated covariance matrix before and after correction respectively.

Data Association. The data association process is shown in Algorithm 1. We found that there are serious occlusions in low score detection boxes, and most of the extracted feature

belong to another object, as shown in Fig. 3. Therefore, we only perform appearance feature matching on high score detections. It can not only avoid false matching, but also reduce the number of feature extraction to speed up network interference.

Detections that unmatched with tracklets and sub high score detections will be matched with motion feature. For low score detections motion feature matching also performed. In fact, the motion feature matching process is adopted from ByteTrack [20].

Fig. 3. Low score detections with serious occlusion.

Interpolation. Interpolation is widely used in trajectories' post processing, to further improve the tracking performance. Linear interpolation is high popular due to its simplicity, however its accuracy is limited. We adopt the GSI model in StrongSORT [6] for trajectories' post processing, which is a lightweight interpolation algorithm that employs Gaussian process regression to model nonlinear motion.

The GSI model is formulated by follows:

$$\boldsymbol{p}_k = f^{(i)}(k) + \varepsilon \tag{13}$$

where $k \in \boldsymbol{F}$ is the frame, $\boldsymbol{p}_k \in \boldsymbol{P}$ is the position coordinate variate at frame k and $\varepsilon \sim N(0, \sigma^2)$ is Gaussian noise. $S^{(i)} = \{k^{(i)}, p_k^{(i)}\}_{t=1}^{L}$ is the trajectories after linear interpolation with length L. The nonlinear motion processing is to fit the function $f^{(i)}$. StrongSORT assumes it obeys a Gaussian process.

Algorithm 1: Pseudo-code of integrated tracker.

Input: A video sequence `V`; object detector `Det`; Kalman Filter `KF`; feature extraction score threshold τ_f; detection score threshold τ_{high}, τ_{low}; tracking score threshold ε; camera motion compensation `CMC`

Output: Tracks T of the video

Initialization: $T \leftarrow \varnothing$
for *frame* f_k *in* `V` **do**
  $D_k \leftarrow \texttt{Det}(f_k)$
  $D_f \leftarrow \varnothing$
  $D_{high} \leftarrow \varnothing$
  $D_{low} \leftarrow \varnothing$
  for d *in* D_k **do**
    if $d.score > \tau_f$ **then**
      $D_f \leftarrow D_f \cup \{d\}$
    end
    else if $d.score < \tau_f$ *and* $d.score > \tau_{high}$ **then**
      $D_{high} \leftarrow D_{high} \cup \{d\}$
    end
    else if $d.score < \tau_{high}$ *and* $d.score > \tau_{low}$ **then**
      $D_{low} \leftarrow D_{low} \cup \{d\}$
    end
  end
  for t *in* T **do**
    $t \leftarrow \texttt{KF}(t)$
    $t \leftarrow \texttt{CMC}(t)$
  end
  Association T and D_f using cosine distance
  $D_{f-re} \leftarrow$ remaining object boxes from D_f
  $T_{f-re} \leftarrow$ remaining tracks from T
  Association T_{f-re} and $D_{high} \cup D_{f-re}$ using IoU distance
  $D_{remain} \leftarrow$ remaining object boxes from $D_{high} \cup D_{f-re}$
  $T_{remain} \leftarrow$ remaining tracks from T_{f-re}
  Association T_{remain} and D_{low} using IoU distance
  $T_{re-remain} \leftarrow$ remaining tracks from T_{remain}
  $T \leftarrow T \setminus T_{re-remain}$ # delete unmatched tracks
  for d *in* D_{remain} **do**
    if $d.score > \varepsilon$ **then**
      $T \leftarrow T \cup \{d\}$
    end
  end
end

4 Experiments

4.1 Setting

We evaluate the integrated tracker on MOT17 [11] datasets under the "private detection" protocol. Both datasets contain training sets and test sets. For ablation studies, we use the last half of each video in the training set of MOT17, same to the ByteTrack [20]. We also use the pretrain weight of YOLOX-X provided by ByteTrack, and the weight of OSNet [23] provided by Torchreid [24]. All the experiments are implemented on a NVIDIA GTX 1080 GPU, using Pytorch.

The feature extraction threshold τ_f is set to 0.85 and matching threshold is 0.13, other parameters is same to ByteTrack.

4.2 Ablation Study

Our ablation study aims to verify the effectiveness of the bounding box contracting during re-ID embedding. We use the MOT17 validation set, i.e. the last half of each video in the training set. First use the Deep SORT [5] for test, and also adopt YOLOX-X as the object detector. Table 1 shows the comparison results, budget is the number of embedding features for each object and percentage is the ratio of bounding box contracting. It can be seen that MOTA is significantly improved by contracting the bounding box during re-ID embedding. Experiments on our integrated method also prove this, as shown in Table 2.

Table 1. Comparison of difference contracting percentage using Deep SORT.

Tracker	Budget	Percentage(%)	MOTA↑	IDF1↑	IDs↓
Deep SORT	5	0	75.3	76.5	249
Deep SORT	10	0	75.3	76.7	250
Deep SORT	5	5	**75.4**	**76.8**	**247**

Table 2. Comparison of difference contracting percentage using our integrated tracker.

Tracker	Budget	Percentage(%)	MOTA↑	IDF1↑	IDs↓
Ours	1	0	76.8	80.6	121
Ours	1	5	**76.9**	80.6	**119**

4.3 MOT Challenge Results

We compare the integrated tracker with state-of-the-art trackers on the test set of MOT17. The evaluate results are shown in Table 3. Our tracker achieves state-of-the-art performance under the "private detector" protocol. We get 80.6 MOTA, 79.4 IDF1 and 64.4

HOTA. Moreover, the number of ID switches and fragments rank first in MOT17 "private detector" protocol due to precise appearance feature matching. We show some visualization results of difficult cases in Fig. 4.

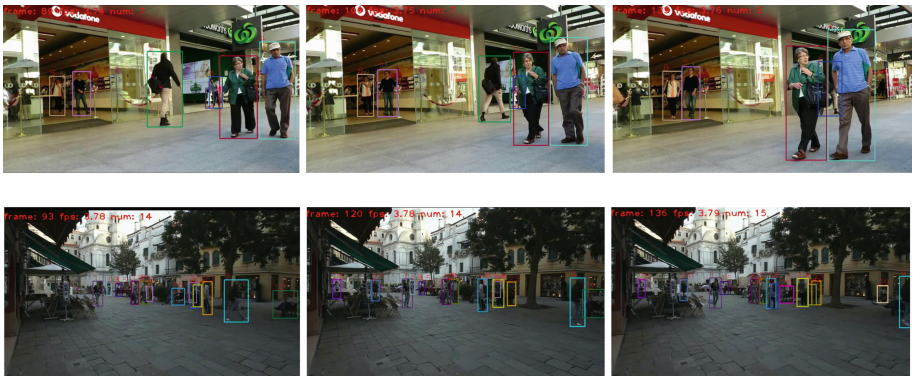

Fig. 4. Visualization results of difficult cases.

Table 3. Comparison of the state-of-the-art methods under the "private detector" protocol.

Tracker	MOTA↑	HOTA↑	IDF1↑	FP↓	FN↓	IDs↓	Frag↓
ReMOT [25]	77.0	59.7	72.0	33204	93612	2853	5304
OCSORT [26]	78.0	63.2	77.5	**15129**	107055	1950	2040
MAA [27]	79.4	62.0	75.9	37320	**77661**	1452	2202
StrongSORT [6]	79.6	64.4	79.5	27876	86205	1194	1866
ByteTrack [20]	80.3	63.1	77.3	25491	83721	2196	2277
BoT-SORT [7]	80.5	**65.0**	**80.2**	22521	86037	1212	1803
Ours	**80.6**	64.4	79.4	22179	86379	**1062**	**1572**

5 Conclusion

In this paper, we present a simple, effective and generic method to alleviate the background noise and motion deformation interference during re-ID embedding. This method can easily be integrated into other tracking by detection trackers. We also integrate a new tracker, which follows the tracking by detection paradigm. The experimental results show that the tracker achieves 80.6 MOTA, 64.4 HOTA, 79.4 IDF1, and the number of ID switches and fragments rank first in MOT17 "private detector" protocol respectively.

Acknowledgement. This work was supported partially by National Natural Science Foundation of China (Grant Nos. 61971156, 61801144), Shandong Provincial Natural Science Foundation (Grant Nos. ZR2019QF003, ZR2019MF035, ZR2020MF141), the Fundamental Research

Funds for the Central Universities, China (Grant No. HIT.NSRIF.2019081) and the Scientific Research Innovation Foundation in Harbin Institute of Technology at Weihai (Grant No. 2019KYCXJJYB06).

References

1. Zheng, L., Tang, M., Chen, Y., Zhu, G., Wang, J., Lu, H.: Improving multiple object tracking with single object tracking. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2453–2462 (2021)
2. Wu, J., Cao, J., Song, L., Wang, Y., Yang, M., Yuan, J.: Track to detect and segment: an online multi-object tracker. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12347–12356 (2021)
3. Saleh, F., Aliakbarian, S., Rezatofighi, H., Salzmann, M., Gould, S.: Probabilistic tracklet scoring and inpainting for multiple object tracking. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14324–14334 (2021)
4. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: 2016 IEEE International Conference on Image Processing, pp. 3464–3468 (2016)
5. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE International Conference on Image Processing, pp. 3645–3649 (2017)
6. Du, Y., Song, Y., Yang, B., Zhao, Y.: StrongSORT: make DeepSORT great again. arXiv preprint arXiv: 2202.13514 (2022)
7. Aharon, N., Orfaig, R., Bobrovsky, B.: BoT-SORT: robust associations multi-pedestrian tracking. arXiv preprint arXiv: 2206.14651 (2022)
8. Kim, C., Fuxin, L., Alotaibi, M., Rehg, J.M.: Discriminative appearance modeling with multi-track pooling for real-time multi-object tracking. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9548–9557 (2021)
9. Pang, J., Qiu, L., Li, X., et al.: Quasi-Dense similarity learning for multiple object tracking. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 164–173 (2021)
10. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: YOLOX: exceeding YOLO series in 2021. arXiv preprint arXiv: 2107.08430 (2021)
11. Milan, A., Leal-Taixe, L., Reid, L., Roth, S., Schindler, K.: MOT16: a benchmark for multi-object tracking. arXiv preprint arXiv: 1603.00831 (2016)
12. Felzenszwalb, P., Mcallester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
13. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2017)
14. Wang, F., Choi, W., Lin, Y.: Exploit all the layers: fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2129–2137 (2016)
15. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
16. Girshick, R.: Fast R-CNN. In: 2015 IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
17. Redmon, J., Farhadi, A.: YOLOv3: an incremental improvement. arXiv preprint arXiv: 1804.02767 (2018)

18. Bochkovskiy, A., Wang, C., Liao, H.M.: YOLOv4: optimal speed and accuracy of object detection. arXiv preprint arXiv: 2004.10934 (2020)
19. Wei, L., Dragomir, A., Dumitru, E., et al.: SSD: single shot multibox detector. In: 14th European Conference on Computer Vision, pp. 21–37 (2016)
20. Zhang, Y., Sun, P., Jiang, Y., et al.: ByteTrack: multi-object tracking by associating every detection box. arXiv preprint arXiv: 2110.06864 (2021)
21. Yan, B., Zhang, X., Wang, D., Lu, H., Yang, X.: Alpha-Refine: boosting tracking performance by precise bounding box estimation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5285–5294 (2021)
22. Luo, H., Jiang, W., Gu, Y., et al.: A strong baseline and batch normalization neck for deep person re-identification. IEEE Trans. Multimed. **22**(10), 2597–2609 (2019)
23. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: 2019 IEEE/CVF International Conference on Computer Vision, pp. 3701–3711 (2019)
24. Zhou, K., Xiang, T.: Torchreid: a library for deep learning person re-identification in pytorch. arXiv preprint arXiv: 1910.10093 (2019)
25. Yang, F., Chang, X., Sakti, S., Wu, Y., Nakamura, S.: ReMOT: a model-agnostic refinement for multiple object tracking. Image Vision Comput. **106**, 104091 (2021)
26. 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)
27. Stadler, D., Beyerer, J.: Modelling ambiguous assignments for multi-person tracking in crowds. In: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, pp. 133–142 (2022)