RETHINKING THE COMPETITION BETWEEN DETECTION AND REID IN MULTI-OBJECT TRACKING

Chao Liang[1], *Zhipeng Zhang*[2], *Yi Lu*[3], *Xue Zhou*[1,4,*], *Bing Li*[2], *Xiyong Ye*[3] *and Jianxiao Zou*[1,4]

[1]School of Automation Engineering, University of Electronic Science and Technology of China(UESTC)
[2]NLPR, Institute of Automation, Chinese Academy of Sciences
[3]Artificial Intelligence Research Institute, Zhejiang Lab
[4]Shenzhen Institute of Advanced Study, UESTC
[*]Corresponding author: zhouxue@uestc.edu.cn

ABSTRACT

Due to balanced accuracy and speed, joint learning detection and ReID-based one-shot models have drawn great attention in multi-object tracking(MOT). However, the differences between the above two tasks in the one-shot tracking paradigm are unconsciously overlooked, leading to inferior performance than the two-stage methods. In this paper, we dissect the reasoning process of the aforementioned two tasks. Our analysis reveals that the competition of them inevitably hurts the learning of task-dependent representations, which further impedes the tracking performance. To remedy this issue, we propose a novel cross-correlation network that can effectively impel the separate branches to learn task-dependent representations. Furthermore, we introduce a scale-aware attention network that learns discriminative embeddings to improve the ReID capability. We integrate the delicately designed networks into a one-shot online MOT system, dubbed CSTrack. Without bells and whistles, our model achieves new state-of-the-art performances on MOT16 and MOT17. Our code is released at https://github.com/JudasDie/SOTS.

***Index Terms*—** Multi-object tracking, Cross-correlation, Residual attention, One-shot model, ReID-based tracker

1. INTRODUCTION

Multi-object tracking (MOT), aiming to estimate the locations and scales of multiple targets in a video sequence, is one of the most fundamental yet challenging tasks in computer vision [1]. The task has numerous applications in practical scenarios, *e.g.*, intelligent driving, human-computer interaction, and pedestrian behavior analysis, etc.

Convolutional Neural Network (CNN) based object detection and re-identification (ReID) have brought unprecedented advances to multi-object tracking in the past few years. The related ReID-based trackers can be summarized into two streams, *i.e.,* two-stage and one-shot structures. The two-stage models follow the tracking-by-detection paradigm [2–6], which divides MOT into two separate tasks, *i.e.,* detection and association. It first obtains bounding boxes of objects in each frame through an off-the-shelf detector, and then matches the extracted re-identification information in each bounding box across frames. Although effective, the two-stage paradigm suffers from computation cost, since the ReID model needs to perform forward inference for each individual bounding box. Alternatively, the one-shot paradigm integrates detection and identification (ID) embedding into a unified system, which is capable of running in real-time speeds [7–9]. Nevertheless, their performance is suppressed by the two-stage methods.

In this paper, we dissect the reasoning process of one-shot tracking first. Our analysis reveals that the performance degradation primarily derives from two aspects: **1) Excessive competition between detection and ReID tasks:** In one-shot methods, the object class confidence, target scale, and ID information are simultaneously derived from a shared embedding. Although efficient, the inherent differences between different tasks are overlooked. The competition between these tasks may lead to ambiguous learning, which means that the pursuit of high performance in one task will lead to performance degradation or stagnation in the other. Specifically, detection requires the embeddings of different objects in the same category (*e.g., pedestrain*) have similar semantics. On the contrary, ReID tends to learn distinguished semantics for two objects, and even they fall into the same category, contradicting the ultimate purpose of detection. **2) Large scale variance in MOT:** The object images in ReID task are normally arranged to a fixed size during searching from gallery, *e.g.* 256×128 in [10]. While in the MOT task [11, 12], features for ReID network are required to have scale-aware capability, since the objects' size may change drastically across frames. However, recent one-shot models only consider the feature maps derived from a single resolution, which lacks the strong ability to represent objects with different scales.

To remedy the competition issue, we propose a novel cross-correlation network to improve the collaborative learning between detection and ReID tasks in the one-shot tracking

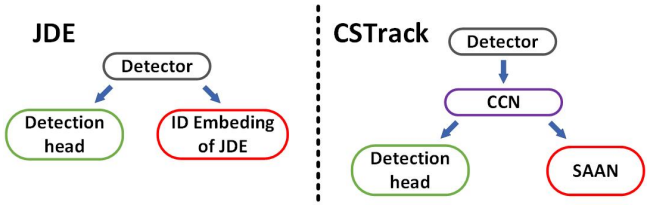

Fig. 1. Framework comparison between JDE and our method.

framework. We first decouple detection and ReID to two separate branches, from which the task-dependent representation can be learned. Then given features corresponded to detection and ReID, the self-relation and cross-relation weight maps are obtained in a self-attention manner. The former impels hidden nodes to learn task-dependent features, and the latter improves the collaborative learning of the two tasks. Moreover, to improve the resilience to scale change, we introduce a scale-aware attention network. Specifically, we apply the spatial and channel attention mechanism to features, which can improve the influence of object-related embeddings in different resolutions. Later, we aggregate the features from different resolutions to a single output, which helps learn scale-aware representations.

The main contributions of our work are three-folds:

- We propose a novel cross-correlation network to enable the model to learn task-dependent representations. It not only effectively mitigate the competence, but also improve the collaborative learning capability between detection and ReID tasks in one-shot MOT methods.
- We introduce a scale-aware attention network that applies spatial and channel attention mechanisms to features from different resolutions. The fused representations improve the resilience of the network to objects with different scales.
- The extensive experiments demonstrate that our method effectively improves the performance of the one-shot MOT method, especially for data associated ability of re-identification features.

2. METHOD

The proposed framework consists of two main components, *i.e.*, cross-correlation network detailed in SubSec. 2.2 and scale-aware attention network described in SubSec. 2.3. Before shedding light on the two essential parts, we first give a brief overview of the whole framework in SubSec. 2.1.

2.1. Overview

We build our framework based on JDE which adopts a hierarchical two-branch architecture to enforce detection and ID embedding tasks in a one-shot model [8], as illustrated in Fig. 1. In JDE, the detection head and ID embedding head take the same features from the detector as input. Due to being ignoring the essential differences between these two tasks, excessive competition causes JDE performance degeneration. To

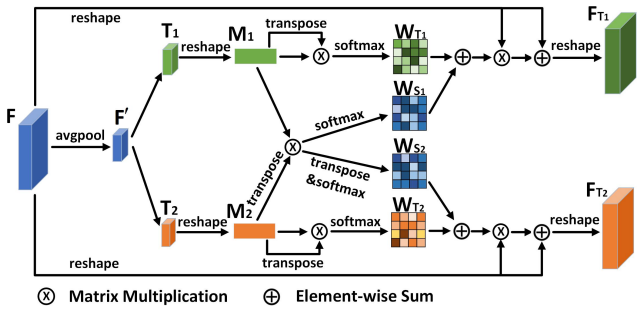

Fig. 2. Diagram of cross-correlation network.

improve the learning of task-dependent representations, we propose a novel cross-correlation network (CCN) to decouple the input. The idea of designing CCN is inspired by recent self-attention [13] and multi-task decoding [14] mechanisms, as shown in Fig. 2, which can enhance representations for each task with only small overheads. For ReID branch, JDE obtains ID embedding through applying a 1x1 convolution layer on each feature map resolution. Although it's simple and useful but only considers the feature maps derived from a single resolution, which lacks the strong ability to represent objects with different scales. In our method we design a scale-aware attention network (SAAN, detailed in Fig. 3) to fuse features from different resolutions. Meanwhile, both spatial and channel-wise attention models [15] are adopted to suppress noisy background for a more concentrated ID embedding.

2.2. Cross-correlation Network

In this section, we propose a cross-correlation network to learn commonalities and specificities of features for detection and ReID tasks. For specificities learning, self-relation reflecting correlations between different feature channels are learned to enhance feature representation for each task. For commonalities learning, shared information between the two tasks can be learned by a elaborately designed cross-relation mechanism.

The structure of our cross-correlation network is presented in Fig. 2. Formally, denote features from the detector as $\mathbf{F} \in R^{C\times H\times W}$. First, we pass it through an *avg-pooling* layer to obtain the statistical information $\mathbf{F'} \in R^{C\times H'\times W'}$. Then, $\mathbf{T_1}$ and $\mathbf{T_2}$ for detection and ReID are respectively generated by passing $\mathbf{F'}$ through different convolution layers. We then reshape them to $\{\mathbf{M_1}, \mathbf{M_2}\} \in R^{C\times N'}$, where $N' = H' \times W'$. Finally, we perform matrix multiplication on $\mathbf{M_1}/\mathbf{M_2}$ (/ indicates or) and its corresponding transpose. A row *softmax* layer is followed to calculate the self-relation weight maps $\{\mathbf{W}_{\mathbf{T_1}}, \mathbf{W}_{\mathbf{T_2}}\} \in R^{\mathrm{C}\times\mathrm{C}}$ for each task, and the calculation is as follows:

$$w_{T_k}^{ij} = \frac{\exp\left(\mathbf{M_k^i} \cdot \mathbf{M_k^j}\right)}{\sum_{j=1}^{C} \exp\left(\mathbf{M_k^i} \cdot \mathbf{M_k^j}\right)}, \mathbf{k} \in \{1, 2\} \qquad (1)$$

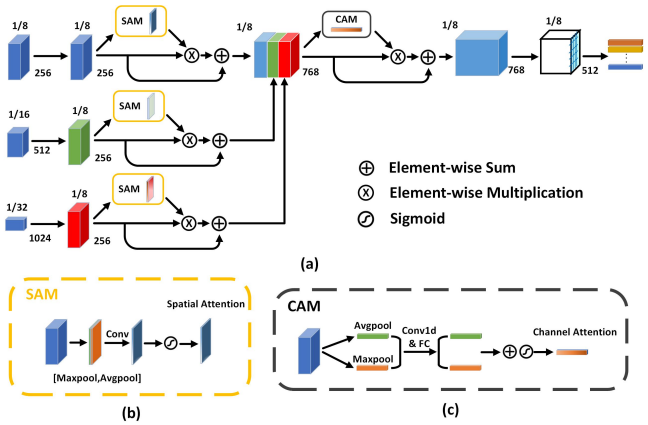

Fig. 3. The details of scale-aware attention network. (a) The overall structure of the network. (b) Diagram of spatial attention module(SAM). (c) Diagram of channel attention module(CAM).

where $w_{T_k}^{ij}$ denotes the relation of the i^{th} and j^{th} channel in $\mathbf{T_k}$. Again, we perform matrix multiplication between $\mathbf{M_1}$ and the transpose of $\mathbf{M_2}$ to learn commonalities between different tasks, and then a row *softmax* layer is followed to generate cross-relation weight maps $\{\mathbf{W}_{S_1}, \mathbf{W}_{S_2}\} \in R^{C\times C}$,

$$w_S^{ij} = \frac{\exp\left(\mathbf{M_{1/2}^i} \cdot \mathbf{M_{2/1}^j}\right)}{\sum_{j=1}^{C} \exp\left(\mathbf{M_{1/2}^i} \cdot \mathbf{M_{2/1}^j}\right)} \tag{2}$$

where w_S^{ij} stands for the effect of the i^{th} channel of task 1/2 on the j^{th} channel of task 2/1. The self-relation and cross-relation weights are finally fused by trainable parameters λ, obtaining $\{\mathbf{W_1}, \mathbf{W_2}\} \in R^{C\times C}$

$$\mathbf{W_{1/2}} = \lambda \times \mathbf{W_{T_1/T_2}} + (1-\lambda) \times \mathbf{W_{S_1/S_2}} \tag{3}$$

We rearrange the original feature map $\mathbf{F}$ to the shape of $R^{C\times N}$, where $N = H \times W$. Then we perform matrix multiplication between the reshaped feature and the learned weight maps to obtain an enhanced representation for each task. The enhanced representation is fused with original $\mathbf{F}$ by residual attention to prevent information loss.

2.3. Scale-aware Attention Network

We build a scale-aware attention network (SAAN, shown in Fig.3) to aggregate features from different resolutions to ensure the ID embedding robustness to different object sizes. The features from the scale of 1/16 and 1/32 (compared to the size of input image) are upsampled to 1/8 firstly. Then 3×3 convolution layers are followed to encode the reshaped feature maps. To enhance the target-related features and suppress background noise, we introduce spatial attention to process the features, as illustrated in Fig. 3 (b). Then, we concatenate the feature maps of different scales and pass it through the channel attention module. The channel attention module is comprised of the *avg-pooling* and *max-pooling* layers, which learn different statistical information of the input features. The outputs of pooling layers are first processed by the shared network consisting of a *1D convolution* layer and a *fully-connected* layer, and then concatenated by element-wise addition. The learned 1D channel attention map is applied on the features by element-wise multiplication. Finally, we use a 3×3 convolution layer to map features to 512 channels, as $\mathbf{E} \in R^{512\times W\times H}$. The re-identification feature $\mathbf{E_{xy}} \in R^{512\times 1\times 1}$ of an object anchor at location (x, y) can be extracted for the subsequent ReID task. The definition of ID loss and training method follow JDE [8].

Table 1. Ablation Studies of CSTrack.

NUM	Method	CCN	SAAN	MOTA↑	IDF1↑	IDs↓	FPS↑
①	JDE(yolo-v3)			68.4	67.0	264	21.7
②	JDE(yolo-v5)			69.8	73.1	116	31.5
③	JDE(yolo-v5)	✓		71.1	74.4	94	30.0
④	JDE(yolo-v5)		✓	68.9	76.8	83	25.1
⑤	JDE(yolo-v5)	✓	✓	70.9	77.1	92	23.6

3. EXPERIMENTS

In this section, we detail the experimental settings in SubSec. 3.1. This is followed by the ablation studies in SubSec. 3.2. We compare our method with state-of-the-art approaches in SubSec. 3.3. Finally, we analyze the data association upper-bound of our framework in SubSec. 3.4.

3.1. Experimental Settings

We design our framework following JDE [8] structure. The network parameters pre-trained on the COCO dataset [16] are used to initialize our model. In our experiments, six publicly available datasets, *i.e.,* ETH [17], CityPerson [18], CalTech [19], MOT17 [12], CUDK-SYSU [20], and PRW [21], are involved to train our network like JDE. We chop off the videos in ETH [17] that are overlapped with testing benchmark, *i.e.,* MOT16 [12]. In our plus version, we introduce CrowdHuman [22] for training, which can further improve the performance. We train our network with the SGD optimizer for 30 epochs on a single RTX 2080 Ti GPU. The batch size is set to 10. The initial learning rate is 5×10^{-4}, and we decay the learning rate to 5×10^{-5} at the 20th epoch. We evaluate our network on MOT16 and MOT17. Following the MOT Challenge [12], the CLEAR metric [23] and IDF1 [24] are employed to evaluate tracking accuracy. FPS is used to measure frame rate of the overall system.

3.2. Ablation Studies

In this section, we study the effectiveness of each component in our tracking framework. All experiments are trained on the MOT17 training set and evaluated on the MOT15 benchmark. It is worth noting that the overlapped videos between MOT17 and MOT15 are removed for fair comparisons. We present the experimental results in Tab. 1.

We first replace the detector in JDE [8] to Yolo-v5 [1] from

[1] https://github.com/ultralytics/yolov5

Table 2. Comparison with the state-of-the-art online MOT systems under the "private detector" protocol on the MOT16 and MOT17 benchmark. ***Bold italic*** font indicates one-shot method. $*$ indicates other joint detection and tracking methods which adopt non-ReID methods for data association. CSTrack++ indicates our model trained with the additional CrowdHuman datasets.

Dataset	Model	MOTA↑	IDF1↑	MT↑	ML↓	IDs↓	FPS↑
MOT16	DeepSORT-2 [4]	61.4	62.2	32.8	18.2	781	<6.7
	RAR16wVGG [5]	63.0	63.8	39.9	22.1	**482**	<1.5
	∗TubeTK [26]	64.0	59.4	33.5	19.4	1117	1.0
	JDE [8]	64.4	55.8	35.4	20.0	1544	**18.8**
	HOGM [6]	64.8	**73.5**	**40.6**	22.0	794	<8.0
	CNNMTT [27]	65.2	62.2	32.4	21.3	946	<5.2
	POI [3]	66.1	65.1	34.0	21.3	805	<5.2
	∗CTrackerV1 [28]	67.6	57.2	32.9	23.1	1897	6.8
	CSTrack(ours)	69.4	69.3	35.0	22.3	958	16.9
	CSTrack++(ours)	**70.7**	71.8	38.2	**17.8**	1071	15.8
MOT17	∗TubeTK [26]	63.0	58.6	31.2	19.9	4137	3.0
	∗CTrackerV1 [28]	66.6	57.4	32.2	24.2	5529	6.8
	∗CenterTrack [29]	67.8	64.7	34.6	24.6	3039	**22.0**
	CSTrack(ours)	67.3	67.9	34.2	24.1	**2994**	16.9
	CSTrack++(ours)	**70.6**	**71.6**	**37.5**	**18.7**	3465	15.8

Yolo-v3 [25], which ensures better performance and faster inference speed (① *vs* ②). The replacement provides a strong baseline for our following design. When equipped with the proposed cross-correlation network (CCN), it achieves 1.3 points gains on MOTA, and the IDs decrease from 116 to 94 (② *vs* ③). This demonstrates the effectiveness of our CCN component. The introduced SAAN module aiming to improve the discriminative ability of embedding ID features have brought 2.7 points gains on IDF1 (③ *vs* ⑤).

3.3. Comparisons with state-of-the-art

We compare our multi-object tracker CSTrack with other start-of-the-art MOT tracking methods on both MOT16 and MOT17. Comparison methods can be divided into three categories. The first are two-stage methods including DeepSORT-2 [4], RAR16wVGG [5], TAP [6], CNNMTT [27] and POI [3]. The second is the one-shot method including JDE [8]. The third are other joint detection and tracking methods which adopt non-ReID methods for data association, including CTrackerV1 [28], TubeTK [26] and CenterTrack [29]. As shown in Tab. 2, we achieve a new state-of-the-art MOTA score for private detections on both MOT16 and MOT17 benchmarks, *i.e.,* 70.7 for MOT16 and 70.6 for MOT17. It is worth noting that our method significantly improves data association ability, *i.e.,* IDF1 +12.4%∼16.0% on MOT16 and IDF1 +6.9%∼14.2% on MOT17 compared to one-shot methods and other joint detection and tracking methods. Moreover, the data association ability of our one-shot model is comparable to two-stage methods while running more faster.

3.4. Upper-bound Analysis

In this section, we study the upper-bound of the data association ability in our model. The tracking performance is greatly affected by the detectors, which is a common wisdom in MOT [23]. To eliminate the influence of the detector, we validate on the MOT16 training set by replacing the detection results with ground-truth bounding boxes. As shown in Tab. 3, compared to JDE [8], our method yields substantial improvements in data association, *i.e.,* 9 points improvement on IDF1. Moreover, the IDF1 score of our method surpasses the widely used two-stage method DeepSORT-2 [4], which further verifies the effectiveness of our framework. Furthermore, we visualize the association comparisons of ID embedding features between JDE and our method, as shown in Fig. 4. It shows that our method is able to obtain a more discriminative ID feature embedding making the same person more similar while keeping different people more distinguishable. This guarantees the robustness of our tracker under crowded scenarios.

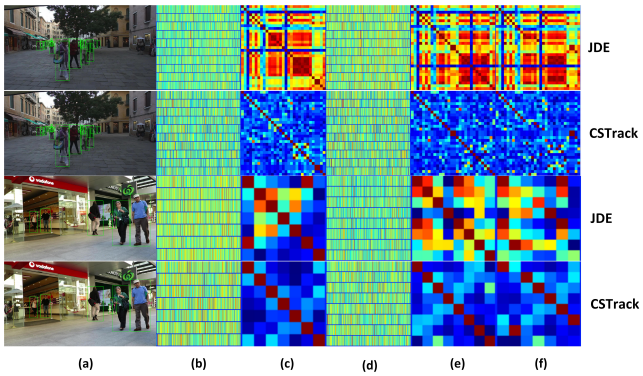

Fig. 4. Visualization about the discriminative ability of ReID features: (a) detection results. (b) ReID features of the current frame. (c) correlation metric matrix between ReID features of the current frame. (d) ReID feature of the template. (e) correlation metric matrix between ReID features of the template. (f) correlation metric matrix between ReID features of the current frame and the template. Red color indicates a higher correlation between two features. Best viewed in color and zoom in.

Table 3. To show the potential of CSTrack. IDP and IDR describe the precision and recall of matching target trajectories.

Model	MOTA↑	IDF1↑	IDP↑	IDR↑	IDs↓
JDE [8]	97.6	87.6	88.3	86.9	871
DeepSORT-2 [4]	**98.9**	95.6	95.9	95.3	**93**
CSTrack(ours)	**98.9**	**96.6**	**97.1**	**96.1**	162

4. CONCLUSION

In this paper, we proposed a one-shot online model CSTrack for the MOT task. A novel cross-correlation network (CCN) and scale-aware attention network (SAAN) are introduced to mitigate the competition and improve the collaboration of detection and ReID subtasks in MOT system. The effectiveness and efficiency of our framework are demonstrated through extensive experiments. Compared with the methods on public benchmarks, our model achieves a new state-of-the-art performance.

5. REFERENCES

[1] W. Luo, J. Xing, et al., "Multi-task learning for dense prediction tasks: A survey," *arXiv preprint arXiv:1409.7618*, 2017.

[2] L. Ott F. Ramos A. Bewley, Z. Ge and B. Upcroft, "Simple online and realtime tracking," in *IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 3464–3468.

[3] F. Yu, W. Li, et al., "Multiple object tracking with high performance detection and appearance feature," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 36–42.

[4] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 3645–3649.

[5] K. Fang, Y. Xiang, X. Li, and S. Savarese, "Recurrent autoregressive networks for online multi-object tracking," in *IEEE Winter Conference on Applications of Computer Vision*, 2018, pp. 466–475.

[6] Z. Zhou, J. Xing, M. Zhang, and W. Hu, "Online multi-target tracking with tensor-based high-order graph matching," in *24th International Conference on Pattern Recognition (ICPR)*, 2018, pp. 1809–1814.

[7] T. Xiao, S. Li, et al., "Joint detection and identification feature learning for person search," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3376–3385.

[8] Z. Wang, L. Zheng, Y. Liu, and S. Wang, "Towards real-time multi object tracking," in *European Conference on Computer Vision (ECCV)*, 2020.

[9] Z. Lu, V. Rathod, R. Votel, and J. Huang, "Retinatrack: Online single stage joint detection and tracking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 14656–14666.

[10] L. Zheng, L. Shen, et al., "Scalable person re-identification: A benchmark," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1116–1124.

[11] L. Leal-Taixe, A. Milan, et al., "Motchallenge 2015: Towards a benchmark for multi-target tracking," *arXiv preprint arXiv:1504.01942*, 2015.

[12] A. Milan, L. Leal-Taixe, et al., "Mot16: A benchmark for multi-object tracking," *arXiv preprint arXiv:1603.00831*, 2016.

[13] J. Fu, J. Liu, et al., "Dual attention network for scene segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3141–3149.

[14] Z. Zhang, Z. Cui, et al., "Pad-net: Multitasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 675–684.

[15] W. Sanghyun, P. Jongchan, L. Joon-Young, and K. I. So, "Cbam: Convolutional block attention module," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[16] T.-Y. Lin, M. Maire, et al., "Microsoft coco: Common objects in context," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 740–755.

[17] A. Ess, B. Leibe, K. Schindler, and L. Van Gool, "A mobile vision system for robust multi-person tracking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[18] S. Zhang, R. Benenson, and B. Schiele, "Citypersons: A diverse dataset for pedestrian detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3213–3221.

[19] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 304–311.

[20] T. Xiao, S. Li, et al., "Joint detection and identification feature learning for person search," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3415–3424.

[21] L. Zheng, H. Zhang, et al., "Person re-identifification in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1367–1376.

[22] S. Shao, Z. Zhao, et al., "Crowdhuman: A benchmark for detecting human in a crowd," *arXiv preprint arXiv:1805.00123*, 2018.

[23] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, no. 1, pp. 1–10, 2008.

[24] E. Ristani, F. Solera, et al., "Performance measures and a data set for multi-target, multi-camera tracking," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 17–35.

[25] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[26] B. Pang, Y. Li, et al., "Tubetk: Adopting tubes to track multi-object in a one-step training model," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6308–6318.

[27] N. Mahmoudi, S. M. Ahadi, and M. Rahmati, "Multi-target tracking using cnn-based features: Cnnmtt," *Multimedia Tools and Applications*, vol. 78, no. 6, pp. 7077–7096, 2019.

[28] J. Peng, C. Wang, et al., "Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking," in *European Conference on Computer Vision (ECCV)*, 2020.

[29] X. Zhou, V. Koltun, and P. Krahenb¨uhl, "Tracking objects as points," in *European Conference on Computer Vision (ECCV)*, 2020.

[30] H. Luo, Y. Gu, et al., "Bag of tricks and a strong baseline for deep person re-identifification," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.