ReIMOT: Rethinking and Improving Multi-object Tracking Based on JDE Approach

2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI) | 978-1-6654-9916-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/PRAI55851.2022.9904121

Haoxiong Hou
Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences
Xi'an, China
University of Chinese Academy of Sciences
Beijing, China
houhaoxiong@opt.ac.cn

Ximing Zhang
Space Optical Research Lab Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences
Xi'an, China
zhangximing@opt.ac.cn

Zhonghan Sun
Space Optical Research Lab Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences
Xi'an, China
sunzhonghan@opt.ac.cn

Wei Gao*
Space Optical Research Lab Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences
Xi'an, China
gaowei@opt.ac.cn

***Abstract*—The multi-object tracking (MOT) algorithms of the joint detection and embedding (JDE) approach estimate bounding boxes and re-identification (re-ID) features of objects with the single network, which balance the tracking accuracy and inference speed. However, when the appearance information between different objects is highly similar, these algorithms are usually easy to cause identity switches, and the comprehensive tracking performance is poor in crowded scenes. Aiming at the above problems, we propose a stronger multi-object tracking algorithm termed as ReIMOT, based on FairMOT. A joint loss function of combining normalized Softmax Loss and the center distance penalty term is designed to supervise the re-ID branch, which increases the intra-class similarity and makes the extracted appearance features more discriminative. To further improve the tracking performance, we introduce coordinate attention to make the encoder-decoder network focus more on features of interest. The experimental results show that the proposed ReIMOT is more effective than the other advanced multi-object tracking algorithms, and decreases the number of ID switches by 13.8% compared to FairMOT on the MOT17 dataset.**

Keywords—multi-object tracking, joint detection and embedding, re-identification, joint loss function, coordinate attention

I. INTRODUCTION

Multi-object tracking (MOT), one of the hottest high-level computer vision tasks, is used in the fields of autonomous driving, video surveillance and even epidemic prevention. The objective of MOT is to locate multiple objects and obtain trajectories in a video while assigning a unique and long-term valid ID number to each object. In crowded scenes, it is a difficult challenge to solve the tracking problems caused by the interference between similar objects and frequent occlusion.

Most of the current MOT algorithms adopt the strategy based on separating detection and embedding approach. Two different neural networks are used to sequentially estimate the bounding boxes and re-ID features of objects. Then these algorithms associate the data by Kalman filtering and Hungarian matching, such as SORT [1], DeepSORT [2]. However, since the extracted features can not be shared between the two independent networks, it leads to the consumption of storage and computing resources. In response to the complex tracking model and slow inference speed, the one-stage MOT algorithms have achieved great results recently. Two tasks of object detection and re-ID feature extraction can be accomplished in a single network. Track-RCNN [3] adds a fully connected layer at the head of the Mask-RCNN [4] network to extract re-ID features for tracking association, enabling object tracking at the pixel level. To achieve faster inference speed, Wang et al. propose the JDE algorithm [5]. By extending the one-stage object detection network named YOLOv3 [6], it becomes the first real-time tracking algorithm. Yifu Zhang et al. propose FairMOT [7] based on the anchor-free object detection network CenterNet [8]. The algorithm verifies that anchors are not suitable for extracting re-ID features when dealing with the MOT task. It achieves better tracking accuracy on multiple datasets and the real-time requirement is achieved.

In crowded scenes, FairMOT still suffers from a noticeable degradation of tracking metrics. In this paper, we identify two factors behind the unstable performance. The one issue is caused by the loss function for re-ID branch. The re-ID branch uses the Softmax Loss function, which can better distinguish different categories of objects with large appearance differences. But the MOT task is often required to distinguish different objects of the same category with large similarity. Training with Softmax Loss function results in that the extracted re-ID features of the same object are not compact enough. The other issue is caused by the encoder-decoder network DLA-34 [9]. The high-frequency fusion inside the component block and across different scale layers, while obtaining a larger global view, weakens the perception of focal features and limits the feature representation.

In this work, we propose a simple algorithm termed as ReIMOT, which addresses the above issues as illustrated. ReIMOT is short for "Rethinking and Improving Multi-Object Tracking". ReIMOT is built on top of CenterNet using FairMOT as a baseline. Similarly, detection and re-ID tasks are integrated in one neural network. We argue that the discriminative re-ID features can facilitate solving the problem of high identity switches (IDs) and keep the continuity of trajectories in crowded scenes. Therefore we propose a joint loss function to train the re-ID branch. On the one hand, the original Softmax Loss is normalized to convert the optimization from inner product into

* Corresponding author.

This work is supported by National Natural Science Foundation (NNSF) of China under Grant 61906186.

angles. This is inherently consistent with the mechanism of using the cosine metric to compute the similarity in the process of data association. On the other hand, the center distance penalty term [10] is added to further shrink the distribution space of the same objects and increase the intra-class similarity. In addition, we also think it becomes critical for the anchor-free detection network to extract more accurate key point features. Attention mechanism enables the convolutional neural network to focus more on local features of interest. We redesign the encoder-decoder network DLA-34 by introducing coordinate attention [11] with the very low computational and storage consumption. This module computes attention masks for high-resolution features on channels, horizontal and vertical coordinate, enhancing the ability to locate object centers. The contributions of this paper are summarized as follows:

- A joint loss function optimizing in the angular space is proposed to supervise the re-ID branch. By setting the center distance penalty term, maximizing intra-class similarity makes the extracted re-ID features more high-quality.
- The feature extraction network DLA-34 is optimized by adding the coordinate attention to focus more on features that are more valuable to the multi-object tracking task.
- The experimental results on the 2DMOT15 and MOT17 datasets show that our proposed algorithm can effectively improve the performance of multi-object tracking, and surpass the other advanced algorithms.

II. The Proposed Method

In this section, we describe technical details of the proposed multi-object tracking algorithm including ReIMOT architecture, re-ID loss function as well as coordinate attention.

A. ReIMOT Architecture

The overview of ReIMOT is shown in Fig. 1, which adopts the joint detection and embedding approach. Inspired by FairMOT, ReIMOT is built on top of CenterNet, an anchor-free detection network. The architecture consists of three parts, an encoder-decoder network for extracting high-resolution features, a detection branch for obtaining the object center heatmap, box size and center offset, and a re-ID branch for extracting appearance features. Detection branch and re-ID branch share features extracted by the encoder-decoder network DLA-34, and the re-ID feature is estimated based on the predicted object center position for data association.

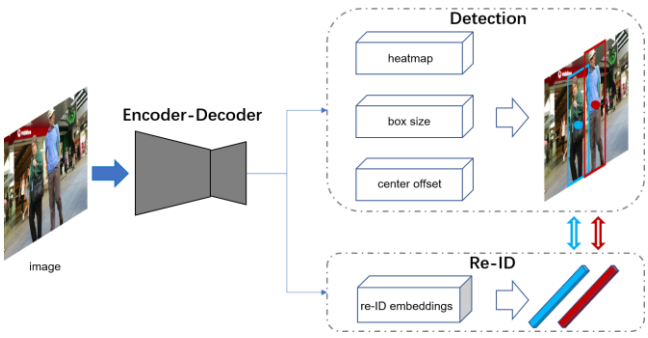

Fig. 1. Overview of our proposed tracker ReIMOT based on JDE approach.

B. Re-ID Loss Function

The role of re-ID branch is to generate the features that can recognize different objects. For the multi-object tracking task, object similarity matching is a fine-grained process. Different objects of the same category have high similarity. In crowded scenes, frequent inter-object interactions and occlusions further impose a higher requirement on the discriminativeness of re-ID features.

In this paper, we propose a joint loss function for supervising re-ID branch training. It mainly includes two parts: (1) Our approach normalizes the feature vector and the weight vector, thus projecting features in the original Euclidean space onto the angular space; (2) Based on the idea of central clustering, we set a center penalty term to reduce the distance within the class.

The Softmax Loss (SL), which is widely used in coarse-grained classification tasks, is defined as follows:

$$L_S = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\Sigma_{j=1}^{n} e^{W_j^T x_i + b_j}}. \quad (1)$$

where N is the batch size of training and n is the number of classes. x_i denotes a feature vector of the i-th sample belonging to the class y_i. W, b denote the weight and bias in the last fully-connected layer of the network. For simplicity of optimization, we set the bias $b = 0$. Thus the exponential term in Eqn. (1) can be transformed from the vector inner to the angular cosine as follows:

$$W^T x = \|W\|\|x\| cos\theta. \quad (2)$$

where θ denotes the angle between weight vectors W and feature vectors x. Further we regularize weight vectors and feature vectors by L_2 normalization. We fix $\|W\| = 1$ and $\|x\| = 1$. During the MOT data association process, the similarity between two re-ID features is computed using the cosine distance as a metric. This suggests that the norm is more in line with the object discrimination. Then normalized SL can be formulated as:

$$L_{NS} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{r\cos(\theta_{y_i,i})}}{\Sigma_{j=1}^{n} e^{r\cos(\theta_{j,i})}}. \quad (3)$$

where r is a hyperparameter. We constrain the feature vectors to a hypersphere of radius r by normalization.

However, the learned features driven by the L_{NS} are divided only by the number of categories, ensuring that the classes are separable but not requiring the intra-class compactness. This is not suitable for the fine-grained classification. To improve discriminability of the features, we add the center penalty term, which is set by calculating the Euclidean distance between a sample and its category center. The proposed joint loss function for re-ID branch is as follows:

$$L_{NS-CL} = -\frac{(1-\alpha)}{N}\sum_{i=1}^{N} log\frac{e^{r\cos(\theta_{y_i,i})}}{\Sigma_{j=1}^{n} e^{r\cos(\theta_{j,i})}} + \frac{\alpha}{2N}\sum_{i=1}^{N}\left\|x_i - c_{y_i}\right\|_2^2. \quad (4)$$

where y_i denotes the category to which the i-th sample belongs and c_{y_i} is the feature center of y_i. Besides, α is used as a hyperparameter to balance the two loss functions. When $\alpha = 0$, L_{NS-CL} degenerates to L_{NS}. The proposed L_{NS-CL} can pull all features of each category towards the corresponding category center. The re-ID features learned are made to have stronger representational ability.

C. Coordinate Attention

The attention mechanism is inspired by the human visual system [12], which enables the model to focus more on valuable information and helps to extract features of interest in computer vision tasks. Coordinate attention computes attention masks for feature maps on channels, horizontal and vertical coordinate, which achieves excellent results with the low computational and storage consumption. It can encode both channel relationships and capture long-range dependencies with precise positional information.

Our encoder-decoder network adopts ResNet-34 [13] as backbone and a modified Deep Layer Aggregation (DLA) for features fusion, as shown in Fig. 2. (a). Where the number denotes the downsampling multiple. The network has more frequent skip connections between low-level and high-level features to expand the receptive field. Notably, the output feature map has 1/4 high resolution of the original image, which facilitates the identification of small objects.

We introduce the coordinate attention to the encoder-decoder network for improving MOT tracking accuracy. Firstly, coordinate attention respectively encodes each channel along the horizontal and vertical directions using average pooling, which can capture long-range interactions spatially. Secondly, the features in the two spatial directions are concatenated and convolved to further fusion. Finally, the above feature map is re-divided into two separate masks along the spatial dimension, and the two attention masks are multiplied onto the input feature maps to enhance the representation ability of the feature map. The whole flow is shown in Fig. 2 (b).

On the 1/8 and 1/16 resolution stages, coordinate attention encodes channel relationships and spatial locations. It can suppress background noise while increasing weights of the salient regions and key point features, which is very effective for anchor-free detection network CenterNet depending on the object center for positioning.

III. Experimental Results

A. Dataset and Metrics

We train on the combination of CrowdHuman dataset and MOT17 half training set. And we only use the bounding box annotations of the CrowdHuman to train detection branch of ReIMOT. We present ablation experiments on the 2DMOT15 dataset and compare the tracking ability with other MOT algorithms on the other half set of MOT17.

We use the CLEAR metrics [14], including multiple object tracking accuracy (MOTA), identification F1 score (IDF1), false positive (FP), false negative (FN), and number of identity switches (IDs) to evaluate different aspects of the tracking performance following FairMOT.

B. The Implementation Details

The experimental environment is a deep learning server with an Intel Xeon CPU Gold 6130 processor and two RTX 2080 Ti GPUs. We evaluate tracking performance on a single GPU. For ReIMOT, we initialize the model by adopting the CenterNet detection model parameters pretrained on the COCO dataset. The input image is resized to 1088 × 608. During data preprocessing, we introduce the standard data augmentation methods including rotation, scaling and color jittering. We train our ReIMOT with the adam optimizer for 40 epochs with a starting learning rate of 10^{-4}. At the 20th epoch and 35th epoch, the learning rate respectively decreases to 10^{-5} and 10^{-6}. The model is trained with batch size of 12.

C. Ablation Experiments

In order to verify the effectiveness of the proposed ReIMOT, we perform an ablation experiment on the 2DMOT15 dataset. The results are shown in Table I.

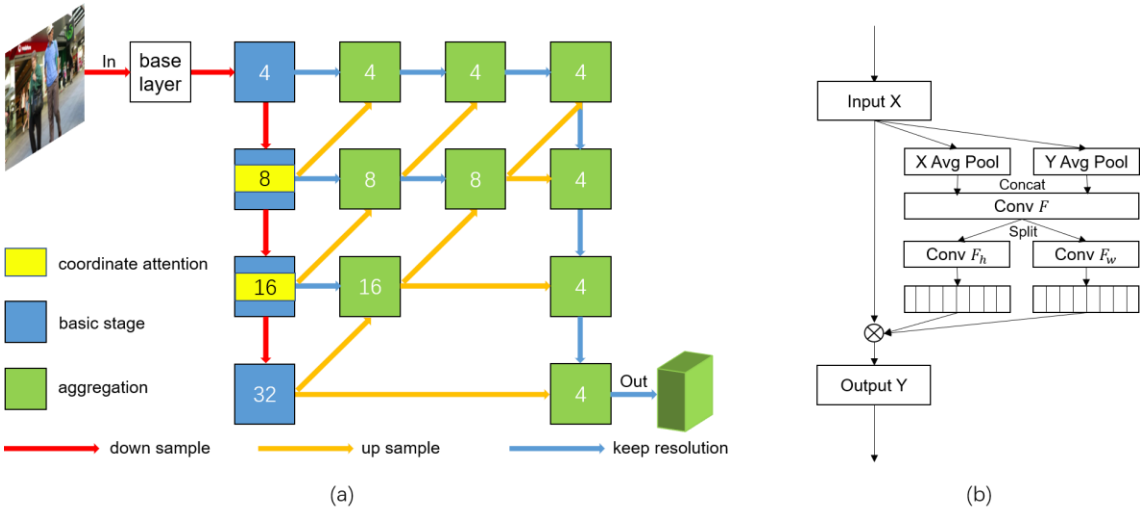

Fig. 2. (a) Architecture of the encoder-decoder network DLA-34. (b) Flow of the coordinate attention.

The best results are shown in **bold**, and ✓ means that the corresponding component is applied. When we adopt the proposed joint loss function L_{NS-CL} to train the re-ID branch, IDF1, which is more sensitive to the trajectories continuity, increases from 74.8 to 75.8. The normalized Softmax Loss with the center distance penalty term can make the re-ID features more discriminative. Besides, when we only introduce the coordinate attention (CA), both IDF1 and MOTA increase by 0.5. This shows that coordinate attention enables the network to focus more on features of interest, which are valuable for the multi-object tracking task. Finally, our algorithm improves the IDF1 from 74.8 to 76.4 and MOTA from 74.3 to 75.0. The results show that our ReIMOT has a better multi-object tracking performance compared to the baseline FairMOT.

TABLE I. RESULTS OF ABLATION EXPERIMENTS

L_{NS-CL}	CA	IDF1	MOTA	FP	FN	IDs
		74.8	74.3	1194	1955	95
✓		75.8	74.6	1232	1936	105
	✓	75.3	74.8	1263	**1935**	104
✓	✓	**76.4**	**75.0**	**1158**	1957	**93**

D. Comparison of Other Advanced Algorithms

To verify the superiority of the proposed algorithm in this paper, we compare our ReIMOT to other advanced multi-object tracking algorithms in Table II. It is worth noting that the MOT17 dataset contains rich crowded scenes. For the results on the MOT17, we can see that ReIMOT significantly outperforms other algorithms on almost all metrics. In addition, ReIMOT outperforms FairMOT on all metrics (i.e. +1.4 IDF1 ,+0.5 MOTA, -309 FP and -210 FN), and decreases the number of ID switches by 13.8%. All indicate that our algorithm achieves very high tracking performance.

TABLE II. COMPARISON OF OTHER ADVANCED ALGORITHMS

Tracker	IDF1	MOTA	FP	FN	IDs
Chained-Track [15]	60.9	63.1	2955	16174	755
JDE [5]	63.6	60.0	2923	18158	473
CenterTrack [16]	64.2	66.1	**2442**	15286	528
QuasiDense [17]	67.8	67.3	2637	14605	377
FairMOT [7]	74.7	71.2	3516	11935	413
ReIMOT (ours)	**76.1**	**71.7**	3207	**11725**	**356**

E. Visualization Results

The visualization results of ReIMOT compared to FairMOT on the set of MOT17-Seq are shown in Fig. 3. Where IDs and FN respectively indicate that the identity of the tracked object is switched and the object is not recognized. The checkmark indicates that the identity of the object is not switched. The difficult cases include severe occlusion and highly similar objects. We can see that ReIMOT can assign the correct identities with the help of high-quality re-ID features when the two objects are completely masked over. In particular, the object with large information loss can be detected correctly.

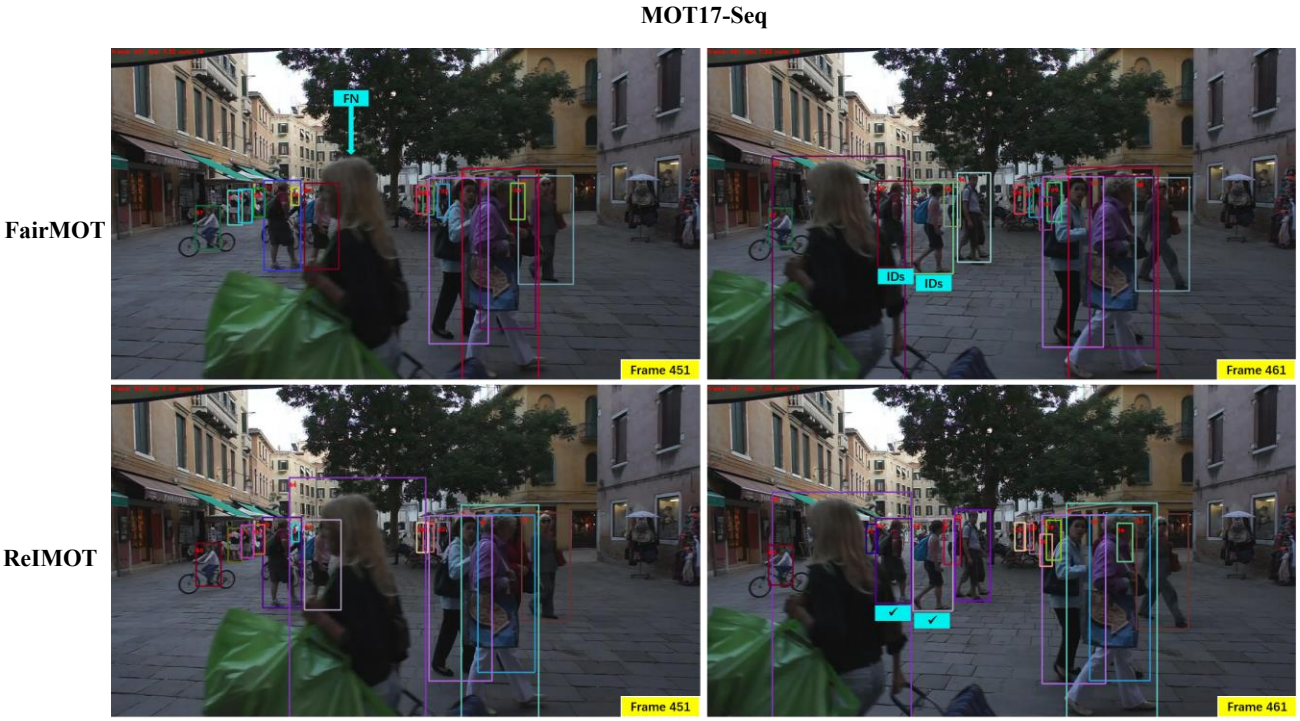

Fig. 3. Robustness of our ReIMOT compared to FairMOT.

IV. Conclusion

In this paper we propose the ReIMOT, which improves multi-object tracking performance based on JDE approach. Aiming at the similar object interference and occlusion in crowded scenes, we propose a joint re-ID loss function that is more suitable for fine-grained classification in the multi-object tracking task. Adding the center distance constraint further makes the extracted re-ID features more high-quality. In addition, we add coordinate attention to the encoder-decoder network, which helps to extract the key point features. The experiments show that ReIMOT outperforms other advanced MOT algorithms on almost all metrics, which verifies the effectiveness of our algorithm. In the future, we would like to continue to explore the JDE-based MOT algorithm which is important for balancing tracking accuracy and inference speed.

References

[1] Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B. "Simple online and realtime tracking". In *2016 IEEE international conference on image processing (ICIP)*, pp. 3464-3468, September, 2016.

[2] Wojke, N., Bewley, A., & Paulus, D. "Simple online and realtime tracking with a deep association metric". In *2017 IEEE international conference on image processing (ICIP)*, pp. 3645-3649, September, 2017.

[3] Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B. B. G., Geiger, A., & Leibe, B. "Mots: Multi-object tracking and segmentation". In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 7942-7951, 2019.

[4] He, K., Gkioxari, G., Dollár, P., & Girshick, R. "Mask r-cnn". In *Proceedings of the IEEE international conference on computer vision*, pp. 2961-2969, 2017.

[5] Wang, Z., Zheng, L., Liu, Y., Li, Y., & Wang, S. "Towards real-time multi-object tracking". In *European Conference on Computer Vision*, pp. 107-122, August, 2020.

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

[7] 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(11), pp. 3069-3087, 2021.

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

[9] Yu, F., Wang, D., Shelhamer, E., & Darrell, T. "Deep layer aggregation". In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2403-2412, 2018.

[10] Wen, Y., Zhang, K., Li, Z., & Qiao, Y. "A discriminative feature learning approach for deep face recognition". In *European conference on computer vision*, pp. 499-515, October, 2016.

[11] Hou, Q., Zhou, D., & Feng, J. "Coordinate attention for efficient mobile network design". In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13713-13722, 2021.

[12] Baiqi, X., Gangwu, J., Jianhui, L., Xin, W., & Peidong, Y. "Aircraft Rotated Boxes Detection Method Based on YOLOv5". In *2021 4th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*, pp. 390-394, August, 2021.

[13] 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.

[14] Bernardin, K., & Stiefelhagen, R. "Evaluating multiple object tracking performance: the clear mot metrics". In *EURASIP Journal on Image and Video Processing*, pp. 1-10, 2008.

[15] Peng, J., Wang, C., Wan, F., Wu, Y., Wang, Y., Tai, Y., ... & Fu, Y. "Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking". In *European conference on computer vision*, pp. 145-161, August, 2020.

[16] Zhou, X., Koltun, V., & Krähenbühl, P. "Tracking objects as points". In *European Conference on Computer Vision*, pp. 474-490. August, 2020.

[17] Pang, J., Qiu, L., Li, X., Chen, H., Li, Q., Darrell, T., & Yu, F. "Quasi-dense similarity learning for multiple object tracking". In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 164-173, 2021.