Multiple object tracking using GAN

Asif Iqbal Rahaman, Steven DeVerteuil, Ting-Chih Chen and Yu-Chung Cheng
Department of Computer Sciences, Virginia Tech
asifiqbal@vt.edu, stedevert33@vt.edu, tingchih@vt.edu, anthonycheng@vt.edu

Abstract

Multiple object tracking is a critical aspect of computer vision with cutting-edge applications such as autonomous driving. Existing methods focus on two dominant approaches: The first entails initial object detection, followed by video-based ID re-identification, exhibiting high performance, but burdened with issues such as compute resource intensiveness and time consumption. Conversely, the second approach, exemplified by models like FairMOT [48], seamlessly integrates both stages, showcasing a notable enhancement in tracking performance. In this project, we propose a GAN-based model to enhance FairMOT's performance given the emergence of GAN architectures in the multiple object tracking domain.

A. Introduction

Multiple Object Tracking (MOT) has been a persistent objective in the field of computer vision [2, 4, 31, 32]. This pursuit involves predicting trajectories for objects of interest within video sequences. Successful MOT can yield advantages across various applications, including intelligent video analysis [5, 10, 20], human-computer interaction [49], human activity recognition [25, 29], and autonomous driving [7, 19, 27, 38].

Current approaches predominantly adopt one of two methodologies. The initial approach [12, 46, 48, 50] involves first recognizing the objects and subsequently making predictions, primarily leveraging the Mask-RCNN model [17]. On the other hand, the second method [11, 16, 21, 34] focuses on predicting trajectories by capturing factors such as speed, direction of motion, and past trajectory, with a primary reliance on the Generative Adversarial Network (GAN) model [15].

Significant advancements have been made in both the first (Mask-RCNN) and second (GAN) approaches. Nonetheless, the primary challenge in the first approach lies in the inherent two-stage nature of the models. Many first-approach models face issues related to computational resource constraints and time consumption. FairMOT [48] has successfully addressed these challenges, yet the two-stage structure remains a concern in Figure 5. We attribute this issue to the necessity of performing predictions after object detection. In light of this issue, our strategy is to adopt the second approach for object tracking. We believe that the second model holds the potential to be more efficient and stable compared to the first model.

B. Related work

B.1. Object Detection Methods

Object detection has been an important task in computer vision since the field's onset. Early successes include feature representation algorithms such as SIFT [26], the Viola–Jones object detection framework [39], Histogram of Oriented Gradients algorithm [9], and Deform-able Part Based Machines [13].

With the onset of deep learning, the state of the art of the field has rapidly shifted approaches to end-to-end algorithms [43]. Most often, these end-to-end neural net algorithms utilize convolutional neural nets (CNN) to automatically discern underlying features in the images. Landmark CNN models include AlexNet [22], R-CNN [14], YOLO [30], ResNet [18], and Mask R-CNN [17] to name a few. These algorithms produce greater accuracies than 'classical' methods, and are fundamental to the detection of objects within tracking algorithms, as discussed in the next section.

B.2. Object Tracking Methods

Multiple object tracking (MOT) is a difficult but important problem in computer vision. This task aims to label each object in a frame and track it across multiple frames. MOT algorithms generally fall into either 'online' or 'offline' categories depending on the use case. Online applications are ones that cannot use future frames, such as autonomous driving, live sports, etc. Offline applications, conversely, can use future frames for tracking, and include applications like video analysis. State-of-the-art MOT algorithms of the past decade generally follow the *Tracking-By-Detection* paradigm which follows two separate steps:

1) detection and 2) data association [40, 41], respectively discussed in the following paragraphs.

First, objects are detected through an object detector, typically composed of one or two stages. The two stage detectors (region of interest proposals) offer high accuracy but at the cost of speed. On the contrary, one-stage detectors are faster but less accurate. One-stage YOLO detectors have been widely used thanks to their speed and accuracy. Further, many improvements have been made to YOLO detectors to address issues of anchor-based and anchor-free detectors, prominently resulting in detectors such as PRB-Net [8].

Second, objects are algorithmically associated between frames. There are many association algorithms, ranging from purely statistical algorithms to generative algorithms. For example, the statistical SORT algorithm [4] used Kalman filtering and Hungarian data association algorithm to achieve a MOT16 benchmark [24]. Several MOT algorithms are based off SORT including DeepSORT [42], ByteTrack [47], BoT-SORT [1], and OC-SORT [6].

This *'Separate Detection Embedding'* (SDE) model architecture can introduce efficiency problems as its efficiency will only ever be the sum of the two steps. Consequently, several authors have explored the joint training of object detection and appearance embedding (aka data association). *'Joint Detection Embedding' (JDE)* [41] is one such approach, with several other successors including Track-RCNN [36], FairMOT [48], and UNICORN [45].

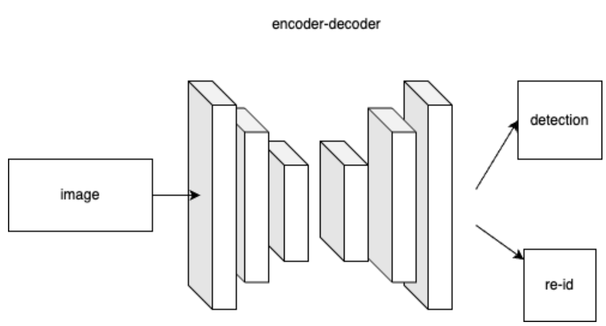

Figure 1. FairMoT architecture

Further strategies for predicting the trajectories and appearances of detected objects include transformer based approaches. Such approaches include Trackformer [23], TransTrack [37], and TransCenter [44]. However, these attention-based transformer approaches are very computationally expensive and are therefore not useful to real-time applications. Finally, there are generative prediction methods as discussed in the next section.

B.3. GAN Tracking Methods

Generative methods focus on generating data from given data distributions. Hence, we can utilize their statistical property to generate future trajectories or sequences of motion. These techniques leverage deep learning to acquire the ability to produce plausible current trajectories or future motion.

Like object detection, generative methods range from classic statistical approaches to modern deep learning approaches. One such example of a classic approach is a trajectory sampling technique named Monte Carlo method. It is used to generate future trajectories by sampling from learned motion distributions. On the other hand, deep learning approaches, such as Variational Autoencoders (VAEs), learn probabilistic representations of motion data and produce diverse trajectories by sampling from learned latent variables. Similarly, Generative Adversarial Networks (GANs) comprise a generator and discriminator network, trained to create realistic motion trajectories by minimizing the difference between generated and real trajectories.

GANs have improved, but not solved, some of the known problems of object tracking, such as object occlusion, as demonstrated in the multiple methods from [35]. Similarly, they have improved the performance of object and motion prediction compared to traditional physics-based approaches [11]. However, GANs historically have not been used for object tracking. We feel that their under-utilization is an opportunity to present a new framework for object tracking.

C. Method

Our tracking method appends a GAN architecture to the existing FairMOT architecture. Given that GANs learn very well under careful architecture and hyperparameters [15, 28], we hypothesize that adding a GAN architecture to the existing FairMOT architecture in the Figure 1 will improve object tracking and reduce speed. We will use the GAN to refine the detected bounding boxes and re-ID features offline, as visible in Figure 2. Then, bounding boxes and re-IDs will be generated quickly and accurately with solely the GAN component for online testing.

Since we modeled the discriminator similar to the loss function in 2, we focus solely on the discriminator's loss function 1. The use of 0 for true data indicates our aim for true data to achieve a loss value close to zero while ensuring that predicted data incurs as large a loss value as possible in the discriminator.

$$\begin{aligned} L_D = &\ \text{BCEWithLogitsLoss}(\text{Truth data}, 0) \\ &+ \text{BCEWithLogitsLoss}(\text{Predicted data}, 1), \end{aligned} \tag{1}$$

$$L_{FairMOT} = Discriminator(\text{Predicted data}), \tag{2}$$

The design of our discriminator is illustrated in 3. We apply convolution to process the 2D data (detection data),

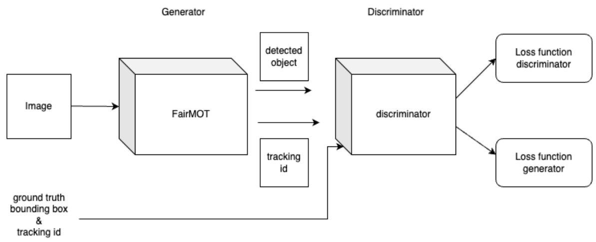

Figure 2. Proposed FairMOT + GAN Architecture

flatten it into a linear layer, and then concatenate it with ReID. The goal for the second layer is to extract features from both predicted and ground truth data. A final layer is used to differentiate these features.

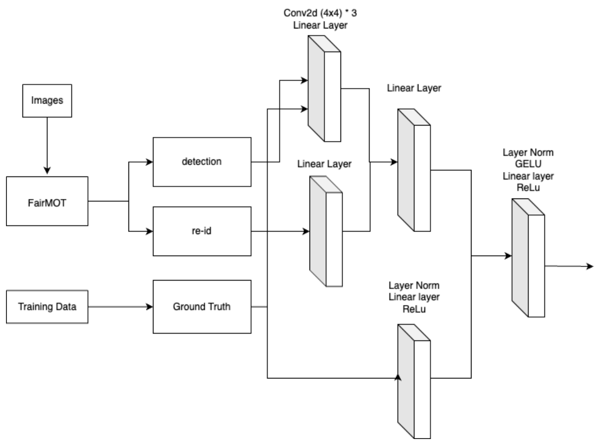

Figure 3. Discriminator architecture

The strategy involves leveraging baseline FairMoT as the foundation of our code. Specifically, we will be directly lifting-and-shifting all of FairMOT as the first component of our system. Then, we will add on our custom GAN architecture.

D. Dataset

For our experiment, we will be using the MOT20 dataset[1]. This dataset has extensive documentation and benchmarks on other architectures, making for easy model comparison.

The dataset contains 4 videos in the training set and 4 videos in the test set. The training videos contain a total of 8931 frames in a period of 357 seconds, whereas the test videos contain 4479 frames in a period of 178 seconds. The data contains bounding boxes coordinate and object ID labels. The total size of the dataset is 5GB, reducing risk for computational overload while training our models.

[1]https://motchallenge.net/data/MOT20/

E. Evaluation

In our evaluation process, we employ four key metrics to assess tracking performance. We gauge detection results using 1) average precision, while overall tracking accuracy is evaluated through the 2) CLEAR metric [3] and 3) IDF1 [33]. These metrics collectively provide a comprehensive assessment of our tracking results.

For the prediction generated by GAN models, we will use 4) IoU (Intersection over Union) metric. This metric calculates the overlap between ground truth (*gt*) and prediction (*pd*) over the area of union. IoU can range from 0 to 1, where 0 implies no overlap and 1 implies complete overlap as calculated by: $IoU = \frac{gt \cap pd}{gt \cup pd}$

F. Results

We did not achieve our desired results. We obtained a training loss of 19.14 across 20 epochs of training on the MOT20 training set. As can be seen in the figure 5, the training loss increased over time.

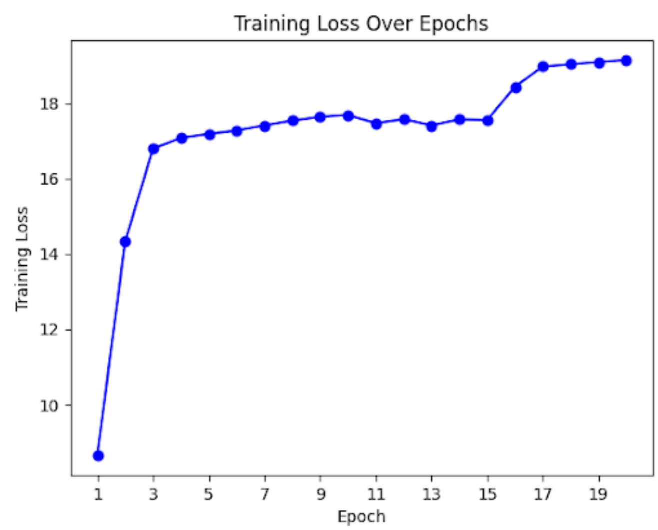

Figure 4. Train Loss

We were unable to obtain a test accuracy as our accuracy depending on submitting our predictions to the MOTA challenge[2] auditors and we did not receive our accuracy figure in time for this paper's submission. Further, we were unable to evaluate using our other aforementioned metrics (CLEAR, IDF1, IoU).

In a test video demonstration of our tracking on MOT20, we found the bounding boxes to appear out-of-frame or partially out-of-frame in most cases. Further, they were stagnant for the duration of the test video. Specifically, they were "jittering" in place the whole time. Additionally, they

[2]https://motchallenge.net/instructions/

were of various shapes and sizes, not nearly encompassing people and/or over-encompassing large frame regions. These errors can be seen in the below figure.

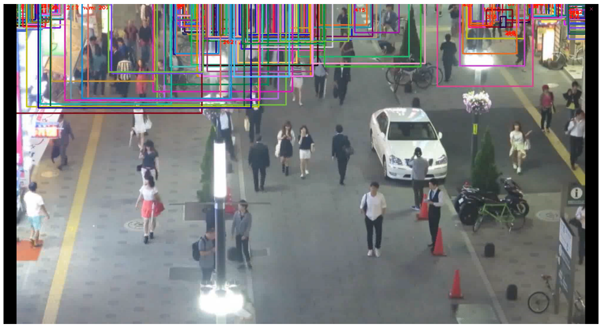

Figure 5. Single Test Demonstration Video Frame

G. Discussion

We have several hypotheses as to why we obtained subpar results. Firstly, we believe that the Discriminator architecture is flawed. Upon examining the inference results, it is evident that the model fails to detect objects, indicating that the employed loss function inadequately guides the model in object tracking. Additionally, looking at the Discriminator architecture, it is very plausible that by passing the ground truth labels to a single separate layer, and passing its output as input to another layer that is separate from the predicted data, the Discriminator was able to immediately discern that the predictions were "fake". As seen in figure 3, the ground truth and predicted data are coming from two separate layers, and the discriminator would be able to easily tell the data apart by looking at which layer the data is coming from.

Second, we believe that something was incorrect with the bounding box coordinate system given that the bounding boxes in the video were all or mostly out of frame. We believe that something went wrong with the data annotation. Specifically, the data for MOT16 and MOT17 were annotated different from MOT20. We believe that we followed the steps required by the original FairMOT code [3] to adjust accordingly, but perhaps there was something we did incorrectly and/or missed. This hypothesis would explain why the bounding boxes were all shifted up to be in the upper half of the frame or beyond. If the issue had solely been the Discriminator architecture, we hypothesize that the bounding boxes would've been at least somewhat more randomly spaced throughout the frames, rather than being grouped in the upper half and beyond.

[3] https://github.com/ifzhang/FairMOT#training

H. Conclusion

Given the success of anchor-free approaches to multi-object tracking, as seen within FairMOT, and the emergence of GAN architectures used within the MOT domain, we hypothesized that appending a GAN architecture on top of the existing anchor-free object detector could enhance performance and potentially minimize some of the common issues associated with MOT (occlusion, re-identification, etc). We found that, without paying careful attention to the data annotation format and the appended Discriminator architecture, such an approach does not work. We were unable to obtain a final MOTA accuracy, but given our test demonstration video, it is safe to say that our accuracy would not be near the accuracies of FairMOT or other MOT methods. Future work would include improving our Discriminator architecture by first basing it off a similar, existing implementation before trying to start from scratch with our own. Further, ensuring that the data is still correct after our lift-and-shift approach would be very beneficial.

References

[1] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking, 2022. 2

[2] Y. Bar-Shalom. *Tracking and Data Association*. Academic Press Professional, Inc., USA, 1987. 1

[3] Keni Bernardin and Rainer Stiefelhagen. Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics. *EURASIP Journal on Image and Video Processing*, 2008(1):246309, May 2008. 3

[4] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, sep 2016. 1, 2

[5] Arjun Ravi Shankar Varun Kejriwal Goel Jocelyn Barker Amer Ghanem Philip Lee Meghan Milecky Natalie Stottler Bokai Zhang, Darrick Sturgeon and Svetlana Petculescu. Surgical instrument recognition for instrument usage documentation and surgical video library indexing. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 11(4):1064–1072, 2023. 1

[6] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking, 2023. 2

[7] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers, 2023. 1

[8] Ping-Yang Chen, Ming-Ching Chang, Jun-Wei Hsieh, and Yong-Sheng Chen. Parallel residual bi-fusion feature pyramid network for accurate single-shot object detection. *IEEE Transactions on Image Processing*, 30:9099–9111, 2021. 2

[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, 2005. 1

[10] Armin Danesh Pazho, Christopher Neff, Ghazal Alinezhad Noghre, Babak Rahimi Ardabili, Shanle Yao, Mohammadreza Baharani, and Hamed Tabkhi. Ancilia: Scalable intelligent video surveillance for the artificial intelligence of things. *IEEE Internet of Things Journal*, 10(17):14940–14951, 2023. 1

[11] Patrick Dendorfer, Aljoša Ošep, and Laura Leal-Taixé. Goal-gan: Multimodal trajectory prediction based on goal position estimation, 2020. 1, 2

[12] Kuan Fang, Yu Xiang, Xiaocheng Li, and Silvio Savarese. Recurrent autoregressive networks for online multi-object tracking, 2018. 1

[13] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010. 1

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation, 2014. 1

[15] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 1, 2

[16] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks, 2018. 1

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2018. 1

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 1

[19] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17853–17862, June 2023. 1

[20] Xin Jin, Ruoyu Feng, Simeng Sun, Runsen Feng, Tianyu He, and Zhibo Chen. Semantically video coding: Instill static-dynamic clues into structured bitstream for ai tasks, 2022. 1

[21] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, S. Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks, 2019. 1

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, may 2017. 1

[23] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers, 2022. 2

[24] Anton Milan, Laura Leal-Taixe, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking, 2016. 2

[25] Taima Rahman Mim, Maliha Amatullah, Sadia Afreen, Mohammad Abu Yousuf, Shahadat Uddin, Salem A. Alyami, Khondokar Fida Hasan, and Mohammad Ali Moni. Gru-inc: An inception-attention based approach using gru for human activity recognition. *Expert Systems with Applications*, 216:119419, 2023. 1

[26] Jean-Michel Morel and Guoshen Yu. Is the " scale invariant feature transform " (sift) really scale invariant ? 2010. 1

[27] Mahyar Najibi, Jingwei Ji, Yin Zhou, Charles R. Qi, Xinchen Yan, Scott Ettinger, and Dragomir Anguelov. Unsupervised 3d perception with 2d vision-language distillation for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8602–8612, October 2023. 1

[28] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016. 2

[29] Abhisek Ray, Maheshkumar H. Kolekar, R. Balasubramanian, and Adel Hafiane. Transfer learning enhanced vision-based human activity recognition: A decade-long analysis. *International Journal of Information Management Data Insights*, 3(1):100142, 2023. 1

[30] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2016. 1

[31] D. Reid. An algorithm for tracking multiple targets. *IEEE Transactions on Automatic Control*, 24(6):843–854, 1979. 1

[32] Seyed Hamid Rezatofighi, Anton Milan, Zhen Zhang, Qinfeng Shi, Anthony Dick, and Ian Reid. Joint probabilistic data association revisited. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 1

[33] Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking, 2016. 3

[34] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, S. Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints, 2018. 1

[35] Kaziwa Saleh, Sándor Szénási, and Zoltán Vámossy. Generative adversarial network for overcoming occlusion in images: A survey. *Algorithms*, 16(3), 2023. 2

[36] Bing Shuai, Andrew G. Berneshawi, Davide Modolo, and Joseph Tighe. Multi-object tracking with siamese track-rcnn, 2020. 2

[37] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer, 2021. 2

[38] Siyu Teng, Xuemin Hu, Peng Deng, Bai Li, Yuchen Li, Yunfeng Ai, Dongsheng Yang, Lingxi Li, Zhe Xuanyuan, Fenghua Zhu, and Long Chen. Motion planning for autonomous driving: The state of the art and future perspectives. *IEEE Transactions on Intelligent Vehicles*, 8(6):3692–3711, 2023. 1

[39] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I, 2001. 1

[40] Yu-Hsiang Wang, Jun-Wei Hsieh, Ping-Yang Chen, Ming-Ching Chang, Hung Hin So, and Xin Li. Smiletrack: Similarity learning for occlusion-aware multiple object tracking, 2023. 2

[41] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking, 2020. 2

[42] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric, 2017. 2

[43] Xiongwei Wu, Doyen Sahoo, and Steven C. H. Hoi. Recent advances in deep learning for object detection, 2019. 1

[44] Yihong Xu, Yutong Ban, Guillaume Delorme, Chuang Gan, Daniela Rus, and Xavier Alameda-Pineda. Transcenter: Transformers with dense representations for multiple-object tracking, 2022. 2

[45] Bin Yan, Yi Jiang, Peize Sun, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Towards grand unification of object tracking, 2022. 2

[46] Fengwei Yu, Wenbo Li, Quanquan Li, Yu Liu, Xiaohua Shi, and Junjie Yan. Poi: Multiple object tracking with high performance detection and appearance feature, 2016. 1

[47] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box, 2022. 2

[48] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. FairMOT: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129(11):3069–3087, sep 2021. 1, 2

[49] Rui Zhen, Wenchao Song, Qiang He, Juan Cao, Lei Shi, and Jia Luo. Human-computer interaction system: A survey of talking-head generation. *Electronics*, 12(1), 2023. 1

[50] Zongwei Zhou, Junliang Xing, Mengdan Zhang, and Weiming Hu. Online multi-target tracking with tensor-based high-order graph matching. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 1809–1814, 2018. 1