Real-time Multi-Person Tracking for Mobile Robotics Using Knowledge Distillation

by

Grace (Sooyeon) Bae

Supervisor: Prof. Goldie Nejat

April 27, 2023

B.A.Sc. Thesis

Division of Engineering Science
UNIVERSITY OF TORONTO

Abstract

This thesis aimed to address the challenge of achieving high-performance multiple person tracking on mobile robots. To this end, the state-of-the-art MOT method, MOTR, was trained on a dataset captured by a moving camera for deployment on mobile robots. Knowledge distillation was applied to maintain tracking performance, enabling real-time operation of the model. After 2 days of training, the distilled model achieved an overall 52.8% MOTA, while the non-distilled model achieved 65.2% MOTA after 4 days of training. While the performance of the distilled model was significantly impacted by poor performance on a subset of the dataset, the results demonstrate the benefits of applying distillation to MOT training, achieving similar performance to the original model while significantly reducing training time.

Acknowledgements

I would first like to thank Professor Goldie Nejat for giving me the opportunity to work on this project and contribute to the Autonomous Systems and Biomechatronics Laboratory. I am grateful to the Ph.D. student, Angus Fung, who was a great help throughout the thesis, giving me useful insights, support, and direction. Finally, I would like to thank my parents for their unwavering love and support, which has been instrumental in my academic journey.

Table of Contents

List of Abbreviations, Figures, and Tables

List of Abbreviations

MOT	Multi-Object Tracking
MOTA	Multi-Object Tracking Accuracy
MOTP	Multi-Object Tracking Precision

List of Figures

List of Tables

1. Introduction

Person tracking is a crucial task in various fields such as navigation in cluttered environments, crowd analysis of movement and behaviour, and human-robot interaction. It involves identifying and tracking individuals within a scene, even in crowded and cluttered environments. However, person tracking is a challenging task due to several factors such as occlusions, background complexity, and data association challenges [1]. Occlusions occur when the target is blocked and not visible, making it difficult to ensure accurate restoration of information when the target reappears [2]. The background complexity, such as changes in brightness, objects in the background, scene transformations, and shadows, adds an extra layer of difficulty to the tracking process. Additionally, data association, the process of matching the tracking target with the trajectory of prediction to get the correct track of the target, becomes more complex when multiple people need to be tracked in crowded environments, and this tends to decrease the efficiency of association algorithms [1]. These challenges make it difficult to implement person tracking algorithms in real-time robotic applications, as accurate methods often lack time efficiency, and efficient methods lack accuracy. Additionally, the time and space complexity of a model is even more critical when applied to mobile robots, as they do not have the same storage and computation capabilities as self-driving cars.

This research aims to advance the existing method of multi-object tracking (MOT) by utilizing knowledge distillation to reduce computation costs while maintaining the high performance of the original model. The state-of-the-art MOTR (Multiple-Object Tracking Transformer) model [12], a transformer-based MOT model, will be leveraged as the foundation. It has demonstrated better performance when compared to other transformer-based models such as TrackFormer[3] and TransTrack[4]. The initial objective is to reproduce the results of the MOTR model on general object tracking datasets, and then apply it to person tracking in mobile robot environments. Finally, student-teacher distillation will be applied in the MOT process. The small, lightweight “student” model is trained to mimic the output of a larger “teacher” model, with the goal of achieving similar performance while having a smaller model size and lower computational requirements. The types of networks for both the student and teacher will be investigated during the design phase. The same dataset will then be used to train and evaluate the distilled model, allowing for a comparison between the distilled and non-distilled models.

The contribution of this thesis will include: 1) proposing a multi-object tracking (MOT) model with knowledge distillation (KD), 2) evaluating the performance of the distilled and non-distilled models, and 3) implementing the model in real-time simulations to verify its capabilities. This could help to address the compatibility issues that existing MOT methods often face when they are implemented in real-world applications.

2. Literature Review

The literature review will first discuss two types of MOT: tracking-by-detection and joint detection and tracking, and their state-of-the-art algorithms. Then, usage of person tracking and knowledge distillation in robotics will be discussed.

2.1 Multi-Object Tracking

Multiple object tracking (MOT) is a continuously researched topic and has been applied to a wide range of tasks in robotics, such as navigation and self-driving of assistive mobile robots. It is typically divided into three main steps: detection, association, and state estimation. Object detection is used to identify multiple objects in each frame, and detections from different frames are compared to determine which detections correspond to which tracks. The tracker then uses the target objects to predict the position of the object in the next frame [5].

MOT is classified into two categories: tracking-by-detection (TBD) and joint detection and tracking (JDT). TBD trains object detection and re-identification (Re-ID) models in two separate systems. Many recent works are based on the SORT[6], DeepSORT[7], and JDE[8] approaches because they promise high accuracy. However, these SORT-like algorithms have difficulty maintaining correct identities over time (IDF1) while still achieving high accuracy. BoT-SORT[9], a state-of-the-art TBD approach, addresses this issue by integrating the Hungarian algorithm into ByteTrack[13], allowing it to handle a large number of objects and detections while still maintaining reasonable computational cost.

However, BoT-SORT still has the fundamental issue of TBD: computation cost. Calculating the global motion of the camera can be time-consuming when large images need to be processed[9]. Separated appearance trackers have relatively lower running speed than joint trackers. Another common limitation among TBD algorithms is that pretrained object detectors may limit the potential classes that the algorithm can detect and track, and it is costly to compute both the detector and identity model during inference[14].

Joint detection and tracking (end-to-end) methods combine detection and feature extraction into a single multi-task network. This can be achieved by adding identity embedding heads on top of existing detection networks such as MOTS[10] and RetinaTrack[11], or using tracking-by-attention, as introduced by TrackFormer. The performance of JDT is generally a bit more limited compared to TBD because it requires training on a larger network with fewer

training samples. However, this method is more preferred for real-time robotic applications because the algorithm has a shorter inference time as it detects and extracts features with a single, backbone-shared multi-task network, avoiding re-computation[27].

An attempt to improve on the existing state-of-the-art TBD method JDE (joint detection and embedding) was proposed in [14]. The author proposes a single network that simultaneously outputs detection results and the corresponding appearance embeddings of the detected boxes, thus increasing efficiency. This method successfully obtained near real-time performance (20.2 FPS runtime on Nvidia Titan xp) while maintaining the same level of accuracy as the original TBD method. However, one limitation of JDE is that it has more ID switches than other methods when multiple people pass by the robot with large overlap, resulting in a lower IDF1 score [14].

2.2 Person tracking in robotics

Mobile robots are employed in person tracking to locate and follow individuals in real-time. Some of the ways that robots are applied include surveillance, human-robot interaction, assistive robotics, and search and rescue. In order to track people, robots utilize a combination of sensors such as cameras, LIDAR, and ultrasonic sensors to detect and locate individuals[18]. Similar to MOT, person tracking methods can also be classified into two-step and end-to-end approaches. The first two-step approach was proposed in [15], where pedestrians are first detected from a scene image and then processed by a Re-ID network, but this method can be slow. On the other hand, end-to-end person tracking methods prioritize improving feature discrimination. The first end-to-end model with Faster R-CNN was introduced in [16], where the Re-ID is directly connected to Faster R-CNN with shared base layers. However, state-of-the-art search methods are not suitable for large-scale video monitoring scenarios where the robot needs to due to computational inefficiency. Therefore, research is being conducted to use knowledge distillation to simplify the process and reduce computation cost to make it more suitable for real-time applications.

2.3 Knowledge Distillation

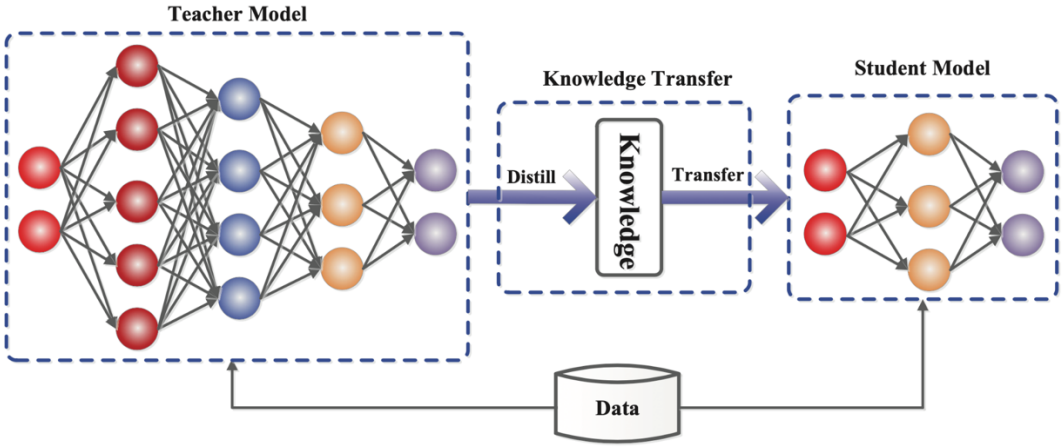

Figure 1. Diagram of Knowledge Distillation [32]

Knowledge distillation (KD) is a form of model compression that was first successfully demonstrated by [17]. It is widely used in neural networks to transfer the knowledge learned by a large set of models to a single smaller model for deployment under real-world constraints. Its properties allow for the deployment of large deep neural network models on edge devices with limited memory and computational capacity. Although various forms of knowledge are defined based on the purpose, one common characteristic of KD is symbolized by the student-teacher framework: the teacher provides knowledge and the student model learns it[24]. The softened probabilities outputted by a pre-trained teacher model contains more information than hard labels (e.g. one-hot encoding for class label) and provides better results when combined with softened and hard labels [19]. In MOT, a pre-trained Re-ID model serves as the teacher and the identity embedding branch acts as the student.

KD can be applied in MOT to reduce the computation cost while maintaining the high performance of the original model. The proposed MOT model with KD can be implemented in real-time simulations to verify its capabilities and address the compatibility issues that existing MOT methods often face in real-world applications.

To date, there has not been any extensive research on the application of KD to the MOT problem for mobile robots. An end-to-end KD framework that does not require ID annotation or sequential connection in training by [19]. It concludes that combining with hard labels does not significantly improve the performance but worsens MOTA (Multi Object Tracking Accuracy) because soften labels already contain sufficient information to teach the student model.

Therefore, ID annotations are not required in training, which allows the model to be trained on static autonomous images.

2.4 Transformer-based MOT method

Many state-of-the-art models use CNNs for object detection and re-identification because it ensures high accuracy, but these methods are computationally intensive and do not scale well to large datasets. In contrast, transformer-based models have shown remarkable success in natural language processing (NLP) and image recognition tasks. They have also been applied to computer vision such as in image segmentation and object detection. The transformer architecture capture long-range dependencies and sequential information using self-attention mechanism. It is particularly well-suited for real-time MOT since it is highly parallelizable and the size is generally lighter than CNN-based models. Recently, MOT systems were based around transformers to better exploit the spatial-temporal relations between the adjacent frames. This includes information about the object's appearance previous extracted being useful in later tracking inference. Two transformer-based MOT systems will be explored: TrackFormer[3] and MOTR[12].

2.4.1 TrackFormer

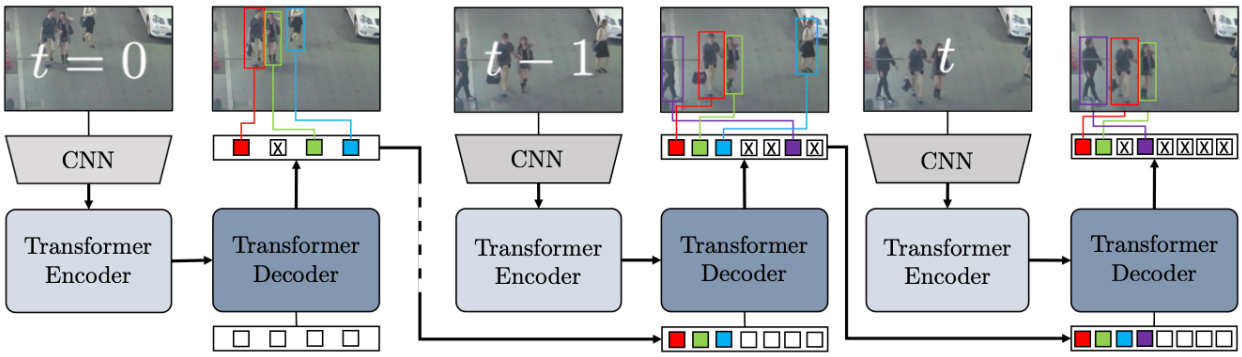

Figure 2. Workflow of TrackFormer

TrackFormer is an end-to-end trainable encoder-decoder Transformer architecture which introduces tracking-by-attention. The video frames are processed by a common CNN backbone (e.g. ResNet50) to extract features for each object in each frame. These features are then fed into a transformer network, which produces a set of object embeddings that capture the spatiotemporal relationships between the objects.

The transformer decoder is composed of several layers of self-attention mechanisms, where each layer consists of a multi-head self-attention and a feedforward neural network. In the self-attention, the visual features are transformed into query, which is used to compute the object embeddings that capture the spatiotemporal relationships between the objects in the frame.
In addition to the self-attention, the Transformer decoder network also utilizes encoder-decoder attention to incorporate information from previous frames. In the encoder-decoder attention, the decoder input queries are compared to the encoder outputs from all previous frames, and the attention map is used to weight the encoder outputs based on their similarity to the queries. This allows the decoder to attend to relevant information from previous frames when producing the object embeddings for the current frame.

The combination of self-attention and encoder-decoder attention allows the Transformer decoder network to model the spatiotemporal relationships between the objects in a video sequence. By attending to relevant information in previous frames, the decoder can produce object embeddings that capture the object interactions and movements over time, which is crucial for accurate object tracking.

One of the key advantages of TrackFormer is its ability to handle occlusion and other challenging scenarios. Self-attention allows the network to learn to attend to relevant objects and ignore irrelevant ones, even in crowded scenes Also, the use of object embeddings allows for more accurate track association, as the embeddings capture more information about the objects than traditional methods.

2.4.2 MOTR

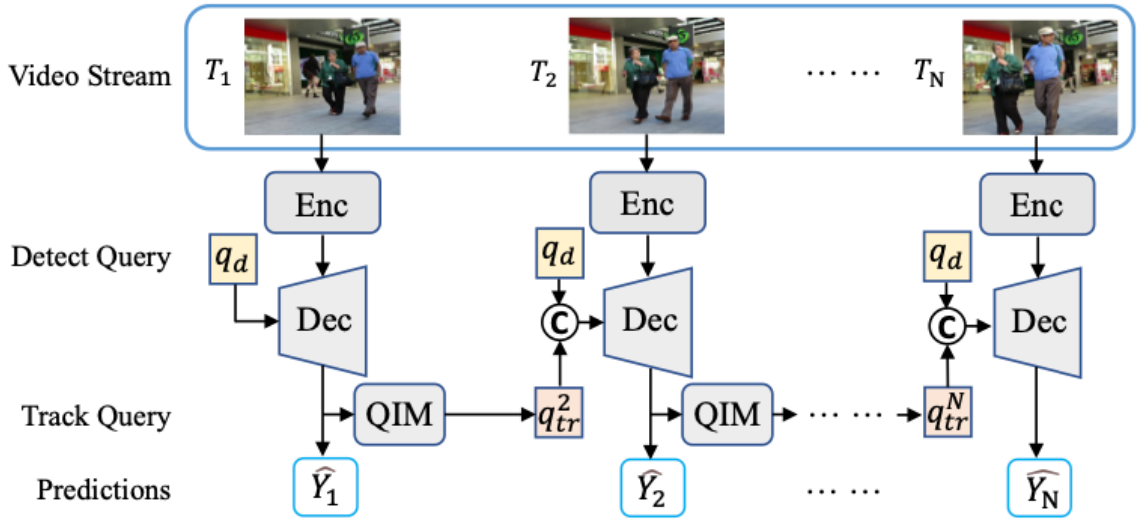

Figure 3. Workflow of MOTR

MOTR is a fully end-to-end MOT framework that can implicitly learn the appearance and position variances in a joint manner. Similar to TrackFormer, the model also consists of a backbone network, and transformer-based encoder-decoder.

The backbone network is a CNN based on the ResNet50 that extracts visual features from the input frames. The encoder is a series of transformer layers that process the visual features extracted by the backbone network. The decoder is responsible for making predictions about the location and trajectory of each object, as well as associating detections with the correct object trajectory using a set of learned attention weights.

While both MOTR and TrackFormer share similar architecture, MOTR achieves better performance on several benchmark datasets compared to TrackFormer. For example, on the MOT17 dataset, MOTR achieves MOTA of 61.2% compared to TrackFormer's MOTA of 59.8%. On the MOT20 dataset, MOTR achieves MOTR of 60.6% compared to TrackFormer's MOTA of 58.8%, proving that MOTR is more effective approach for MOT [3][12].

2.5 Evaluation Metrics

Two sets of metrics are typically used to evaluate the performance of MOT algorithms: the CLEAR metrics [29] and the VACE metrics [30]. The CLEAR metrics aim to measure the overall performance for all predicted trajectories and include metrics such as MOTA (multi-object tracking accuracy) and MOTP (multi-object tracking precision). On the other hand, the

VACE metrics are used to describe individual metrics from different aspects, such as FP (false positives), FN (false negatives), FAF (false alarm per frame), MT (mostly tracked), ML (mostly lost), IDS (number of ID switches) and Frag (number of fragments). Among all these metrics, MOTA is considered to be the most reliable metric for evaluating MOT performance compared to the others [30], which is defined as the following:

$$\mathrm{MOTA} = 1 - \frac{\sum \mathrm{FN+FP+IDSW}}{\sum \mathrm{GT}} \text{ [29]}$$

IDF1, or Identification F1, is a commonly used metric for evaluating the accuracy of tracking identification in multi-object tracking. Unlike MOTA, which focuses on the accuracy of object detection and association, IDF1 emphasizes the accuracy of track association. It accomplishes this by using a bijective mapping between predicted trajectories and ground truth trajectories to determine which trajectories are present.

IDF1 is calculated by combining two other metrics: IDP (ID Precision) and IDR (ID Recall). IDP measures the percentage of correct track associations out of all track associations made by the algorithm, while IDR measures the percentage of correct track associations out of all possible tracks [1].

$$\mathrm{IDR} = \frac{\mathrm{IDTP}}{\mathrm{IDTP} + \mathrm{IDFN}}$$

$$\mathrm{IDP} = \frac{\mathrm{IDTP}}{\mathrm{IDTP} + \mathrm{IDFP}}$$

$$\mathrm{IDF1} = \frac{\mathrm{IDTP}}{\mathrm{IDTP} + 0.5\,\mathrm{IDFN} + 0.5\,\mathrm{IDFP}}$$

IDTP: Identity True Positives, IDFN: Identity False Negatives, IDFP: Identity False Positives [1]

A high IDF1 score indicates that the algorithm can accurately identify and track objects over time, while a low IDF1 score suggests that the algorithm is struggling to make correct track associations. This is often used alongside other metrics such as MOTA and MOTP to assess the overall performance of a tracking algorithm.

3. Methods

3.1 Datasets

The ideal person detection dataset for this research must be collected indoors using a moving camera to fit the scope of the research. A moving camera is preferred over a static camera, as the research involves a mobile robot which is not static. A smaller dataset is preferred as the mobile robot has limited storage capacity and therefore cannot train on a large set. Three datasets were examined: the RGB-D Person dataset, Mobility Aids dataset, and InOutDoorPeople dataset.

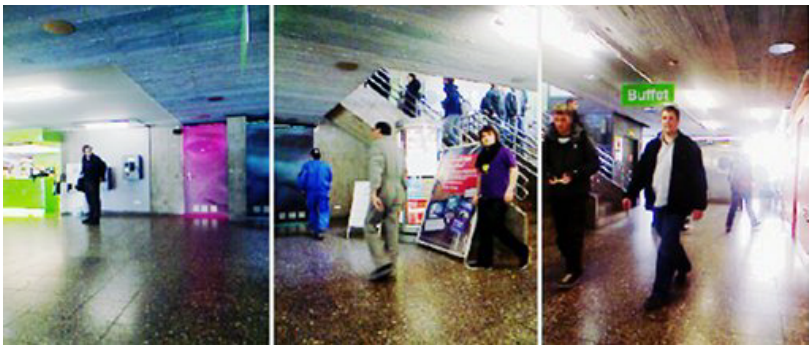

Figure 4. Sample images of RGB-D Person Dataset

The RGB-D Person dataset[20] contains 3000 RGB-D images of walking and standing people acquired in a university hall from three static Kinect sensors. Despite the size of the dataset being the smallest of the three, it was not chosen because the data contains different levels of occlusions and since three cameras were used to take image of the same scene, tracking people across the three views will be needed, making it difficult to ensure higher accuracy. These were beyond the scope of this thesis, and therefore, this data is inappropriate and may affect the performance of MOT.

Figure 5. Sample images of Mobility Aids Dataset

The MobilityAids dataset[21] contains over 17,000 RGB-D images collected in a hospital and facilities of the University of Freiburg, using a dynamic mobile robot equipped with a Kinect sensor. The purpose of the dataset was to recognize people with different mobility aids, it categorizes people according to the mobility aids they use such as people in wheelchairs, people with clutches, and people using walking frames. This dataset was not chosen because the aids can affect the detection since they are blocking the people, and multi-class detection is not the scope of this thesis.

Figure 6. Sample images of InOutDoorPeople Dataset

Finally, the InOutDoorPeople dataset[22] was collected by a moving mobile robot as it transitioned from indoor to outdoor environments in a single take. This dataset was chosen to evaluate the MOT model even though the transition of environments is not required, as the size was reasonable to be trained on a mobile robot, and most importantly, the images were collected by a moving camera which is a key component of this thesis.

3.2 Data Preparation

In order to prepare the InOutDoorPeople dataset for training the model, several pre-processing steps were necessary to ensure that the data was in the appropriate format. The process was modified to fit the custom dataset, following MOTR GitHub repository [14].

Firstly, both the images and annotation files were reorganized into four distinct sequences, following the order specified in the sequence list. Then, empty frames with no people or annotations were removed to avoid high false positives. These sequences were then renamed in ascending order of integers, with sequence 0 through 2 designated as the train set, and sequence 3 as the test set.

Next, the annotations were processed in accordance with the desired format for the model, which followed the MOT Challenge's submission guidelines. This involved saving the annotations in a file named "gt.txt" for each sequence, which required some degree of automation to generate. However, a manual check was necessary to ensure that the IDs were correctly assigned to each person. Label files were then generated using this ground truth information.

Finally, the "labels_with_ids" directory was created, following the FairMOT style annotation. The directory consists of text files for each image frame with information on the size of bounding boxes. The file names match the image names so that the model can detect the correct people in the frame.

3.3 Choosing state-of-the-art MOT Model

There are a few MOT models that are transformer-based: TrackFormer, TransTrack, MOTR, and TraDeS. Of these, a state-of-the-art model was selected based on the performance, size, and whether the model uses JDT or not. The models were first sorted by which model is the most recent, and then comparing the performance on MOT17 [25], which was a common benchmark dataset.

TrackFormer was the focus of the initial research. The model's performance was first reproduced to ensure accurate implementation. However, while reproducing the model's training to ensure accurate implementation, I decided to stop using this model because the codebase was hard to work with due to lack of documentations. Then, I moved forward to MOTR which was the latest transformer-based MOT model that has been published.

3.4 Training TrackFormer and MOTR.

TrackFormer's training result was reproduced on MOT17 using the python implementation provided by the author. The model was first pre-trained with the CrowdHuman dataset [26] using the provided checkpoint. Some modifications were made to the training process due to computation cost and compatibility issue with the GPU. The feature extraction model was changed from ResNet50 to ResNet18, and the number of epochs was also reduced from 50 to 15.

MOTR was evaluated also in the same process as TrackFormer. Instead of MOT17, MOT15 [25] train set was used because MOT17 does not have test set available for public use. After successfully reproducing the performance, the model was trained on the InOutDoorPeople dataset. Here, training hyperparameters were adjusted to cope with number of available GPUs.

3.4 Implementing Knowledge Distillation to MOTR

KD can be applied to several steps of MOT including object detection, tracking, and feature extraction. The teacher model can be a complex model that is trained on a large dataset. The student model can be then trained to mimic the teacher's behaviours, but with fewer parameters and faster inference times. Of these, I decided to apply KD to feature extraction because it tends to be more computationally expensive and time-consuming compared to the other two processes. Feature extraction requires analyzing large amounts of data to extract relevant features for subsequent analysis and becomes more complex when working with high-resolution images. Therefore, it is possible to reduce the computational resources required for the task by using a smaller and distilled model, while maintaining the accuracy of the features. On the other hand, distillation may not be as effective for object detection because modifying these may lead to changing the main features of the model, which is undesirable.

A simple student-teacher distillation was implemented. To create a smaller model of MOTR, the backbone network of feature extraction was changed from ResNet50 to ResNet18. As the name implies, ResNet18 has 18 layers while ResNet50 has 50 layers, making ResNet18 a shallower and less deep network compared to ResNet50. Then, the student model was trained with the pre-trained model from teacher network.

4. Results and Discussion

4.1 TrackFormer on MOT17

The table shows the results of experiments to reproduce TrackFormer on MOT17. Full training results of MOT17 is shown in Appendix. The same tuning parameters were used for both experiments.

MOT17	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	ID SW↓
Original	74.2%	71.7%	849	177	7431	78057	1449
Obtained	54.6%	58.0%	445	535	4709	147126	1258

Table 1. Comparison of training results

The overall performance was poorer than the original work. High FN represents the people present in the frame were not identified correctly, which led to low MOTA and IDF1.

This was expected since these results were obtained under different constraints. The original work had stronger computational power (7 x32GB GPUs vs. 1 x 8GB GPU) and our model was trained with lesser epochs which significantly reduced the performance. Since the purpose of this experiment was to investigate the capability of the model, further fine-tuning was not conducted.

To visualize the training results the demo was also reproduced with a provided video sequence. The MP4 file was first converted into image frames and evaluated on the same model.

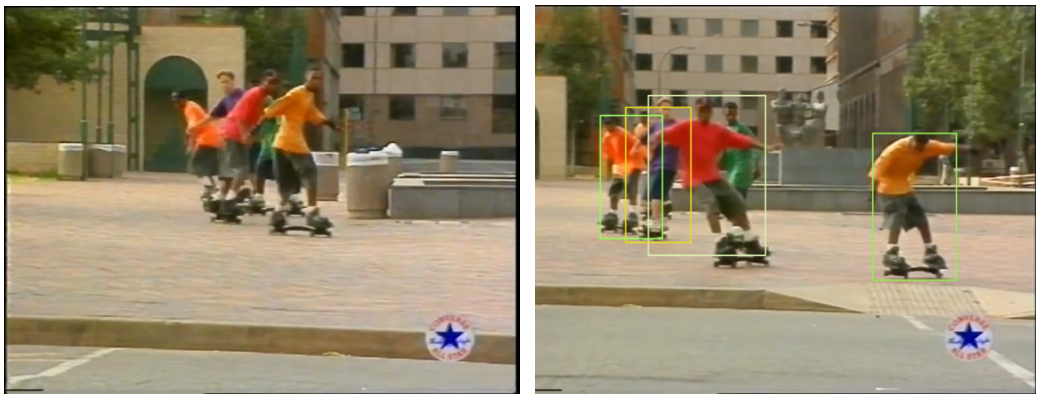

Figure 7. Visualization of TrackFormer performance

Regardless of the low accuracy, bounding box was assigned for each person and followed well even when occlusion occurs.

4.2 MOTR on MOT15

	IDF1	IDP	IDR	Rcll	Prcn	GT	MT	PT	ML	FP	FN	IDs	FM	MOTA	MOTP	IDt	IDa	IDm
ADL-Rundle-6	88.8%	85.7%	92.1%	96.8%	90.0%	24	23	1	0	540	162	4	7	85.9%	0.154	3	3	2
ETH-Bahnhof	56.8%	54.5%	59.2%	66.5%	61.1%	223	104	47	72	3240	2571	27	153	23.9%	0.184	47	15	38
KITTI-13	51.2%	39.6%	72.3%	73.4%	40.2%	42	23	19	0	1013	247	4	6	-36.1%	0.191	3	3	2
PETS09-S2L1	77.3%	76.7%	77.9%	93.9%	92.5%	19	17	2	0	356	283	4	52	86.2%	0.268	3	2	1
TUD-Stadtmitte	79.2%	75.5%	83.3%	92.0%	83.4%	10	10	0	0	212	93	6	7	73.1%	0.289	2	4	1
ADL-Rundle-8	65.1%	49.8%	93.9%	94.2%	50.0%	28	25	3	0	6386	394	8	97	-0.1%	0.226	4	3	2
KITTI-17	50.2%	39.1%	70.2%	78.4%	43.7%	9	5	4	0	791	169	13	21	-24.4%	0.215	5	7	1
ETH-Pedcross2	86.3%	88.0%	84.7%	86.4%	89.8%	151	108	12	31	664	920	12	34	76.4%	0.163	10	6	6
ETH-Sunnyday	78.0%	68.7%	90.3%	92.1%	70.1%	30	25	3	2	746	150	5	17	52.6%	0.160	1	4	1
TUD-Campus	66.0%	53.3%	86.6%	96.7%	59.4%	8	8	0	0	237	12	6	6	29.0%	0.280	4	2	1
Venice-2	55.0%	38.1%	98.3%	98.4%	38.2%	26	25	1	0	11378	111	1	8	-60.9%	0.231	0	1	0
OVERALL	67.7%	56.8%	83.7%	88.1%	59.8%	570	373	92	105	25563	5112	90	408	28.7%	0.206	82	50	55

Table 2. Results of MOTR on MOT15

The table shows the training result on MOT15 train set using the pretrained model. The author did not provide their result on this dataset so direct comparison could not be done. However, low MOTA was expected when comparing the results obtained by other users [28]. The result looks promising regardless of the accuracy since the overall recall rate is very high (88.1%), which implies the model can detect most of the people in the frame.

Below. demo was also regenerated with the same pretrained model which shows bounding boxes and its accuracies even when the people are concentrated at one place (top-left). The model can also handle occlusion well because the bounding boxes recover with very high accuracy as soon as the two people walk by.

Figure 8. Bounding boxes tracked by MOTR on the demo sequence

4.3 MOTR on InOutDoorPeople Dataset

Two tests were conducted: first was to evaluate InOutDoorPeople dataset on MOTR, and second was to train and evaluate on the custom dataset.

	IDF1	IDP	IDR	Rcll	Prcn	GT	MT	PT	ML	FP	FN	IDs	FM	MOTA	MOTP	IDt	IDa	IDm
seq0	59.6%	46.4%	83.5%	95.1%	52.8%	14	14	0	0	3322	193	19	25	9.7%	0.214	12	11	5
seq1	66.5%	57.0%	79.9%	97.1%	69.2%	11	11	0	0	2950	197	11	32	53.8%	0.217	2	8	0
seq2	81.9%	73.0%	93.2%	99.7%	78.1%	14	14	0	0	928	11	6	1	71.5%	0.204	3	2	0
seq3	48.0%	40.3%	59.4%	93.2%	63.2%	19	18	1	0	1404	175	25	11	37.9%	0.213	15	16	6
OVERALL	64.6%	54.1%	80.2%	96.5%	65.1%	58	57	1	0	8604	576	61	69	44.5%	0.213	32	37	11

Table 3. Pretrained model evaluated on InOutDoorPeople dataset

	IDF1	IDP	IDR	Rcll	Prcn	GT	MT	PT	ML	FP	FN	IDs	FM	MOTA	MOTP	IDt	IDa	IDm
seq0	76.8%	74.0%	79.8%	97.8%	90.8%	14	13	1	0	387	85	22	16	87.4%	0.106	11	14	4
seq1	68.2%	63.8%	73.3%	98.5%	85.8%	11	10	0	1	1115	102	20	17	81.9%	0.104	2	19	1
seq2	71.1%	64.3%	79.6%	98.4%	79.4%	14	14	0	0	844	53	13	11	72.5%	0.096	1	12	0
seq3	54.9%	49.4%	61.8%	91.4%	73.2%	19	17	2	0	867	221	22	21	57.0%	0.171	7	18	3
OVERALL	68.6%	63.7%	74.3%	97.2%	83.4%	58	54	3	1	3213	461	77	65	77.5%	0.112	21	63	8

Table 4. MOTR trained and evaluated on InOutDoorPeople dataset

Sequence 0 to 2 were used as train sets and sequence 3 was used as a test set. Training MOTR on InOutDoorPeople dataset took approximately 4 days to train 100 epochs using one or two Nvidia GeForce RTX 3070 GPUs depending on the availability.

The InOutDoorPeople dataset is a new dataset that the author's model has not been trained on. Therefore, it is expected that the model's performance on this dataset will be lower than its performance on the MOT17 and CrowdHuman datasets that it was trained on.
The high overall recall rate in both results indicates that the model is good at detecting people in the images. However, it is important to note that the author's model was trained n images collected by static cameras, which means it is more prone to tracking people at a stationary frame since these cameras capture images of people from a fixed position. This makes it easier for the model to detect and track people who are not moving or moving slowly.

As shown in the above tables, the visualization of the results shows that the model sometimes fails to detect people when their full body is not present in the image or when they are standing at the edge of the frame, leading to high FP. This could be due to the dataset did not have enough images that capture such scenarios. Therefore, the model may not have learned to detect people in these situations as well as it could have. Further optimization and training it on more diverse datasets could improve its performance.

High MOTA scores in Table 4 can be further indicative of the effectiveness of MOTR in tracking people. The relatively low FP and FN rates suggest that the model can accurately detect

people in the images, and it can re-track people successfully when occlusion occurs. This is important because occlusions are common in crowded scenes, where multiple people are moving near each other and can cause problems for tracking algorithms. High IDP and IDR indicate that the model accurately associated person detections over time and tracked all possible people in the scene.

Below are some visualizations of tracking results with bounding boxes.

Figure 9. Top: Pretrained model Bottom: trained on InOutDoorPeople dataset.

Since the top model was trained on more data, it can detect people better although their full bodies are not present.

4.4 Implementing Knowledge Distillation on MOTR

	IDF1	IDP	IDR	Rcll	Prcn	GT	MT	PT	ML	FP	FN	IDs	FM	MOTA	MOTP	IDt	IDa	IDm
seq0	67.1%	69.8%	64.5%	88.0%	95.4%	14	9	4	1	167	471	40	39	82.7%	0.126	4	37	1
seq1	43.2%	37.7%	50.6%	83.5%	62.2%	11	5	5	1	3468	1127	37	40	32.2%	0.125	1	36	0
seq2	51.2%	53.6%	49.0%	84.3%	92.3%	14	9	5	0	232	520	28	33	76.4%	0.115	2	26	0
seq3	31.1%	49.9%	22.6%	39.4%	87.0%	19	1	8	10	152	1567	53	68	31.4%	0.206	2	52	1
OVERALL	48.7%	48.2%	49.2%	77.9%	76.3%	58	24	22	12	4019	3685	158	180	52.8%	0.129	9	151	2

Table 5. Distilled MOTR evaluated after 100 epochs

The non-distilled model (teacher) from section 4.3 was used as the pretrained model for the distilled model (student). The training process for the distilled model took approximately 2 days to train 100 epochs, using two Nvidia GeForce RTX 3070 GPUs.

Although the overall performance of the distilled model was slightly poorer than that of the non-distilled model, some sequences preserved its performance when comparing sequence by sequence with Table 4. Sequence 0 and 2 maintained almost the same level of tracking accuracy as the teacher model. The student model identified and tracked correct people in the scene, which led to lesser FP and FN, as indicated by high IDR and IDP. However, for Sequence 1, there was a delay in detecting and following a newly appeared person, resulting in high false positive and false negative detections, depicted in Figure 10. It is possible that the visual characteristics of Sequence 1 were different from other sequences since more occlusions and a person's full body was not present all the time as mentioned in Section 4.3.

Figure 10. Detection results of teacher model (left) and student model (right) at frame 607 of Sequence 1

Further investigation is needed to understand the poor performance of the distilled model on the test set, particularly in Sequence 3 where detection failed even in well-lit environments with full body presence. One potential solution is to use a larger dataset for the teacher model since the current limited number of samples may not be representative of the diverse range of scenes that the model could encounter. By training the teacher model on a larger dataset, it could learn a wider range of appearance variations, which can in turn improve the distilled model's

ability to detect and track people in new and unfamiliar scenes. Detection results for the test set are compared in Figures 11 and 12.

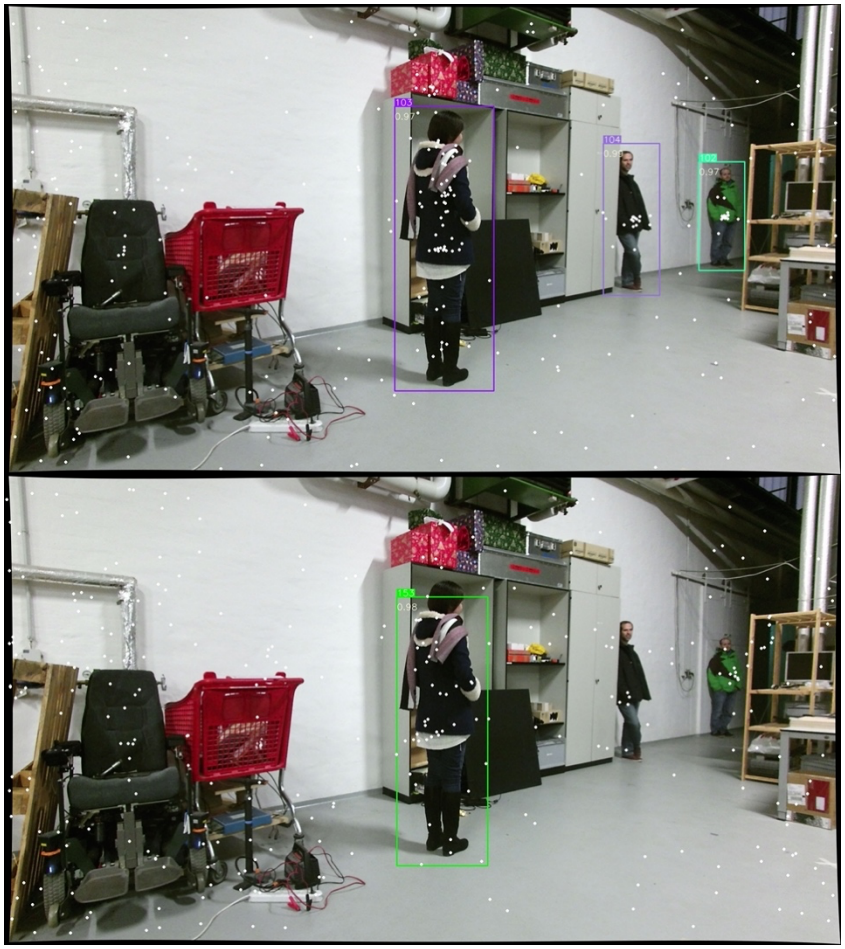

Figure 11. Detection results of teacher model (top) and student model (bottom) at frame 0 (left) of Sequence 3

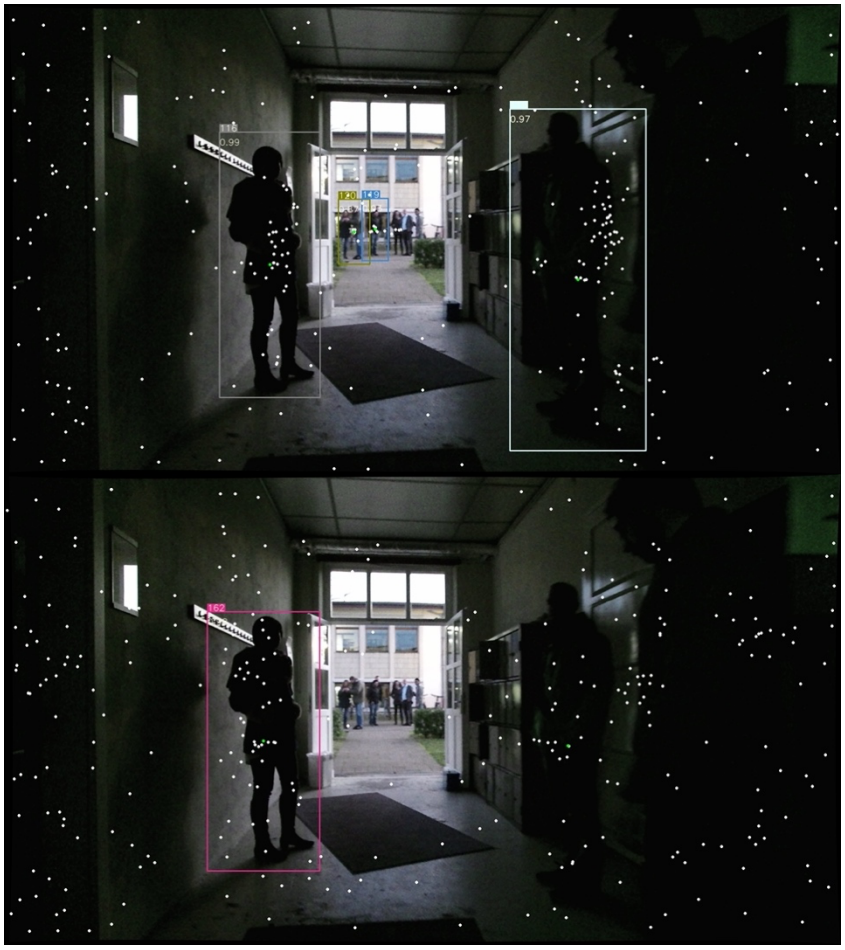

Figure 12. Detection results of teacher model (top) and student model (bottom) at frame 357 of Sequence 3

In Figure 11, only one bounding box was detected despite three people with full bodies being present in the frame. In Figure 12, the student model only detected 2 people while the teacher model detected 4 people. This discrepancy may be due to a change in illumination as the camera moved from indoors to outdoors, and the student model may not have been trained extensively enough to detect people at greater distances.

4.5 Key Claims

While the distilled model did not perform as well as the on-distilled model overall, it is important to note that the goal of distillation is not necessarily to surpass the original model's performance, but rather to create a smaller and more efficient model that can be deployed in resource-constrained settings. Therefore, the slight drop in performance may be an acceptable trade-off for the benefits of model compression.

5. Conclusion

The thesis successfully demonstrated that distillation in multi-person tracking can help create smaller and more efficient models while still preserving the main performance metrics such as MOTA. These findings are particularly important for motile robot applications, where lightweight and computationally efficient models are often required to operate in real-time with limited resources.

However, the challenges in tracking newly appeared people in crowds suggest that further analysis and improvements are needed to address these issues. To improve the performance of the distilled models, future work could involve fine-tuning the training parameters and using larger dataset to train the teacher model. Additionally, further experiments with different distillation methods are required to find the best approach for creating efficient and effective multi-person tracking models for mobile robots.

To validate the performance of these models, it is necessary to test them in real-time scenarios, such as on mobile robot platforms operating in real-world environments. Further research could also explore the potential for integrating multi-person tracking with other applications such as human-robot interaction and social robotics.

References

[1]J. Abawajy, K.-K. R. Choo, R. Islam, Z. Xu, and M. Atiquzzaman, "A survey of multi-object video tracking algorithms," in International conference on applications and techniques in cyber security and intelligence ATCI 2018 applications and techniques in cyber security and Intelligence, Cham: Springer International Publishing, 2019, p. 351.

[2] I. Ahmed, S. Din, G. Jeon, F. Piccialli, and G. Fortino, "Towards Collaborative Robotics in Top View Surveillance: A Framework for Multiple Object Tracking by Detection Using Deep Learning," IEEE/CAA Journal of Automatica Sinica, vol. 8, no. 7, pp. 1253–1270, Jul. 2021, doi: 10.1109/jas.2020.1003453.

[3] T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer, "Trackformer: Multi-object tracking with Transformers," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[4] P. Sun, Y. Jiang, R. Zhang, E. Xie, J. Cao, X. Hu, T. Kong, Z. Yuan, C. Wang, and P. Luo, "Transtrack: Multiple-object tracking with trans-former," CoRR, vol. abs/2012.15460, 2020.

[5] I. Ahmed, S. Din, G. Jeon, F. Piccialli, and G. Fortino, "Towards Collaborative Robotics in Top View Surveillance: A Framework for Multiple Object Tracking by Detection Using Deep Learning," IEEE/CAA Journal of Automatica Sinica, vol. 8, no. 7, pp. 1253–1270, Jul. 2021, doi: 10.1109/jas.2020.1003453.

[6] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple Online and Realtime Tracking," in2016 IEEE International Conference on Image Processing (ICIP), Sep. 2016, pp. 3464–3468. doi: 10.1109/ICIP.2016.7533003.

[7] N. Wojke, A. Bewley, and D. Paulus, "Simple Online and Realtime Tracking with a Deep Association Metric." arXiv, Mar. 21, 2017. Accessed: Oct. 11, 2022. [Online]. Available: http://arxiv.org/abs/1703.07402

[8] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, "Towards Real-Time Multi-Object Tracking," in Computer Vision – ECCV 2020, vol. 12356, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 107–122. doi: 10.1007/978-3-030-58621-8_7.

[9] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, "BoT-SORT: Robust Associations Multi-Pedestrian Tracking." [Online]. Available: https://arxiv.org/pdf/2206.14651.pdf

[10] P. Voigtlaender et al., “MOTS: Multi-Object Tracking and Segmentation.” arXiv, Apr. 08, 2019.Accessed: Oct. 11, 2022. [Online]. Available: http://arxiv.org/abs/1902.03604

[11] Z. Lu, V. Rathod, R. Votel, and J. Huang, “RetinaTrack: Online Single Stage Joint Detection and Tracking.” arXiv, Mar. 30, 2020. Accessed: Oct. 11, 2022. [Online]. Available: http://arxiv.org/abs/2003.13870

[12]F. Zeng, D. 1⋆, Y. Zhang, T. Wang, X. Zhang, and Y. Wei, “MOTR: End-to-End Multiple-Object Tracking with Transformer.” Accessed: Jan. 28, 2023. [Online]. Available: https://arxiv.org/pdf/2105.03247.pdf

[13] Y. Zhang, P. Sun, Y. Jiang, D. Yu, Z. Yuan, P. Luo, W. Liu, and X. Wang. Bytetrack: Multi-object tracking by associat- ing every detection box. arXiv preprint arXiv:2110.06864, 2021.

[14]Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, “Towards Real-Time Multi-Object Tracking,” Computer Vision – ECCV 2020, pp. 107–122, 2020, doi: 10.1007/978-3-030-58621-8_7.

[15] Y. Xu, B. Ma, R. Huang, and L. Lin, “Person Search in a Scene by Jointly Modeling People Commonness and Person Uniqueness,” Proceedings of the 22nd ACM international conference on Multimedia, Nov. 2014, doi: 10.1145/2647868.2654965.

[16]T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, “Joint Detection and Identification Feature Learning for Person Search,” arXiv:1604.01850 [cs], Apr. 2017, Accessed: Jan. 28, 2023. [Online]. Available: https://arxiv.org/abs/1604.01850

[17] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, “Model compression,” Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06,2006.

[18]C. Schlegel, J. Illmann, H. Jaberg, M. Schuster, and R. Wörz, “Vision Based Person Tracking with a Mobile Robot £.” Accessed: Jan. 30, 2023. [Online]. Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=19562fbef0e0593e3fa7267ae55e1d7be3348dfe

[19]W. Zhang et al., “Boosting End-to-end Multi-Object Tracking and Person Search via Knowledge Distillation,” Proceedings of the 29th ACM International Conference on Multimedia, Oct. 2021, doi: 10.1145/3474085.3481546.

[20]O. Mees and A. Eitel, “Choosing Smartly: Adaptive Multimodal Fusion for Object Detection in Changing Environments.” Accessed: Jan. 28, 2023. [Online]. Available: http://ais.informatik.uni-freiburg.de/publications/papers/mees16iros.pdf

[21]M. Kollmitz, A. Eitel, A. Vasquez, and W. Burgard, “Deep 3D perception of people and their mobility aids,” Robotics and Autonomous Systems, vol. 114, pp. 29–40, Apr. 2019, doi: 10.1016/j.robot.2019.01.011.

[22]O. Mees and A. Eitel, “Choosing Smartly: Adaptive Multimodal Fusion for Object Detection in Changing Environments.” Accessed: Jan. 28, 2023. [Online]. Available: http://ais.informatik.uni-freiburg.de/publications/papers/mees16iros.pdf

[23]A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” CoRR, vol. abs/1706.03762, 2017.

[24]J. Gou, B. Yu, S. Maybank, and D. Tao, “Knowledge Distillation: A Survey,” Int. J. Comput. Vis., 2021, doi: 10.1007/s11263-021-01453-z.

[25]P. Dendorfer, A. Osep, A. Milan, K. Schindler, D. Cremers, I. D. Reid, S. Roth, and L. Leal-Taix´e, “Motchallenge: A benchmark for single-camera multiple target tracking,” CoRR, vol. abs/2010.07548, 2020.

[26]S. Shao et al., ‘CrowdHuman: A Benchmark for Detecting Human in a Crowd’, arXiv preprint arXiv:1805. 00123, 2018.

[27]S. K. Pal, A. Pramanik, J. Maiti, and P. Mitra, “Deep learning in multi-object detection and tracking: state of the art,” Applied Intelligence, vol. 51, no. 9, pp. 6400–6429, Apr. 2021, doi: 10.1007/s10489-021-02293-7.

[28]“Reproduced results on validation set MOT15 with the provided weights(motr_final_0915.pth), MOTA: 29.1% IDF1: 67.1%. · Issue #32 · megvii-research/MOTR,” GitHub. https://github.com/megvii-research/MOTR/issues/32 (accessed Jan. 30, 2023).

[29] Kasturi, R., Goldgof, D., Soundararajan, P., et al.: ‘Framework for performance evaluation of face, text, and vehicle detection and tracking in video: data, metrics, and protocol’, IEEE Trans. Pattern Anal. Mach. Intell., 2009, 31, (2), pp. 319–336

[30] Li, Y., Huang, C., Nevatia, R.: ‘Learning to associate: hybrid boosted multitarget tracker for crowded scene’. Proc. IEEE Conf. Computer Vision and Pattern Recognition, Miami, FL, USA, 2009, pp. 2953–2960

[31] Leal-Taixe, L., Milan, A., Schindler, K., et al.: ‘Tracking the trackers: an analysis of the state of the art in multiple object tracking’, arXiv preprint arXiv:1704.02781, 2017

[32] J. Gou, B. Yu, S. J. Maybank, and D. Tao, “Knowledge distillation: A survey,” arXiv.org, 20-May-2021. [Online]. Available: https://arxiv.org/abs/2006.05525. [Accessed: 27-Mar-2023].

Appendix A. Full result of TrackFormer evaluated on MOT17

	IDF1	IDP	IDR	Rcll	Prcn	GT	MT	PT	ML	FP	FN	IDs	FM	MOTA	MOTP	IDt	IDa	IDm
MOT17-02-DPM	36.0%	74.6%	23.8%	31.4%	98.6%	62	8	19	35	84	12746	46	54	30.7%	0.151	20	38	12
MOT17-02-FRCNN	36.5%	73.5%	24.3%	32.6%	98.5%	62	8	22	32	90	12530	58	67	31.8%	0.154	31	39	12
MOT17-02-SDP	38.1%	71.3%	26.0%	35.9%	98.3%	62	8	25	29	114	11919	65	82	34.9%	0.159	33	43	11
MOT17-04-DPM	64.2%	83.3%	52.2%	62.1%	99.2%	83	31	24	28	244	18014	22	42	61.6%	0.125	21	15	14
MOT17-04-FRCNN	67.1%	87.3%	54.4%	61.7%	98.9%	83	32	27	24	319	18230	19	39	61.0%	0.127	16	14	11
MOT17-04-SDP	71.5%	88.3%	60.1%	67.4%	99.1%	83	36	26	21	307	15483	24	42	66.7%	0.126	20	15	11
MOT17-05-DPM	59.3%	79.5%	47.3%	57.3%	96.4%	133	29	60	44	147	2952	69	75	54.2%	0.177	44	42	17
MOT17-05-FRCNN	59.4%	79.4%	47.4%	57.2%	95.8%	133	30	61	42	173	2962	74	83	53.6%	0.186	41	48	16
MOT17-05-SDP	62.2%	79.5%	51.1%	61.7%	96.1%	133	34	69	30	172	2648	91	100	57.9%	0.181	46	59	15
MOT17-09-DPM	55.5%	68.6%	46.6%	66.5%	97.8%	26	13	12	1	80	1786	36	33	64.3%	0.118	11	29	4
MOT17-09-FRCNN	56.1%	69.9%	46.9%	65.5%	97.8%	26	13	11	2	79	1835	34	31	63.4%	0.116	9	28	3
MOT17-09-SDP	55.8%	68.4%	47.1%	67.4%	97.8%	26	14	11	1	79	1738	36	34	65.2%	0.118	11	29	4
MOT17-10-DPM	46.9%	67.0%	36.1%	51.6%	95.7%	57	8	31	18	297	6217	64	123	48.8%	0.197	48	34	19
MOT17-10-FRCNN	47.9%	64.7%	38.0%	55.8%	94.9%	57	13	30	14	383	5678	75	137	52.2%	0.201	58	39	24
MOT17-10-SDP	49.1%	65.6%	39.2%	57.0%	95.3%	57	15	27	15	363	5524	71	139	53.6%	0.201	57	36	23
MOT17-11-DPM	57.6%	72.4%	47.8%	64.2%	97.2%	75	18	27	30	172	3381	19	31	62.1%	0.106	6	18	5
MOT17-11-FRCNN	62.3%	75.3%	53.1%	68.7%	97.4%	75	21	33	21	170	2951	23	40	66.7%	0.113	8	21	6
MOT17-11-SDP	63.2%	73.6%	55.3%	73.4%	97.6%	75	27	30	18	168	2512	32	51	71.3%	0.118	9	28	5
MOT17-13-DPM	48.6%	77.4%	35.4%	42.8%	93.5%	110	24	36	50	345	6663	128	156	38.7%	0.212	114	46	33
MOT17-13-FRCNN	56.7%	78.5%	44.4%	52.7%	93.1%	110	32	41	37	454	5509	149	193	47.5%	0.227	128	54	35
MOT17-13-SDP	54.3%	77.6%	41.7%	49.8%	92.5%	110	31	36	43	469	5848	123	181	44.7%	0.229	110	46	36
OVERALL	58.0%	79.2%	45.7%	56.3%	97.6%	1638	445	658	535	4709	147126	1258	1733	54.6%	0.147	841	721	316