TrafficMOT: A Challenging Dataset for Multi-Object Tracking in Complex Traffic Scenarios

Lihao Liu
University of Cambridge
Cambridge, United Kingdom
ll610@cam.ac.uk

Yanqi Cheng
University of Cambridge
Cambridge, United Kingdom
yc443@cam.ac.uk

Zhongying Deng
University of Cambridge
Cambridge, United Kingdom
zd294@cam.ac.uk

Shujun Wang
Hong Kong Polytechnic University
Hong Kong, China
sjwang8020@gmail.com

Dongdong Chen
Microsoft
Seattle, United States
cddlyf@gmail.com

Xiaowei Hu
Shanghai AI Laboratory
Shanghai, China
huxiaowei@pjlab.org.cn

Pietro Liò
University of Cambridge
Cambridge, United Kingdom
pl219@cam.ac.uk

Carola-Bibiane Schönlieb
University of Cambridge
Cambridge, United Kingdom
cbs31@cam.ac.uk

Angelica Aviles-Rivero
University of Cambridge
Cambridge, United Kingdom
ai323@cam.ac.uk

Abstract

Multi-object tracking in traffic videos is a crucial research area, offering immense potential for enhancing traffic monitoring accuracy and promoting road safety measures through the utilisation of advanced machine learning algorithms. However, existing datasets for multi-object tracking in traffic videos often feature limited instances or focus on single classes, which cannot well simulate the challenges encountered in complex traffic scenarios. To address this gap, we introduce `TrafficMOT`, an extensive dataset designed to encompass diverse traffic situations with complex scenarios. To validate the complexity and challenges presented by `TrafficMOT`, we conducted comprehensive empirical studies using three different settings: fully-supervised, semi-supervised, and a recent powerful zero-shot foundation model Tracking Anything Model (TAM). The experimental results highlight the inherent complexity of this dataset, emphasising its value to drive advancements in the field of traffic monitoring and multi-object tracking. Code and data are available at the project page: https://lihaoliu-cambridge.github.io/trafficmot

CCS Concepts

• **Computing methodologies → Scene understanding**; **Tracking**.

Keywords

Traffic Video Dataset, Multi-object Tracking, Foundation Model

1 Introduction

Multi-object tracking (MOT) aims to detect the trajectories of multiple targets simultaneously along video sequences. It is a fundamental but indispensable task in diverse realistic application scenarios and has recently achieved significant advances in visual surveillance [19, 33], autonomous driving [4, 18, 25], AI city [21, 22, 30], and sports analysis [6, 11, 20, 29, 37]. Among these diverse applications, multi-object tracking for traffic stands out, as it could unlock the potential for Intelligent Traffic Systems (ITS), which could enhance transportation efficiency and road safety [22].

Large-scale benchmarks are crucial for most artificial intelligence-related applications. Such benchmarks provide shared datasets for model training and fair competition in evaluation, e.g., the accident analysis dataset [27] and vehicle re-identification dataset [17], thus prompting the ITS development. Importantly, the development of effective multi-object tracking algorithms heavily also relies on the availability of high-quality datasets [9, 27, 30]. Despite the availability of multiple traffic MOT datasets, there are still significant limitations that hinder the progress in traffic analysis. For example, many existing datasets lack the complexity and diversity that is crucial for accurately reflecting real-world traffic scenarios, such as occlusions, varying lighting conditions, diverse traffic patterns, and importantly dense scenes. Consequently, the need for a new dataset arises, one that investigates the challenges encountered in complex traffic environments.

Current Datasets in Traffic Analysis and Their Limitations. Two widely used datasets in the field of traffic analysis are the VERI-Wild dataset [17] by Lou *et al.* and the MOTS dataset [31] by Voigtlaender *et al.* The VERI-Wild dataset focuses on vehicle re-identification in urban scenes and consists of data from 174 cameras. While it provides a comprehensive view of vehicle tracking scenarios, it is not specifically designed for multi-object tracking in intricate traffic scenarios. Similarly, the MOTS dataset extends
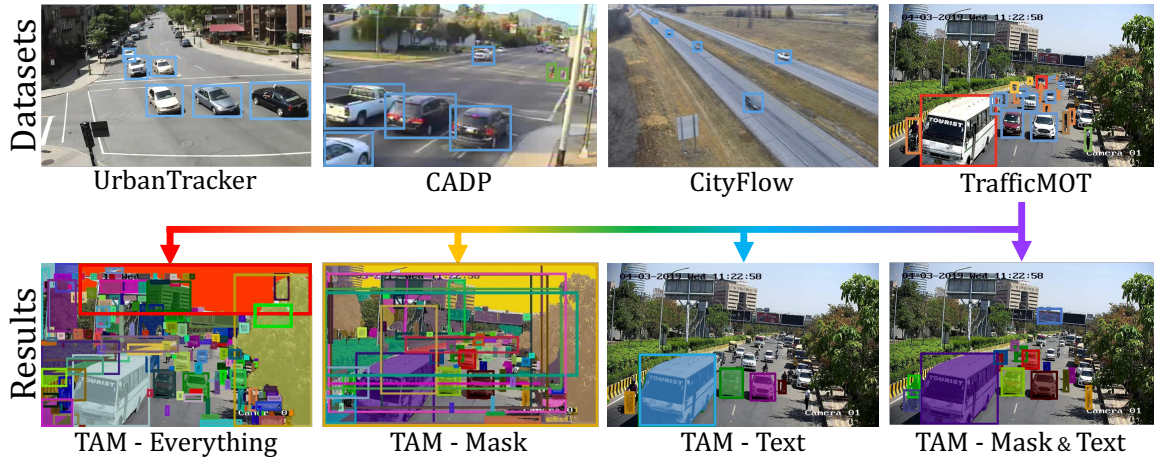
Figure 1: Comparison of `TrafficMOT` with other traffic multi-object tracking datasets, and performance evaluations of the recent powerful zero-shot foundation model (Tracking Anything Model). The first row illustrates the enhanced complexity of our `TrafficMOT`, with more classes and instances. The second row presents the visual results in different settings of Tracking Anything Model. The visuals indicate that even the advanced model struggles to manage the complexities of our dataset.

multi-object tracking by incorporating object segmentation, but it also lacks a specific focus on complex traffic scenarios. Moreover, both datasets are not specifically designed for multi-object tracking in the context of intricate traffic scenarios.

Figure 1 and Table 1 provide the comparisons between our `TrafficMOT` and other datasets in the field of traffic multi-object tracking. Note that the Urban Tracker [9] is introduced in 2014, focusing on multiple object tracking in urban mixed traffic. However, this dataset consists of only five collected video sequences, limiting its diversity and scalability. Later, Shah *et al.* [27] present a CADP dataset for traffic accident analysis using CCTV traffic cameras. While the dataset contain 1,416 video segments with full spatio-temporal annotations, it is specifically tailored to accident analysis and does not cover the broader challenges of multi-object tracking in traffic. Recently, CityFlow [30] is introduced for vehicle tracking and identification, providing a valuable dataset for research in the field. However, the number of classes is limited and it is unable to handle dense traffic scenarios. Most recently, improved versions of CityFlow are released in [21], where a series of AI city challenges are held for multi-camera people tracking, tracked vehicle retrieval by natural language, and traffic safety [22]. While the above datasets are specifically tailored for MOT in traffic scenes, significant limitations still persist, including limited class numbers and traffic density.

Contributions. In this paper, we present `TrafficMOT`, a novel benchmark dataset that introduces new challenges in the field of multi-object tracking in traffic analysis. Acquired from fixed CCTV cameras across a wide range of cities, `TrafficMOT` stands out among existing datasets due to its unique and challenging characteristics. Notably, our dataset offers an extensive number of object classes, surpassing any other existing dataset. Additionally, `TrafficMOT` provides dense scenes, mainly capturing the intricate dynamics and congestion found in traffic scenarios. To highlight the complexity and challenges inherent in `TrafficMOT`, we also conduct an empirical study with three representative settings: fully

Table 1: Comparison of characteristics between existing traffic multi-object tracking datasets and `TrafficMOT`.

DATASET	Year	#Video	#Mins	#Class	#Object
UrbanTracker	2014	5	18.59	3	5.4
CADP	2018	1,416	86.37	6	3.6
CityFlow	2022	40	**195**	1	8.2
⋆ `TrafficMOT`	2023	**2,102**	105.1	**10**	**22.8**

supervised, semi-supervised, and a recent powerful zero-shot foundation model [10] named as Tracking Anything Model (TAM) [35]. Our experimental results underscore the inherent complexity of this dataset, demonstrating its potential to drive advancements in the field of traffic monitoring and multi-object tracking. Our contributions are summarised as follows:

- We introduce `TrafficMOT`, a novel benchmark dataset specifically designed for multi-object tracking in complex traffic scenarios. This dataset offers a larger number of object classes and encompasses intricate traffic situations, including densely populated scenarios.
- We conduct empirical studies with three representative settings: fully-supervised, semi-supervised, and a recent powerful zero-shot foundation model called Tracking Anything Model (TAM). Experiment results highlight the intrinsic challenges posed by this dataset, thus emphasising its potential to spur advancements in traffic monitoring and road safety.

2 TrafficMOT: Unlocking Multi-Object Tracking in Complex Traffic

In this section, we delve into a comprehensive exploration of our dataset, providing detailed insights into its characteristics and statistical properties. We first introduce the data collection, annotation, and partitioning process. Then, we conduct statistical analyses to reveal the complexity of dataset.
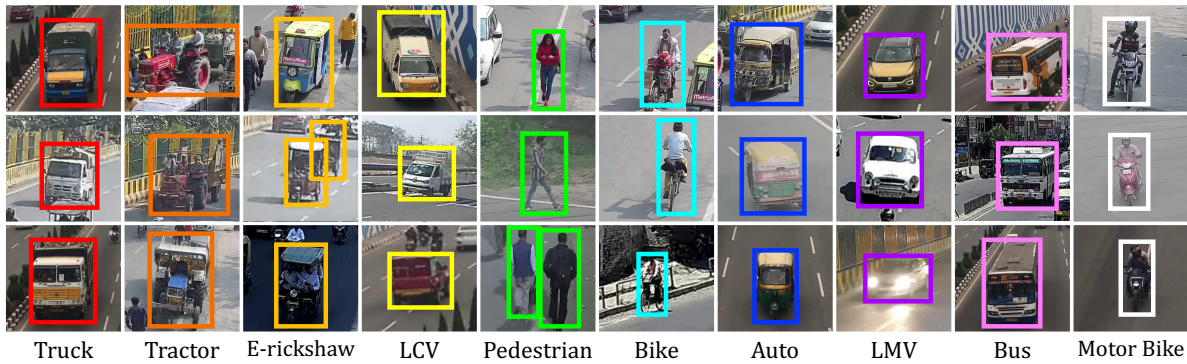
Truck Tractor E-rickshaw LCV Pedestrian Bike Auto LMV Bus Motor Bike

Figure 2: Visualisation of annotated instances on different classes in `TrafficMOT`.

2.1 TrafficMOT Dataset Construction

Video Collection. We collected videos from eight cities spanning different regions in India, including northern, southern, eastern, western, and central parts of the country. The video acquisition process involved the use of fixed CCTV cameras installed on roads and highways. These cameras provided a comprehensive view of the traffic scenarios in different urban and suburban settings.

Our dataset consists of a total of 2,102 videos, where each video comprises 30 frames, allowing for a significant amount of temporal information to be extracted. To capture the heterogeneity of real-world surveillance systems, the videos were recorded using different cameras, resulting in varying frame resolutions, including 352 × 288, 1056 × 864, 1280 × 720, and 1920 × 1080. This diversity in frame resolutions reflects the practical challenges encountered in real-world traffic surveillance scenarios.

Instance Annotation. The annotation process for our proposed `TrafficMOT` dataset involved providing instance bounding box (bbox) labels and categorising each instance into one of the ten classes illustrated in Figure 2 for each annotated frame. Out of the total 2,102 videos in the dataset, 78 videos were fully annotated, meaning that all 30 frames in each of these videos were annotated. In these fully annotated videos, `TrafficMOT` provides tracking ids for each instance, enabling the tracking of objects across frames within the video. For the remaining 2,024 videos, annotations were only provided for the first frame. This resulted in a total of 8,689 unannotated frames across the dataset. The presence of this large number of unlabelled frames presents an opportunity to explore and advance semi-supervised learning techniques that leverage a limited number of annotations.

To ensure the quality of the annotations, a professional company was engaged to carry out the annotation process. Following the completion of the annotations, a dedicated quality control group meticulously reviewed all the annotations. Any annotations identified as low quality were sent back to the company for refinement. This iterative process of annotation refinement was conducted multiple times to ensure a high level of annotation quality in the `TrafficMOT` dataset.

Official Partition. To facilitate the training and validation of multiple object tracking task on `TrafficMOT` dataset, the dataset was divided into training and test sets at the video level. Out of the 78 fully annotated videos, 51 videos were selected as the testing set.
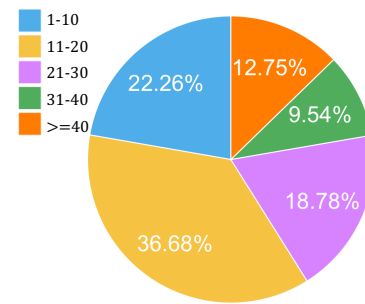

Figure 3: Comparison of instance numbers per frame.

These videos contain instance-level annotations (bounding box + category) for every frame and provide instance tracking IDs that persist throughout the video.

During the fully-supervised training process, we employed the remaining 27 fully annotated videos in conjunction with the 2,024 videos that only had annotations on the first frame. In contrast, the semi-supervised setting incorporated an additional 8,689 unlabelled frames as part of the training set. This inclusion of unlabelled data presents a chance to explore semi-supervised learning techniques that leverage both labelled and unlabelled data for improved performance. Partitioning the dataset into separate train and test sets ensures a rigorous evaluation of multi-object tracking performance and enables fair comparisons for different methods.

2.2 TrafficMOT Statistics & Key Features

Diverse Classes. Figure 2 presents annotation examples for each class, illustrating that certain classes exhibit a notable resemblance. For instance, distinguishing between the e-rickshaw and auto, as well as between the bike and motorbike, requires careful examination. Accurately identifying these classes is vital to ensure road safety, particularly considering the potential risks associated with the interaction between motor vehicles and non-motor vehicles. This underscores the importance of exploring complex traffic scenarios that involve interclass correlations.

In Figure 4, we present the correlation matrix depicting the presence of different classes within the same video. We divide the ten classes into two groups: motor and non-motor vehicles. The motor vehicle classes exhibit a strong correlation with each other
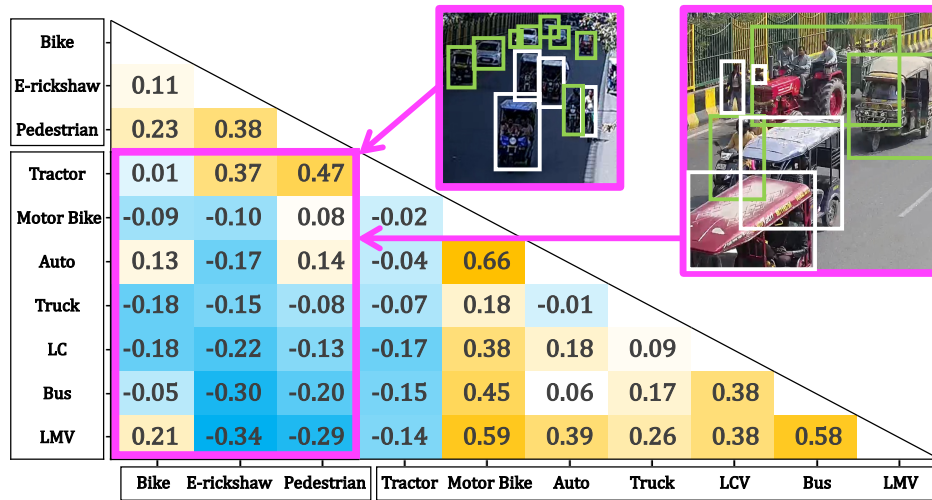
Figure 4: Correlation matrix of classes built on the appearance of class tracks within the same video.

indicating that different types of motor vehicles commonly co-occur in our dense and complex `TrafficMOT` dataset. Moreover, the correlation between motor and non-motor vehicles is shown in the pink part in Figure 4. It is evident that pedestrian and e-rickshaws classes (0.37 and 0.47) frequently appear alongside heavy vehicles such as tractors, raising safety concerns for citizens. Hence, our `TrafficMOT` dataset is beneficial for analysing dense traffic conditions and vehicle behaviour to mitigate the risks of accidents.

Crowded Scenes in Multi-Object Tracking. Our dataset stands out from existing datasets in terms of traffic density, offering a unique opportunity to explore and tackle the challenges of multi-object tracking in dense traffic flow scenarios. While existing datasets are often limited in terms of the number of objects (as indicated in Table 1, and detailed in Figure 3), our dataset presents a significant advancement by providing a rich and challenging environment with densely populated traffic scenes.

A key characteristic of our dataset is the high average number of objects per frame. On average, our dataset contains 22.8 objects per frame, surpassing the ranges found in other datasets which average from 5.4 to 8.2 objects per frame, as shown in Table 1. This notable difference in object density creates complex tracking scenarios, requiring advanced algorithms and techniques to accurately track and analyse multiple objects amidst congested traffic conditions.

Why is Traffic Density Relevant? The heightened object density in our dataset imposes greater demands on tracking algorithms, requiring them to exhibit enhanced robustness, accuracy, and efficiency in handling a larger number of objects simultaneously. Furthermore, the dense scene characteristic exposes tracking algorithms to challenges such as occlusions, partial object visibility, and overlapping trajectories, all of which are prevalent in crowded traffic scenarios. This realistic representation of traffic conditions provides a more meaningful and representative benchmark for evaluating the performance of multi-object tracking algorithms in practical traffic analysis applications.

Weather Conditions in TrafficMOT. Our dataset comprises traffic videos captured under diverse weather conditions, which
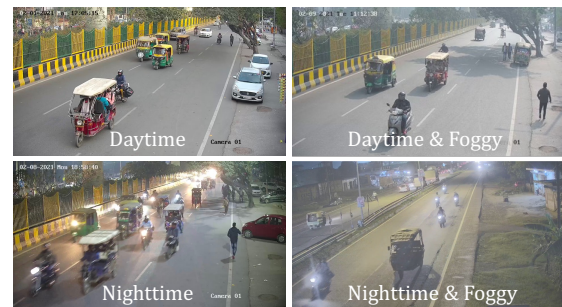

Figure 5: Visualisation of different weather scenes.

provides a realistic representation of real-world scenarios. It encompasses 4 diffrent weather condition: sunny (daytime) weather (1,761 videos), foggy conditions (218 videos), and low-light conditions during nighttime (123 videos). This deliberate inclusion of various weather conditions enhances the training difficulty and exposes the model to a wider range of scenarios that it may encounter in the real world.

The presence of videos captured under sunny weather conditions reflects typical daytime scenarios, enabling the model to learn patterns and behaviours under optimal lighting conditions. This serves as a baseline for evaluating the model's performance in ideal weather conditions. Incorporating videos captured in foggy conditions introduces challenges associated with reduced visibility, as shown in Figure 5. The model learns to detect and track objects amidst the obscuring effects of fog, enhancing its ability to handle adverse weather conditions. The inclusion of videos recorded in low-light conditions during nighttime presents additional challenges, such as limited illumination and potential occlusions. This exposure allows the model to adapt and improve its performance in scenarios where lighting conditions are less favourable. By exposing the model to a diverse range of environmental conditions, our dataset increases the training difficulty and enables the model to learn robust representations that generalise well across different scenarios. This prepares the model to handle real-world challenges,

contributing to the development of more reliable and versatile multi-object tracking systems.

3 Benchmarking TrafficMOT on Multiple Object Tracking

We conduct a comprehensive benchmark on multi-object tracking (MOT) by evaluating three types of methods: fully-supervised MOT methods (Section 3.1), semi-supervised MOT methods (Section 3.2), and a recent powerful zero-shot foundation model, the Tracking Anything Model (Section 3.3). This extensive evaluation allowed us to assess the performance of various MOT approaches on our `TrafficMOT` dataset and gain valuable insights into their respective strengths. It is worth noting that, to the best of our knowledge, *our work is the first to apply large-scale foundation models to multi-object tracking tasks in complex traffic scenarios.*

3.1 Fully-Supervised MOT Methods

Fully-supervised deep learning techniques have proven to be powerful in recent years [14, 36]. We first train fully-supervised MOT models using existing baseline MOT methods. Most baselines integrate an object detector followed by a tracker, so the training process is formulated as

$$\mathcal{L} = \mathcal{L}_{det} + \mathcal{L}_{tra}, \tag{1}$$

where $\mathcal{L}_{det}$ represents the object detection and classification loss, typically employed by methods like Faster R-CNN [26]. $\mathcal{L}_{tra}$ denotes object association or tracking loss, such as the loss used for learning instance similarity [23]. In our study, we specifically focus on four fully-supervised MOT methods: DeepSORT [32], ByteTrack [38], QDTrack [23], and OC-SORT [2].

DeepSORT includes appearance information to SORT [1] to better track occluded objects. SORT, which uses the overlap of bounding boxes as a metric for object association, cannot effectively track through occlusions due to the simple association metric. DeepSORT thus integrates the appearance and motion information into the association metric to further improve SORT.

ByteTrack makes full use of both the high- and low-score detection boxes for object association while most other MOT methods only utilise the high-score ones. It first associates the high-score detection boxes to certain tracklets based on their IoU similarity. After that, it matches the low-score detection boxes to the remaining unmatched tracklets.

QDTrack (Quasi-Dense Tracking) adopts a two-step approach to enhance feature embedding and object association. Firstly, it employs quasi-dense contrastive learning, which involves densely matching extensive regions of interest in one image with corresponding regions in another. This process aims to learn a more robust feature embedding. Secondly, QDTrack uses bi-directional softmax to associate objects, ensuring bi-directional consistency. This means that the feature embeddings of two associated objects should be the nearest neighbors of each other.

OC-SORT (Observation-Centric SORT) addresses noise accumulation in SORT [1], which employs the Kalman filter (KF) and an object association matrix. OC-SORT proposes two strategies: Observation-centric Re-Update (ORU) and Observation-Centric Momentum (OCM). ORU rechecks untracked tracks associated with object tracklets and updates KF parameters accordingly. OCM introduces a cost matrix for object association, enforcing consistent motion direction among objects.

3.2 Semi-Supervised MOT Training

In semi-supervised training setting, we incorporate the common pseudo-labelling mechanism for object detection to implement multi-object tracking. We apply three semi-supervised strategies across all four MOT baseline methods, namely STAC [28], SoftTeacher [34], and MotionPrior [15].

STAC operates in two stages. In the first stage, an MOT model is trained using labelled training data. In the second stage, the trained MOT model is utilized to generate pseudo-labels for unlabelled data. Subsequently, the MOT model is re-trained using both the labelled and pseudo-labelled data, improving its performance through the inclusion of additional unlabelled samples.

SoftTeacher shares a similar setting with STAC but introduces a single-stage strategy to perform the two-stage pseudo-labelling mechanism in an end-to-end manner.

MotionPrior extends STAC by incorporating simple motion prior present in traffic datasets. Therefore, it can filter out noisy pseudo labels, thereby enhancing the quality of the training dataset.

Tracking Branches in Semi-Supervised Training. The three aforementioned pseudo-labelling strategies are primarily designed for object detection, which may result in inaccurate pseudo-labels for tracking IDs. Consequently, in the context of semi-supervised MOT, it is necessary to ensure the consistency of pseudo-labels across the time series before utilizing them for training purposes.

To this end, given a predicted object at a certain position specified by bounding box $\hat{b}_i$, we will find the nearest object to $\hat{b}_i$ in the next frame, which is denoted as $\hat{b}_{i+1}$. Then, we check the consistency C of these two objects via

$$C = \begin{cases} 1, & \text{if } \text{IoU}(\hat{b}_i, \hat{b}_{i+1}) > \tau_c \text{ and } (\hat{y}_i == \hat{y}_{i+1}) \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $\text{IoU}(\cdot, \cdot)$ is the function computing the intersection over the union score of two inputs, and τ_c is a threshold. If the predicted bounding boxes of these two objects are close to each other and their object identities are the same, then $C = 1$. In this case, we regard them as the same tracking object and keep their predictions as pseudo-labels. In this way, we can obtain the pseudo-labels for all the unlabelled videos, for both detection and tracking results. Subsequently, we incorporate these pseudo-labelled videos alongside the labelled ones to train another MOT model, employing the same loss functions as defined in Equation (1).

3.3 Zero-Shot Foundation Model for MOT

Additionally, we conduct empirical studies using the Tracking Anything Model (TAM) [35], a recent and powerful zero-shot foundation model [10]. TAM demonstrates potential applicability in the context of semi-supervised MOT, further enhancing the versatility of our research.

Track Anything Model, also named Segment and Track Anything (SAM-Track), is a powerful zero-shot foundation model for general MOT tasks such as unsupervised MOT, semi-supervised MOT, and interactive MOT. It adopts the latest Segment Anything

Table 2: Numerical comparison on supervised settings. Best results are marked in bold.

Techniques		Evaluation Metrics			
Method	Year	mAP	mAP$_{50}$	IDF1	MOTA
DeepSORT	2017	0.436	0.621	0.498	0.213
ByteTrack	2021	**0.463**	**0.651**	**0.637**	0.326
QDTrack	2021	0.440	0.633	0.592	**0.414**
OC-SORT	2022	0.460	0.646	0.562	0.118

Figure 6: Visual comparison of different methods in fully-supervised Settings. For visualisation purposes, an object is displayed along with the identified error over time. The pink bounding box signifies a tracking error, indicated by a tracking ID that differs from the one assigned to the white bounding box.

Model (SAM) [10] to generate a segmentation mask for the first frame of videos with the help of user interactive clicks. Then it exploits XMem [3], a semi-supervised MOT method requiring a precise mask to initialise, to output a predicted mask. If the mask is of poor quality, SAM is further used to refine the mask. Otherwise, the mask is accepted as the final predicted tracking mask.

We consider four different ways to apply TAM for zero-shot multi-object tracking, as shown in Figure 7. (a) We directly adopt TAM to track everything on `TrafficMOT` using default settings. (b) We make full use of the first-frame annotations in our dataset to guide the learning process of TAM to track traffic-related objects. (c) We input the class names in our dataset as the text prompts to guide the TAM to learn to track the objects of these classes. (d) We use both text prompts and first-frame annotations to better

adapt TAM to our dataset. Through these empirical studies, we illustrate that even a large-scale MOT model such as TAM struggles to effectively handle the MOT task as defined in our dataset. These comprehensive studies serve as a testament to the challenges posed by complex traffic scenarios within our dataset.

4 Experimental Results

4.1 Implementation Details, Preprocessing, & Evaluation Metrics

Implementation Details. We implemented the MOT architecture based on MMTracking [5] framework using the PyTorch [24] library. For fair comparison with all the fully-supervised MOT methods [2, 23, 32, 38], we implemented the object detection module
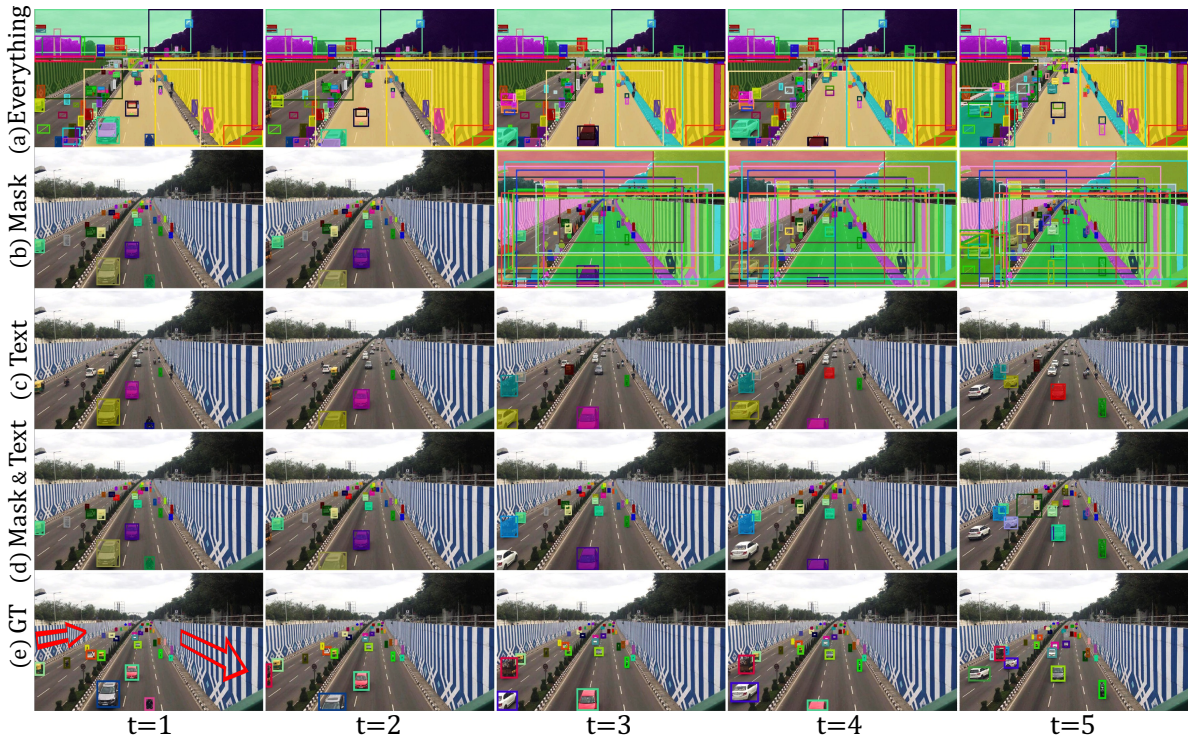
Figure 7: Visual comparison using different information on TAM: (a) Everything, (b) Mask, (c) Text, (d) Mask & Text. For better comparisons of the outputs, the ground truth (GT) is displayed in (e). In the first frame of GT, we emphasise the two contrary road directions using red arrows.

with the same architecture: Faster R-CNN (Region-based Convolutional Neural Network) [26] incorporating with region proposal network (RPN) [12] on ResNet-50 [8] backbone with ImageNet [7] pre-trained weights. The models were trained for 15 epochs with the AdamW optimiser [16] by step decay initialling with a learning rate of 2×10^{-4}. For the training of semi-supervised frameworks, we initialised their weights with the corresponding trained fully-supervised models. Further, we tuned the initial learning rate on AdamW optimiser to 5×10^{-5} and trained an additional five epochs. The empirical studies on fully-supervised and semi-supervised methods are all trained on an NVIDIA A100 GPU with 80GB RAM for approximately six hours and twlve hours respectively. For the TAM experiments, we set the new object updating parameters as default in all four track anything settings following [35].

Preprocessing. We followed the data augmentation strategy in MMTracking to enhance the variety of the dataset. Specifically, we first resized the images to $1,088 \times 1,088$ while preserving the aspect ratio, then performed photometric distortion, random flipping, normalisation, and padding.

Detection Evaluation. We measure the detection performance with the mAP (mean Average Precision) and mAP_{50} metrics. mAP calculates the precision-recall curve by varying the detection threshold and averages the AP values for all classes, while mAP_{50} takes the average precision at an IoU threshold of 0.5 for better evaluating moderate overlap objects.

Tracking Evaluation. We evaluate the tracking performance with Multi-Object Tracking Accuracy (MOTA) and Identity F1 score (IDF1). MOTA can assess the overall tracking accuracy by considering false positive, false negative, and tracking ID switches. IDF1 measures the accuracy of tracking identity by the F1 score based on the overlap between predicted and ground truth tracks.

4.2 Benchmark Results

Fully-Supervised Results. We benchmark four fully-supervised MOT methods, namely DeepSORT [32], ByteTrack [38], QDTrack [23], and OC-SORT [2]; as shown in Table 2 and Figure 6.

DeepSORT demonstrates relatively poor performance, mainly due to its challenges in handling occlusions, which are common complexities in traffic scenarios. On the other hand, ByteTrack [38] achieves the best performance in terms of mAP, mAP_{50}, and IDF1, with scores of 0.463, 0.651, and 0.637, respectively. In Figure 6, Byte-Track accurately detects bounding boxes but faces difficulties in establishing consistent connections across frames for tracking. Despite its limitations in detection, QDTrack outperforms ByteTrack in terms of MOTA by 0.088, indicating its superior ability to process tracking information. However, OC-SORT, being a newer method, exhibits below-average performance in terms of MOTA, suggesting it may not be well-suited for traffic scenarios. Figure 6 demonstrates instances where OC-SORT loses track. These results are obtained by re-training the networks with unified training parameters to ensure fair comparisons across different methods.

Table 3: Numerical comparison between baselines on semi-supervised Settings. The results display four metrics, with the best results highlighted in bold. "ST" stands for "SoftTeacher", and "MT" denotes "MotionPrior" [13].

(a) Evaluation results on DeepSORT.

	Results on DeepSORT			
	mAP	mAP_{50}	IDF1	MOTA
STAC	0.378	0.565	0.504	0.333
ST	0.367	0.552	0.501	0.324
MP	**0.390**	**0.569**	**0.511**	**0.340**

(b) Evaluation results on ByteTrack

	Results on ByteTrack			
	mAP	mAP_{50}	IDF1	MOTA
STAC	0.440	0.624	**0.636**	0.452
ST	0.436	0.601	0.601	0.450
MP	**0.455**	**0.627**	0.635	**0.455**

(c) Evaluation results on QDTrack.

	Results on QDTrack			
	mAP	mAP_{50}	IDF1	MOTA
STAC	0.393	0.575	0.549	0.388
ST	0.393	0.569	0.530	0.346
MP	**0.404**	**0.590**	**0.571**	**0.399**

(d) Evaluation results on OC-SORT.

	Results OC-SORT			
	mAP	mAP_{50}	IDF1	MOTA
STAC	0.422	0.610	0.603	0.385
ST	0.415	0.600	0.555	0.207
MP	**0.444**	**0.629**	**0.610**	**0.397**

Semi-Supervised Results. In addition to benchmarking supervised methods, we also evaluate three semi-supervised pseudo-labelling strategies across all four baseline MOT models, the results of which are presented in Table 3.

Upon initial analysis, it is observed that MotionPrior generally outperforms the other two strategies, and STAC marginally surpasses SoftTeacher. When compared to the fully-supervised setting, all semi-supervised strategies yield lower scores in terms of mAP and mAP_{50}, which indicates reduced detection accuracy. However, models like DeepSORT, ByteTrack, and OC-SORT exhibit improved IDF1 and MOTA scores, which indicates improved tracking performance. Yet, even though the semi-supervised setting can help in improving tracking accuracy, the overall performance remains relatively low due to the dataset's complexity. Consequently, the design and development of further algorithms tailored to manage such challenging datasets is imperative.

TAM Results. Although TAM has demonstrated superior performance in interactive object tracking and segmentation in some videos [35], it yields unsatisfactory results on our TrafficMOT dataset, as depicted in Figure 7. We analyse the experimental results for five different TAM settings, as explained in Sec. 3.3. In the (a) track Everything mode, TAM outputs trajectories for all objects, including background and irrelevant objects, leading to a loss of focus on traffic-related objects. The second row in Figure 7 displays the tracking results when using the first frame annotation to guide the tracking of subsequent objects. It is evident that this model performs well only on the initial frames, and then transitions towards the (a) Everything mode for the remaining frames. From the third row, we observe that using class names as prompts allows TAM to generate stable tracking outputs for vehicles and pedestrians, but with numerous missing objects in crowded areas. When integrating the (d) first frame annotation and class name to guide TAM, the output outperforms all other settings. However, it may still generate

unusual objects, as depicted by the large green bounding box in the last column. Moreover, new objects, such as the white car in the bottom-left corner visible in the third and fourth frames, are not tracked until the fifth frame. A critical observation across all four settings is that while objects can be tracked, they are not accurately recognised. All instances are generalised as 'objects' without specific class designations. Despite TAM's promising pixel-level annotation, it falls short in providing precise class information.

5 Conclusion

This study addresses the vital area of multi-object tracking in traffic videos, recognising its potential to enhance traffic monitoring and promote road safety through advanced machine learning algorithms. We acknowledge the limitations of current datasets, which often oversimplify the task and lack adequate instances or diverse class representation. To tackle these challenges, we present TrafficMOT, a comprehensive and diverse dataset specifically designed to cover a wide range of complex traffic scenarios. Our research includes three empirical studies: fully-supervised, semi-supervised, and the recent zero-shot foundation model (TAM). These studies underscore the inherent complexity and difficulties of our TrafficMOT dataset. Our experimental findings highlight the inherent difficulty of our dataset while also showcasing its potential to drive significant advancements in the fields of traffic monitoring and multi-object tracking. With its high level of complexity and diversity, we envision that TrafficMOT will serve as a challenging yet realistic benchmark for the development and refinement of future multi-object tracking algorithms.

Acknowledgments

References

[1] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. 2016. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*. IEEE, 3464–3468.

[2] Jinkun Cao, Xinshuo Weng, Rawal Khirodkar, Jiangmiao Pang, and Kris Kitani. 2022. Observation-Centric SORT: Rethinking SORT for Robust Multi-Object Tracking. *arXiv preprint arXiv:2203.14360* (2022).

[3] Ho Kei Cheng and Alexander G Schwing. 2022. XMem: Long-Term Video Object Segmentation with an Atkinson-Shiffrin Memory Model. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*. Springer, 640–658.

[4] Hsu-kuang Chiu, Jie Li, Rareş Ambruş, and Jeannette Bohg. 2021. Probabilistic 3d multi-modal, multi-object tracking for autonomous driving. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 14227–14233.

[5] MMTracking Contributors. 2020. MMTracking: OpenMMLab video perception toolbox and benchmark. https://github.com/open-mmlab/mmtracking.

[6] Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. 2023. SportsMOT: A Large Multi-Object Tracking Dataset in Multiple Sports Scenes. *arXiv preprint arXiv:2304.05170* (2023).

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[9] Jean-Philippe Jodoin, Guillaume-Alexandre Bilodeau, and Nicolas Saunier. 2014. Urban tracker: Multiple object tracking in urban mixed traffic. In *IEEE Winter Conference on Applications of Computer Vision*. IEEE, 885–892.

[10] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643* (2023).

[11] Pengze Li, Lihao Liu, Carola-Bibiane Schönlieb, and Angelica I Aviles-Rivero. 2024. Optimised ProPainter for Video Diminished Reality Inpainting. *arXiv preprint arXiv:2406.02287* (2024).

[12] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2117–2125.

[13] Lihao Liu, Yanqi Cheng, Dongdong Chen, Jing He, Pietro Liò, Carola-Bibiane Schönlieb, and Angelica I Aviles-Rivero. 2023. Traffic Video Object Detection using Motion Prior. *arXiv preprint arXiv:2311.10092* (2023).

[14] Lihao Liu, Qi Dou, Hao Chen, Jing Qin, and Pheng-Ann Heng. 2019. Multi-task deep model with margin ranking loss for lung nodule analysis. *IEEE transactions on medical imaging* 39, 3 (2019), 718–728.

[15] Lihao Liu, Jean Prost, Lei Zhu, Nicolas Papadakis, Pietro Liò, Carola-Bibiane Schönlieb, and Angelica I Aviles-Rivero. 2023. SCOTCH and SODA: A Transformer Video Shadow Detection Framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10449–10458.

[16] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[17] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Lingyu Duan. 2019. Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3235–3243.

[18] Chenxu Luo, Xiaodong Yang, and Alan Yuille. 2021. Exploring simple 3d multi-object tracking for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10488–10497.

[19] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. 2022. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8844–8854.

[20] Sungwon Moon, Jiwon Lee, Dowon Nam, Howon Kim, and Wonjun Kim. 2017. A comparative study on multi-object tracking methods for sports events. In *2017 19th International Conference on Advanced Communication Technology (ICACT)*. IEEE, 883–885.

[21] M. Naphade, S. Wang, D. C. Anastasiu, Z. Tang, M. Chang, Y. Yao, L. Zheng, M. Shaiqur Rahman, A. Venkatachalapathy, A. Sharma, Q. Feng, V. Ablavsky, S. Sclaroff, P. Chakraborty, A. Li, S. Li, and R. Chellappa. 2022. The 6th AI City Challenge. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE Computer Society, 3346–3355. https://doi.org/10.1109/CVPRW56347.2022.00378

[22] Milind Naphade, Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Yue Yao, Liang Zheng, Mohammed Shaiqur Rahman, Meenakshi S. Arya, Anuj Sharma, Qi Feng, Vitaly Ablavsky, Stan Sclaroff, Pranamesh Chakraborty, Sanjita Prajapati, Alice Li, Shangru Li, Krishna Kunadharaju, Shenxin Jiang, and Rama Chellappa. 2023. The 7th AI City Challenge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[23] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. 2021. Quasi-dense similarity learning for multiple object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 164–173.

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[25] Akshay Rangesh and Mohan Manubhai Trivedi. 2019. No blind spots: Full-surround multi-object tracking for autonomous vehicles using cameras and lidars. *IEEE Transactions on Intelligent Vehicles* 4, 4 (2019), 588–599.

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).

[27] Ankit Shah, Jean Baptiste Lamare, Tuan Nguyen Anh, and Alexander Hauptmann. 2018. CADP: A Novel Dataset for CCTV Traffic Camera based Accident Analysis. *arXiv preprint arXiv:1809.05782* (2018). First three authors share the first authorship..

[28] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. 2020. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757* (2020).

[29] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. 2022. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20993–21002.

[30] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasiu, and Jenq-Neng Hwang. 2019. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8797–8806.

[31] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. 2019. Mots: Multi-object tracking and segmentation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. 7942–7951.

[32] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. 2017. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*. IEEE, 3645–3649.

[33] Yu Xiang, Alexandre Alahi, and Silvio Savarese. 2015. Learning to track: Online multi-object tracking by decision making. In *Proceedings of the IEEE international conference on computer vision*. 4705–4713.

[34] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. 2021. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3060–3069.

[35] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. 2023. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968* (2023).

[36] Lipei Zhang, Yanqi Cheng, Lihao Liu, Carola-Bibiane Schönlieb, and Angelica I Aviles-Rivero. 2024. Biophysics Informed Pathological Regularisation for Brain Tumour Segmentation. *arXiv preprint arXiv:2403.09136* (2024).

[37] Tianzhu Zhang, Bernard Ghanem, and Narendra Ahuja. 2012. Robust multi-object tracking via cross-domain contextual information for sports video analysis. In *2012 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 985–988.

[38] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. 2022. Bytetrack: Multi-object tracking by associating every detection box. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*. Springer, 1–21.