2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

METEOR: A Dense, Heterogeneous, and Unstructured Traffic Dataset With Rare Behaviors

Rohan Chandra[1*], Xijun Wang[1*], Mridul Mahajan[2], Rahul Kala[2], Rishitha Palugulla[3], Chandrababu Naidu[3], Alok Jain[3], and Dinesh Manocha[1,4]

Dataset, Code, and Video at https://gamma.umd.edu/meteor

***Abstract*— We present a new traffic dataset, METEOR, which captures traffic patterns and multi-agent driving behaviors in unstructured scenarios. METEOR consists of more than 1000 one-minute videos, over 2 million annotated frames with bounding boxes and GPS trajectories for 16 unique agent categories, and more than 13 million bounding boxes for traffic agents. METEOR is a dataset for rare and interesting, multi-agent driving behaviors that are grouped into traffic violations, atypical interactions, and diverse scenarios. Every video in METEOR is tagged using a diverse range of factors corresponding to weather, time of the day, road conditions, and traffic density. We use METEOR to benchmark perception methods for object detection and multi-agent behavior prediction. Our key finding is that state-of-the-art models for object detection and behavior prediction, which otherwise succeed on existing datasets such as Waymo, fail on the METEOR dataset. METEOR is a step towards developing more sophisticated perception models for dense, heterogeneous, and unstructured scenarios.**

I. INTRODUCTION

Recent research in learning-based techniques for robotics, computer vision, and autonomous driving has been driven by the availability of datasets and benchmarks. Several traffic datasets have been collected from different parts of the world to stimulate research in autonomous driving, driver assistants, and intelligent traffic systems. These datasets correspond to highway or urban traffic, and are widely used in the development and evaluation of new methods for perception [1], prediction [2], behavior analysis [3], and navigation [4].

Many initial autonomous driving datasets were motivated by computer vision or perception tasks such as object recognition, semantic segmentation or 3D scene understanding. Recently, many other datasets have been released that consist of point-cloud representations of objects captured using LiDAR, pose information, 3D track information, stereo imagery or detailed map information for applications related to 3D object recognition and motion forecasting. Many large-scale motion forecasting datasets such as Argoverse [5], and Waymo Open Motion Dataset [6], among others, have been used extensively by researchers and engineers to develop robust prediction models that can forecast vehicle trajectories. However, existing datasets do not capture the rare behaviors or heterogeneous patterns. Therefore, prediction models trained on these existing datasets are not very robust in terms of handling challenging traffic scenarios that arise in the real world.

A major challenge currently faced by research in autonomous driving is the *heavy tail problem* [5], [6], which refers to the challenge of dealing with rare and interesting instances. There are several ways in which existing datasets currently address the heavy tail problem:

1) **Mining:** The Argoverse and Waymo datasets use a mining procedure that includes scoring each trajectory based on its "interestingness" to explicitly search for difficult and unusual scenarios [5], [6].
2) **Diversifying the taxonomy:** Train the prediction and forecasting models to identify the unknown agents at the time of testing. This approach necessitates annotating a diverse taxonomy of class labels. Argoverse and nuScenes [9] contain 15 and 23 classes, respectively.
3) **Increasing dataset size:** This approach is to simply collect more data with the premise that collecting more traffic data will likely also increase the number of such scenarios in the dataset.

In spite of many efforts along these lines, existing datasets manage to collect only a handful of such instances, due to the infrequent nature of their occurrence. For example, the Waymo Open Motion dataset [6] contains only atypical interactions and diverse scenarios while the Argoverse dataset [5] contains only atypical interactions. There is clearly a need for a different approach to addressing the heavy tail problem. Our solution is to build a traffic dataset from videos collected in India, where the inherent nature of the traffic is dense, heterogeneous, and unstructured. The traffic patterns and surrounding environment in parts of India are more challenging. than those in other parts of the world. This includes high congestion and traffic density. Some of these roads are unmarked or unpaved. Moreover, the traffic agents moving on these roads correspond to vehicles, buses, trucks, bicycles, pedestrians, auto-rickshaws, two-wheelers such as scooters and motorcycles, etc.

A. Main Contributions

1) We present a novel dataset, METEOR, corresponding to the dense, heterogeneous, and unstructured traffic in India. METEOR is the first large-scale dataset containing annotated scenes for rare and interesting instances and multi-agent driving behaviors, broadly grouped into:
   a) Traffic violations—running traffic signals, driving in the wrong lanes, taking wrong turns).
   b) Atypical interactions—cut-ins, yielding, overtaking, overspeeding, zigzagging, lane changing.

*Denotes equal contribution. This work was supported in part by Semiconductor Research Corporation (SRC).

[1]Department of Computer Science, University of Maryland, College Park, USA. Corresponding email: rchandr1@umd.edu
[2] Centre for Intelligent Robotics, Indian Institute of Information Technology, Allahabad, India.
[3] NavAjna Technologies Pvt. Ltd.
[4]Department of Electrical and Computer Engineering, University of Maryland, College Park, USA.

TABLE I: **Characteristics of Traffic Datasets:** We compare METEOR with state-of-the-art autonomous driving datasets that have been used for trajectory tracking, motion forecasting, semantic segmentation, prediction, and behavior classification. METEOR is the largest (in terms of number of annotated frames) and most diverse in terms of heterogeneity, scenarios, varying behaviors, densities, and rare instances. Darker shades represent a richer collection in that category. Best viewed in color.

										Rare and Interesting Behaviors[‡]		
Datasets	Location	Bad weather	Night	Road type	Het.[⋆]	Size	Density	Lidar	HD Maps	Traffic Violations	Atypical Interactions	Diverse Scenarios
Argoverse [5]	USA	✓	✓	urban	10	22K	Medium	✓	✓	✗	✓	✗
Lyft Level 5 [7]	USA	✗	✗	urban	9	46K	Low	✓	✓	✗	✗	✗
Waymo [6]	USA	✓		urban	4	200K	Medium	✓	✓	✗	✓	✓
ApolloScape [8]	China	✗	✓	urban, rural	5	144K	High	✓	✓	✗	✗	✗
nuScenes [9]	USA/Sg.	✓	✓	urban	13	40K	Low	✓	✓	✗	✓	✓
INTERACTION [10]	International	✗	✗	urban	1	–	Medium	✓	✓	✗	✗	✗
CityScapes [11]	Europe	✗	✗	urban	10	25K	Low	✗	✗	✗	✗	✗
IDD [12]	India	✗	✗	urban, rural	12	10K	High	✗	✗	✗	✗	✗
HDD [13]	USA	✗	✗	urban	–	275K	Medium	✓	✗	✗	✓	✓
Brain4cars [14]	USA	✗	✗	urban	–	2000K	Low	✗	✓	✗	✗	✗
D2-City [15]	China	✓	✗	urban	12	700K	Medium	✗	✗	✗	✗	✓
TRAF [16]	India	✗	✓	urban, rural	8	72K	High	✗	✗	✗	✗	✗
BDD [17]	USA	✓	✓	urban	8	**3000K**	High	✗	✗	✗	✗	✓
ROAD [18]	UK	✓	✓	urban	7	122K	Low	✓	✗	✗	✗	✓
METEOR	**India**	✓	✓	**urban, rural[†]**	**16[††]**	2027K	**High[§]**	✗	✗	✓	✓	✓

[‡] Rare instances can be broadly grouped into (i) traffic violations, (ii) atypical interactions, and (iii) difficult scenarios.
[†] Includes roads without lane markings. Roads in other datasets with rural roads may contain lane markings.
[⋆] Heterogeneity. We indicate the classes corresponding to moving traffic agents only, excluding static objects such as poles, traffic lights, etc.
[§] Up to 40 agents per frame.
[††] Up to 9 unique agents per frame.

c) Diverse scenarios—intersections, roundabouts, and traffic signals.

2) METEOR has more than 2 million labeled frames and 13 million annotated bounding boxes for 16 unique traffic agents, and GPS trajectories for the ego-agent.
3) Every video in METEOR is tagged using a diverse range of factors including weather, time of the day, road conditions, and traffic density.
4) We use METEOR to extract new insights in perception tasks such as 2D object detection and multi-agent behavior recognition in unstructured traffic. Additionally, we present a novel, fine-grained analysis on the relationship between traffic environments (traffic density, mixture of agents, area, time of the day, and weather conditions) and 2D object detection.

B. Applications and Benefits

We list some promising directions in which METEOR can contribute towards autonomous driving research:

- **Towards Robust Perception:** We observe that perception tasks like 2D object detection and multi-agent behavior recognition fail in challenging Indian traffic scenarios, compared to their performance on existing datasets captured in the US, Europe, and other developed nations. METEOR can be a useful benchmark for research in perception in unstructured traffic environments and developing nations.
- **Towards Risk-Aware Planning and Control:** METEOR can aid the development of risk-aware motion planners by predicting the behaviors of surrounding agents. Motion planners can compute controls that guarantee safety around aggressive drivers who are prone to overtaking and overspeeding.
- **Towards Fine-grained Traffic Analysis:** With METEOR, researchers can study the causality relationship between traffic patterns, static scene elements, and dynamic agent behaviors resulting in novel ADAS for unstructured traffic environments.

II. Comparison with Existing Datasets

A. Tracking and Trajectory Prediction Datasets

Datasets such as the Argoverse [5], Lyft Level 5 [7], Waymo Open Dataset [6], ApolloScape [8], nuScenes dataset [9] are used for trajectory forecasting [16], [19], [20], [21], [22] and tracking [1]. Several of these datasets use mining procedure [6], [5] that heuristically searches the dataset for rare and interesting scenarios. The resulting collection of such scenarios and behaviors, however, is only a fraction of the entire dataset. METEOR, by comparison, exclusively contains such scenarios due to the inherent nature of the unstructured traffic in India.

METEOR has many additional characteristics with respect to these datasets. For instance, METEOR's 2.02 million annotated frames are more than $10\times$ the current highest number of annotated frames with respect to other dataset with high density traffic (ApolloScape). Furthermore, METEOR consists of 16 different traffic-agents that include only on-road moving entities (and not static obstacles). This is by far, the most diverse in terms of class labels. In comparison, Argoverse and nuScenes both contain 10 and 13 traffic-agents, respectively. METEOR is the first motion forecasting and behavior prediction dataset with traffic patterns from rural and urban areas that consist of unmarked roads and high-density traffic. In contrast, traffic scenarios in Argoverse, Waymo, Lyft, and nuScenes have been captured on sparse to medium density traffic with well-marked structured roads in urban areas.

B. Semantic Segmentation Datasets

CityScapes [11] is widely used for several tasks, primarily semantic segmentation. It is based on urban traffic data collected from European cities with structured roads and low traffic density. In contrast, the Indian Driving Dataset (IDD) [12] is collected in India with both urban and rural areas with high-density traffic. A common aspect of both these datasets (CityScapes and IDD), however, is the relatively low

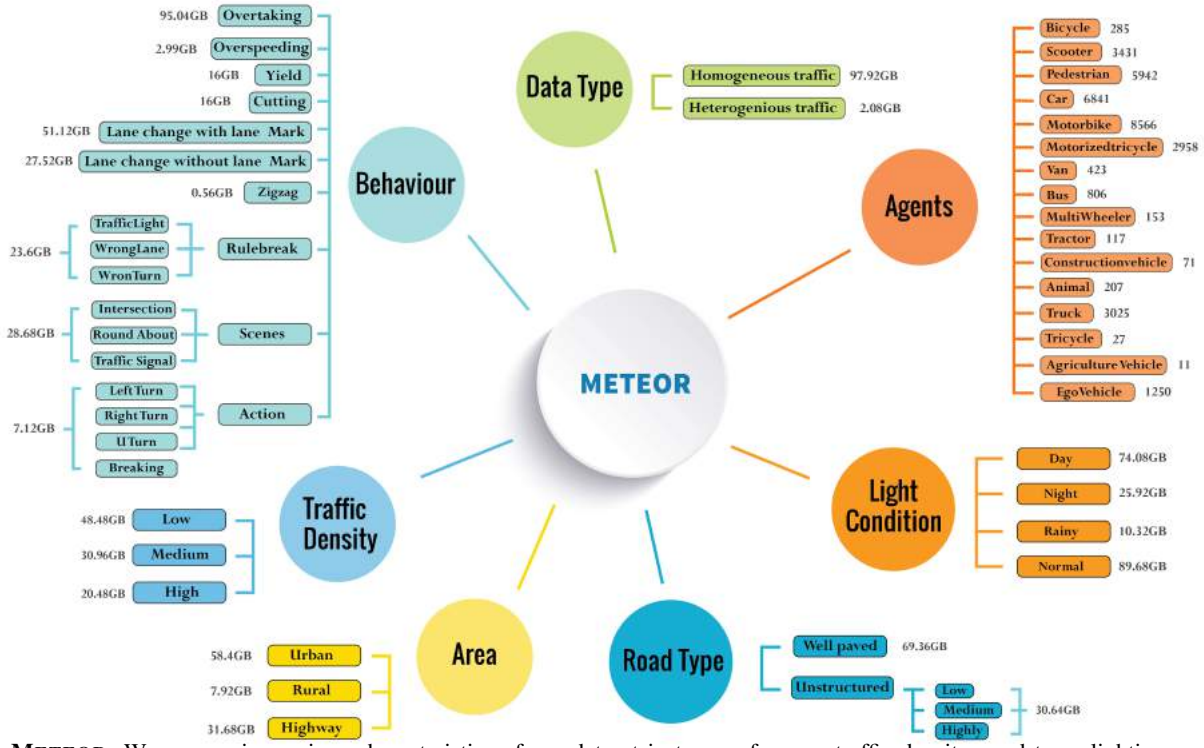


Fig. 1: **METEOR:** We summarize various characteristics of our dataset in terms of scene: traffic density, road type, lighting conditions, agents (we indicate the total count of each agent across 1250 videos), and behaviors, along with their size distribution (in GB). The total size of the current version of the dataset is around 100GB, and it will continue to expand. Our dataset can be used to evaluate the performance of current and new methods for perception, prediction, behavior analysis, and navigation based on some or all of these characteristics. Details of the organization of our dataset are given at **https://gamma.umd.edu/meteor**.

annotated frame count (25K and 10K, respectively). This is probably due to the effort involved with annotating every pixel in each image. IDD also contains high-density traffic scenarios in rural areas, similar to METEOR. However, our dataset has 200× the number of annotated frames and 1.6× the number of traffic-agent classes. Similar to TRAF, the IDD does not contain behavior data.

C. Behavior Prediction

Behavior prediction corresponds to the task of predicting turns (right, U-turn, or left), acceleration, merging, and braking in addition to driver-intrinsic behaviors such as over-speeding, overtaking, cut-ins, yielding, and rule-breaking. The two most prominent datasets for action prediction include the Honda Driving Dataset (HDD) [13] and the BDD dataset [17]. Some of the major distinctions between METEOR and the HDD in terms of size (approximately 10×), the availability of scenes with night driving and rainy weather, and the inclusion of unstructured environments in low-density traffic. The BDD dataset [17] contains more annotated samples than METEOR, however, the BDD dataset contains 100K videos while METEOR contains 1K videos. So the number of annotated samples per video is 66× higher for METEOR. The annotations in prior datasets are limited to actions and do not contain the rare and interesting behaviors contained in METEOR.

III. METEOR DATASET

Our dataset is summarized in Figure 1 and visually shown in Figure 2. Below, we present some details of the data collection process and discuss some of the salient characteristics of METEOR.

A. Dataset Collection and Organization

The data was collected in and around the city of Hyderabad, India within a radius of 42 to 62 miles. Several outskirts were chosen to cover rural and unstructured roads. Our hardware capture setup consists of two wide-angle Thinkware F800 dashcams mounted on an MG Hector and Maruti Ciaz. The camera sensor has 2.3 megapixel resolution with a 140° field of view. The video is captured in full high definition with a resolution of 1920×1080 pixels at a frame rate of 30 frames per second. The dashcam is embedded with an accurate positioning system that stores the GPS coordinates, which were processed into the world frame coordinates. The sensor synchronizes between the camera and the GPS. Recordings from the dashcam are streamed continuously and are clipped into 1 minute video segments.

The dataset is organized as 1250 one-minute video clips. Each clip contains static and dynamic XML files. Each static file summarizes the meta-data of the entire video clip including the behaviors, road type, scene structure etc. Each dynamic file describes frame-level information such as bounding boxes, GPS coordinates, and agent behaviors. Our dataset can be searched using helpful filters that sort the data according to the road type, traffic density, area, weather, and behaviors. We also provide many scripts to easily load the data after downloading.

(a) Cut-ins/Jaywalking. (b) Yielding/Cut-ins.

(c) Overtaking/Overspeeding. (d) Driving in wrong lane.

(e) Running red traffic lights. (f) Ignoring lane signs/wrong lane driving.

(g) High density. (h) Rainy weather. (i) Night time. (j) Rural areas.

Fig. 2: **Annotations for rare instances:** One of the unique aspects of METEOR is the availability of explicit labels for rare and interesting instances including atypical interactions, traffic violations, and diverse scenarios. These annotations can be used to benchmark new methods for object detection and multi-agent behavior prediction.

B. Annotations

We manually annotated the videos using the Computer Vision Annotation Tool (CVAT) and provide the following labels: (i) bounding boxes for every agent, (ii) agent class IDs, (iii) GPS trajectories for the ego-vehicle, (iv) environment conditions including weather, time of the day, traffic density, and heterogeneity, (v) road conditions with urban, rural, lane markings, (vi) road network including intersections, roundabouts, traffic signal, (vii) actions corresponding to left/right turns, U-turns, accelerate, brake, $(viii)$ rare and interesting behaviors, and (ix) the camera intrinsic matrix for depth estimation to generate trajectories of the surrounding vehicles. This set of annotations is the most diverse and extensive compared prior datasets.

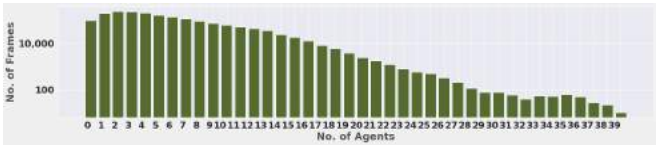
A diverse and rich taxonomy of agent categories is necessary to ensure that autonomous driving systems can detect different types of agents in any given scenario. Towards that goal, datasets for autonomous driving are designed or captured to achieve two goals: (a) capture as many different types of agent categories as possible; (b) capture as many instances of each category as possible. In both these aspects, METEOR outperforms all prior datasets. We annotate 16 types of moving traffic entities with rare and interesting behaviors. Note specifically that the percentages of pedestrians, motorbikes, and bicycles are higher than the percentage of passenger vehicles. This is particularly useful as the former categories are known as "vulnerable road users" (VRUs) [23], and it is important for autonomous driving systems to be able to detect them–necessitating many instances of these VRUs in any dataset.

C. Rare and Interesting Behaviors

We provide a total of 17 different types of rich collection of rare and interesting cases that are unique to our dataset. They are be visually shown in Figure 2. They can be summarized in terms of the following groups:

1) Atypical Interactions: Atypical interactions correspond to pairwise interactions among traffic agents that are not often observed in regular traffic scenarios. Some examples of atypical interactions include yielding to, and cutting across, pedestrians, zigzagging through traffic, pedestrian jaywalking, overtaking, sudden lane changing, and overspeeding. We describe these in more detail below:

- *Overtaking (OT)*: When an agent overtakes another agent with sudden or aggressive movement.
- *Overspeeding (OS)*: If the vehicle over-speeds (based on speed limits) due to any reason.
- *Yield (Y)*: A pedestrian, bicycle, or any slow-moving agent trying to cross the road in front of another agent. If the latter slows down or stops, letting them cross the road then such behavior is labeled as yield.
- *Cutting (C)*: When pedestrians, bicycles, or any slow-moving agents trying to cross the road is interrupted by another agent. Yielding and cutting can also be re-labeled as instances of jaywalking. In a majority of these

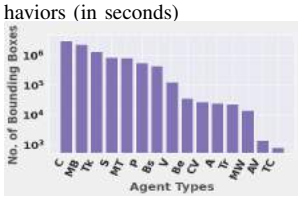


(a) **High traffic density:** METEOR has up to 40 agents per frame.

(b) Number of scenes in which behaviors occur

(c) Average scene duration of behaviors (in seconds)

(d) **High heterogeneity:** Up to 9 unique agents in a single frame.



(e) **Rich features:** Up to 13 million boxes.

Fig. 3: We highlight the high traffic density, heterogeneity, and the richness of behavior information in METEOR. Abbreviations correspond to various behavior categories and are explained in Section III-C

cases, one of the agents involved is a pedestrian crossing the road in the middle of traffic.

- *Lane change w. lane markings (LC(m))*: Agents aggressively change lanes on roads with clear lane markings.
- *Lane change w/o. lane markings (LC)*: Agents aggressively change lanes on roads without lane markings.

The above two annotations can be used to identify videos in the dataset that contain roads without lane markings for relevant applications.

- *Zigzagging (ZM)*: If any of the agent of interest undergoes a zigzag movement in the traffic, the agent behavior is classified as zigzagging.

2) Traffic Violations: In addition to the above driving behaviors, we also annotate traffic agents breaking traffic rules. These are particularly unique since rule breaking scenarios are rare.

- *Running a traffic light (RB TL)*: Passing through an intersection even though the traffic signal is red.
- *Wrong Lane (RB WL)*: A road may not be divided for inbound and outbound traffic by a physical barrier, making it possible for the motorists to use the inbound lane for the outbound traffic and vice versa. This behavior identifies all such cases.
- *Wrong Turn (RB WT)*: When an agent makes an illegal turn (including U-turns).

3) Diverse Scenarios: Finally, we provide annotations for challenging scenarios that include intersections, roundabouts, traffic signals, executing left turns, right turns, and U-turns.

D. Dataset statistics

We analyze the dataset statistics and distribution of agents and their behaviors in terms of total count, uniqueness, and duration (in seconds). Figures 3a and 3d show that METEOR is very dense and highly heterogeneous, respectively; the total number of agents in a single frame can reach up to 40 and up to 9 unique agents can exist in a single frame. Figure 3b represents the distribution of behaviors across videos and Figure 3c shows the distribution of each behavior's average duration. In particular, we note that the average duration can reach up to 3 seconds which, at 30 frames per second, corresponds to approximately 90 frames that contain visual, contextual, and semantic information that can inform behavior prediction algorithms for more accurate perception and prediction.

TABLE II: **Training Details for Object Detection** (BS: Batch size, Mom: Momentum, WD: Weight decay, MGN: Max Gradient Norm)

Method	Backbone	BS	Opt.	LR	Mom.	WD (L_2)	MGN
DETR [24]	ResNet-50	2	AdamW	1e−4	–	1e−4	0.1
Def. DETR [25]	ResNet-50	2	AdamW	2e−4	–	1e−4	0.1
YOLOv3 [26]	Darknet-53	8	SGD	1e−3	0.9	5e−4	35
CenterNet [27]	ResNet-18	16	SGD	1e−3	0.9	5e−4	35

IV. USING METEOR TO EXTRACT NEW INSIGHTS IN PERCEPTION IN UNSTRUCTURED TRAFFIC

We provide the pre-trained models for object detection and behavior prediction at **https://gamma.umd.edu/meteor**.

A. 2D Object Detection

Our main goal is to leverage the unique aspects of METEOR to address the following questions:

1) How do static scene features (location, traffic density, traffic composition, weather, time of the day etc.) affect 2D object detection?
2) How do state-of-the-art 2D object detectors for structured traffic compare to unstructured traffic?

We use the MMDetection [29] toolbox to train the following 2D object detection models—DETR [24], Deformable DETR [25] (with iterative bounding box refinement), YOLOv3 [26] (with scale 608), CenterNet [27] (with normal convolutions), and Swin-T [30]. We provide the training details in Table II and report the mAP, mAP_{50}, mAP_{75}, mAP_{S}, mAP_{M}, and mAP_{L} metrics [31].

Results: Unstructured traffic scenarios are richly diverse in terms of scene-specific features and it is important to understand the impact of static features such as the road conditions, location, weather conditions, time of the day, and types of agents on 2D object detection. METEOR facilitates extensive empirical analysis along these lines. In Table III, **bolded** attributes observed high detection accuracy. We immediately spot some expected trends such as object detection being better in the day time and in clear weather. We also, however, note some new observations: modern deep learning-based 2D object detection is susceptible to both low- and high-density traffic in rural areas with uniform agents.

Furthermore, in Table IV we observe that the most widely used 2D object detectors, that perform well on the Waymo Open Motion Dataset [6] and the KITTI dataset [28], do not perform well on METEOR. More specifically, the detectors achieve $37\% - 65\%$ and $23\% - 81\%$ mAP on the Waymo and KITTI datasets, respectively, while the same methods achieve $8\% - 31\%$ mAP on the METEOR dataset. In other words, the best possible result on METEOR is $\frac{1}{2}\times$ and $\frac{1}{3}\times$ the

TABLE III: **Effect of meta features on object detection:** We analyze how meta features such as traffic density, type of agents, location, time of the day, and weather play a role in 2D object detection using the DETR, Deformable DETR, YOLOv3 and CenterNet object detectors. **Bold** indicates the type of meta feature that is the most effective for object detection.

DETR and Deformable DETR (in parentheses)											
	Density			Agents		Environment		Time		Weather	
	Low	**Medium**	High	**Mixed**	Uniform	**Urban**	Rural	**Day**	Night	**Clear**	Rainy
mAP	19.00 (22.70)	27.00 (38.30)	19.30 (28.10)	27.00 (38.30)	14.80 (31.30)	27.00 (38.30)	14.20 (25.70)	27.00 (38.30)	12.00 (20.60)	27.00 (38.30)	12.00 (20.90)
mAP_{50}	33.33 (36.80)	48.40 (61.80)	32.40 (41.40)	48.40 (61.80)	31.80 (44.30)	48.40 (61.80)	23.40 (34.90)	48.40 (61.80)	22.70 (36.10)	48.40 (61.80)	21.90 (32.70)
mAP_{75}	21.50 (22.10)	28.10 (41.50)	20.40 (31.30)	28.10 (41.50)	11.70 (37.00)	21.80 (41.50)	16.30 (28.40)	28.10 (41.50)	12.20 (20.50)	28.10 (41.50)	12.60 (22.90)
mAP_S	2.60 (7.10)	1.20 (12.10)	0.20 (2.50)	1.20 (12.10)	0.30 (12.80)	1.20 (12.10)	2.00 (10.30)	1.20 (12.10)	0.10 (0.30)	1.20 (12.10)	1.80 (9.50)
mAP_M	7.40 (25.20)	8.30 (22.50)	10.50 (16.90)	8.30 (22.50)	7.20 (34.30)	8.30 (22.50)	11.70 (28.10)	8.30 (22.50)	3.30 (12.50)	8.30 (22.50)	6.20 (19.90)
mAP_L	25.60 (24.90)	45.90 (54.10)	24.70 (35.60)	45.90 (54.10)	40.30 (57.80)	45.90 (54.10)	26.30 (35.60)	45.90 (54.10)	16.70 (27.80)	45.90 (54.10)	15.10 (23.80)
YOLOv3 and CenterNet (in parentheses)											
	Density			Agents		Environment		Time		Weather	
	Low	**Medium**	High	**Mixed**	Uniform	**Urban**	Rural	**Day**	Night	**Clear**	Rainy
mAP	19.20 (22.90)	30.40 (32.90)	21.10 (23.30)	30.40 (32.90)	19.10 (30.20)	30.40 (32.90)	13.80 (13.60)	30.40 (32.90)	13.30 (15.90)	30.40 (32.90)	13.40 (14.00)
mAP_{50}	36.90 (34.80)	52.50 (55.40)	36.30 (32.50)	52.50 (55.40)	35.10 (43.40)	52.50 (55.40)	22.00 (22.70)	52.50 (55.40)	25.00 (25.70)	52.50 (55.40)	25.00 (22.50)
mAP_{75}	16.10 (28.10)	32.30 (33.40)	23.20 (26.70)	32.30 (33.40)	19.70 (37.30)	32.30 (33.40)	15.70 (13.20)	32.30 (33.40)	13.40 (27.00)	32.30 (33.40)	13.60 (15.50)
mAP_S	2.70 (8.40)	2.40 (13.10)	0.60 (2.90)	2.40 (13.10)	7.90 (19.30)	2.40 (13.10)	5.20 (5.40)	2.40 (13.10)	0.00 (0.90)	2.40 (13.10)	1.30 (10.90)
mAP_M	14.10 (26.20)	13.10 (30.50)	11.70 (17.60)	13.10 (30.50)	19.10 (38.80)	13.10 (30.50)	22.50 (25.80)	13.10 (30.50)	7.50 (11.60)	13.10 (30.50)	11.60 (17.40)
mAP_L	23.70 (29.50)	48.70 (44.60)	27.30 (27.90)	48.70 (44.60)	38.90 (40.00)	48.70 (44.60)	21.20 (21.40)	48.70 (44.60)	18.50 (21.70)	48.70 (44.60)	16.40 (14.30)

TABLE IV: **Object detection on Waymo and KITTI:** We report the standard mAP for many widely used methods on autonomous driving datasets.

	DETR [24]	CenterNet	YOLO v3	Def. DETR	Swin-T
KITTI [28]	23.00	80.40	81.60	42.20	–
Waymo [6]	65.31	64.83	56.93	65.31	37.20
METEOR	8.30	12.10	14.30	15.80	32.50

TABLE V: **Swin-T on Waymo and METEOR:** We present a more detailed analysis of Swin-T, one of the state-of-the-art object detection approaches, on Waymo and METEOR.

Dataset	mAP	mAP_{50}	mAP_{75}	mAP_S	mAP_M	mAP_L
Waymo [6]	37.20	70.60	52.00	17.20	41.80	67.20
METEOR	32.50	46.00	36.20	22.30	35.20	49.40

best result on the Waymo and KITTI datasets, respectively. In Table V, we compare METEOR in depth with the Waymo dataset using the Swin-T method [30], which is currently one of the top performing methods on the standard COCO 2D object detection benchmark leaderboard [32]. The Swin-T method performs 14% better on the Waymo Dataset.

We reiterate that our goal in this paper is *not* to improve object detection on the METEOR dataset, rather, it is to show that perception performance degrades significantly in unstructured traffic scenarios. Investigating the causes of this degradation and exploring ways to improve object detection in unstructured environments is beyond the scope of this work, but is a promising direction of future work.

B. Multi-Agent Behavior Recognition

The METEOR dataset is ideal for spatio-temporal multi-agent behavior recognition due to the availability of bounding box annotations and their corresponding behavior labels for more than 1231 one-minute video clips and over 2 million annotated frames. We use 1000 video clips for training and 231 video clips for testing. As the guidelines of the benchmarks, we evaluate 16 behavior classes with mean Average Precision (mAP) as the metric, using a frame-level IoU threshold of 0.5. We use the ActorContext-Actor Relation Network (ACAR-Net) [33] which builds upon a novel high-order relation reasoning operator and an actor-context feature bank for indirect relation reasoning for spatio-temporal action localization.

TABLE VI: **ACAR-Net on METEOR:** PT: pre-train, BS: batch size, Opt.: Optimization, LR: learning rate, WD: weight decay, FR(RX-101): Faster R-CNN (ResNeXt-101), Kin.-700: Kinetics-700, CR (Swin-T): Cascade R-CNN (Swin-T)

Dataset	Detector	PT	BS	Opt.	LR	WD	mAP
METEOR	CR(Swin-T)	Kin.-700	32	*SGD*	0.008	1e−7	6.10

Results: In Table VI, we show that ACAR-Net yields 6.1% mAP on METEOR. We hypothesize that several reasons cause this degradation: (i) METEOR consists of 16 different categories of agents from vehicles to animals, most of which are novel for most detectors and therefore hard to detect, (ii) the movements of the agents on the road are very fast, making them hard to capture, and (iii) different agents have different motion patterns; pedestrians move differently than vehicles and buses move differently than motorbikes. All of these factors collectively contribute to the complexity of multi-agent behavior recognition in dense, heterogeneous, and unstructured traffic scenarios. Our experiments show that there is much room for improvement and our hope with METEOR is that it provides the research community the resources it needs to tackle this important problem.

V. Conclusion, Limitations and Future Work

We present a new dataset, METEOR, for autonomous driving applications in dense, heterogeneous, and unstructured traffic with rare and interesting scenarios. We found that current object detection and behavior prediction models fail on the METEOR, necessitating development of sophisticated and robust perception for unstructured scenarios.

Our dataset has some limitations. Currently, we do not provide trajectory information from a fixed reference frame. One would have to use depth estimation techniques to extract such trajectories. Furthermore, our dataset does not contain HD maps and pointcloud data, which are used in many applications. For future work, we hope that our dataset can benefit in terms of design and evaluation of new motion forecasting and behavior prediction algorithms in dense and heterogeneous traffic.

REFERENCES

[1] R. Chandra, U. Bhattacharya, T. Randhavane, A. Bera, and D. Manocha, "Roadtrack: Tracking road agents in dense and heterogeneous environments," 2019.

[2] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 961–971.

[3] R. Chandra, A. Bera, and D. Manocha, "Stylepredict: Machine theory of mind for human driver behavior from trajectories," *arXiv preprint arXiv:2011.04816*, 2020.

[4] A. Sadat, S. Casas, M. Ren, X. Wu, P. Dhawan, and R. Urtasun, "Perceive, predict, and plan: Safe motion planning through interpretable semantic representations," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 414–430.

[5] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan *et al.*, "Argoverse: 3d tracking and forecasting with rich maps," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[6] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. Qi, Y. Zhou *et al.*, "Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset," *arXiv preprint arXiv:2104.10133*, 2021.

[7] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platinsky, W. Jiang, and V. Shet, "Level 5 perception dataset 2020," https://level-5.global/level5/data/, 2019.

[8] Y. Ma, X. Zhu, S. Zhang, R. Yang, W. Wang, and D. Manocha, "Trafficpredict: Trajectory prediction for heterogeneous traffic-agents," in *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 33, 2019, pp. 6120–6127.

[9] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," *arXiv preprint arXiv:1903.11027*, 2019.

[10] W. Zhan, L. Sun, D. Wang, H. Shi, A. Clausse, M. Naumann, J. Kummerle, H. Konigshof, C. Stiller, A. de La Fortelle *et al.*, "Interaction dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps," *arXiv preprint arXiv:1910.03088*, 2019.

[11] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[12] G. Varma, A. Subramanian, A. Namboodiri, M. Chandraker, and C. Jawahar, "Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019.

[13] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko, "Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7699–7707.

[14] A. Jain, H. S. Koppula, S. Soh, B. Raghavan, A. Singh, and A. Saxena, "Brain4cars: Car that knows before you do via sensory-fusion deep learning architecture," *arXiv preprint arXiv:1601.00740*, 2016.

[15] Z. Che, G. Li, T. Li, B. Jiang, X. Shi, X. Zhang, Y. Lu, G. Wu, Y. Liu, and J. Ye, "D2-city: A large-scale dashcam video dataset of diverse traffic scenarios," *arXiv preprint arXiv:1904.01975*, 2019.

[16] R. Chandra, U. Bhattacharya, A. Bera, and D. Manocha, "Traphic: Trajectory prediction in dense and heterogeneous traffic using weighted interactions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8483–8492.

[17] X. W. W. X. Y. Chen, F. L. V. M. T. Darrell, F. Yu, and H. Chen, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," *arXiv preprint arXiv: 1805.04687*, 2018.

[18] G. Singh, S. Akrigg, M. Di Maio, V. Fontana, R. J. Alitappeh, S. Khan, S. Saha, K. Jeddisaravi, F. Yousefi, J. Culley *et al.*, "Road: The road event awareness dataset for autonomous driving," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 1036–1054, 2022.

[19] R. Chandra, U. Bhattacharya, C. Roncal, A. Bera, and D. Manocha, "Robusttp: End-to-end trajectory prediction for heterogeneous road-agents in dense traffic with noisy sensor inputs," in *ACM Computer Science in Cars Symposium (CSCS)*, 2019, pp. 1–9.

[20] R. Chandra, T. Guan, S. Panuganti, T. Mittal, U. Bhattacharya, A. Bera, and D. Manocha, "Forecasting trajectory and behavior of road-agents using spectral clustering in graph-lstms," *IEEE Robotics and Automation Letters (RA-L)*, vol. 5, no. 3, pp. 4882–4890, 2020.

[21] H. Zhao, J. Gao, T. Lan, C. Sun, B. Sapp, B. Varadarajan, Y. Shen, Y. Shen, Y. Chai, C. Schmid *et al.*, "Tnt: Target-driven trajectory prediction," *arXiv preprint arXiv:2008.08294*, 2020.

[22] Y. Chai, B. Sapp, M. Bansal, and D. Anguelov, "Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction," *arXiv preprint arXiv:1910.05449*, 2019.

[23] A. Constant and E. Lagarde, "Protecting vulnerable road users from injury," *PLoS medicine*, vol. 7, no. 3, p. e1000228, 2010.

[24] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision (ECCV)*. Springer, 2020.

[25] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

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

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

[28] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR))*, 2012.

[29] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[30] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 10 012–10 022.

[31] R. Padilla, S. L. Netto, and E. A. B. da Silva, "A survey on performance metrics for object-detection algorithms," in *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2020.

[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision (ECCV)*. Springer, 2014.

[33] J. Pan, S. Chen, M. Z. Shou, Y. Liu, J. Shao, and H. Li, "Actor-context-actor relation network for spatio-temporal action localization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 464–474.