

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

Rohan Chandra<sup>1\*</sup>, Xijun Wang<sup>1\*</sup>, Mridul Mahajan<sup>2</sup>, Rahul Kala<sup>2</sup>, Rishitha Palugulla<sup>3</sup>,  
Chandrababu Naidu<sup>3</sup>, Alok Jain<sup>3</sup>, and Dinesh Manocha<sup>1,4</sup>

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 marks the first step towards the development of 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 [7] 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 In-

\*Denotes equal contribution.

<sup>1</sup>Department of Computer Science, University of Maryland, College Park, USA. Corresponding email: [rchandr1@umd.edu](mailto:rchandr1@umd.edu)

<sup>2</sup> Centre for Intelligent Robotics, Indian Institute of Technology, Allahabad, India.

<sup>3</sup> NavAjna Technologies Pvt. Ltd.

<sup>4</sup>Department of Electrical and Computer Engineering, University of Maryland, College Park, USA.

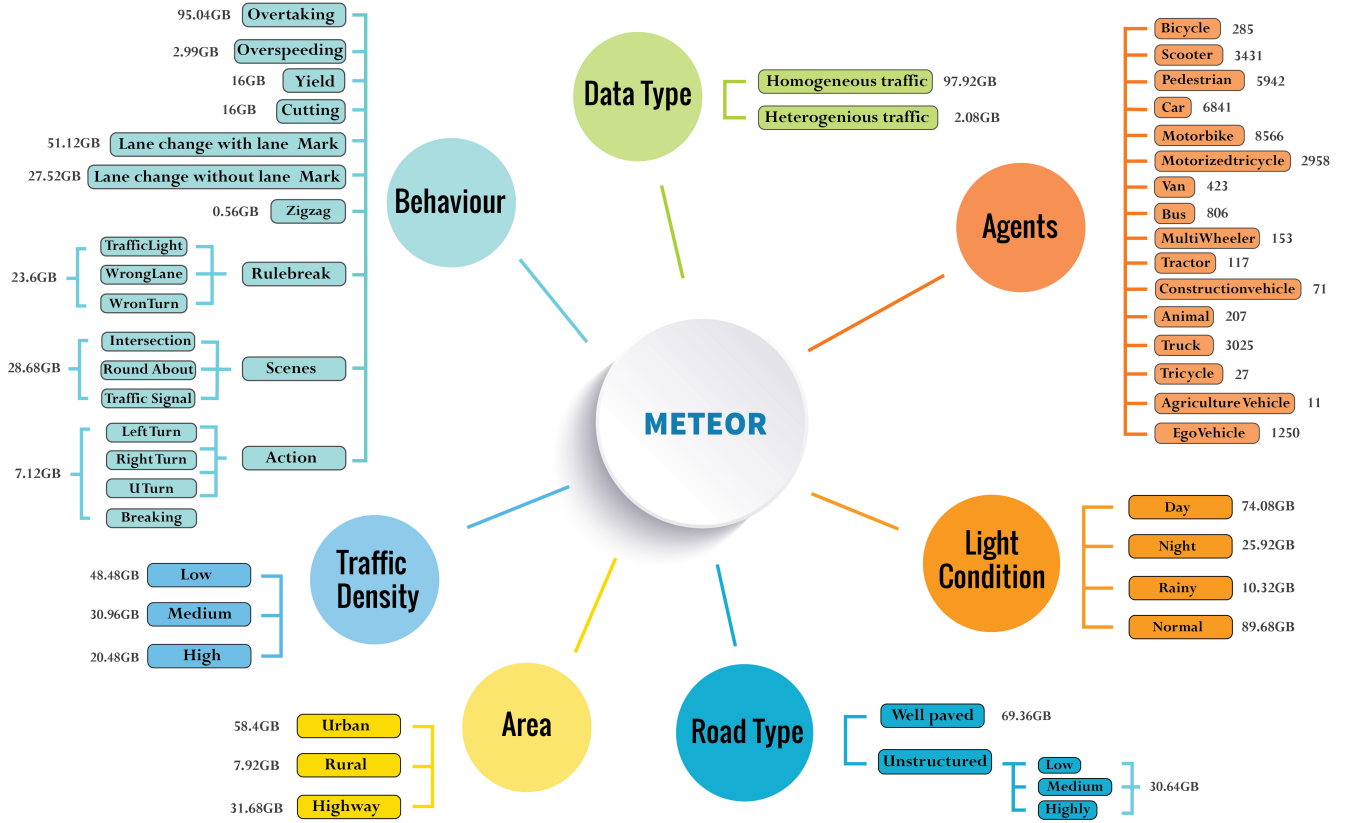


Fig. 1: **METEOR**: We summarize various characteristics of our dataset in terms of scene: traffic density, road type, lighting conditions, agents (we indicate the total count of each agent across 1250 videos), and behaviors, along with their size distribution (in GB). The total size of the current version 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>.

dia. METEOR is the first large-scale dataset containing annotated scenes for rare and interesting instances and multi-agent driving behaviors, broadly grouped into:

- Traffic violations—running traffic signals, driving in the wrong lanes, taking wrong turns).
  - Atypical interactions—cut-ins, yielding, overtaking, overspeeding, zigzagging, lane changing.
  - Diverse scenarios—intersections, roundabouts, and traffic signals.
- 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.
  - Every video in METEOR is tagged using a diverse range of factors including weather, time of the day, road conditions, and traffic density.
  - We evaluate state-of-the-art methods for object detection and multi-agent behavior prediction on METEOR.
  - We present a novel, fine-grained analysis on the relationship between traffic environments and perception. Specifically we study the effect of 2D object detection in varying traffic density, mixture of agents, area, time of the day, and weather conditions.

## B. Applications and Benefits

- Towards Risk-Aware Planning and Control:** Our multi-agent behavior prediction benchmark 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 Robust Perception:** We observe that these models fail in challenging Indian traffic scenarios, compared to their performance on existing datasets captured in the US, Europe, and other developed nations. As a result, METEOR can be a useful benchmark for research in perception in unstructured traffic environments and developing nations.
- Towards Fine-grained Traffic Analysis:** Our novel analysis studying the relationship between traffic patterns and 2D object detection can lead to more informed research in perception for autonomous driving.

## II. COMPARISON WITH EXISTING DATASETS

### A. Tracking and Trajectory Prediction Datasets

Datasets such as the Argoverse [5], Lyft Level 5 [8], Waymo Open Dataset [6], ApolloScape [9], nuScenes



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.

Datasets	Location	Bad weather	Night	Road type	Het.*	Size	Density	Lidar	HD Maps	Rare and Interesting Behaviors <sup>‡</sup>		
										Traffic Violations	Atypical Interactions	Diverse Scenarios
Argoverse [5]	USA	✓	✓	urban	10	22K	Medium	✓	✓	✗	✓	✗
Lyft Level 5 [8]	USA	✗	✗	urban	9	46K	Low	✓	✓	✗	✗	✗
Waymo [6]	USA	✓	✓	urban	4	200K	Medium	✓	✓	✗	✓	✓
ApolloScape [9]	China	✗	✓	urban, rural	5	144K	High	✓	✓	✗	✗	✗
nuScenes [7]	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	Low	✗	✗	✗	✗	✓
<b>METEOR</b>	<b>India</b>	<b>✓</b>	<b>✓</b>	<b>urban, rural<sup>†</sup></b>	<b>16<sup>††</sup></b>	<b>2027K</b>	<b>High<sup>‡</sup></b>	<b>✗</b>	<b>✗</b>	<b>✓</b>	<b>✓</b>	<b>✓</b>

<sup>‡</sup> Rare instances can be broadly grouped into (i) traffic violations, (ii) atypical interactions, and (iii) difficult scenarios.

<sup>†</sup> 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.

<sup>‡</sup> Up to 40 agents per frame.

<sup>††</sup> Up to 9 unique agents per frame.

dataset [7] are used for trajectory forecasting [16], [18], [19], [20], [21] 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× 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 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 the behavior data that is provided by METEOR.

### 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 visually shown in Figure 1. Below, we present some details of the data collection process and discuss some of the salient features and characteristics of METEOR.

### A. Dataset Collection

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



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.

Thinkware F800 dashcams mounted on an MG Hector and Maruti Ciaz. The camera sensor has 2.3 megapixel resolution with a  $140^\circ$  field of view. The video is captured in full high definition with a resolution of  $1920 \times 1080$  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.

### B. Dataset organization

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.

### C. Annotations

We provide the following annotations in our dataset: (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 (See Section III-D), 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, not including static obstacles listed in Figure 1 along with their distribution. 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) [22], and it is important for autonomous driving systems to be able to detect them—necessitating many instances of these VRUs in any dataset.

#### D. 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 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 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.

### E. 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

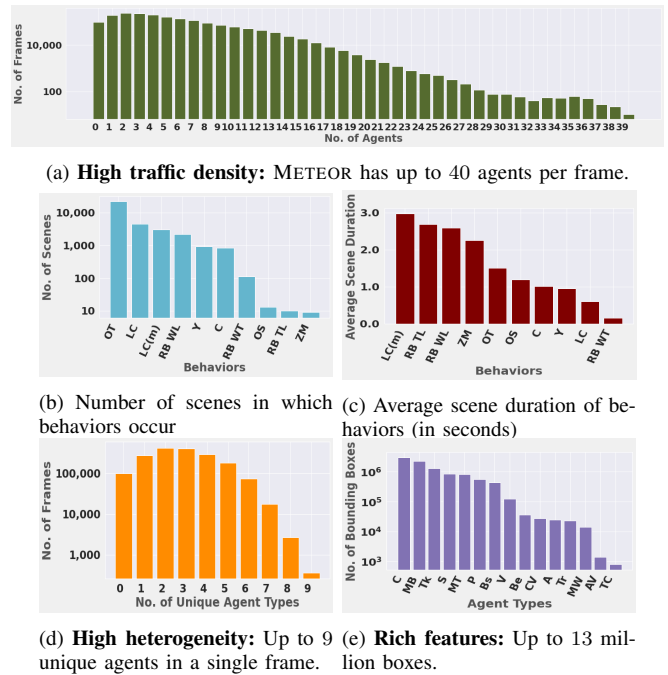


Fig. 3: We highlight the high traffic density, heterogeneity, and the richness of behavior information in METEOR. Abbreviations correspond to various behavior categories and are explained in Section III-D.

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.

## IV. EXPERIMENTS AND ANALYSIS

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

### A. Analyzing Object Detection in Unstructured Scenarios

Existing datasets have helped develop sophisticated and robust 2D detection methods. We use the MMDetection [28] toolbox to train the following 2D object detection models—DETR [23], Deformable DETR [24] (with iterative bounding box refinement), YOLOv3 [25] (with scale 608), CenterNet [26] (with normal convolutions), and Swin-T [29]. The models are pre-trained on the COCO dataset [30] and fine-tuned on METEOR. We provide the training details in Table III and report results using the standard mAP, mAP<sub>50</sub>, mAP<sub>75</sub>, mAP<sub>S</sub>, mAP<sub>M</sub>, and mAP<sub>L</sub>. We refer the reader to [31] for a primer on these metrics.

In Table IV, we report the mAP for the 2D object detectors listed above. We observe that the most widely used 2D object detectors, that perform well on the state-of-the-art



TABLE II: **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	Normal	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 <sub>50</sub>	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 <sub>75</sub>	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 <sub>s</sub>	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 <sub>M</sub>	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 <sub>L</sub>	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	Normal	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 <sub>50</sub>	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 <sub>75</sub>	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 <sub>s</sub>	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 <sub>M</sub>	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 <sub>L</sub>	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 III: **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 [23]	ResNet-50	2	AdamW	1e-4	—	1e-4	0.1
Def. DETR [24]	ResNet-50	2	AdamW	2e-4	—	1e-4	0.1
YOLOv3 [25]	Darknet-53	8	SGD	1e-3	0.9	5e-4	35
CenterNet [26]	ResNet-18	16	SGD	1e-3	0.9	5e-4	35

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

	DETR [23]	CenterNet	YOLO v3	Def. DETR	Swin-T
KITTI [27]	23.00	80.40	81.60	42.20	—
Waymo [6]	65.31	64.83	56.93	65.31	37.20
<b>METEOR</b>	<b>8.30</b>	<b>12.10</b>	<b>14.30</b>	<b>15.80</b>	<b>32.60</b>

autonomous driving datasets, like the Waymo Open Motion Dataset [6] and the KITTI dataset [27], 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 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 [29], which is currently one of the top performing methods on the standard COCO 2D object detection benchmark leaderboard [30]. The Swin-T method performs 14% better on the Waymo Dataset.

There are two possible reasons for performance degradation on METEOR. First, 2D detectors are typically pre-trained on MS COCO [30] and ImageNet [32], which contain only up to 9 categories of the commonly occurring traffic agents. This was not an issue for detectors on existing datasets like Waymo and KITTI since those datasets contain a subset of those 9 classes. METEOR, on the other hand, contains 16 agent categories that are approximately equally distributed. The approximately 7 – 8 traffic agent categories that are contained in METEOR but do not appear in MS COCO are

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.

	mAP	mAP <sub>50</sub>	mAP <sub>75</sub>	mAP <sub>s</sub>	mAP <sub>M</sub>	mAP <sub>L</sub>
Waymo [6]	37.20	70.60	52.00	17.20	41.80	67.20
<b>METEOR</b>	<b>32.60</b>	<b>46.90</b>	<b>36.20</b>	<b>20.50</b>	<b>35.40</b>	<b>54.70</b>

novel to these 2D object detectors and are not classified correctly.

The other reason why object detection deteriorates on METEOR is due to the challenging traffic environments in METEOR. More specifically, METEOR contains many challenging scenarios such as bad weather, nighttime traffic, rural area, high density traffic, etc. (see Figure 2). We analyze the effect of meta-features such as traffic conditions (density and heterogeneity), road conditions, weather, and time-of the day on 2D object detection and present this analysis in Table II. For this analysis, we form separate test sets corresponding to each label in a meta-feature (for example, we have two test sets for day and night). Most datasets contain videos of medium density traffic. In Table II, we see that the performance of the DETR, Deformable DETR, YOLOv3, and CenterNet suffers as the traffic density increases from medium to high. Similar reasoning can be made for other factors—object detection is less effective for homogeneous traffic, in rural areas, at nighttime, and in rainy weather. In most datasets, the number of annotated data samples with these adverse and challenging factors are a fraction of the entire dataset, which partly explains why 2D detectors are more successful on those datasets. The analysis in this section empirically validates the difficulty that the heavy-tail problem poses to perception tasks in autonomous driving.

### B. Multi-Agent Behavior Recognition

Multi-agent behavior recognition (MABR) is the task of first localizing agents in a video followed by classifying their behaviors. This task has drawn attention in recent years and plays an important role in autonomous driving.



TABLE VI: **ACAR-Net on AVA and METEOR:** We applied currently the state-of-the-art multi-agent action recognition approach on AVA to our METEOR dataset. (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
AVA [33]	FR(RX-101)	Kin.-700	32	<i>SGD</i>	0.008	$1e-7$	30.0
<b>METEOR</b>	<b>CR(Swin-T)</b>	Kin.-700	32	<i>SGD</i>	0.008	$1e-7$	<b>6.10</b>

Unlike object detection, which can be accomplished solely by observing visual appearances, MABR reasons about the actors’ interactions with the surrounding context, including environments, other people and objects.

**Dataset Preparation:** The METEOR dataset is ideal for spatio-temporal MABR due to the availability of bounding box annotations and their corresponding behavior labels for more than 1231 video clips, each lasting one minute in duration, 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.

**Framework:** We use the ActorContext-Actor Relation Network (ACAR-Net) [34] 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. This framework is composed of an object detector, backbone network, and ACAR components.

**Object Detector:** For the object detection step, we use the Swin-T detector, generated by combining a Cascade R-CNN [35] with a Swin-T [29] backbone. The model is pre-trained on ImageNet and MS COCO, and fine-tuned on METEOR using the same settings as Swin-T [29]: multi-scale training [36] (resizing the input with the shorter side between 480 and 800 and the longer side at most 1333), AdamW [37] optimizer (initial learning rate of  $1e-4$ , weight decay of 0.05, and batch size of 16), and  $1\times$  schedule (12 epochs).

**Backbone Network:** Following ACAR-Net [34], we use SlowFast networks [38] as the backbone in the localization framework and double the spatial resolution of res5. We conduct experiments using a SlowFast R-101  $8\times 8$ , pre-trained on the Kinetics-700 dataset [39], without non-local blocks. The inputs are 64-frame clips, where we sample  $T = 8$  frames with a temporal stride  $\tau = 8$  for the slow pathway, and  $\alpha T (\alpha = 4)$  frames for the fast pathway.

**Training Settings:** We train ACAR-Net using synchronous SGD with a batch size of 16. For the first 3 epochs, we use a base learning rate of 0.008, which is then decreased by a factor of 10 at iterations 4 epochs and 5 epochs. We use a weight decay of  $1e-7$  and Nesterov momentum of 0.9. We use both ground-truth boxes and predicted object boxes for training. For inference, we scale the shorter side of input frames to 384 pixels and use detected object boxes with scores greater than 0.85 for final behavior classification.

**Results:** We compare METEOR with the AVA dataset [33] as the latter is the state-of-the-art in multi-agent action recognition. In Table VI, we show that the current state-

of-the-art approach, ACAR, achieves 30.0% mAP on AVA but yields 6.1% mAP on METEOR. There are several reasons why ACAR performs better on AVA. AVA focuses exclusively on only one target, humans, a category which most state-of-the-art object detectors can detect with ease. Furthermore, the videos in the AVA dataset consist of high-definition movies, in which agents (actors) are clearly visible, the background is simple, and the movements performed are also exaggerated and easier to identify. METEOR, on the other hand consists of 16 different categories of agents from vehicles to animals, most of which are novel for most detectors and therefore hard to detect. Moreover, the movements of the agents on the road are very fast, making them hard to capture. Finally, different agents have different motion patterns; for example, pedestrians move differently than vehicles and buses move differently than motorbikes. All of these factors collectively contribute to the complexity of MABR in dense, heterogeneous, and unstructured traffic scenarios. Our experiments and analysis 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 scenarios. rain consists of more than 1000 one-minute video clips, over 2 million annotated frames with 2D and GPS trajectories for 16 unique agent categories, and more than 13 million bounding boxes for traffic agents. We found that current models for object detection and multi-agent behavior prediction fail on the METEOR dataset. METEOR marks the first step towards the development of more sophisticated and robust perception models for dense, heterogeneous, and unstructured scenarios.

Our dataset has some limitations. While METEOR contains bounding box information for the surrounding agents, we currently 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. Finally, we hope to include semantic segmentation capability as part of METEOR by providing pixel labels for each object.

## 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] 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.
- [8] 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.
- [9] 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.
- [10] W. Zhan, L. Sun, D. Wang, H. Shi, A. Clausse, M. Naumann, J. Kummerle, H. Königshof, 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, "Trophic: 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. 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] 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.
- [19] 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.
- [20] 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.
- [21] Y. Chai, B. Sapp, M. Bansal, and D. Anguelov, "Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction," *arXiv preprint arXiv:1910.05449*, 2019.
- [22] A. Constant and E. Lagarde, "Protecting vulnerable road users from injury," *PLoS medicine*, vol. 7, no. 3, p. e1000228, 2010.
- [23] 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.
- [24] 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.
- [25] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.
- [26] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.
- [27] 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.
- [28] 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.
- [29] 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. 10012–10022.
- [30] 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.
- [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] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 248–255.
- [33] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar *et al.*, "Ava: A video dataset of spatio-temporally localized atomic visual actions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6047–6056.
- [34] 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.
- [35] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6154–6162.
- [36] 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, pp. 213–229.
- [37] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.
- [38] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 6202–6211.
- [39] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A short note on the kinetics-700 human action dataset," *arXiv preprint arXiv:1907.06987*, 2019.