Research Article

DAVE: Diverse Atomic Visual Elements Dataset with High Representation of Vulnerable Road Users in Complex and Unpredictable Environments

Xijun Wang[1], Pedro Sandoval-Segura[1], Chengyuan Zhang[1], Junyun Huang[1], Tianrui Guan[1], Ruiqi Xian[1], Fuxiao Liu[1], Rohan Chandra[2], Boqing Gong[3], Dinesh Manocha[1]

1. University of Maryland, College Park, United States; 2. University of Virginia, United States; 3. Boston University, United States

Most existing traffic video datasets e.g. the Waymo Open Motion Dataset (WOMD)[1], are collected in Western countries and consist of simple, structured, and predictable traffic. Most Asian scenarios, however, are far denser, more unstructured, and more heterogeneous, with many road-agents routinely disobeying common traffic rules. Consequently, state-of-the-art computer vision and autonomous driving perception algorithms trained on existing datasets do not transfer to traffic in Asian countries. Addressing this gap, we present a new dataset, DAVE, designed for evaluating perception methods with high representation of Vulnerable Road Users (VRUs: e.g. pedestrians, animals, motorbikes, and bicycles) in complex and unpredictable environments. DAVE is a manually annotated dataset encompassing 16 diverse actor categories (spanning animals, humans, vehicles, etc.) and 16 action types (complex and rare cases like cut-ins, zigzag movement, U-turn, etc.), which require high reasoning ability. DAVE densely annotates over 13 million bounding boxes (bboxes) actors with identification, and more than 1.6 million boxes are annotated with both actor identification and action/behavior details. The videos within DAVE are collected based on a broad spectrum of factors, such as weather conditions, the time of day, road scenarios, and traffic density. DAVE can benchmark video tasks like Tracking, Detection, Spatiotemporal Action Localization, Language-Visual Moment retrieval, and Multi-label Video Action Recognition. Given the critical importance of accurately identifying VRUs to prevent accidents and ensure road safety, in DAVE, vulnerable road users constitute 41.13% of instances, compared to 23.14% in WOMD. DAVE provides an invaluable resource for the development of more sensitive and accurate visual perception

 doi.org/10.32388/7O9XQ1

algorithms in the complex real world. Our experiments show that existing methods suffer degradation in performance when evaluated on DAVE, highlighting its benefit for future video recognition research.

Corresponding author: Xijun Wang, xijun@umd.edu

1. Introduction

Accurate perception of road users—including vehicles, bicycles, pedestrians, animals, and more—is a critical challenge in autonomous driving[2][3]. Video recognition serves as the cornerstone of perception systems in autonomous driving vehicles. Video recognition research has made significant progress in recent years, enabling successful applications such as autonomous driving[4], surveillance systems, and human-computer interaction. At the core of these advancements lies the development of comprehensive and challenging datasets that facilitate the training, evaluation, and benchmarking of novel algorithms[5][6]. However, the focus has predominantly been on western traffic[1], structured environments[7], featuring human-centric activities[8] and relatively simplistic scenes[9] that, while beneficial, do not encapsulate the breadth of complexities inherent in natural environments[7][9]. This dissimilarity between existing training datasets and the real-world distribution hinders the generalization capabilities of video recognition models, ultimately limiting their effectiveness when applied to multifaceted and unpredictable real-world situations for autonomous driving.

Limitations of existing video recognition datasets include:

- *Few Vulnerable Road Users:* As shown in Table 1, low representation of vulnerable road users in terms of both quantity and diversity.
- *Lack of Unstructured Environments:* Most existing traffic video datasets are structured, focusing predominantly on Western traffic, which hinders global applicability. This lack of real-world complexity, such as cluttered scenes, occlusions, and the variety of interactions, hinders the development of robust perception models.
- *Limited Scope:* For behavior recognition, most existing datasets primarily focus on human actors performing isolated actions (one action in one clip) in simplistic and controlled settings. This narrow scope restricts the ability of models to generalize to diverse scenarios with varying object categories, environmental factors, and complex interactions.

- *Sparse Annotations:* As shown in Table 2, the lack of fine-grained information about object locations, interactions, and temporal relationships hinders the evaluation of various tasks like Spatiotemporal Action Localization and Video Moment Retrieval, which require detailed temporal and spatial annotations.

Name	VRUs Category	Geographical Location	VRUs #	Total #
DoTA[10]	Person, Bike, Rider, Motor	Various Countries (YouTube)	22,117	170,739
ROAD[11]	Pedestrians, Cyclist	United Kingdom	92,347	122,908
TITAN[12]	Pedestrians, 2-wheeled vehicles	Tokyo	498,544	645,384
Waymo[1]	Pedestrians, cyclists	United States	1,808,771	7,817,150
DAVE (Ours)	Animal, Bicycle, MotorBike, MotorizedTricycle,			
MultiWheeler, Pedestrian, Scooter, TriCycle	India	5,352,711	13,012,635	

Table 1. The existing traffic datasets with Vulnerable Road Users and bbox. Total # includes all labeled instances.

Dataset	Action Annotation			Tube Annotation		
	Pedestrain	Vehicle	O-VRUs	Pedestrain	Vehicle	O-VRUs
SYNTHIA[13]	-	-	-	-	-	-
SemKITTI[14]	-	-	-	-	-	-
Cityscapes[15]	-	-	-	-	-	-
A2D2[16]	-	-	-	-	-	-
Waymo[1]	-	-	-	✓	✓	-
Apolloscape[17]	-	-	-	✓	✓	-
PIE[18]	✓	-	-	✓	-	-
TITAN[12]	✓	✓	-	✓	✓	-
KITTI360[19]	-	-	-	-	-	-
A*3D[20]	-	-	-	-	-	-
H3D[21]	-	-	-	✓	✓	-
Argoverse[22]	-	-	-	✓	✓	-
NuScense[23]	-	-	-	✓	✓	-
DriveSeg[24]	-	-	-	-	-	-
ROAD[11]	✓	✓	-	✓	✓	-
Spatiotemporal action detection datasets						
UCF24[25]	✓	-	-	✓	-	-
AVA[7]	✓	-	-	✓	-	-
Multisports[26]	✓	-	-	✓	-	-
DAVE (Ours)	✓	✓	✓	✓	✓	✓

Table 2. Comparison of datasets with respect to pedestrian, vehicle, and other Vulnerable Road Users (O-VRUs) action and tube annotations.

Property	Values
Basic Information	Location: India (urban and semi-urban settings)
Action Types (16)	NormalDriving, Yield, Cutting, LaneChanging(m), OverSpeeding, WrongTurn, TrafficLight, WrongLane,
	ZigzagMovement, LaneChanging, OverTaking, Keep, LeftTurn, RightTurn, UTurn, Breaking
Action Statistics	Max action num per frame: 40, Average action num per frame: 6.7
	Max unique action num per frame: 6, Average unique action num per frame: 2.0
Types of Actors (16)	AgricultureVehicle, Animal, Bicycle, Bus, Car, ConstructionVehicle, EgoVehicle, MotorBike,
	MotorizedTricycle, MultiWheeler, Pedestrian, Scooter, Tractor, TriCycle, Truck, Van
Actor Statistics	Max actor num per frame: 40, Average actor num per frame: 6.5
	Max unique actor num per frame: 10, Average unique actor num per frame: 3.9

Table 3. DAVE Characteristics: We annotate 16 types of actions performed by 16 types of actors. We highlight the maximum and average number of actions and actors per frame. LaneChanging(m) denotes lane changing on roads with clear lane markings.

In order to address the above limitations, we introduce a new dataset, DAVE (Diverse Atomic Visual Elements), which offers more scenarios with less predictable and dense environments. We collected this dataset in India because it provides a richer variety of interactions, particularly with vulnerable road users who share the roads with cars. DAVE introduces new challenges through its inclusion of unique scenes (e.g., more vulnerable road users are on the road, some unique agents like motorized tricycles, tricycles, and animals) not commonly found in datasets from the US, UK, or Europe. Our analysis shows that the data is denser and includes more instances of road users not following traffic rules, occlusions, and other complexities, making it more challenging and harder to predict. These challenges are valuable for enhancing perception models' ability to handle unpredictable environments[27].

In DAVE, every visible object is annotated and considered an atomic visual element. It is specifically designed to evaluate perception methods in unstructured environments that are more indicative of real-world scenarios. The unstructured environments in DAVE cover different geographical landforms, diverse actors (not only humans but also animals, vehicles, etc.), and complex actions (cut-in, overtaking, u-turn, etc.). As shown in Fig. 2, DAVE prioritizes replicating the richness and complexities encountered in real-world situations. We highlight its applicability to various video recognition tasks as shown in Fig. 1, including Tracking, Detection, Video Moment Retrieval, Spatiotemporal Action Localization, and Multi-label Video Action Recognition. In each case, DAVE has its distinctive features and novel challenges. Some key characteristics of DAVE include:

- *Vulnerable Road Users (VRUs):* DAVE has a higher representation of vulnerable road users (VRUs), constituting 41.13% compared to 23.14% in Waymo[1]. This is a precious property to prevent accidents and ensure road safety.
- *Less predictable and Dense Environments:* DAVE features videos captured in diverse real-world settings, encompassing various weather conditions, times of day, road scenarios, and traffic densities. This inherent complexity better reflects the challenges encountered in practical applications.
- *Diverse Actor Categories:* DAVE extends beyond human-centric datasets, incorporating 16 diverse actor categories. This diversity fosters the development of models capable of generalizing beyond a limited set of actor types.
- *Rich Annotations:* DAVE provides dense annotations, including over 13 million bounding boxes (bboxes) for actors and over 1.6 million bboxes encompassing both actor and action details (Table. 3). We also offer actors' GPS information and the keyframe for the action. This comprehensive annotation allows for the evaluation of a wider range of potential tasks.
- **Complex Actions:** Compared with human-centric simple actions (e.g stand, watch, sit, walk), DAVE has more complex actions (e.g. cut-in, overtaking, u-turn, zigzag movement), which require higher reasoning ability for perception models.

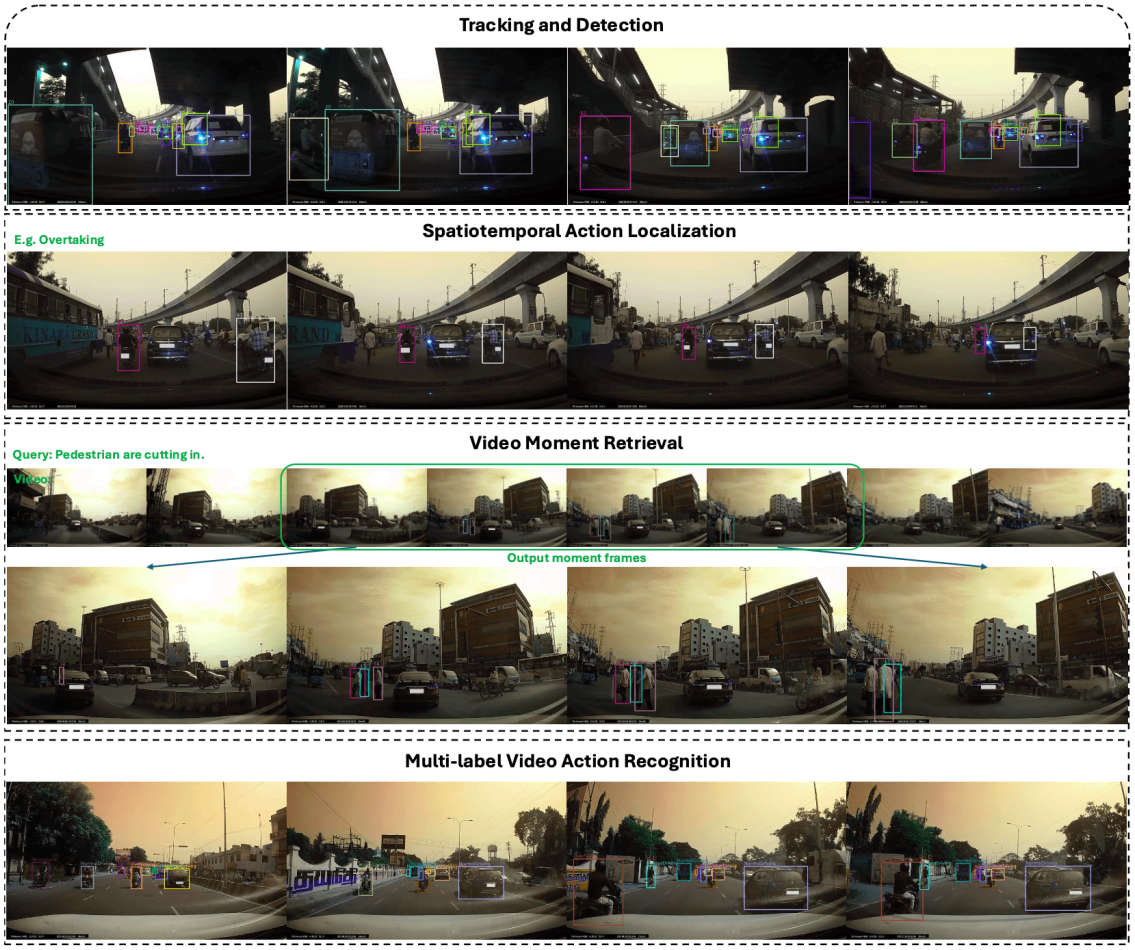

Figure 1. Tasks Overview. We use DAVE for various video recognition tasks, including Tracking, Detection, Video Moment Retrieval, Spatiotemporal Action Localization, and Multi-label Video Action Recognition. Our large-scale dataset is made up of complex environments that are densely annotated. Each bounding box (bbox) corresponds to an actor, and the text above each bbox serves as either the tracking ID or indicates the associated action.

doi.org/10.32388/7O9XQ1

Figure 2. Challenging Characteristics of DAVE: These videos correspond to different times of the day with different brightness, different geographical landforms from city and rural areas, high density and unpredictable road conditions, diverse actors including humans, animals, vehicles, etc.

We highlight the advantages of DAVE for five video tasks:

Tracking: Compared to datasets like MOT17[28], which primarily focus on tracking pedestrians and vehicles in controlled settings, DAVE's diverse actors occur under a variety of illumination conditions and provide a more significant challenge for tracking algorithms[29][30][31]. This allows for the evaluation of robust tracking methods capable of handling occlusions, cluttered scenes, and dynamic environments. Other large instance tracking benchmarks include GOT-10k[32] and VastTrack[33]. GOT-10k contains over 10k video segments to reflect the diversity of real life while VastTrack contains more than 50k videos of more than 2k object classes. But traffic involving high numbers of VRUs is only a small fraction of both datasets. For example, in GOT-10k there are fewer than 200 videos of bicycles, while in VastTrack non-motor vehicles (a type of VRU) make up only 1.5% of the dataset. From our experiments, ARTrack[34] performs 23.7% worse on DAVE than GOT-10k, which highlights the complexity of DAVE as compared to other datasets.

Detection: Datasets like COCO[35] and Pascal VOC[36] have been instrumental in advancing object detection methods[37]. While these datasets include a variety of object categories, they often lack the contextual complexity and scene diversity found in DAVE (e.g. intricate street-scapes at different times of day, higher representation of VRUs, such as pedestrians, animals, motorbikes, and bicycles, compared to vehicles). With its extensive annotations encompassing over 13 million bounding boxes,

DAVE offers a unique challenge to detection algorithms, pushing the boundaries of what these models can recognize and how well they can adapt to diverse and unstructured environments. In our experiments, Swin-T[38] outperforms by 18% on the COCO dataset, as compared to DAVE. This highlights the complexity of DAVE.

Spatiotemporal Action Localization (STAL): Spatiotemporal action localization requires algorithms to not only recognize specific actions but also pinpoint their occurrence within both the spatial and temporal domains of video content. Datasets like AVA[7] have laid the groundwork for this task. It is, however, a movie-human-centric dataset, meaning the video clips in AVA are sourced from movies, which might not perfectly reflect the full diversity of real-world scenarios. This could potentially limit the generalizability of models trained on this dataset. In contrast, DAVE introduces a richer layer of complexity by featuring the actions performed by different actor categories in unstructured settings. This complexity is important for developing models that can understand and interpret actions in a manner that is similar to human perception. In our experiments, ACAR-Net[39] gets 6.3% mAP accuracy on DAVE versus 33.3% on AVA v2.2, which highlights the challenging scenarios in DAVE.

Video Moment Retrieval (VMR): Moment retrieval involves identifying specific moments within a video that correspond to given queries, often described in natural language. While datasets such as DiDeMo[40] are widely used for this task, DAVE consists of videos of more complicated and cluttered environments. These scenarios not only demand accurate video understanding but also necessitates sophisticated language processing capabilities to interpret the queries and localize the relevant moments within real-world video content. In our experiments, CG-DETR[41] obtains 5.1 R1@0.5 on DAVE (versus 58.4 on Charades-STA). This implies that that video moment retrieval is still a challenging problem in the unstructured environment.

Multi-label Video Action Recognition (M-VAR): Multi-label video action recognition is a task that demands the identification of multiple actions within a single video clip. Existing datasets like Charades[42] have been widely used for this video task. DAVE's video segments with multiple actions occurring within the densely populated and unstructured scenes offer a challenging testbed for algorithms. In our experiments, SlowFast[43] gets 41.0 mAP accuracy on DAVE, while achieving 4.2% higher performance on Charedes.

Overall, DAVE offers a valuable resource for researchers aiming to develop robust and generalizable video recognition models that can work well in real-world scenarios. DAVE's rich annotations make it

suitable for evaluating various video recognition tasks.

2. DAVE Dataset

2.1. Data Collection

To meet the requirement, data collection was meticulously executed within a defined geographic perimeter encompassing the urban and suburban zones of India. The selection of numerous suburban locations was strategic, aiming to encompass a broad spectrum of road environments, including both rural pathways and those lacking structured design or layout. To capture this data, our equipment consisted of two wide-angle Thinkware F800 dashcams. These devices were installed on two vehicles, specifically an MG Hector and a Maruti Ciaz, chosen for their operational reliability in diverse road conditions. The dashcams are equipped with sensors boasting a resolution of 2.3 megapixels, alongside a comprehensive 140-degree field of view, ensuring wide coverage of the surrounding environment. Video capture was conducted at a high-definition quality, with a resolution of 1920$\times$ 1080 pixels, and a smooth playback of 30 frames per second was maintained to accurately document the dynamic road conditions.

An integral component of our capture system was the dashcam's embedded positioning technology, which provided precise GPS coordinates. This functionality was essential for the transformation of these coordinates into world frame references, facilitating a coherent geographical mapping of the data collected. Additionally, the system's synchronization capability ensured seamless integration of video and GPS data, enhancing the reliability of the spatial information.

The resulting dataset comprises 1231 video clips, each spanning one minute in duration. These clips are accompanied by corresponding information such as the behaviors observed, the type of road, and the overall scene structure. For granular details at the frame level, we offer bounding boxes, precise GPS coordinates, and the behaviors of moving agents within the frame.

DAVE is methodically organized to support efficient querying, facilitated by a range of filters. Users can refine searches based on criteria such as road type, traffic density, geographic area, prevailing weather conditions, and observed behaviors.

2.2. Annotations

In our research, we undertook a meticulous process of manually annotating video data using the Computer Vision Annotation Tool (CVAT) [44], a widely recognized tool for video and image annotation in the field of computer vision. Our annotation process was comprehensive, covering a broad spectrum of labels that are crucial for the development and evaluation of autonomous driving systems. These labels include:

- **Bounding Boxes:** For each agent visible in the video footage, we provided bounding boxes. These are essential for object detection tasks, enabling algorithms to identify and track the location and dimensions of various agents within the scene.
- **Actions and Maneuvers:** The dataset catalogues specific vehicle actions and maneuvers, including left/right turns, U-turns, overtaking, braking, etc. This is critical for predicting vehicle behavior and for training systems in decision-making.
- **Actor Class IDs:** We classified each agent into distinct categories, assigning a unique class ID to facilitate the differentiation and identification of various types of agents, such as vehicles, pedestrians, and bicycles.
- **Rare and Interesting Behaviors:** We have specifically noted instances of rare and unusual behaviors among traffic participants. Capturing these scenarios is important for preparing autonomous systems to handle edge cases safely.
- **GPS Trajectories for the Ego-Vehicle:** The dataset includes precise GPS trajectories for the ego-vehicle, providing valuable data on its movement and position over time.
- **Environmental Conditions:** Annotations in this category encompass weather conditions, time of day, traffic density, and the diversity of traffic participants. This information is crucial for testing and developing autonomous systems that can operate under a wide range of environmental scenarios.
- **Road Conditions:** We have annotated various aspects of road conditions including whether the environment is urban or rural, the presence and visibility of lane markings, and more. This aids in assessing how different road conditions affect the performance of autonomous driving technologies.
- **Road Network Features:** Detailed annotations of road network features such as intersections, roundabouts, and traffic signals are included. These are vital for navigation algorithms and for understanding traffic flow and driving behaviors in complex road networks.

- **Camera Intrinsic Matrix:** For depth estimation and generating accurate trajectories of surrounding vehicles, we include the camera intrinsic matrix. This technical detail enables the conversion of 2D images into 3D representations, essential for spatial understanding and accurate positioning of objects in relation to the ego-vehicle.

As shown in Fig. 3, our dataset stands out with its wide-ranging and rich taxonomy of agent and action categories. This diversity is crucial for ensuring perception systems can operate safely and efficiently in varied and unpredictable environments[45][46]. Furthermore, our dataset is meticulously designed to capture a wide variety of action categories and a high number of instances within each category. This dual focus on the breadth of agents, action types, and depth of instances allows for more robust and effective training of video recognition models.

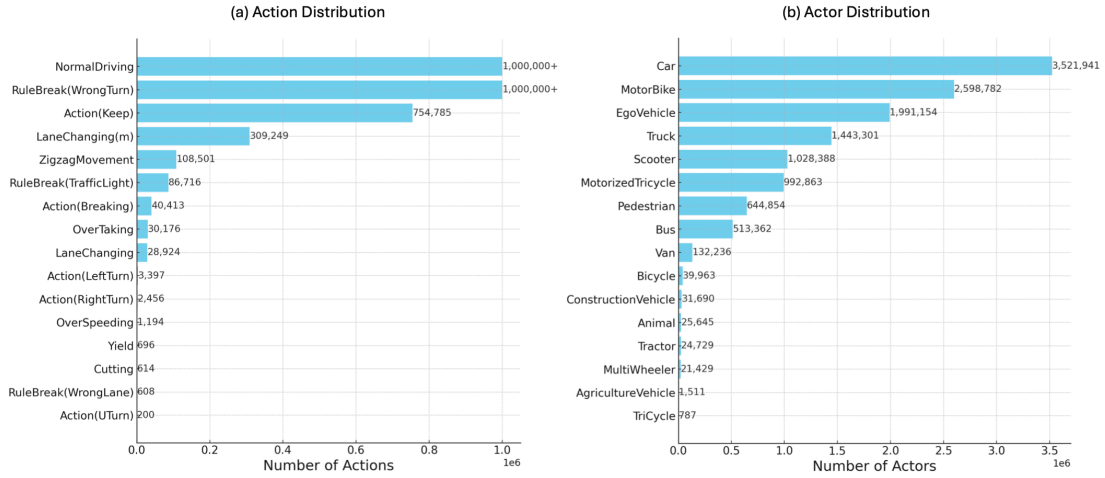

Figure 3. Annotation Statistic. The actor and action distribution for DAVE, includes a wide-ranging and rich taxonomy of 16 agents and 16 action categories. This dual focus on both the breadth of agent and action types and the depth of instances allows for more robust and effective training of video recognition models.

Following the popular Waymo[1] dataset, we obey the widely used data collection and use similar rules. We collected this data for Non-commercial Purposes including the use of the Dataset to perform benchmarking for purposes of academic or applied research publication. To protect privacy and ensure that the identities of pedestrians and other cars are not discernible, we will blur the faces of persons (using Retinaface[47]) and blur the license plates of vehicles[48].

3. Datasets for Different Tasks and Experiments

3.1. Tracking

Dataset Structure: DAVE contains annotations for multiple objects, so we can construct sequences of frames in which the same object is present. Of DAVE's 1231 videos, we can construct 44.8k frame sequences suitable for tracking.

Experiment Setting: To assess visual object tracking on DAVE, we use Autoregressive Visual Tracking (ARTrack)[34], which boasts SOTA performace on GOT-10k[32], TrackingNet[49], LaSOT[50], and $\text{LaSOT}_{\text{ext}}$[51]. We utilize a publicly released "ARTrack-256" checkpoint, pretrained on COCO[52], GOT-10k, LaSOT, and TrackingNet. ARTrack handles single object tracking as a coordinate sequence interpretation task using a template region from an initial frame. ARTrack does not determine when a tracking ID is visible, so we only use sequences of frames in which the same object is present. From the 231 videos in the DAVE validation split, we filter 5227 frame sequences in which one tracking ID is continuously present for at least 60 frames. This filtering of sequences gives ARTrack a slight advantage because it is a harder task to both detect visibility and track over time. Bounding box predictions from ARTrack-256 are compared to ground truth using average area overlap (AO), success rate at 0.5 IoU ($SR_{0.5}$), and success rate at 0.75 IoU ($SR_{0.75}$).

Results: We find that DAVE is comparable to GOT-10k in AO but more challenging for both success rate metrics. For $SR_{0.75}$, ARTrack performs 23.7% worse on DAVE than GOT-10k, despite our preprocessing to keep the same object present in each frame sequence. While ARTrack performs well on the AO metric, the degradation in SR implies that the tracker may generate bboxes larger than the actual object or that it has increased sensitivity to object appearance changes. For example, illumination variations or pose changes can cause inaccurate predictions in some frames even when average overlap remains decent. We believe DAVE becomes even more challenging when one considers the entire video sequence, requiring the tracking of multiple objects as they move in and out of the frame.

Dataset	Sequence number	Annotation	SOTA Performance		
			65.8@HOTA	**81.0@MOTA**	**81.1@IDF1**
MOT17[53]	14	Manual	65.8@HOTA	81.0@MOTA	81.1@IDF1
TAO[54]	2.9k	Manual	47.2@TETA	66.2@LocA	46.2@AssocA
LaSOT[50]	1.4k	Manual	74.0@AUC	82.8@PNor	81.1@P
TrackingNet[49]	30.0k	Semi-auto	86.1@AUC	90.4@PNor	86.2@P
GOT-10k[32]	10.0k	Manual	79.5@AO	87.8@SR50	79.6@SR75
DAVE	**44.8k**	**Manual**	**72.6@AO**	**70.2@SR50**	**47.2@SR75**

Table 4. Comparison of Various Tracking Datasets. DAVE is comparable to GOT-10k in AO but more challenging for both success rate metrics. For SRO.75, ARTrack performs 23.7% worse on DAVE than GOT-10k, despite our preprocessing to keep the same object present in each frame sequence.

3.2. Detection

Dataset Structure: For detection, we have 13 million annotated bounding boxes with identifying actors in 16 categories. We prepare them in COCO format.

3.2.1. Vulnerable Road Users Detection

Experiment Setting: We used the YOLOv8[55] for Vulnerable Road User detection, training three models on three different datasets: Waymo[1], DAVE, and a combined Waymo + DAVE dataset. The models were trained for 30 epochs with batch sizes of 32, an image size of 640, all default on all other parameters as in[55]. All models were evaluated on the DAVE validation set.

Results: As shown in Table 5, the model trained on our DAVE training set outperforms the one trained on the Waymo training set by 20.8% in terms of mAP50. When combining the Waymo and DAVE training sets, the model achieves a 24.0% improvement over Waymo alone and a 3.2% improvement over DAVE alone. The results demonstrate that DAVE is significantly more challenging than Waymo in terms of Vulnerable Road User detection. Looking ahead, we foresee efforts to integrate datasets like

Waymo and DAVE to build a more globally representative traffic dataset. DAVE is a crucial step in this direction.

Dataset	Training Instances Number	Precision	Recall	mAP50	mAP50-95
Waymo[1]	1,808,771	0.00772	0.0236	0.00266	0.00194
DAVE	5,352,711	0.606	0.245	0.235	0.159
Dave + Waymo	7,161,482	0.604	0.261	0.267	0.179

Table 5. Comparison of VRUs Datasets. Compared with Waymo and Waymo + DAVE with the same setting, DAVE training set outperforms the Waymo training set by 20.8% in terms of mAP50. When combining the Waymo and DAVE training sets, the model achieves a 24.0% improvement over Waymo alone and a 3.2% improvement over DAVE alone. The results show that our DAVE dataset is more challenging than Waymo.

3.2.2. Full Dataset Detection

Experiment Setting: For the object detection step, we use the Swin-T detector, generated by combining a Cascade R-CNN[56] with a Swin-T[38] backbone. The model is pre-trained on ImageNet and MS COCO, and fine-tuned on DAVE using the same settings as Swin-T[38]: multi-scale training[57] (resizing the input with the shorter side between 480 and 800 and the longer side at most 1333), AdamW optimizer (initial learning rate of $1e-4$, weight decay of 0.05, and batch size of 16), and $1\times$ schedule (12 epochs).

Results: In this paper, our objective is not to enhance object detection within the DAVE dataset. Instead, we aim to demonstrate the decline in perception performance in unstructured situations. Delving into the reasons behind this performance drop and identifying methods to better object detection in these chaotic environments is not covered in our current research community. The results show that our DAVE dataset is more challenging than the existing datasets.

Dataset	Bbox #	Size	Frame #	Annotation	Weather	Country	SOTA (mAP)
COCO[52]	2.5M	Variable	330K images	Manual	Various	/	66.0
Pascal VOC[36]	20K	Variable	11K images	Manual	/	/	89.3
Waymo[1]	11M	Variable	/	Manual/Auto	Various	USA	41.6
COCO-Swin-T[52]	2.5M	Variable	330K images	Manual	Various	/	50.5
DAVE	13M	1920x1280	2M images	Manual	Has Bad weather	India	32.5

Table 6. Comparison of Various Detection Datasets. Compared with COCO, with the same setting, Swin-T performs 18% better on the COCO Dataset. The results show that our DAVE dataset is more challenging than the existing datasets.

3.3. Video Moment Retrieval

Dataset Structure: For the Video Moment Retrieval task, we annotated 26863 queries, 21,477 for training, and 5,386 for testing. Our query is like "Car is doing lane changing with clear lane markings.", "MotorBike runs in the wrong lane.", "Motorized Tricycle is overtaking.". Those queries are very challenging since some actors are not usual in most visual encoder training data. The actions require the reasoning of the actor, the nearby agents, and the environment.

Experiment Setting: Following CG-DETR[41] on Charades-STA, we utilize slowfast and CLIP backbone features. The model is trained with a batch size of 32 over 200 epochs, employing a learning rate of 2×10^{-4} without any learning rate drop. To accommodate adaptive cross-attention mechanisms, 45 dummy tokens are utilized. The selection process for moment-representative saliency involves pooling 10 candidates, from which 2 are chosen. The architecture includes 3 transformer encoder layers, 3 transformer decoder layers, and 2 layers each for adaptive cross-attention and dummy encoding. Additionally, there is 1 layer each dedicated to moment and sentence encoding. The loss function coefficients are set uniformly to 1 for most, except for highlight detection and distillation where they are increased to 4 and 10 respectively, to emphasize their importance in the training process. These settings are meticulously chosen to enhance the model's ability to understand and generate accurate moment retrievals.

Results: As shown in Table 7, R1@0.5 refers to a metric that evaluates the model's ability to rank the most relevant moment within the top 1 results, with a minimum overlap of 50% between the predicted and ground-truth moment durations. The CG-DETR method only gets 5.1 R1@0.5 on DAVE, the perception performance degrades significantly illustrating that Video Moment Retrieval is still a challenging problem in the unstructured environment.

Dataset	Videos #	Queries #	Duration	Domain	Source	R1@0.5
DiDeMo[40]	10,464	40,543	30s	Open	Flickr	33.4
TACOS[58]	127	18,818	296s	Cooking	Lab Kitchen	41.5
ActivityNet-Captions[59]	19,209	71,957	180s	Open	YouTube	60.6
Charades-STA (CG-DETR)[60]	9,848	16,128	31s	Daily activities	Homes	58.4
DAVE (CG-DETR)	1,231	26,863	60s	Open	Self-collected	5.1

Table 7. Statistics of datasets for Video Moment Retrieval task. The CG-DETR method only gets 5.1 R1@0.5 on DAVE (58.4 on Charades-STA), and the perception performance degrades significantly illustrating that Video Moment Retrieval is still a challenging problem in the unstructured environment.

3.4. Spatiotemporal Action Localization

Dataset Structure: The DAVE dataset stands out as a premier choice for Spatiotemporal Action Localization, thanks to its comprehensive provision of bounding box annotations and associated behavior labels, encompassing more than 2 million annotated frames. For Spatiotemporal Action Localization, we set the allocation as 1000 video clips for the training phase and 231 clips designated for the testing process. Adhering to established benchmark protocols, our evaluation encompasses 16 distinct behavior classes, employing the mean Average Precision (mAP) as the evaluation metric, predicated on a frame-level Intersection over Union (IoU) threshold set at 0.5.

Experiment Setting: The spatiotemporal action localization pipeline includes detections and recognition. For the object detection, we use the Swin-T detector in Section 3.2. For recognition network, following ACAR-Net[39], we conduct experiments using a SlowFast R-101, pre-trained on the Kinetics-700 dataset[61], 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. 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 1×10^{-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: As shown in Table 8, ACAR-Net gets 6.3% mAP on DAVE versus 33.3% on AVA v2.2, which shows DAVE is a very challenging dataset and has tremendous room to improve. DAVE's complexity arises from diverse agents (16 categories VS 1 category of other human-centric datasets), fast and varied motion patterns, and dense traffic. It offers valuable resources to improve multi-agent behavior recognition.

Dataset	Bbox #	Instance #	Video #	Actor class	Action class	Resource	SOTA (mAP)
UCF101-24[62]	574k	4,458	3,207	-	24	YouTube	90.3
J-HMDB[63]	32k	928	928	-	21	Movies, YouTube	83.8
AVA v2.2[7]	426k	386k	430	1	80	Movies, YouTube	45.1
AVA v2.1[7]	426k	386k	430	1	80	Movies, YouTube	41.7
MultiSports[26]	902k	37,701	3,200	1	66	YouTube	8.8
AVA v2.2 (ACAR)[7]	426k	386k	430	1	80	Movies, YouTube	33.3
DAVE	1,600k	/	1,231	16	16	self-collected	6.3

Table 8. Spatiotemporal Action Localization. ACAR-Net gets 6.3% mAP on DAVE, which shows DAVE is a very challenging dataset and has tremendous room to improve.

3.5. Multi-label Video Action Recognition

Dataset Structure: DAVE for Multi-label Video Action Recognition dataset is composed of 10,083 videos clips, involving interactions with 16 actors classes in 16 types of driving behavior action classes.

Following the standard split, it has 8,166 training video and 1,917 validation video.

Experiment Setting:

Following SlowFast[43], for the temporal domain, we randomly sample a clip from the full-length video. For the spatial domain, we randomly crop 224×224 pixels from a video, or its horizontal flip, with a shorter side randomly sampled in [256, 320] pixels. Performance is measured in mean Average Precision (mAP).

Results: As shown in Table 9, SlowFast[43] gets 41.0 mAP when using Kinetics-600 pre-trained model on DAVE. SlowFast achieves 4.2% more performance on Charedes, which means DAVE is harder in terms of Multi-label Video Action Recognition task.

Dataset	Size	Video #	Actions per video	Labelled instances	domain	SOTA (mAP)
Charades[42]	/	9,848	6.8	67k	Daily Activities	66.3
Charades (SlowFast) [42]	/	9,848	6.8	67k	Daily Activities	45.2
DAVE (SlowFast)	1920× 1080	10,083	1-13	1.6M	Outdoor Actions	41.0

Table 9. Multi-label Video Action Recognition. SlowFast achieves 4.2% more performance on Charedes than DAVE, which means DAVE is harder.

4. Conclusion

We present a new video dataset, DAVE, which provides a new benchmark for video recognition research on autonomous driving tasks. It is a robust platform for developing, testing, and refining algorithms capable of handling the complexity of real-world environments. DAVE provides more effective/hard data points and more unpredictable scenarios for training models to better recognize and protect vulnerable road users. We believe this contributes to improving the robustness of perception models in complex and unpredictable environments. Through DAVE's diverse actor categories, range of actions, and unstructured nature of its video content, DAVE represents a

significant step forward in the quest for models that can truly understand and interpret the visual world around an ego-car.

Appendix A.

A.1. More Related Datasets

A.1.1. Traffic Dataset

In recent years, numerous traffic datasets have been developed to address perception challenges in autonomous driving systems. These datasets vary in focus, spanning action recognition, tracking, and object detection tasks. Below, we highlight key datasets that contribute to the field, categorized by their annotations and complexity.

SYNTHIA[13] and Cityscapes[15] focus on semantic segmentation for urban driving. SYNTHIA generates synthetic images with pixel-level annotations, offering scalability for training models. In contrast, Cityscapes provides real-world images from over 50 cities, annotated for both pixel-level and instance-level semantic labeling, under diverse illumination and environmental conditions. SemKITTI[14] and A2D2[16] are centered around 3D semantic segmentation using LiDAR data. SemKITTI annotates dense 360° LiDAR scans for scene understanding, while A2D2 combines RGB and LiDAR data to offer large-scale annotations of 3D point clouds in outdoor environments. Waymo[1], KITTI-360[19], and H3D[21] are notable multimodal datasets for object detection and tracking. Waymo includes synchronized LiDAR and camera data with bounding box annotations for pedestrians and vehicles. KITTI-360 extends KITTI with richer 2D and 3D semantic instance annotations, providing a full 360° field of view. H3D focuses on crowded traffic scenes with detailed annotations for multi-object detection and tracking using LiDAR data. Apolloscape[17] and Argoverse[22] provide detailed annotations and high-definition maps to enable self-localization and 3D scene understanding. Apolloscape includes dense semantic point cloud labels, stereo annotations, and precise location data, while Argoverse integrates sensor fusion data and 3D bounding boxes for tasks like tracking and trajectory forecasting. PIE[18] and TITAN[12] specializes in pedestrian behavior prediction and trajectory forecasting. PIE focuses on pedestrian crossing intentions and future motion estimation, while TITAN includes hierarchical annotations for pedestrian and vehicle actions, offering insights into interactive urban traffic scenarios. NuScenes[23] and A*3D[20] are large-scale multimodal

datasets designed for autonomous driving in challenging conditions. NuScenes includes 3D bounding boxes for 23 object classes, with data from cameras, LiDAR, and radar under varying weather and nighttime conditions. Similarly, A*3D addresses highly diverse environments, featuring dense annotations with significant nighttime scenes. DriveSeg[24] and ROAD[11] stand out for their focus on dynamic scene understanding. DriveSeg provides pixel-level semantic labeling for continuous driving scenes, capturing amorphous objects like road construction and vegetation. ROAD introduces road event awareness by annotating agent-action-location triplets for vehicles, pedestrians, and vulnerable road users (VRUs), supporting spatiotemporal action detection.

In comparison to these datasets, DAVE (Ours) is specifically designed to address the challenges of unstructured, high-density traffic environments, with a strong emphasis on vulnerable road users. By annotating complex interactions, rare behaviors, and diverse actor categories, DAVE provides a benchmark for developing robust and generalizable perception models for autonomous driving systems.

A.1.2. Tracking

The field of object tracking has significantly advanced with the development and introduction of various benchmark datasets, which are crucial for evaluating the performance of tracking algorithms. One of the earliest and most widely used datasets is the OTB dataset, introduced by Wu et al.[64], which has played a pivotal role in benchmarking the accuracy and robustness of trackers. The OTB dataset provides comprehensive ground truth for various objects across numerous videos, allowing for a detailed analysis of tracking algorithms under different conditions. Following the OTB, the VOT[65] challenge has introduced datasets annually since 2013, with each iteration presenting new challenges and advancements over the previous versions. The VOT challenge datasets are known for their rigorous annotation protocols and have introduced several innovations in evaluation methodologies, such as the no-reset evaluation protocol and real-time tracking evaluations. Another significant contribution to the field is TrackingNet[49], which provides a large-scale dataset covering a wide variety of objects and scenarios. The LaSOT dataset by Zhan et al.[50] further extends the boundaries by offering a large-scale, high-quality dataset with lengthy video sequences and is aimed at evaluating the long-term capabilities of tracking algorithms. LaSOT provides detailed annotations and a diverse set of challenges, making it an invaluable resource for developing and testing long-term trackers. The GOT-10k dataset by Huang et al.[32] introduces a unique approach by focusing on a wide

variety of object classes with a zero-shot evaluation protocol. This dataset challenges trackers to perform well on previously unseen objects, pushing the boundaries of generalization in object tracking. PoseTrack[66] and GDTM[67] focus more on specialized datasets. PointOdyssey[68] is a synthetic dataset specifically designed for long-term point tracking, addressing the limitation of short temporal context in existing datasets.

Compared with those datasets, DAVE's diverse actors allow for the evaluation of robust tracking methods capable of handling occlusions, cluttered scenes, and dynamic environments. It broadens the scope of tracking scenarios, facilitating the development of algorithms capable of operating under a wider range of real-world conditions.

A.1.3. Detection

In the realm of object detection, except for Pascal VOC challenge[36] and the MS COCO dataset[52], there are some specific applications such as autonomous driving[1][69], and dedicated datasets have been created to address the unique challenges of this domain. Waymo Open Dataset[1] represents a significant leap forward in scale and diversity for autonomous driving datasets. It encompasses a vast array of sensor data, including high-resolution LiDAR and camera footage, across a wide range of driving conditions and scenarios. This dataset has been instrumental in pushing the boundaries of perception algorithms in terms of scalability, robustness, and accuracy. The NuScenes dataset[23] is another pivotal dataset for autonomous vehicle perception, offering a rich set of sensor modalities, including RADAR, which is less common in other datasets. NuScenes provides detailed annotations for a variety of object classes in complex urban environments, making it a valuable resource for multi-modal perception systems.

Compared with those datasets, DAVE has more challenges in terms of the mixture of agents, area, time of the day, traffic density, and weather conditions.

A.1.4. Spatiotemporal Action Localization

Spatiotemporal action localization is a crucial task in computer vision that involves identifying both the temporal and spatial boundaries of actions within videos. This task enables the understanding of complex video content by pinpointing where and when specific actions occur. Over the years, several datasets have been introduced to facilitate research and development in this area. Here, we review some of the key datasets that have significantly contributed to advancing spatiotemporal action

localization research. UCF101-24[62] is one of the earliest datasets tailored for spatiotemporal action localization. Derived from the UCF101 dataset, it includes 24 sports categories with temporal annotations and bounding boxes around the action instances. Despite its relatively small size, UCF101-24 has been pivotal in early methodological developments. The J-HMDB dataset[63] is another fundamental resource that consists of 21 different action classes with 928 video clips. Each action instance is annotated with a bounding box across all frames, providing detailed spatial and temporal information. The dataset's focus on human actions makes it particularly valuable for human-centered action localization research. Furthermore, MEVA[70] and VIRAT[71] focus on unmanned aerial vehicles and surveillance activity detection.

More recently, the MultiSports dataset[26] has been introduced, focusing on multi-person and multi-action scenarios within sports videos. It contains annotations for 133 action classes across more than 20 different types of sports, with precise spatiotemporal bounding boxes for each action instance. This dataset is particularly challenging due to the dynamic nature of sports, which include frequent occlusions and interactions between athletes. Our DAVE dataset makes the progression from relatively simple, single-action instances in constrained environments to complex, multi-action scenarios in uncontrolled environments and challenging scenarios.

A.1.5. Video Moment Retrieval

The task of Video Moment Retrieval (VMR) involves identifying specific moments within a video that correspond to a textual query. This area has seen significant interest due to its applications in video understanding, search, and interaction. Various datasets have been introduced to facilitate research in VMR, each with its unique characteristics and challenges. This section reviews some of the key datasets that have been influential in advancing VMR research. One of the earliest and most widely used datasets in this domain is the Charades dataset by Sigurdsson et al.[60]. It consists of videos of daily activities annotated with descriptions and temporal intervals. The dataset has been instrumental in developing early VMR models due to its rich annotations and the naturalistic setting of the videos. Building on the foundations laid by Charades, the ActivityNet Captions dataset[72] offers a larger scale and diversity of activities. This dataset features dense temporal annotations with corresponding natural language descriptions, making it a staple for training and evaluating VMR systems. Another significant contribution to the field is the TVR dataset[73]. This dataset stands out for its focus on television show episodes, providing a mix of dialogue, action, and interaction that is more complex

than daily activities. The TVR dataset is particularly noted for its challenging queries that require deep understanding of both the video content and the textual descriptions. The DiDeMo dataset[40] offers a different approach by focusing on describing distinct moments in a video with a single sentence. Its unique structure facilitates research into more granular moment retrieval and alignment between video content and textual descriptions. These datasets have collectively contributed to the progress in VMR by providing diverse challenges and enabling the development of advanced models capable of understanding complex video-text relations. However, the unstructured videos in DAVE add more sophistication and increase the complexity of tasks that models are expected to perform.

A.1.6. Multi-label Video Action Recognition

In the field of computer vision, multi-label video action recognition has become increasingly important for applications ranging from surveillance to content analysis and retrieval. Unlike single-label action recognition, where each video is associated with a single action, multi-label video action recognition involves identifying multiple actions that occur simultaneously or sequentially within a video.

The Charades dataset by Sigurdsson et al.[42] is the most popular and is specifically designed for multi-label video action recognition. It contains 9,848 videos with an average length of 30 seconds, annotated with 157 action labels. The dataset stands out for its focus on everyday activities, with videos featuring multiple actions performed by the actors. Charades facilitates the development and evaluation of models capable of recognizing multiple simultaneous actions, making it a cornerstone in multi-label video action recognition research.

Given that Charades focuses on daily activities, it primarily includes indoor scenarios. This focus may limit the applicability of derived models for outdoor activities or other contexts not covered by the dataset. Our DAVE dataset makes up for the indoor limitation and introduces more complex actions, leading to the advancement of more sophisticated and accurate recognition models.

Statements and Declarations

Author contributions

Xijun Wang: Conceptualization, Methodology, Validation, Writing- Original draft preparation. Pedro Sandoval-Segura: Methodology, Writing- Reviewing and Editing. Chengyuan Zhang: Software, Data

curation. Junyun Huang: Data curation, Investigation. Tianrui Guan: Validation. Ruiqi Xian: Validation. Fuxiao Liu: Validation. Rohan Chandra: Conceptualization, Writing- Reviewing and Editing. Boqing Gong: Conceptualization. Dinesh Manocha: Conceptualization, Writing- Reviewing and Editing, Supervision.

References

1. a, b, c, d, e, f, g, h, i, j, k, l *Sun P, Kretzschmar H, Dotiwalla X, Chouard A, Patnaik V, Tsui P, Guo J, Zhou Y, C hai Y, Caine B, et al. Scalability in perception for autonomous driving: Waymo open dataset. CVPR. 202 0.*
2. ^ *Chandra R, Bhattacharya U, Bera A, Manocha D (2019). "Traphic: Trajectory prediction in dense and h eterogeneous traffic using weighted interactions". In: Proceedings of the IEEE/CVF Conference on Comp uter Vision and Pattern Recognition. pp. 8483--8492.*
3. ^ *Chandra R, Guan T, Panuganti S, Mittal T, Bhattacharya U, Bera A, Manocha D (2020). "Forecasting tr ajectory and behavior of road-agents using spectral clustering in graph-lstms". IEEE Robotics and Auto mation Letters. 5(3): 4882--4890.*
4. ^ *Parikh C, Mishra RS, Chandra R, Sarvadevabhatla RK (2024). "Transfer-LMR: Heavy-Tail Driving Beh avior Recognition in Diverse Traffic Scenarios". arXiv preprint arXiv:2405.05354. Available from: https://arxiv.org/abs/2405.05354.*
5. ^ *Chandra R, Wang M, Schwager M, Manocha D. Game-Theoretic Planning for Autonomous Driving am ong Risk-Aware Human Drivers. In: 2022 International Conference on Robotics and Automation (ICR A). 2022. p. 2876-2883.*
6. ^ *Chandra R, Manocha D (2022). "Gameplan: Game-theoretic multi-agent planning with human driver s at intersections, roundabouts, and merging". IEEE Robotics and Automation Letters. 7 (2): 2676–268 3.*
7. a, b, c, d, e, f, g *Gu C, Sun C, Ross DA, Vondrick C, Pantofaru C, Li Y, Vijayanarasimhan S, Toderici G, Ricco S, Sukthankar R, et al. AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions. In: Pr oceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018. p. 6047-6056.*
8. ^ *Liu F, Yacoob Y, Shrivastava A. "Covid-vts: Fact extraction and verification on short video platforms". arXiv preprint arXiv:2302.07919. 2023.*

9. a, b *Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Nat sev P, et al. The kinetics human action video dataset. arXiv preprint arXiv:1705.06950. 2017.*
10. ^ *Yao Y, Wang X, Xu M, Pu Z, Atkins E, Crandall D (2020). "When, where, and what? A new dataset for a nomaly detection in driving videos". arXiv preprint arXiv:2004.03044.*
11. a, b, c *Singh G, Akrigg S, Di Maio M, Fontana V, Alitappeh RJ, Khan S, Saha S, Jeddisaravi K, Yousefi F, Cul ley J, et al. Road: The road event awareness dataset for autonomous driving. IEEE transactions on patter n analysis and machine intelligence. **45**(1):1036-1054 (2022).*
12. a, b, c *Malla S, Dariush B, Choi C. "Titan: Future forecast using action priors." In: Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition. 2020. pp. 11186–11196.*
13. a, b *Ros G, Sellart L, Materzynska J, Vazquez D, Lopez AM. "The synthia dataset: A large collection of synt hetic images for semantic segmentation of urban scenes." In: Proceedings of the IEEE conference on co mputer vision and pattern recognition. 2016:3234-3243.*
14. a, b *Behley J, Garbade M, Milioto A, Quenzel J, Behnke S, Stachniss C, Gall J (2019). "Semantickitti: A data set for semantic scene understanding of lidar sequences". In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9297--9307.*
15. a, b *Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B. The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on co mputer vision and pattern recognition. 2016. p. 3213-3223.*
16. a, b *Geyer J, Kassahun Y, Mahmudi M, Ricou X, Durgesh R, Chung AS, Hauswald L, Pham VH, Mühlegg M, Dorn S, Fernandez T, Jänicke M, Mirashi S, Savani C, Sturm M, Vorobiov O, Oelker M, Garreis S, Schu berth P. A2D2: Audi Autonomous Driving Dataset. arXiv [cs.CV]. 2020. Available from: https://arxiv.org/ abs/2004.06320.*
17. a, b *HuangXinyu W, WangPeng, et al. TheApolloscape opendatasetforautonomousdrivinganditsapplicati on. IEEE Transactions on Pattern Analysis and Machine Intelligence. **42**(10): 2702 (2020).*
18. a, b *Rasouli A, Kotseruba I, Kunic T, Tsotsos JK (2019). "Pie: A large-scale dataset and models for pedestr ian intention estimation and trajectory prediction." In: Proceedings of the IEEE/CVF International Confe rence on Computer Vision. pp. 6262–6271.*
19. a, b *Liao Y, Xie J, Geiger A (2022). "Kitti-360: A novel dataset and benchmarks for urban scene understa nding in 2d and 3d". IEEE Transactions on Pattern Analysis and Machine Intelligence. **45**(3): 3292–331 0.*

20. a, b Pham QH, Sevestre P, Pahwa RS, Zhan H, Pang CH, Chen Y, Mustafa A, Chandrasekhar V, Lin J. "A* 3 d dataset: Towards autonomous driving in challenging environments." In: 2020 IEEE International conf erence on Robotics and Automation (ICRA). IEEE; 2020. p. 2267-2273.

21. a, b Patil A, Malla S, Gang H, Chen YT. "The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes." In: 2019 International Conference on Robotics and Automation (ICR A). IEEE; 2019. p. 9552-9557.

22. a, b Chang MF, Lambert J, Sangkloy P, Singh J, Bak S, Hartnett A, Wang D, Carr P, Lucey S, Ramanan D, e t al. Argoverse: 3d tracking and forecasting with rich maps. In: Proceedings of the IEEE/CVF conference o n computer vision and pattern recognition. 2019. pp. 8748-8757.

23. a, b, c Caesar H, Bankiti V, Lang AH, Vora S, Liong VE, Xu Q, Krishnan A, Pan Y, Baldan G, Beijbom O (202 o). "nuScenes: A multimodal dataset for autonomous driving". CVPR.

24. a, b Ding L, Terwilliger J, Sherony R, Reimer B, Fridman L (2020). "MIT driveseg (manual) dataset for dy namic driving scene segmentation". Tech. Rep., Technical report, Massachusetts Institute of Technolog y.

25. ^ Idrees H, Zamir AR, Jiang YG, Gorban A, Laptev I, Sukthankar R, Shah M. The thumos challenge on acti on recognition for videos \\"in the wild\\". Computer Vision and Image Understanding. **155**:1-23 (2017).

26. a, b, c Li Y, Xu J, Qiu Z, Tian Y, Hu T, Wang J, Li J, Xiong H, Hauptmann AG. "MultiSports: A Multi-Person Video Dataset for Action Spotting, Localization, and Detection". arXiv preprint arXiv:2103.07514. 2021.

27. ^ Chandra R. Towards autonomous driving in dense, heterogeneous, and unstructured traffic. Ph.D. thesi s, University of Maryland, College Park; 2022.

28. ^ Milan A, Leal-Taixé L, Reid I, Roth S, Schindler K (2016). "MOT17: A Benchmark for Multi-Object Trac king". arXiv preprint arXiv:1603.00831.

29. ^ Chandra R, Bhattacharya U, Bera A, Manocha D. Densepeds: Pedestrian tracking in dense crowds using front-rvo and sparse features. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Sys tems (IROS). IEEE; 2019. p. 468-475.

30. ^ Chandra R, Bhattacharya U, Randhavane T, Bera A, Manocha D. RoadTrack: Realtime Tracking of Roa d Agents in Dense and Heterogeneous Environments. In: 2020 IEEE International Conference on Roboti cs and Automation (ICRA). IEEE; 2020. p. 1270-1277.

31. ^ Chandra R, Bhattacharya U, Roncal C, Bera A, Manocha D. Robusttp: End-to-end trajectory prediction for heterogeneous road-agents in dense traffic with noisy sensor inputs. In: Proceedings of the 3rd ACM Computer Science in Cars Symposium. 2019. p. 1-9.

32. a, b, c, d Huang L, Zhao X, Huang K (2019). "Got-10k: A large high-diversity benchmark for generic objec t tracking in the wild". IEEE Transactions on Pattern Analysis and Machine Intelligence. 43(5): 1562–15 77.

33. ^ Peng L, Gao J, Liu X, Li W, Dong S, Zhang Z, Fan H, Zhang L. VastTrack: Vast Category Visual Object Tra cking. In: The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchm arks Track; 2024. Available from: https://openreview.net/forum?id=5Lo5sLRIlQ.

34. a, b Wei X, Bai Y, Zheng Y, Shi D, Gong Y (2023). "Autoregressive visual tracking." In: Proceedings of the I EEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9697–9706.

35. ^ Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft COCO: Com mon Objects in Context. European Conference on Computer Vision. 2014. Springer.

36. a, b, c Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A (2010). "The Pascal Visual Object Cla sses (VOC) Challenge". International Journal of Computer Vision. 88 (2): 303–338.

37. ^ Guan T, Wang J, Lan S, Chandra R, Wu Z, Davis L, Manocha D. M3detr: Multi-representation, multi-sc ale, mutual-relation 3d object detection with transformers. In: Proceedings of the IEEE/CVF winter conf erence on applications of computer vision. 2022. p. 772–782.

38. a, b, c Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin transformer: Hierarchical vision transf ormer using shifted windows. In: IEEE International Conference on Computer Vision (ICCV). 2021. pp. 10 012-10022.

39. a, b Pan J, Chen S, Shou MZ, Liu Y, Shao J, Li H. Actor-context-actor relation network for spatio-tempora l action localization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2021. pp. 464–474.

40. a, b, c Hendricks LA, Wang O, Shechtman E, Sivic J, Darrell T, Russell B (2017). "Localizing moments in vi deo with natural language". In: Proceedings of the IEEE International Conference on Computer Vision. p pp. 5803–5812.

41. a, b Moon W, Hyun S, Lee S, Heo JP (2023). "Correlation-guided Query-Dependency Calibration in Video Representation Learning for Temporal Grounding". arXiv. **cs.CV**. Available from: arXiv:2311.08835.

42. a, b, c, d Sigurdsson GA, Varol GV, Wang X, Farhadi A, Laptev I, Gupta A. Hollywood in homes: Crowdsour cing data collection for activity understanding. In: European Conference on Computer Vision. Springer; 2016. p. 510-526.

43. a, b, c Feichtenhofer C, Fan H, Malik J, He K. Slowfast networks for video recognition. In: IEEE Internation al Conference on Computer Vision (ICCV). 2019. p. 6202-6211.

44. ^CVAT.ai Corporation. Computer Vision Annotation Tool (CVAT) [software]. Version 2.8.2. Nov 2023. Available from: https://github.com/opencv/cvat.

45. ^Kothandaraman D, Chandra R, Manocha D. BoMuDANet: unsupervised adaptation for visual scene understanding in unstructured driving environments. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021. pp. 3966–3975.

46. ^Kothandaraman D, Chandra R, Manocha D. SS-SFDA: Self-supervised source-free domain adaptation for road segmentation in hazardous environments. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021. pp. 3049-3059.

47. ^Deng J, Guo J, Ververas E, Kotsia I, Zafeiriou S. Retinaface: Single-shot multi-level face localisation in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020. pp. 5203-5212.

48. ^Yan J, Yu J, Xiaoxiao. "HyperLPR3 - High Performance License Plate Recognition Framework". In: https://github.com/szad670401/HyperLPR (2023).

49. a, b, cMuller M, Bibi A, Giancola S, Alsubaihi S, Ghanem B. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In: Proceedings of the European conference on computer vision (ECCV). 2018. p. 300–317.

50. a, b, cZhan X, Wu Q, Wang M, Sun J, He T (2019). "LaSOT: A Large-Scale High-Resolution Benchmark for Visual Object Tracking." In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4641--4650.

51. ^Fan H, Bai H, Lin L, Yang F, Chu P, Deng G, Yu S, Harshit, Huang M, Liu J, et al. Lasot: A high-quality large-scale single object tracking benchmark. International Journal of Computer Vision. 129:439–461 (2021).

52. a, b, c, dLin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft COCO: Common Objects in Context. In: European Conference on Computer Vision. Springer; 2014. p. 740–755.

53. ^Sun S, Akhtar N, Song H, Mian A, Shah M (2019). "Deep Affinity Network for Multiple Object Tracking". arXiv. arXiv:1810.11780.

54. ^Dave A, Khurana T, Tokmakov P, Schmid C, Ramanan D (2020). "TAO: A Large-Scale Benchmark for Tracking Any Object". arXiv. arXiv:2005.10356.

55. a, bJocher G, Chaurasia A, Qiu J. Ultralytics YOLOv8 [software]. 2023. Available from: https://github.com/ultralytics/ultralytics. ORCID: 0000-0001-5950-6979, 0000-0002-7603-6750, 0000-0003-3783-7069. Version 8.0.0. License: AGPL-3.0.

56. ^Cai Z, Vasconcelos N. "Cascade r-cnn: Delving into high quality object detection." In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018:6154-6162.

57. ^Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with t ransformers. In: European Conference on Computer Vision (ECCV). Springer; 2020. p. 213–229.

58. ^Regneri M, Rohrbach M, Wetzel D, Thater S, Schiele B, Pinkal M (2013). "Grounding action description s in videos". Transactions of the Association for Computational Linguistics. 1: 25–36.

59. ^Wang J, Jiang W, Ma L, Liu W, Xu Y (2018). "Bidirectional Attentive Fusion with Context Gating for Den se Video Captioning". arXiv. **cs.CV**. Available from: arXiv:1804.00100.

60. a, bSigurdsson GA, Varol G, Wang X, Farhadi A, Laptev I, Gupta A. Hollywood in homes: Crowdsourcing d ata collection for activity understanding. European Conference on Computer Vision. 2016.

61. ^Carreira J, Noland E, Hillier C, Zisserman A (2019). "A short note on the kinetics-700 human action da taset". arXiv preprint arXiv:1907.06987.

62. a, bSoomro K, Zamir AR, Shah M (2012). "UCF101: A dataset of 101 human actions classes from videos in the wild". arXiv preprint arXiv:1212.0402.

63. a, bJhuang H, Gall J, Zuffi S, Schmid C, Black MJ. Towards understanding action recognition. In: Proceedi ngs of the IEEE International Conference on Computer Vision. 2013. p. 3192-3199.

64. ^Wu Y, Lim J, Yang MH. "Object tracking benchmark". IEEE Transactions on Pattern Analysis and Machi ne Intelligence. 37(9): 1834-1848 (2013).

65. ^Kristan M, Matas J, Leonardis A, Felsberg M, Cehovin L, Fernandez G, Vojir T, Hager G, Nebehay G, Pflu gfelder R. "The visual object tracking vot2015 challenge results." In: Proceedings of the IEEE internation al conference on computer vision workshops. 2015:1-23.

66. ^Zhang X, Yu F, He B, Yang M, Li X, Zhu J, Wang J. PoseTrack: A Dataset for Person Search, Multi-Object Tracking and Multi-Person Pose Tracking. In: Proceedings of the IEEE Conference on Computer Vision a nd Pattern Recognition (CVPR). 2021. pp. 14077-14086.

67. ^Liu Z, Yang Z, Xu X, Jiang C, Li X, He X. GDTM: An Indoor Geospatial Tracking Dataset with Distributed Multimodal Sensors. arXiv preprint arXiv:2402.1413. 2024.

68. ^Zheng Z, Tang S, Guo Y, Liu Y, Zhou S, Li H. PointOdyssey: A Large-Scale Synthetic Dataset for Long-T erm Point Tracking. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 202 3. p. 5216–5225.

69. ^Chandra R, Wang X, Mahajan M, Kala R, Palugulla R, Naidu C, Jain A, Manocha D. Meteor: A dense, he terogeneous, and unstructured traffic dataset with rare behaviors. In: 2023 IEEE International Conferen

ce on Robotics and Automation (ICRA). IEEE; 2023. p. 9169-9175.

70. ^*Corona K, Osterdahl K, Collins R, Hoogs A. Meva: A large-scale multiview, multimodal video dataset fo r activity detection. In: Proceedings of the IEEE/CVF winter conference on applications of computer visio n. 2021. pp. 1060-1068.*

71. ^*Oh S, Hoogs A, Perera A, Cuntoor N, Chen CC, Lee JT, Mukherjee S, Aggarwal J, Lee H, Davis L, et al. A l arge-scale benchmark dataset for event recognition in surveillance video. In: CVPR 2011. IEEE; 2011. p. 3 153-3160.*

72. ^*Krishna R, Hata K, Ren F, Fei-Fei L, Carlos Niebles J. Dense-captioning events in videos. In: Proceeding s of the IEEE International Conference on Computer Vision; 2017.*

73. ^*Lei J, Li L, Zhou L, Gan Z, Berg TL, Bansal M, Luo J. TVR: A large-scale dataset for video-subtitle mome nt retrieval. In: European Conference on Computer Vision. Springer; 2020.*

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.