DivNEDS: Diverse Naturalistic Edge Driving Scene Dataset for Autonomous Vehicle Scene Understanding

Datasets Available

Publisher: IEEE Cite This PDF

John Owusu Duah ; Armstrong Aboah ; Stephen Osafo-Gyamfi **All Authors**

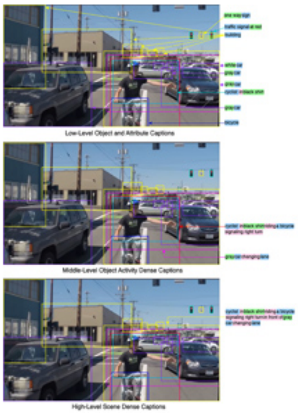
1040
Full
Text Views

PDF
Help

Open Access Comment(s)

Abstract

Abstract:
The safe implementation and adoption of Autonomous Vehicle (AV) vision models on public roads requires not only an Institute of Electrical and Electronics Engineers nprising pedestrians and other vehicles but also the ability to reason about edge situations such as unpredictable maneuvers by other drivers, impending accidents, erratic movement of pedestrians, cyclists, and motorcyclists, animal crossings, and cyclists using hand signals. Despite advances in complex tasks such as object tracking, human behavior modeling, activity recognition, and trajectory planning, the fundamental challenge of interpretable scene understanding, especially in out-of-distribution environments, remains evident. This is highlighted by the 84% of AV disengagements attributed to scene understanding errors in real-world AV tests. To address this limitation, we introduce the Diverse Naturalistic Edge Driving Scene Dataset (DivNEDS), a novel dataset comprising 11,084 edge scenes and 203,000 descriptive captions sourced from 12 distinct locations worldwide, captured under varying weather conditions and at different times of the day. Our approach includes a novel embedded hierarchical dense captioning strategy aimed at enabling few-shot learning and mitigating overfitting by excluding irrelevant scene elements. Additionally, we propose a Generative Region-to-Text Transformer, with a baseline embedded hierarchical dense captioning performance of 60.3mAP, a new benchmark for AV scene understanding models trained on dense captioned data sets. This work represents a significant step toward improving AVs' ability to comprehend diverse, real-world edge and complex driving scenarios, thereby enhancing their safety and adaptability in dynamic environments. The dataset and instructions are available at https :// github . com / johnowusuduah / DivNEDS.


one way sign
traffic signal at red
building
white car
gray car
gray car
cyclist in black shirt
gray car
bicycle
Low-Level Object and Attribute Captions
cyclist in black shirt riding a bicycle signaling right turn
gray car changing lane
Middle-Level Object Activity Dense Captions
cyclist in black shirt riding a bicycle signaling right turn in front of gray car changing lane
High-Level Scene Dense Captions

Deconstruction of novel embedded hierarchical dense captioning strategy showing three (3) levels of annotation utilized in each image of DivNEDS. Our embedded hierarchica... Show More

Published in: IEEE Access (Volume: 12)

Page(s): 60628 - 60640

DOI: 10.1109/ACCESS.2024.3394530

Date of Publication: 29 April 2024

Publisher: IEEE

Electronic ISSN: 2169-3536

SECTION I.

Introduction

In recent years, the performance of Autonomous Vehicle (AV) driving systems in tasks related to visual scene understanding has witnessed significant improvement, owing to the emergence of various benchmark datasets [1], [2], [3], [4], [5], [6], [7], [8] and rapid advances in deep learning [9], [10], [11].

Eighty-four percent (84%) of AV disengagements during tests on public roads in California, the only state that mandates AV manufacturers to report instances of AV disengagements on public roads, are attributed to errors in scene understanding in unexpected environments [12]. This indicates that despite advancements in human behavior modeling, pedestrian prediction, activity recognition, trajectory planning, object detection, tracking, and other downstream tasks, the fundamental task of timely interpretable scene understanding, particularly in out-of-distribution scenarios, remains a challenge. For example, in a fatal accident involving Uber's AV on March 18, 2018 in Tempe, Arizona [13], Uber's AV initially identified the victim of the accid who was pushing a bicycle across a poorly lit four-lane road at an undesignated location, as an unknown object, then as a vehicle, and finally as a bicycle, all within a 6-second time span. Enabling Autonomous Vehicles (AVs) to rapidly grasp edge scenarios involving a diverse array of road users under varying conditions is imperative for their safe operation in dynamic environments. This capability is crucial for ensuring that AVs meet the stringent safety requirements necessary to gain widespread acceptance and adoption. To address this challenge, this study introduces a novel dataset comprising diverse naturalistic edge scenarios. These scenarios are annotated using a novel **embedded hierarchical dense captioning strategy** that effectively localizes and describes salient regions within driving scenes. Furthermore, the study introduces a benchmark Generative Region-to-Text Transformer, known as the *Diverse Naturalistic Edge Driving Scenes Transformer (DivNET)*, designed specifically for enhancing scene understanding in these challenging contexts.

PDF
Help

A significant body of research embraces graph-based frameworks to model the relationships between various elements in traffic scenarios [14], [15], [16], [17]. However, these graph-based methods, while effective at capturing relationships, often lack interpretability for downstream tasks, particularly in complex and edge scenarios. Understanding why an AV agent makes specific decisions based on a graph structure remains challenging. In parallel, research on natural language and AV vision grounding has emerged, with a focus on interpretable AV action understanding and AV navigation with natural language. *DRAMA* [18], for instance, connects risk localization to explanations in driving scenes, wheras *HDD* [9] deconstructs driving scenes into layers to understand the causal relationships between human driver actions and traffic situations. *BDD-X* [19] predicts AV control commands and provides rationalization and introspective explanations of actions. *HAD* [20] and *Talk2Nav* [21] explore the use of natural language for AV guidance, with *HAD* assessing behavior, comprehending user language, and explaining its internal state, wheras *Talk2Nav* focuses on enhancing human-AV agent interaction through verbal navigational instructions. However, these studies suffer from poor generalization as a result of relying on datasets with narrow scope and diversity, and single scene annotations that include irrelevant background information.

Whereas previous research has made significant contributions to AV scene understanding, it falls short in addressing the timely and thorough comprehension of diverse naturalistic edge driving scenes for the following reasons:

1. Existing datasets typically comprise a narrow scope, idealistic environments and homogeneous driving scenes that are specific to particular cities, thereby constraining their capacity to be generalized to diverse edge and complex scenarios. For instance, *DRAMA* [18] exclusively captures scenes in which an ego-driver must engage brakes in Tokyo, Japan. *Oxford RobotCar* [22] consists of 100 repetitions of a consistent route through Oxford, United Kingdom. Similarly, both *HDD* [9] and *HAD* [20] were exclusively captured in San Francisco, USA.

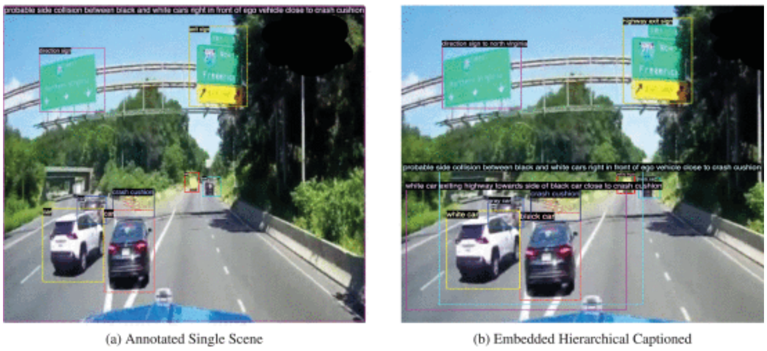
2. Current research beyond semantic scene understanding [9], [18], [20] relies on single scene bounding boxes and descriptive annotations, including background elements irrelevant to specific annotations. This dilutes the feature spaces and leads to poor performance on unseen data with similar backgrounds. The ultimate result is that AV systems learn extraneous information and perform poorly when applied to unseen data with similar background. Fig. 1 illustrates the difference between the single scene description bounding box annotation strategy used in contemporary research and our embedded hierarchical dense captioning strategy. We observe that the majority of the receptive field captured by the single scene bounding box is irrelevant to its description.

3. Traditional approaches to comprehensive scene understanding rely on extensive multi modal sensor data, including video feeds, LiDAR point clouds, radar and GPS. However, LiDAR and radar sensors have limited operational ranges, introducing failure modes [1], including latency and overfitting, during adverse driving conditions. However, human drivers can make informed decisions with less precise representations of the real world, highlighting the need for AVs to infer relationships, attributes, and actions from low-level data like images [23]. These limitations have also been highlighted in avoidable

safety incidents involving AVs on public roads [13], underscoring the need for more robust and generalizable approaches. Our current study focuses on filling these gaps using a *Diverse Naturalistic Edge Driving Scene Dataset (DivNEDS)*. The main contributions of this study is summarized as follows:

1. We present 11,084 scenarios comprising a wide range of diverse naturalistic edge situations annotated with 203,000 descriptive captions. These scenarios involve sudden and unpredictable maneuvers by other road users, impending accidents, animals and debris within right of way, hand signals from cyclists, and unusual interactions involving pedestrians, cyclists, and motorcyclists in right-hand and left-hand traffic. The dataset was captured in various locations, including New York, San Francisco, Seattle, Minneapolis (United States), Toronto (Canada), Madhepur, Mumbai (India), Johannesburg (South Africa), Jakarta (Indonesia), Melbourne (Australia), London (England), and Lagos (Nigeria) to capture variations in road standards and behaviors. The data collection occurred under different lighting conditions such as direct sunlight, overcast, snowy, foggy, and rainy conditions. *DivNEDS* contains an equal distribution of images captured during both daytime and nighttime. Unlike other related datas *DivNEDS* includes scenes from rural areas and images with varying resolutions and qualities making robust representation of real-world scenarios. To the best of our knowledge, this dataset stands is the most diverse and comprehensive resource for understanding complex and edge scenes in the context of AV research.

2. We introduce a novel **embedded hierarchical dense captioning strategy** that enables few-shot learning and improves conventional dense captioning schemes. Conventional dense captioning methods utilize a flat annotation strategy that requires multiple images and frames to capture contextual information, which, in contrast, can be efficiently captured using an embedded hierarchical dense captioning scheme within a single image. We used hierarchical annotations with the lowest level approximating object detection annotations with attributes embedded within middle-level annotations. In Fig. 2, the low-level bounding box is captioned, "cyclist in black shirt," and is embedded within a middle-level bounding box with the caption, "cyclist in black shirt riding a bicycle signaling right turn." The middle-level annotations describe relationships and actions of objects in a road scene. The highest annotation layer embeds low-level and middle-level annotations and describes the scenes captured within the images. As illustrated in Fig. 2, the highest level scene caption, "cyclist in black shirt riding a bicycle signaling right turn in front of gray car changing lane," embeds all relevant low and middle-level information describing the scene. This approach allows the definition of multiple contextual feature spaces within each image, leading to a reduced dataset comprising a relatively smaller number of images while maintaining a comparable feature space. Each level of annotation contributes to different aspects of the scene understanding, thereby creating various layers of context. The outcome is a feature-rich space with 203,000 descriptive annotations that highlight salient regions in all 11,084 images. Embedded hierarchical dense captioning requires a computer vision system to identify and describe multiple salient regions within images in natural language. This approach is necessary because image captioning often falls short of effectively describing multiple activities in a singular complex and edge scene. In addition, in light of the existing gaps in AV scene understanding and the potential impact of enhancing safety measures, our primary focus in this paper is on addressing the need for interpretable scene understanding.


probable side collision between black and white cars right in front of ego vehicle close to crash cushion
direction sign
exit sign
crash cushion
car
highway exit sign
direction sign to north virginia
probable side collision between black and white cars right in front of ego vehicle close to crash cushion
white car exiting highway towards side of black car close to crash cushion
crash cushion
gray car
white car
black car

(a) Annotated Single Scene

(b) Embedded Hierarchical Captioned

FIGURE 1.
Single Scene Annotation vs Embedded Hierarchical Dense Captioning (a): We can observe that single scene bounding boxes, when paired with scene descriptions, encompass background features that are essentially irrelevant to the given description. In this instance, approximately less than 40% of the receptive field of the single scene or image caption is salient to the scene description. (b): On the other hand, in (b), embedded hierarchical dense captioning showcases its ability to capture a more effective feature space that eliminates irrelevant background regions. Zoom in for best viewing.

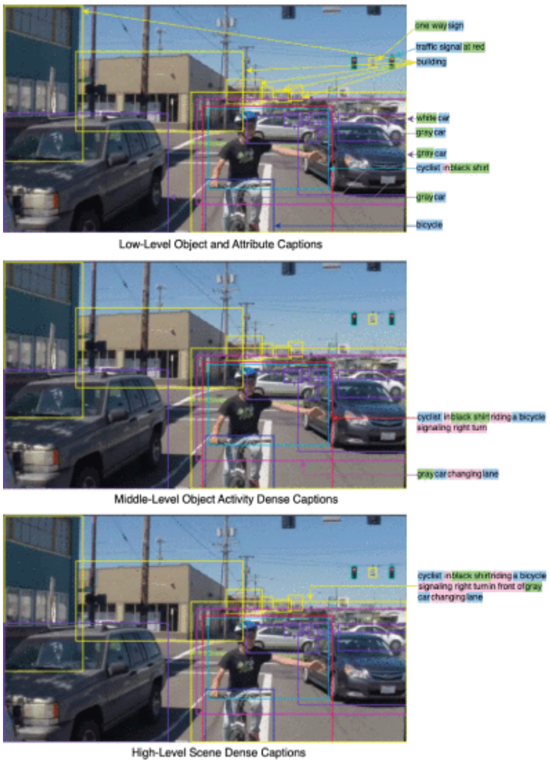

one way sign
traffic signal at red
building
white car
gray car
gray car
cyclist in black shirt
gray car
bicycle
Low-Level Object and Attribute Captions
cyclist in black shirt riding a bicycle signaling right turn
gray car changing lane
Middle-Level Object Activity Dense Captions
cyclist in black shirt riding a bicycle signaling right turn in front of gray car changing lane
High-Level Scene Dense Captions

FIGURE 2.
Deconstruction of embedded hierarchical dense captioning strategy showing three (3) levels of annotation utilized in each image. Objects, attributes and actions are highlighted in blue, green and pink, respectively. The lowest level of annotation deviates from conventional object detection annotations by including object attributes. Our embedded hierarchical dense captioning strategy allows our transformer to learn associative relationship between the receptive fields of low, middle and high level captions.

Our work includes experiments aimed at exploring the performance of pre-trained schemes and zero-shot understanding of *DivNET*. We also performed ablations studies that evaluated multiple backbones of the vision transformer used in *DivNET*. The remainder of the paper is organized as follows: Section II provides a detailed discussion of related works. Sections III and IV discuss data statistics and the architecture of *DivNET*, respectively. Experiments, results and applications are discussed in Section V. Section VI provides the applications of *DivNEDS* and *DivNET*, and conclusions are outlined in Section VII.

SECTION II.
Related Works

A. Graph-Based Driving Scene Understanding

Several studies have utilized graph-based frameworks to model relationships among traffic participants, as well as spatial and temporal information to understand the current and future state of AV environments for autonomous navigation [14], [15], [16], [17]. In [14], the authors decompose egocentric interactions into two types: *ego-thing* and *ego-stuff* interactions, using two graph convolutional networks. They also introduce a novel *MaskAlign* operation to extract features of irregular objects in ego-stuff interactions. Mylavarapu et al. [14] employ a multi-relational graph with bidirectional edges to encode the spatio-temporal relations between nodes. This representation is used to represent active and passive objects in driving scenes. In [16], the authors combine a Multi-Relational Graph with a Long Short-Term Memory (*LSTM*) Neueral Network and attention layers. This combination helps model the risk of driving maneuvers, which is formulated as a supervised scene classification problem. *RSG-Net* [11] is proposed to simulate human-level understanding of dynamic road events by predicting potential semantic relationships among objects in a road scene. *RSG-Net* [11] proposes the use of scene graphs as a more effective approach to tackle the intricacies of real-world scenarios. Positioned between model-based and end-to-end deep network models, *RSG-Net* relies on a Road Scene Graph dataset to model the behaviors of vehicles, pedestrians, and obstacles. Although these graph-based methods excel at capturing relationships, they lack clear interpretability for downstream tasks, such as decision-making, especially in complex scenarios. Understanding why an AV agent makes a particular choice based on a graph structure can be challenging.In addition, these methods inherit overfitting problems inherent in the datasets on which they are trained.

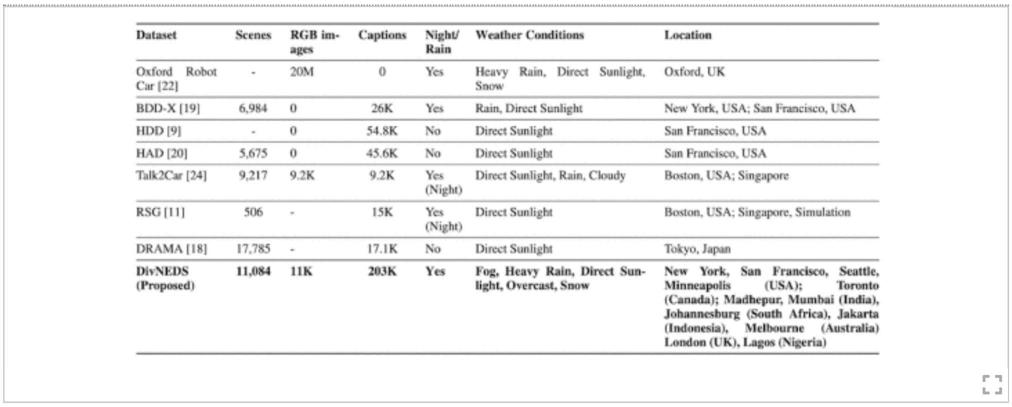
B. Natural Language and AV Vision Grounding

Safe cooperative and autonomous driving relies on an accurate understanding of risk during navigation and easy-to-interpret introspective explanations of AV behavior. Research in this domain is further classified into interpretable AV action understanding [9], [18], [19] and AV navigation with natural language [20], [21]. *DRAMA* [18] proposes a method that connects risk localization to explanations in driving scenes by using 17,785 interactive driving scenarios collected in Tokyo, Japan. Each video clip in the *DRAMA* dataset captures an ego-driver's reaction to the perceived risk, often resulting in vehicle braking. Although risk is often attributed to individual objects in *DRAMA*, it is important to note that in complex environments, risk rarely stems solely from individual objects. For example, consider a scenario in which an AV follows another vehicle that suddenly brakes owing to a pedestrian crossing the roadway at an undesignated location up ahead. In this case, the risk is not solely assigned to the braking vehicle. Instead, it involves a causal relationship between the action of the pedestrian and the driver of the braking vehicle. *HDD* [9] introduces a novel annotation framework that deconstructs driving scenes into layers of goal-oriented action, stimulus-driven action, ca and attention to understand the interactions and causal relationships between human driver actions and corresponding traffic scene situations. *BDD-X* [19] proposes novel methods to predict AV control commands given scenes and extracts rationalization and introspective explanations for such actions. This provides end-users with an understanding of what triggered a particular behavior. The strategic approach of *ROAD* [10] views agents, actions, and their locations as essential components for understanding road events. However, *ROAD* has limitations. It confines road events to permutations of 30 distinct action classes, 12 distinct agent classes, and 15 unique location classes. *ROAD* inherits the poor generalizability associated with the *Oxford RobotCar* dataset [22]. Although these studies have advanced beyond scene understanding, their use of single scene captions encompasses the entire background and irrelevant scene parts, failing to effectively capture multiple disparate salient regions in complex and edge scenes. Another limitation is the limited scope of AV actions in their datasets. For instance, *DRAMA* captures only the ego-driver's braking in reaction to perceived risk. *HAD* [20] introduces an innovative driving model that accepts natural language inputs from end-users. It focuses on two types of guidance: goal-oriented advice, which assists the vehicle in navigation tasks, and stimulus-driven advice, which directs the vehicle's attention to visual cues. *HAD* possesses the capability to assess behavior, comprehend user language, and explain its internal state for communication with the vehicle. *Talk2Nav* [21] is designed to enhance the intuitiveness of human-AV agent interaction by training a model to navigate an interactive visual environment using verbal navigational instructions. Inspired by spatial cognition research, *Talk2Nav* utilizes a soft dual-attention mechanism to extract two partial instructions from segmented language instructions; one for matching upcoming visual landmarks and the other for matching local directions. Additionally, a spatial memory scheme is introduced to encode local directional transitions. These models inherit limitations, including a lack of diversity and a limited scope, from the datasets on which they are trained on.

C. Datasets

A common limitation of graph-based driving scene understanding [24], natural language and vision grounding [9], [18], [19], [20] is that they are often captured in singular locations. For instance, *Oxford RobotCar* comprises 100 repetitions of a consistent route through Oxford, United Kingdom, and both *HDD* and *HAD* are captured in San Francisco, USA. Consequently, models developed using these datasets may not be generalizable to other environments. In addition, the use of single scene captions in these datasets fails to effectively capture per-pixel salient features for AV training. Finally, these datasets do not capture complex edge traffic scenes, leading to overfitted models that perform poorly on unseen edge driving scenes. As illustrated in Table 1, *DivNEDS* presents the most diverse dataset reported thus far, collected from 12 different locations, across all weather conditions involving diverse edge scene types. This dataset is annotated using a novel embedded hierarchical annotation strategy [25], [26] that maps per-pixel salient regions to natural descriptions of scenes, addressing some of the limitations observed in existing datasets.

TABLE 1 Comparative Summary of DivNEDS Dataset in Relation to Other Related Datasets Exploring Varied Annotations, Environmental Conditions and Locations Worldwide



Dataset	Scenes	RGB images	Captions	Night/ Rain	Weather Conditions	Location
Oxford Robot Car [22]	-	20M	0	Yes	Heavy Rain, Direct Sunlight, Snow	Oxford, UK
BDD-X [19]	6,984	0	26K	Yes	Rain, Direct Sunlight	New York, USA; San Francisco, USA
HDD [9]	-	0	54.8K	No	Direct Sunlight	San Francisco, USA
HAD [20]	5,675	0	45.6K	No	Direct Sunlight	San Francisco, USA
Talk2Car [24]	9,217	9.2K	9.2K	Yes (Night)	Direct Sunlight, Rain, Cloudy	Boston, USA; Singapore
RSG [11]	506	-	15K	Yes (Night)	Direct Sunlight	Boston, USA; Singapore, Simulation
DRAMA [18]	17,785	-	17.1K	No	Direct Sunlight	Tokyo, Japan
DivNEDS (Proposed)	**11,084**	**11K**	**203K**	**Yes**	**Fog, Heavy Rain, Direct Sunlight, Overcast, Snow**	**New York, San Francisco, Seattle, Minneapolis (USA); Toronto (Canada); Madhepur, Mumbai (India), Johannesburg (South Africa), Jakarta (Indonesia), Melbourne (Australia) London (UK), Lagos (Nigeria)**

D. Embedded Hierarchical Dense Captioning

Several general scene understanding datasets utilize a flat dense captioning scheme, wherein single bounding boxes with descriptive captions are leveraged to capture contextual information about image contents [25], [27], [28], [29]. This necessitates the annotation of numerous images to construct representations from which computer vision models can learn. We address this challenge through our novel embedded hierarchical dense captioning approach, which enables few-shot learning via nested multilevel scene descriptions.

SECTION III.
Data Statistics, Crowdsourcing Strategies and Annotation Process

A. Data Statistics

The curation of scenes for *DivNEDS* was informed by situations that would be critical for AVs to understand in order to interact efficiently with the real world. We decided to source images from diverse origins in distributions that mimic how often ego vehicles are anticipated to encounter such environments. Fig. 3 outlines a representative sample from each geographic location, weather condition, resolution, demography and time of day. *DivNEDS* comprises 11,084 images from 12 distinct locations. In total, there are 203,619 bounding boxes accompanied by descriptive captions.

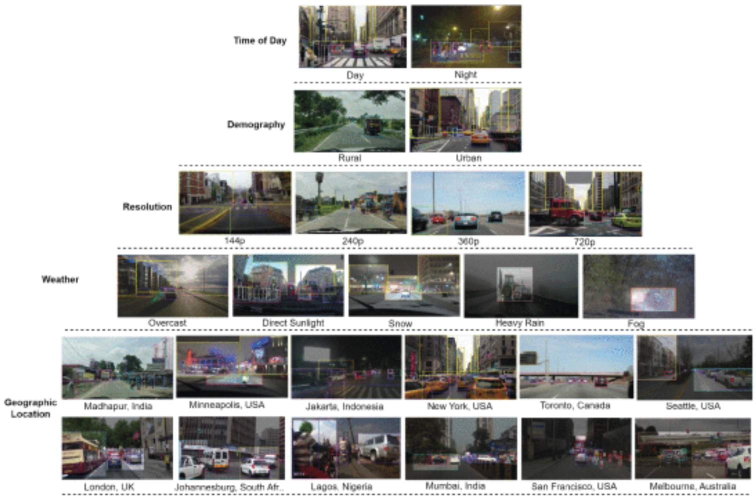


FIGURE 3.
Grid layout representation of DivNEDS showing diverse characteristics.

The average number of captions in a scene is 18. A significant majority of the images, constituting 19.5% of *DivNEDS*, are sourced from London, UK. New York, USA, and Melbourne, Australia contribute 16.6% and 14.2% of the images, respectively. The lowest proportion of images is derived from Minneapolis, USA, at 0.8%. Driving scenes from Jakarta, Indonesia, and Madhepur, India, comprise 1.1% and 1.4% of the images, respectively. Scenes captured in Minneapolis and Seattle were captured under snowy and overcast weather conditions, respectively. Scenes from each geographic location are randomly split into training, validation and test sets, with an approximate split of 80%-10%-10%. *DivNEDS* edge scenarios comprise 44% of congested corridors with motorcyclists and cyclists, 31% involve unexpected maneuvers of other vehicles including impending accidents, debris within the roadway, and animal crossings, and 21% involve unusual interactions between the ego-vehicle and pedestrians and cyclists. Nighttime snowy conditions, rural corridors, and cyclists using hand signals each representing 1% of *DivNEDS*. The object classes and scene descriptions in the images sourced for this study were not artificially balanced to ensure an even distribution across classes. As a result, the uneven caption count distribution observed in Fig. 4 reflects real-world frequencies. For example, although animals and objects obstructing roadways represent an important edge case, such scenarios arise less frequently in practice than pedestrian jaywalking incidents. Retaining an authentic representative sample of real-world edge scenarios strengthens *DivNET's* real-world capability.

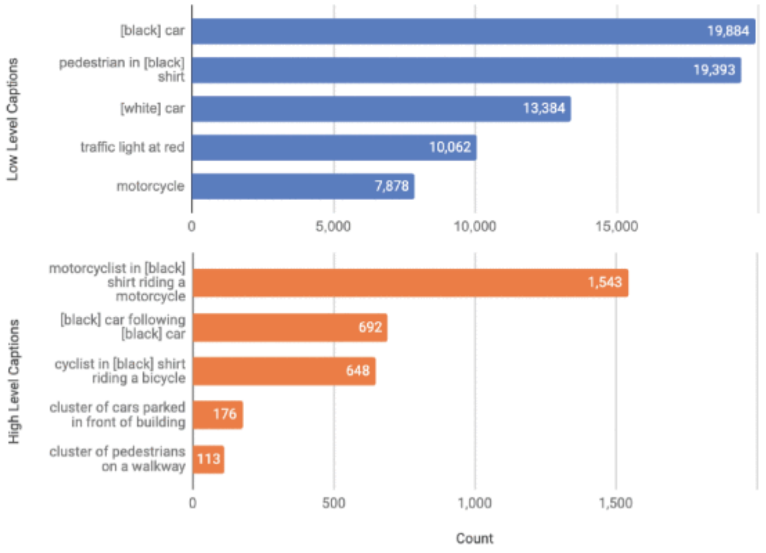


FIGURE 4.
Distribution of the Top 5 High-Level and Low-Level Captions. Square brackets [], as employed in [black], serve as a placeholder for an object attribute, as demonstrated by the example of color in this case.

B. Crowdsourcing Strategy

To capture diverse representations of naturalistic autonomous vehicle environments of edge case scenarios, *DivNEDS* comprises both original captures by the authors and publicly available images under Creative Commons licenses. We captured scenes in New York, Minneapolis, Seattle and San Francisco in person. We sourced scenes from Madhapur, Mumbai, Johannesburg, Melbourne, London, and Jakarta from Wikimedia Commons. Supplementary images portraying unexpected vehicle maneuvers, animal crossings, and cyclist hand signals were sourced from *Pixabay* using Creative Commons Zero public domain dedication. Images displaying identifiable vehicle license plate information were discarded from the final dataset as an additional extra data privacy step. Fig. 5 and Fig. 6 provide multidimensional analysis of the composition of *DivNEDS*.

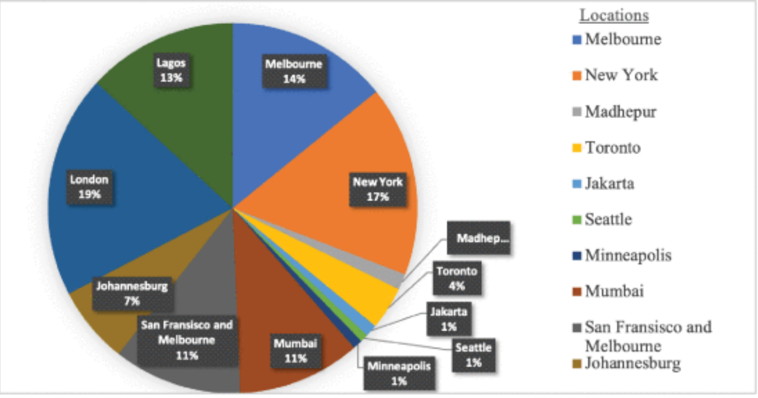


FIGURE 5.
Composition of *DivNEDS* by geographic location.

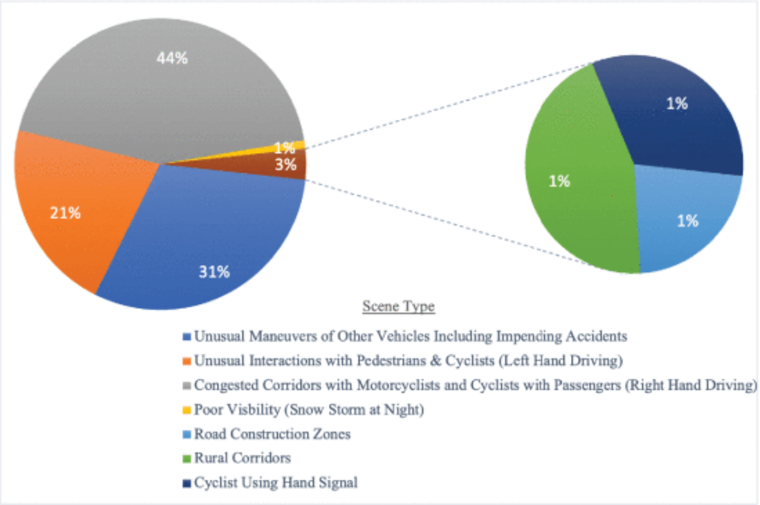


FIGURE 6.
Composition of *DivNEDS* by edge scenario type.

C. Annotation Process

Four (4) human workers with qualifications in transportation engineering annotated the images, and the authors verified annotations. The dataset was collected and annotated over 12 months, following a 3-month period of experimentation. During the experimental stage, we honed in on a sequence of drawing low-level bounding boxes around objects first. Bounding boxes were drawn to cover entire objects and were as tight as possible around objects mentioned in the description. We drew middle-level bounding boxes with captions and then drew high-level bounding boxes and captions. We found that having more than four (4) human workers resulted in long turnarounds. One of the critical requirements of the workflow was to ensure that objects and attributes occurring in multiple images across origins had consistent descriptions. To achieve this, the authors annotated 20 diverse images from each origin to ensure the consistency and quality of the natural language across all human workers. A video recording of the annotation process was shared with the human workers with documentation. The first annotation stage involved human workers annotating the first batch of 30% of images from each origin and a joint verification and revision. The annotation documentation was updated with the most common errors of the first stage annotation. At this stage, approximately 40% of the images had typographical errors, grammatical errors or inconsistent syntax of labels across multiple images. The second stage also involved joint verification and revision of the second batch of 30% of the images from each origin.Fifteen percent of the images were identified as having annotation errors at this stage. The final stage of the annotation process was implemented in two (2) phases. The first phase entailed each human worker’s peer reviewing the final 60% of the images from each origin annotated by another human worker. The authors performed the final verification and revision of all annotations. The final error rate after the two phases was 9%.

SECTION IV.
DivNET

Since *GRiT* [26] outperforms conventional dense captioning models such as *FCLN* [25], *JIVC* [30], *ImgG* [31], *COCD* [31], *COCG* [31], *CAG-Net* [32], and *TDC* + *ROCSU* [33] on the Visual Genome dataset, we modify *GRiT’s* architecture for *DivNET* and make two (2) modifications to the *ViT* [34] backbone to allow few-shot learning on our embedded hierarchical dense captioned dataset. First, we instantiate separate learned position embeddings for each caption level in the *ViT* backbone. Next, we add separate positional embeddings in the forward pass based on the appropriate caption target level for that output, guiding *DivNET* to learn contextual hierarchical representations. Fig. 7 illustrates the high level architecture of *DivNET*.

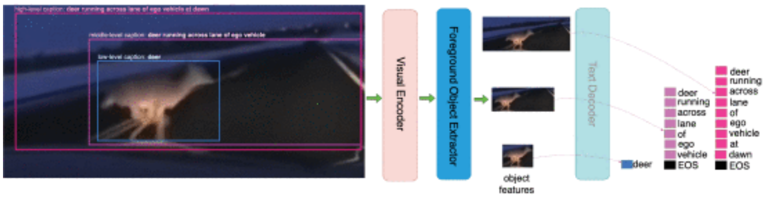

high-level caption: deer running across lane of ego vehicle at dawn
middle-level caption: deer running across lane of ego vehicle
low-level caption: deer
Visual Encoder
Foreground Object Extractor
object features
Text Decoder
deer
deer
running
across
lane
of
ego
vehicle
EOS
deer
running
across
lane
of
ego
vehicle
at
dawn
EOS

FIGURE 7.
High-Level Overview of *DivNET* Architecture: The visual encoder extracts object features defined heirarchies of nested captions then the foreground object extractor predicts object boxes.

A. Visual Encoder

ViT-L serves as the evaluated backbone for the visual encoder, with layer-wise learning rate decays of 0.8. Coswin-H is included in the backbone of the DivNET scheme. A 16×16 input image patch size is used in the training process. Feature maps at scales of 1/64 and 1/128 are generated by downsampling from 1/32.

B. Foreground Object Extractor

CenterNet [35] employs a proposal generator, which generates 2000 proposal boxes during training and during testing. The RoI head is constructed using a 3-stage Cascade R-CNN [36]. For each stage classifier both foreground and background classes are defined, and the object box used by the text decoder is predicted in the final stage. The objectness score is computed by averaging the foreground scores from all three stages.

C. Text Decoder

The text decoder is a component of *DivNET* responsible for comprehending and describing objects, attributes, relationships, and activities in natural language. In this process, words are transformed into text tokens using *BERT's WordPiece tokenizer* [37]. The text decoder is built with a 6-layer transformer, augmented with a beginning token *[task]*. Object descriptions are generated using text tokens individually in an autoregressive manner until an end token *[EOS]* is reached [26].

D. Training

The training loss of *DivNET* consists of losses computed by the foreground object extractor and text decoder, L_o , and language modeling loss, L_t . L_t is imposed on the foreground objects predicted by the foreground object extractor.

Language modeling (LM) loss is computed as follows:

$$L_t = \frac{1}{N+1} \sum_{i=1}^{N+1} CE(y_i, p(y_i|o, y_o, \ldots, i-1)), \tag{1}$$

View Source

where $p(y_i|o, y_o, \ldots, i-1)$ is the predicted score for the i-th text token given object features o and the previously generated text tokens y_o N is the number of text tokens in the given object description. y_o and y_{N+1} are the start and end token, respectively. *CE* is the cross-entropy loss with label smoothing of 0.1. L_t is computed for the foreground objects predicted by the foreground object extractor.

SECTION V.
Experiments

Although *DivNET* stands out as the first AV scene understanding dataset annotated with a dense captioning strategy, making it unique within its category without a comparable model for architectural comparison, we conducted ablation studies as well as pre-trained and zero-shot understanding experiments. Our ablation experiments aimed to dissect *DivNET's* backbone and assess which option achieved the best performance. Furthermore, we conducted pre-trained experiments to determine which pre-training scheme was most effective for our specific use-case. Additionally, evaluating the performance of a scene understanding model in edge and complex scenarios necessitates assessing its capabilities on out-of-distribution data. Therefore, we conducted a zero-shot experiment to evaluate the performance of *DivNET* in previously unseen edge situations.

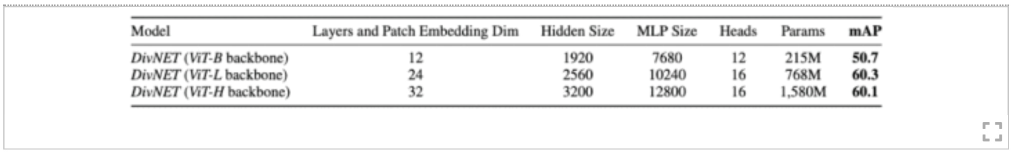
A. Ablation Studies

We ran ablations with all three (3) *ViT* configurations to determine which backbone was best suited for *DivNET*. Unlike [34] where the input image size was 1024, we resized the input images to 416 to improve training and inference speed. Thus, we defined relatively larger patch embedding dimensions for all three *ViT* configurations to compensate for the information loss from down sampling. As summarized in Table 2, the

PDF
Help

base model (*ViT-B*) uses 12 layers, a hidden size of 1920, an MLP size of 7680, 12 heads, and 215M parameters. The large model (*ViT-L*) uses 24 layers, 2560 hidden size, 10240 MLP size, 16 heads, and 768M params. Finally, the huge model (*ViT-H*) has 32 layers, 3200 hidden size, 12800 MLP size, 16 heads, and 1580M parameters. We measured the mAP for each iteration of *DivNET* on the test data.

TABLE 2 Results of DivNET Backbone Ablation Study



Model	Layers and Patch Embedding Dim	Hidden Size	MLP Size	Heads	Params	mAP
DivNET (*ViT-B* backbone)	12	1920	7680	12	215M	**50.7**
DivNET (*ViT-L* backbone)	24	2560	10240	16	768M	**60.3**
DivNET (*ViT-H* backbone)	32	3200	12800	16	1,580M	**60.1**

As shown in Table 2, the large model with *ViT-L* achieved the best performance with an mAP of 60.3. It is counter-intuitive that *DivNET* with a relatively smaller backbone in *ViT-L* outperforms the larger iteratio with a *ViT-H* backbone. We attribute this to the fact that, with a larger number of parameters, the iteratio *DivNET* with the *ViT-H* backbone is prone to overfitting. The number of parameters of *DivNET* with a *ViT-L* backbone aligned better with the number of captions present in *DivNEDS*.

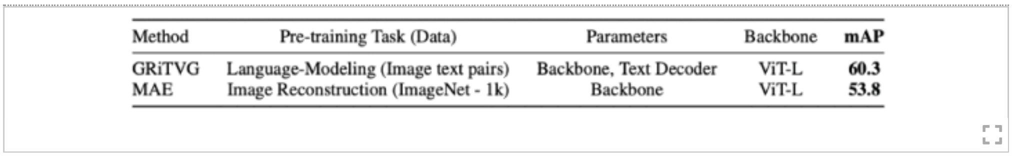
B. Pre-Trained Experiment

All images were resized to 416 × 416 pixels to expedite the training and inference and training were conducted using eight (8) NVIDIA T4 GPUs. A minimum of 90% of the annotated images were enforced to ensure that the models learned to recognize instances where objects were absent. This reduction in false negatives leads to an increase in recall. No data augmentation was applied to the dataset, and all Exchangeable Image File Format rotations were discarded, with standardized pixel ordering. The primary evaluation metric for dense captioning tasks is the mean average precision (mAP), which is calculated across various thresholds for both localization and caption description accuracy. For localization, IoU thresholds of 0.3, 0.4, 0.5, 0.6, and 0.7 were employed, for language description, and a *METEOR* score with thresholds of 0, 0.05, 0.1, 0.15, 0.2, and 0.25. The final mAP metric was derived as the mean of the average precisions (APs) calculated across all pairwise combinations of these threshold types.

Pre-Training: We explore two pre-training schemes. GRiT Visual Genome (GRiTVG pre-training): The ViT backbone and text decoder were initialized from the pre-trained GRiT (ViT-L, which was trained on Visual Genome [27]. We selected this scheme to leverage learned the object descriptions related to pedestrians and vehicles in the Visual Genome dataset.

MAE [38] pre-training: The ViT-L backbone was initialized from the self-supervised MAE model, which was trained on ImageNet-1K [39]. All the other parameters were randomly initialized. We chose MAE for this experiment to evaluate the performance of a model designed to mitigate overfitting. MAE was also chosen for its ability to recover masked image patches, which may lead to superior localization performance. We fine-tuned the model initialized from GRiTVG on DivNEDS for 100k iterations with a training batch size of 32. In contrast, the pre-trained MAE model was fine-tuned for 180k iterations with a batch size of 64. We utilized the AdamW optimizer [40] with a learning rate of 8×10^{-5} and cosine learning rate decay schedule. As shown in Table 3, the GRiTVG pre-trained model outperforms MAE pre-trained model with a mAP of 60.3. We performed inferences on the sample test images using DivNET pretrained on GRiTVG and the results are shown in Fig. 8.

TABLE 3 Result of Pre-Trained Experiment



Method	Pre-training Task (Data)	Parameters	Backbone	mAP
GRiTVG	Language-Modeling (Image text pairs)	Backbone, Text Decoder	ViT-L	**60.3**
MAE	Image Reconstruction (ImageNet - 1k)	Backbone	ViT-L	**53.8**

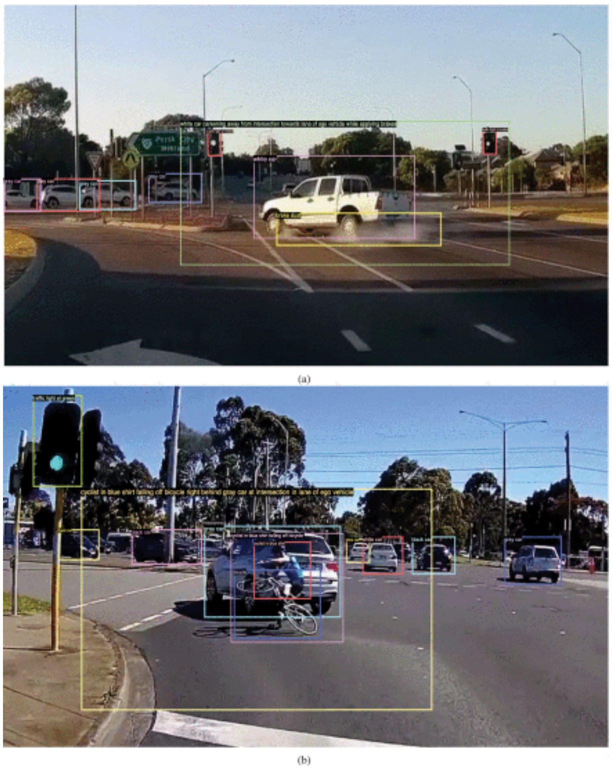

white car careening away from intersection towards lane of ego vehicle while applying brakes
white car
(a)
traffic light at green
cyclist in blue shirt falling off bicycle right behind gray car at intersection in lane of ego vehicle
gray car
black car
white car
(b)

FIGURE 8.

Sample inference results on test data. In (a), *DivNET* accurately understands dust emanating from the car as it careens toward the intersection ahead of the ego vehicle. (b) shows *DivNET's* ability to understand an edge scene involving a cyclist in blue shirt falling of a bicycle behind gray car at intersection in the lane of ego vehicle.

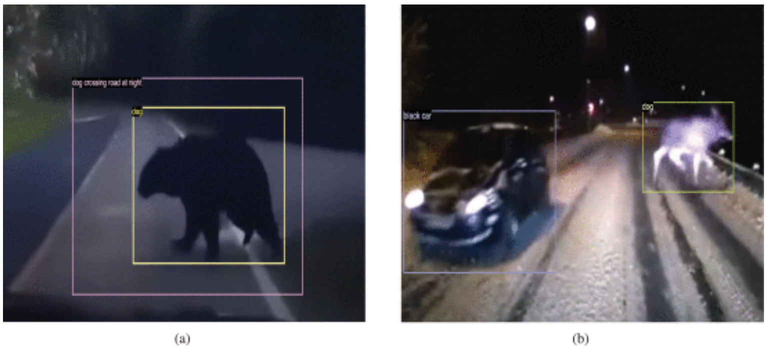


FIGURE 9.

Zero-shot object understanding predictions We observe that *DivNET* has poor zero-shot object understanding of animals crossing the roadway. Here we observe inaccurate predictions of dog instead of bear and deer. Zoom in for the best viewing.

C. Zero-shot Understanding

It is useful for AVs to describe out-of-distribution edge scenarios. Thus, we investigated whether *DivNET* could achieve zero-shot understanding of animal crossing scenes. To accomplish this, we initialized *GRiTVG* and fine-tuned it on a version of *DivNEDS* that excluded animal crossing scenes. We observed that the trained model generates diverse dense captioned descriptions for different regions within the same image. However, as shown in Fig. 8, there were certain images in which the zero-shot scheme did not yield satisfactory results. The results of this experiment revealed that *DivNET* identifies animals crossing roadways. However, this method fails to accurately classify animals in zero-shot scenes. Deers and bears in driving scenes captured at night were wrongly captioned as dogs. Consequently, *DivNET* fails to generate high-level dense captions and some middle-level dense captions. The mAP of the zero-shot understanding experiment is 20.1

SECTION VI.

Applications

As outlined in the Related Works section, current datasets for AV scene understanding often fall short in capturing edge cases and complex scenarios addressed by *DivNEDS*. The comprehensive nature of *DivNEDS*, with its scenes available for open-source use, offers significant potential when integrated into existing AV scene understanding datasets. Beyond enhancing the diversity of scenarios, *DivNEDS* can serve as a foundational resource for integrating AV commands, facilitating risk analysis, and conducting causal and effect studies. These capabilities make *DivNEDS* a valuable asset for various applications within the AV field. In the following sections, we discuss the additional specific applications and benefits of *DivNEDS*.

A. Safety Assessment and Validation

DivNEDS provides a vital benchmark for evaluating AV systems' comprehension of naturalistic edge environments. By leveraging DivNEDS to test AV perception models under diverse and nuanced conditions, researchers can validate the safety, reliability and readiness of these systems for real-world operation. The edge cases represented in DivNEDS are critical for highlighting potential deficiencies in state-of-the-art A technologies and identifying areas that require additional research and development before broad deploy can be undertaken responsibly.

B. Natural Language Interfaces

The dense image captions from *DivNEDS* could enable natural language human-computer interfaces for autonomous vehicles. This allows passengers to query the vehicle's perceptual systems using free-form speech and receive detailed natural language descriptions of the car's surroundings. Such an interface improves the trust and transparency between humans and AVs.

C. Model Interpretability and Regulatory Compliance

Researchers and engineers can harness *DivNEDS* to develop and appraise models designed specifically for interpreting edge driving scenarios. This entails not merely recognizing objects and situations, but also generating explanatory captions that elucidate the rationale behind the model's decisions, providing insights into its inferential process. Such interpretability is often a prerequisite for AV regulatory approval. *DivNEDS* enabled us to demonstrate that AV models can provide lucid, human-understandable clarifications for their perceptual surroundings.

D. Driver Assistance Systems

The dense captions generated by the models trained on *DivNEDS* could be deployed in Advanced Driver Assistance Systems (ADAS) to provide enhanced environmental awareness and hazard avoidance for human drivers. By describing nuanced edge cases, the system can alert drivers to objects and situations that pose imminent danger. This application leverages *DivNEDS* to improve driving safety without requiring complete vehicle autonomy.

E. Simulation and Virtual Testing

Diverse edge driving scenarios in *DivNEDS* can be used to construct challenging simulated environments for virtual testing and validation of autonomous vehicle systems. By augmenting existing simulation platforms with novel edge cases from *DivNEDS*, researchers could thoroughly vet AV technologies without real-world road testing. Such simulations facilitate rapid design iterations and provides crash-free assessment.

SECTION VII.
Conclusion

With its large volume of diverse, naturalistic and embedded hierarchical dense captioned images from a range of edge driving scenarios, *DivNEDS* enables the development, assessment and benchmarking of novel methods for interpretable and timely understanding of complex edge scenes. A mAP of 60.3% demonstrates the strong baseline performance of *DivNET* even as a baseline model. Going forward, harnessing datasets such as *DivNEDS* will prove key to instilling AVs with human-like abilities to instantly make sense of unusual driving environments. By spurring innovations in this direction, the *DivNEDS* resource marks an important step toward next-generation AV agents that are dependably competent in comprehending and responding to even the most difficult real-world edge case.

ACKNOWLEDGMENT

The authors acknowledge the support and mentorship of Prof. Armstrong Aboah and Andrew Phillips, Senior Director, Cloud Software Group (CSG). They also acknowledge the contributions of Drew Mashburn,

DevSecOps Entineer with CSG and Henry Ankomah of the Ministry of Local Government and Rural Development in Ghana. *(John Owusu Duah and Armstrong Aboah contributed equally to this work.)*

Authors

Figures

References

Keywords

Metrics

Code & Datasets

PDF

Help

IEEE Personal Account

CHANGE USERNAME/PASSWORD

Purchase Details

PAYMENT OPTIONS

VIEW PURCHASED DOCUMENTS

Profile Information

COMMUNICATIONS PREFERENCES

PROFESSION AND EDUCATION

TECHNICAL INTERESTS

Need Help?

US & CANADA: +1 800 678 4333

WORLDWIDE: +1 732 981 0060

CONTACT & SUPPORT

Follow

PDF

Help

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | IEEE Ethics Reporting | Sitemap | IEEE Privacy Policy