Evaluating Computer Vision Techniques for Urban Mobility on Large-Scale, Unconstrained Roads

*Harish Rithish¹, *Raghava Modhugu¹, *Ranjith Reddy¹, *Rohit Saluja¹, C.V. Jawahar¹

Abstract—Conventional approaches for addressing road safety rely on manual interventions or immobile CCTV infrastructure. Such methods are expensive in enforcing compliance to traffic rules and do not scale to large road networks. This paper proposes a simple mobile imaging setup to address several common problems in road safety at scale. We use recent computer vision techniques to identify possible irregularities on roads, the absence of street lights, and defective traffic signs using videos from a moving camera-mounted vehicle. Beyond the inspection of static road infrastructure, we also demonstrate the mobile imaging solution’s applicability to spot traffic violations. Before deploying our system in the real-world, we investigate the strengths and shortcomings of computer vision techniques on thirteen condition-based hierarchical labels. These conditions include different timings, road type, traffic density, and state of road damage. Our demonstrations are then carried out on 2000 km of unconstrained road scenes, captured across an entire city. Through this, we quantitatively measure the overall safety of roads in the city through carefully constructed metrics. We also show an interactive dashboard for visually inspecting and initiating action in a time, labor and cost-efficient manner. Code, models, and datasets used in this work will be publicly released.

I. INTRODUCTION

The ever-increasing urban population on roads has increased the number of accidents and deaths [1], many of which could have been avoided with the effective use of technology. The escalating traffic also poses technical challenges in the urban mobility-related tasks to (i) driver assistance, (ii) reliable audit and maintenance of the infrastructure, and (iii) detection and prevention of traffic violations. This is especially true for unstructured driving situations and unconstrained roads, which are common in developing countries.

Unconstrained roads suffer from uneven quality of infrastructures, such as low lighting, inadequate signage, and damaged roads, making road safety a challenging problem [2]. Fig. 1 (top) gives a glimpse of the potential problems seen in an unstructured situation. These include (i) traffic violations, (ii) missing, rusted or non-functional-signage, street lights, and traffic lights, (iii) uneven/partly damaged roads, (iv) illegal movement patterns. Therefore, tasks such as regular audit and maintenance of road infrastructure and traffic violation monitoring have become critical for safety.

Current approaches for addressing road safety, however, rely on manual interventions or large camera networks. This is not scalable or economical for massive road lengths in ever-expanding urban settlements [3], [4]. Thus, we propose to use a simple mobile imaging system to address several common problems in road safety at scale using recent computer vision algorithms. As shown in Fig. 1, the mobile imaging system captures the data across the city, and the proposed road safety system is used to generate the city-level maps for different irregularities. The proposed system is highly automated, affordable, and scalable. We demonstrate our system on 2000 km of unconstrained road scenes, captured across an entire city. Through this, we quantitatively assess the overall safety of roads in the city through carefully constructed metrics. We also develop an interactive dashboard for visually inspecting and initiating action in a time, labor, and cost-efficient manner.

However, vision-based detection algorithms in unstructured settings can be affected by continuously changing backgrounds, different road types (shape, color), variable
lighting conditions, and weather. Therefore, we investigate the behavior of computer vision algorithms in diverse settings before deploying them in the real-world. For this purpose, we collect data in unconstrained road scenes and annotate them with thirteen condition-based hierarchical labels. These conditions include different timings, road type, traffic density, and state of road damage. We study the strengths and shortcomings of recent computer vision techniques in identifying irregularities in road infrastructure and spotting traffic violations on diverse and unconstrained roads through this evaluation.

We select tasks that are critical for road safety and reflect the overall compliance to traffic rules in a city. As part of road infrastructure inspection, we identify possible irregularities in streets (including missing lane markings and potholes), absence of street lights, and defective traffic signs. Such infrastructure provides essential visual cues for navigation and ensures safe mobility. In the case of traffic violations, we focus on two-wheeled motor vehicle riders, as they account for nearly 30% of road accident deaths worldwide [5]. Among them, around 74% of riders were found to be not wearing protective helmets [6]. Therefore, we focus on spotting helmet violations.

To summarize, the main contributions of this paper are:

- We investigate the behavior of recent computer vision techniques in inspecting road infrastructure and traffic violations under diverse and unconstrained conditions.
- We demonstrate our system at a city-scale and assess the safety of roads.
- We release a 75-hour dataset of long video sequences captured in unstructured settings. These sequences are classified into 13 different conditions of time, road type, state of road damage, and traffic.

II. RELATED WORK

Many previous works have demonstrated the applications of computer vision models for object detection, semantic segmentation, and vehicle tracking in urban mobility [7], [8], [9], [4]. The challenges like Nvidia AI city[10], CityFlow[11], and AutoNUE[12] have popularized city-scale mobility-related tasks.

A. Road Safety Systems

Early work on road safety [13], [14] demonstrated the effective collaboration between perception and control for increasing safety. Pedestrian detection [15], driver state and behavior characterization, etc., became an integral part of ADAS system over the years. Zhang et al. [16] demonstrated a stereo vision-based training-free approach to detect road regions accurately and robustly. Comprehensive surveys on vision-based traffic monitoring and vision applications in urban analytics are conducted regularly [17], [18], [19].

Dynamic objects around a vehicle also impact the safety and performance of the vehicles. Billones et al. [20] measure a vehicle’s speed by counting the frames it takes to pass a region of interest or reference lines. Singh et al. [21] present a framework for visual big data analytics that automatically detects bike-riders without helmets. They further discuss the challenges associated with city-scale surveillance for traffic control. Recent works have shown CNN to produce better results for helmet detection and classification [22]. Frossard et al. [23] detect driver’s intent by detecting turn signals and emergency flashers in video sequences. Datasets and robust models on the pedestrian intent prediction are also present in the literature [24], [25], [26], [27], [28]. In comparison to these works, we use recent vision techniques (refer Section IV) to inspect helmet violations at scale on the unconstrained road scenes under different conditions.

B. Road Inspection Systems

The soundness and suitable perceptibility of road infrastructure form the essential elements of driver safety. Burschka et al. [29] introduce a vision-based system for traffic sign detection and ego-motion estimation. Another early work involves detecting cracked regions using a weakly supervised superpixel classifier trained on 220 images [30].

To reduce asphalt pavement distresses, Kanza et al. [31] detect and localize pothole in 120 cropped road images using histogram of oriented gradients (HOG) features and Naive Bayes classifier. Zhang et al. [32] employ Convolutional Neural Network (CNN) to classify image patches of the road as defective or non-defective and evaluate it on 500 images. Yerram et al. [8] train a multi-step ERFRNet based model on around 1000 unconstrained road scenes to segment road pixels into 9 different kinds of defects. However, all these road inspection methods are applied at a small scale compared to the approach in this paper.

The work closest to ours in this direction are i) Ma et al. [3] who utilize Fisher vectors with Convolutional Neural Networks (CNN) to classify 700K images from 70K street

---

**TABLE I**: 2000 km of video data is classified into following categories in different conditions at a resolution of 1 second. Criterion to classify in each condition is provided in the second row. The condition to determine a category is provided within the brackets. v denotes visual discretion.

| Condition | Road Type | Traffic Density | Road Damage | Time | Capture time (24 hr) |
|-----------|-----------|-----------------|-------------|------|---------------------|
| Criteria  | #Lanes    | #Vehicles       | #Potholes   |      |                     |
| Categories| Narrow (1) Standard (2-3) Highway (>3) Bridge (v) | Sparse (<4) Moderate (5-8) Dense (>8) | Low (<2) Moderate (3-4) High (>4) | Morning (7-12) Noon (12-16) Evening (16-19) |
Fig. 2: Distribution of the 13 condition-based hierarchical frame-level labels. These conditions represent diversity in road type, traffic density, time of capture, and state of road damage. The annotations are carried out 2000 km of road scenes captured in unstructured settings.

Fig. 3: Dangerous road conditions. Clockwise from top-left: bumpy and muddy road, low-hanging wires hazardous for heavy vehicles, unguarded under construction structure and narrow street with two-way traffic.

Fig. 4: Traffic participants and driving behavior. Clockwise from top-left: diversity in participants, a pedestrian (bottom-left) jaywalking during heavy traffic, wrong-side driving and illegal parking on the road, and chaotic unmanned junctions.

segments into poor, fair, and good, and ii) Modhugu et al. [33] who demonstrates an automatic model for road infrastructure audit concerning traffic signs, street lights, and lane markings. In contrast to these works, we extend the range of functionalities, volume of training data and test the system’s scalability extensively on a larger dataset spanning 2000 km of unconstrained roads.

III. DATASET
A. Comparison with Other Datasets

Datasets like KITTI [34], Cityscapes [35], Argoverse [36], and NuScenes [37] are captured in developed countries where infrastructure is well-developed and road activity is structured. The results obtained from these datasets are often not directly applicable in unstructured road situations prevalent in large parts of the world. Verma et al. [38] introduced the India Driving Dataset (IDD), a novel dataset for road scene understanding in unstructured environments to meet these needs. It consists of 10,000 images, with 34 classes collected over 182 drive sequences on roads in various parts of India. They use a rich label set with a four-level hierarchy to represent the nuances of real driving behaviors in semantic segmentation. Our attempt is evaluated on situations similar to that of IDD, but at a much large scale in road length and the number of driving hours. In contrast to the previous attempts in IDD, we do not focus on semantic segmentation or the detection of common road objects. We focus on irregularities on the road surface and traffic infrastructure. Fig. 2 demonstrates the diversity in the conditions we have used in the evaluation.

B. Challenges

Constant wear and tear, heavy rains, and laying of underground cables necessitate routine maintenance of roads. Economic and regulatory hindrances, lack of maintenance, weak accountability, and unreliable funding have resulted in unconstrained road conditions. As shown in Figure 3, rough terrains including muddy, bumpy roads, and defects such as potholes, waterlogs and hazardous road objects, are common. The problem is compounded due to varying levels of occlusions, unstructured motion, and cluttered background. Other road infrastructure components that aid navigation is found in a similar state of poor planning and maintenance. Lane markings either do not exist or are covered in dust, traffic signs are worn out or are concealed amidst the background clutter, and street lights are obscured by vegetation.
As shown in Figure 4, there is also a large variation in the traffic participants and their behavior. Two and three-wheeler vehicles are very prominent and contribute immensely to the unstructured motion on the road. Stray animals and jaywalking pedestrians complicate this problem even further. Moreover, unsustainable population and vehicle density in cosmopolitan cities, combined with weak enforcement of traffic rules, has brought about disarray in driving behavior. Non-adherence to road lanes, traffic congestion, and chaotic unmanned junctions are widely prevalent. Additionally, violation of traffic rules such as wrong-side driving, swerving, vehicle overloading, and illegal parking cause further confusion.

C. Mobile Imaging System and Data

Fig. 5 shows our setup for data acquisition. We capture 1920×1080 resolution videos at 15 frames per second. In total, we record 257 video sequences amounting to 75 hours of capture and covering a distance of over 2000 km. The sequences have an average duration of 17 minutes and are up to 73 minutes long. The videos are synchronized to a GPS device polling every second. The collected videos are annotated at 1-second resolution for road type, traffic density, state of road damage, and time of capture. We present this distribution in Fig. 2 and the corresponding criterion for classification in Table I.

IV. ROAD SAFETY MONITORING PIPELINE

We are interested in identifying possible irregularities in streets (including missing lane markings and potholes), absence of street lights, defective traffic signs, and traffic violations from videos. We achieve this by constructing a simple and efficient pipeline using recent computer vision techniques, as shown in Fig. 6. In particular, we select computer vision models that achieve (i) high performance and are widely adopted by the community and (ii) fast inference to facilitate evaluation on large data. We, however, note that the network selection is not a focus of this work and can be replaced with other architectures.

A. Identification of Road Objects

Lane markings are continuous and require strong structural priors for segmentation. We employ Spatial Convolutional Neural Network (SCNN) [39] to segment lane markings as it also extrapolates the lane markings occluded by other objects on the road surface.

Traffic signs, street lights, potholes, and traffic participants are directly detected from the incoming video feed. We employ YOLOv4 by Bochkovskiy et al. [9] as it provides a good trade-off between detection performance and speed. As part of traffic participants, we detect riders and motorcycles separately and associate them using simple heuristics.

Helmet violations are computed on the detected crops of riders and motorcycles. We use the WPOD-NET by Silva et al. [40], which detects and geometrically transforms the license plates to a planar front-view. For helmet detection on riders, we employ YOLOv4 for the same reasoning mentioned earlier.

B. Tracking Identified Objects

The detected objects viz. traffic signs, street lights, potholes, and traffic participants are passed to a tracker. By using a tracker, we (i) avoid redundant counting, (ii) temporally fuse predictions, and (iii) smoothen detections across the frames. We use Simple Online and Realtime Tracking (SORT) by Bewley et al. [41] since it is extremely fast and provides relatively high accuracy. SORT tracks objects by employing Kalman filter to handle motion prediction and the Hungarian method for frame-by-frame data association. They

![Identification of Road Objects](image1)

![Tracking Identified Objects](image2)
use bounding box overlap of detections as the association metric.

C. Temporal Fusion

Since subsequent frames provide a different viewpoint of the same object to the network, we combine individual frames’ predictions and smoothen any aberration. We use majority voting [42], a standard fusion technique, to combine frame-level predictions of traffic sign classification, rider-motorcycle association and helmet classification.

For initial training and quantitative analysis, we annotate 4.3K images which are sampled randomly from our data capture. We use a 80:20 train-test split for this purpose. We report the detection results for all the tasks in Table II. Our entire pipeline runs at ∼10 frames per second, making it feasible to evaluate on large data.

V. Evaluation under Diverse Conditions

The road scene around us is visually diverse and constantly in-transition, with highly-dynamic traffic participants riding on infrastructure that varies by locality and maintenance-levels, and environmental conditions. Studies have shown that such diverse and dynamic settings affect computer vision algorithms [2]. However, prior works that assess road safety do not consider this possible impact on their model performance. This raises concerns regarding the robustness and reliability of computer vision models while deploying them in the real-world. Therefore, we evaluate our proposed framework under varying conditions of the environment (time), road infrastructure (road damage, road type), and road activity (traffic).

For this evaluation, we use the hierarchical condition-based labels described in Section III-C. We sub-sample 100 images from each sub-category and annotate them with bounding-boxes for the detection components in our pipeline. We consider two different design choices for sampling images for a category of interest (say, traffic density). In the first option, we select and fix the conditions in the other categories (e.g., morning-time, narrow-road, low-road damage) while sampling images in the category of interest (i.e., low, medium, and high-traffic density). In the second option, we sample from a uniform distribution without considering the labels of other categories. While the first option allows interpreting results in a conflict-less manner, they would be biased to the exact set of conditions chosen by us. Contrastingly, while the second option makes it slightly tricky for interpreting results, it captures the natural variation in data and avoids being biased towards a chosen set of conditions. Thus, we proceed with the latter option and rely additionally on qualitative results to strengthen our reasoning.

We present the results of this experiment in Table III. Analyzing the sub-category values at different times of the day, we notice uniformity in performance, row-wise, while detecting traffic signs, traffic participants, and helmets. However, there is a noticeable drop in mAP scores of street light and pothole detection during the afternoon session. We observe that this is due to false-negatives in the case of street lights and false-positives with regards to potholes. We attribute the false-negatives to the sunlight’s glare on the street lights, whose intensity is at its peak in the afternoon (Fig. 7a). Interestingly, this glare does not affect the performance of other detection modules. On the other hand, the incidence...
of the sun’s rays is orthogonal to the objects on the road at this time. This produces shadows that are visually similar to potholes and leads to false positives (Fig. 7b). However, we filter out such false-positive cases by removing the tracks covering few (≤3) frames.

We now look at the results under the traffic density category. Since street lights are situated much above the road surface, their visibility is not affected by the traffic participants. However, occlusion arising from dense traffic plays a significant role while detecting traffic signs, traffic participants, and potholes. Since street signs are mounted around vehicle-height, they are severely impacted by heavy traffic activity on the ground (Fig. 7c). In the case of potholes, traffic participants and their shadows significantly reduce the potholes’ visibility (Fig. 7d). As for traffic participants, we observe that, often, nearby riders are confused as a single entity by the model, leading to false negatives (Fig. 7e).

While looking at the state of road damage, we are mainly interested in the detection of potholes. From Table III, we notice a drop in mAP scores when the damage to the road is medium or high. However, while inspecting the qualitative outputs, we observe the model outputs to cover most potholes even when the road is highly damaged (Fig. 7f). We notice that the discrepancy between quantitative and qualitative outputs is due to nearby potholes being detected as a single entity.

To summarize, we notice that the performance of models is acceptable in most settings. We identify the specific (task, condition) combinations for which computer vision models struggle. We observe occlusion in dense traffic and sunlight in the afternoon as the major factors influencing performance and back our reasoning with qualitative outputs. Moreover, we find that the impact of these factors is task-specific, highlighting the need and importance of this study.

VI. CITY-SCALE ROAD SAFETY ASSESSMENT

We deploy our scalable system in the real-world and assess the safety of roads at a city-scale. First, we quantitatively measure the overall safety through carefully constructed metrics. We then show our interactive dashboard for visually inspecting the outputs at an individual and global level. This dashboard serves as a platform for users to analyze and initiate action in a time, labor, and cost-efficient manner. Our demonstrations are carried out on the 2000 km of data captured on unconstrained roads across the city.

A. Quantitative Assessment

Metrics for Evaluation: Traffic signs pre-inform riders of impending road scenes, and hence their range of visibility plays a crucial role. We measure the visibility range of a traffic sign as the distance (calculated using GPS location) traced by the ego-vehicle from the traffic sign’s first frame of detection to its last frame of detection. Additionally, we also classify traffic signs as defective or normal. For lane markings and street lights, we expect them to occur all along the road periodically. Therefore, in these cases, the measure of interest is the road stretches without lane markings and street lights. We tag the detected street lights geographically and calculate the average distance between street lights along the route. For lane markings, we use the normalized lane regularity score described in [33]. This score considers the percentage of pixels identified as lane markings along a road stretch covering 50 meters. Based on this score, we classify the lane markings as fair, faded, or absent. To measure the quality of the road surface, we calculate the number of road defects for every 100 meters. We then classify these road stretches as poor, average, or fair based on the frequency of potholes. For helmet violations, we calculate the percentage of riders not wearing results since it portrays road participants’ compliance level to traffic rules.

Results: In Table IV, we present the quantitative measure of road safety in the city. The results shed light on the lack of maintenance of roads and road objects. Traffic signs are visible only at a close range of 10m, and a significant proportion (37.5%) of them are either rusted or faded out. Further, on average, only one street light is present every 165 meters, making night-driving dangerous. Lane markings are mostly absent or faded (60.3%), especially on side roads and streets. Seldom road repairs result in 4.0% of road stretches with potholes. Finally, we observe that 45.9% of riders violate helmet rules, indicating low compliance with traffic rules. These results indicate the need for urgent actions to ensure the safety of traffic participants.

B. Qualitative Inspection

Since computer vision models do not work perfectly in the real-world, we develop an interactive dashboard (Fig. 8) for humans to verify outputs, perform analysis and initiate action. While we run our framework to identify the objects of interest, we also log the corresponding metadata (time and global coordinates) of capture. Utilizing this information, we construct heatmaps and route maps that show the state of road infrastructure and compliance-level to traffic rules.

| Traffic signs | Street lights | Lanes | Potholes | Helmet |
|---------------|---------------|-------|----------|--------|
| Visibility-Range | Defective | Avg. pair distance | No markings | Percentage of stretches | Percentage of Violating riders |
| 9.7m | 37.5% | 165m | 60.3% | 4.0% | 45.9% |
at a city-scale (Fig. 1 (iii) and Fig. 8a)\(^1\). Such global visualizations can also help in instantly narrowing down on hotspots in the city, where road safety is in a critical state. The dashboard provides functionality to zoom-in on these hotspots and inspect the individual outputs, along with image and video evidence. We can also configure the system to automatically provide warnings when the safety-level falls below a certain threshold. We provide an additional interface for inspecting traffic violations that makes it effortless to issue tickets after reviewing the image and video evidence (Fig. 8b). Upon verification, the automatically recognized license-plate number is searched on the vehicle registration database, and a ticket can be issued to the vehicle owner instantly. We believe this dashboard would serve as a powerful tool to analyze outputs and efficiently initiate action.

### VII. Conclusion

In this paper, we propose a simple mobile imaging setup to address a number of common problems in road safety at scale. We use recent computer vision techniques to identify possible irregularities on roads, the absence of street lights, defective traffic signs. Besides inspecting road infrastructure, we also spot traffic violations. We then investigate the strengths and shortcomings of computer vision techniques on thirteen condition-based hierarchical labels for different timings, road type, traffic density, and state of road damage. We demonstrate our system on 2000 km of unconstrained road scenes and quantitatively measure the overall safety of roads in the city. We also show an interactive dashboard for visually inspecting and initiating action in a time, labour and cost-efficient manner. We release the trained models, code, and annotations to encourage future work in this direction.

### References

[1] D. Tran, E. Tadesse, D. Osipychev, J. Du, W. Sheng, Y. Sun, and H. Chen, “A collaborative control framework for driver assistance systems,” in 2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2017, pp. 6038–6043.

[2] C. Jawahar and V. Padmanabhan, “Technology interventions for road safety and beyond.” ACM, 2019.

[3] K. Ma, M. Hoai, and D. Samaras, “Large-scale Continual Road Inspection: Visual Infrastructure Assessment in the Wild.” in BMVC, 2017.

---

\(^1\)As discussed earlier, we filter out the false positives (second image from left in Fig. 1 (iv) and Fig. 7b) by removing the tracks covering ≤ 3 frames.
machine eyes: A review of deep computer vision in urban analytics," 
*Cities*, vol. 96, p. 102481, 2020.

[20] R. K. C. Biliones, A. A. Bandala, E. Sybingco, L. A. G. Lim, and E. P. Dadios, “Intelligent system architecture for a vision-based contactless apprehension of traffic violations,” in 2016 IEEE Region 10 Conference (TENCON). IEEE, 2016, pp. 1871–1874.

[21] D. Singh, C. Vishnu, and C. K. Mohan, “Visual Big Data Analytics for Traffic Monitoring in Smart City,” in 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), 2016, pp. 886–891.

[22] C. Vishnu, D. Singh, C. K. Mohan, and S. Babu, “Detection of motorcyclists without helmet in videos using convolutional neural network,” in 2017 International Joint Conference on Neural Networks (IJCNN). IEEE, 2017.

[23] D. Frossard, E. Kee, and R. Urtasun, “Deepsignals: Predicting Intent of Drivers Through Visual Signals,” in 2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019, pp. 9697–9703.

[24] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert, “Activity Forecasting,” in European Conference on Computer Vision. Springer, 2012, pp. 201–214.

[25] C. G. Keller and D. M. Gavrila, “Will the Pedestrian Cross? A Study on Pedestrian Path Prediction,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 2, pp. 494–506, 2013.

[26] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila, “Context-Based Pedestrian Path Prediction,” in *European Conference on Computer Vision*. Springer, 2014, pp. 618–633.

[27] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, “Are They Going to Cross? A Benchmark Dataset and Baseline for Pedestrian Crosswalk Behavior,” in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 206–213.

[28] B. Liu, E. Adeli, Z. Cao, K.-H. Lee, A. Shenoi, A. Gaidon, and J. C. Niebles, “Spatiotemporal Relationship Reasoning for Pedestrian Intent Prediction,” *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3485–3492, 2020.

[29] D. Burschka and G. D. Hager, “Vision-based 3d scene analysis for driver assistance,” in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*. IEEE, 2005, pp. 812–818.

[30] S. Varadhakarajan, S. Jose, K. Sharma, L. Wander, and C. Mertz, “Vision for Road Inspection,” in *IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2014, pp. 115–122.

[31] K. Azhar, F. Murtaza, M. H. Youasf, and H. A. Habib, “Computer Vision Based Detection and Localization of Potholes in Asphalt Pavement Images,” in *2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*. IEEE, 2016, pp. 1–5.

[32] L. Zhang, F. Yang, Y. D. Zhang, and Y. J. Zhu, “Road crack detection using deep convolutional neural network,” in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 3708–3712.

[33] M. Raghava, R. Reddy, and C. Jawahar, “Dear commissioner, please fix these: A scalable system for inspecting road infrastructure,” *NCVPRIPG*, 2019.

[34] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3354–3361.

[35] M. Cords, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[36] M.-P. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, et al., “Argoverse: 3d tracking and forecasting with rich maps,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8748–8757.

[37] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, “nuscenes: A multimodal dataset for autonomous driving,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11621–11631.

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