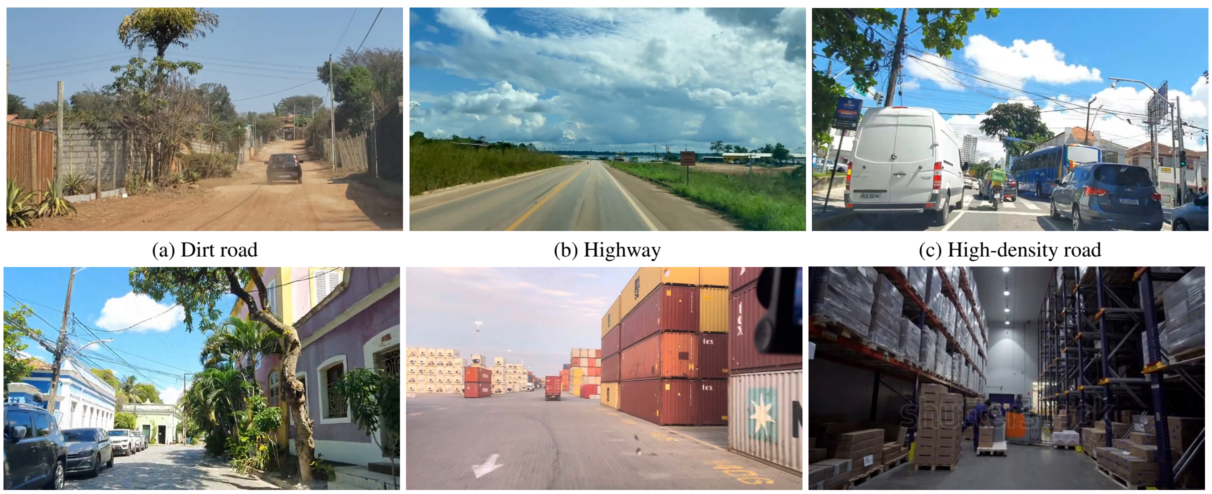
VWise: A novel benchmark for evaluating scene classification for vehicular applications

Pedro Azevedo[1] Emanuella Araújo[1] Gabriel Pierre[1] Willams de Lima[1]
João Marcelo Teixeira[1,2] Valter Ferreira[3] Roberto Jones[4] Veronica Teichrieb[1]

[1]Voxar Labs, Centro de Informática, Universidade Federal de Pernambuco
[2]Departamento de Eletrônica e Sistemas, Universidade Federal de Pernambuco
[3]Volkswagen Trucks and Bus [4]Eyeflow.AI



(a) Dirt road (b) Highway (c) High-density road

(d) Low-density road (e) Port or loading dock (f) Factory access area

Figure 1. An overview of the VWise dataset, containing six classes from urban, industrial, and rural settings.

Abstract

In this work, we propose VWise, a novel dataset for road-type classification and scene classification focused on external contexts related to vehicular applications in Latin America. Existing datasets today lack geographic diversity and representation of road contexts outside North America and Europe. To address this problem, we collect over 520 video clips and 4,500 images covering diverse urban and rural environments across Latin American countries, annotated with six classes of road types. We evaluate several state-of-the-art image and video classification models in baseline experiments on this dataset, obtaining over 84% accuracy. We aim to support training and evaluating models for tasks such as road classification to enhance research on vehicular tasks in Latin America.

1. Introduction

With the current results in fields related to vehicular applications, such as autonomous driving and driver-assistance systems, interest in these fields of applications has been rising over the past years. Overall, new applications are made available and being deployed on multiple vehicles in regions such as North America or Europe [20], such as the detection of pedestrians, vehicles, and traffic lights [18], the identification of pathways in streets and highways [6], as well as evaluating their performance under different climatic conditions [10].

Multiple datasets have been proposed in the past to facilitate the arrival of these technologies; however, there is a clear gap in deploying models trained and evaluated in these datasets to other regions of the world, especially Latin America, which is caused mostly by the different contexts between these regions. A clear example of this difference is how roads in Latin America are mostly unmaintained, while roads in North America and Europe have a higher rate of maintenance. Therefore, collecting data and building datasets that would expand the geographical representation of these studies could lead to higher performance in these regions.

In light of this gap, we propose VWise, a novel dataset mainly focused on external context classification tasks, especially road classification, collected in various regions in Latin America. We expect that VWise will support training and evaluating state-of-the-art models, enriching research in tasks related to vehicular applications, such as autonomous driving or advanced driving assistance systems, and opening paths for deployment in this region.

2. Related works

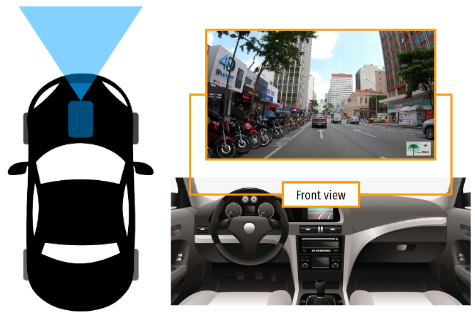
In automotive research, datasets play a pivotal role, dividing into categories such as those utilizing static cameras, exemplified by CityFlow [17], which gathers images from traffic security cameras and annotates them. Another significant category includes datasets with onboard cameras like CityScape [2], offering driver's perspective views of streets, pedestrians, and road obstacles.

Focusing on vehicle onboard-captured images, we distinguish between two main categories: Indoors, showcasing the vehicle's interior, and Outdoors, displaying the vehicle's exterior. Indoor datasets observe actions within vehicles, focusing on drivers' or passengers' emotions, attention, and situational awareness. Examples like 100-Driver [19] and MDAD [9] examine driver distraction with a focus on posture, while others delve into distraction based on head or eye movements [3, 14].

For scene recognition, capturing onboard, outdoor images is essential. Public datasets geared towards autonomous driving, such as KITTI [5] and Cityscape [2], primarily focus on traffic environment perception without providing specific road type annotations. While CityScape [2] and KITTI [5] offer a populous dataset with detailed descriptions of buildings, objects, vehicles, and humans, they lack traffic information. Similarly, nuScene [1], Waymo Open [15], ApolloScape [8], and ONCE [11], although large-scale and real-world, supporting 3D object detection, tracking, and activity predictions, share the limitation of not situating the vehicle in a specific scenario, like a rural dirt road or a metropolitan avenue, potentially leading to inaccurate activity predictions.

Public datasets like the Road Surface Classification Dataset (RSCD) [21] label certain road features but fail to specify the type of road the vehicle is on. Moreover, these data have a noticeable geographic concentration in North America and Europe [20]. To address these gaps, we propose a new extensive dataset for classifying images and videos by road types aimed at assisting driving in various scenarios, with a focus on Latin America. This effort is designed to overcome existing limitations by offering a resource that enhances the development and applicability of autonomous vehicle technologies across diverse geographical and situational contexts.

3. The VWise Dataset



(a) (b)

Figure 2. Desired camera setup for VWise. A point-of-view camera setup allows us to have a similar view to the driver.

To tackle the lack of geographic representativeness in today's datasets, we collect the VWise dataset based on public access videos from YouTube and stock video libraries. Queries related to countries (e.g., Brazil, Colombia, Peru...) and cities (e.g., Quito, Recife, São Paulo...) were combined with queries that would filter down to videos from point-of-view camera placements, as we exemplify in Figure 2. Although this seems a naive approach, it allows us to quickly capture data from diverse regions of Latin America, and the lack of quality in some videos, which are related to different cameras used in recordings (from GoPros to smartphones), actually contributes to a diverse dataset.

We have selected a diverse and comprehensive representation of urban and rural environments in multiple regions of Latin America, related to maintenance levels, suburban regions, or downtown regions, for example. We selected a total of 45 videos, from which we have cropped 521 video clips usually containing 10 seconds. We also processed a static version of the dataset for tasks that uses images as input, generating a total of 4,500 images by sampling these videos. The dataset is divided into *train*, *test*, and *val* splits following the well-accepted 70-20-10 split.

3.1. Data annotation

The data annotation process involves four students doing research on vehicular applications. Two are undergraduates, one is a master's student, and the other is a doctoral candidate, all from computer science and with experience in deep learning applications. The annotation process was performed blindly and independently. We have also employed CLIP [13] to enhance the efficiency of annotations. We have classified the data on six classes, as we show on Table 1.

Class	Description	Videos
Low-density urban street	Streets with low vehicle and pedestrian flow (residential areas and suburbs).	97
High-density urban avenue	Avenues with high traffic density, vehicles, pedestrians, and traffic sign complexity.	59
Highway	Fast tracks and highways, focusing on high-speed situations.	124
Factory interior	Internal factory scenes, including assembly lines, storage, and logistics areas.	24
Port or dock area	Port areas with intense loading and unloading activities, ships, and containers.	86
Dirt or service road	Dirt or service roads, challenging navigation on irregular and less structured terrain.	131

Table 1. Dataset classes, with their descriptions and number of videos in each class.

First, we generated descriptions from the list of classes to serve as input for CLIP, which bulk annotated the entire dataset. After this, multiple annotators verified the labels attributed by CLIP and attributed a new label based on a majority voting rule. Overall, we have seen that based on its zero-shot capabilities, CLIP has annotated most videos with significant accuracy.

3.2. Data curation

We have assessed our dataset using common data curation tools such as Cleanlab[1] and Encord Active[2] to check for the distribution of various image metrics such as area, ratio, brightness and distribution of RGB values to fix outliers and check if more data should be collected to distribute the dataset better. Overall, we have performed several iterations of checking the data and collecting more videos to reach a better data distribution.

4. Experimental baseline

We have also evaluated several state-of-the-art techniques in our dataset to establish an experimental baseline. This experimentation allows other researchers to have a still base of comparison in the future. Specifically, in this evaluation, we set a baseline for the task of road-type classification.

4.1. Image classification

We have selected common image classification architectures, namely EfficientNet-V2 [16], MobileNet-V3 [7], and Vision Transformers (ViT-B-16) [4], for the task of road-type classification.

Implementation details. We have used public implementations of these models as available for the PyTorch framework. They were trained on a single RTX 3050 6GB Laptop GPU using CrossEntropy and SGD with a learning rate of 0.001.

4.2. Zero-shot classification

We have deployed two models for the task of zero-shot classification: CLIP [13] for images and X-CLIP [12] for videos. In this case, as suggested by the good practices in the field, we have used the validation set of our dataset to learn a set of input prompts that would lead to a better result. We show the input prompts for each class in Table 2. We do not discuss the implementation details of these models since we follow the default implementation's training procedures.

Class	Input prompt
Dirt or service road	A rural dirt road or service road
Factory interior	Factory interior
High-density urban avenue	High-density urban avenue
Highway	A highway suitable for road trips
Low-density urban street	Low-density small urban road
Port or dock area	Dock or port area

Table 2. Prompts used for input in CLIP and X-CLIP.

4.3. Results and discussion

For the image classification models, MobileNet-V3 has displayed the higher accuracy of the three models, with an accuracy of 95.02%. EfficientNet-v2 was ranked second, with 91.16% accuracy, and ViT-B-16 ranked third with 88.40%. We show the results from this quantitative evaluation in Table 3. Overall, we can see that the three architectures perform very well on this task and are able to learn strong representations from this data.

Architecture	Accuracy (%)
VisionTransformer (Vit-b-16)	88.40
EfficientNet-V2	91.16
MobileNet-V3	**95.02**

Table 3. Baseline results on common image classification architectures, as trained and evaluated on VWise.

For zero-shot classification, CLIP and X-CLIP have reached similar accuracy levels; it is important to remind, though, that CLIP uses images and inputs, while X-CLIP

[1] Available at https://cleanlab.ai/

[2] Available at https://github.com/encord-team/encord-active

expects videos. For this evaluation, CLIP have reached an accuracy of 84.70% and X-CLIP, 86.52%, as we show in Table 4.

Architecture	Accuracy (%)
CLIP	84.70
X-CLIP	86.58

Table 4. Baseline results on common zero-shot classification architectures, as trained and evaluated on VWise.

5. Conclusion

In this work, we propose VWise, a novel dataset collected in Latin America for multiple vehicular tasks, but focused on road-type classification. With this dataset, we expect to open paths for deploying automotive systems in this region, since nowadays these models are mostly trained and evaluated on datasets sampled in North America or Europe. We also perform a baseline study that will serve as base of comparison for future works that will use VWise.

5.1. Future works

For future works, we plan on enhancing VWise with further annotations related to context with the objective of allowing other applications to leverage it. Examples of annotations include ground-truth levels of road maintenance quality, ground-truth lane annotations, and other features related to the road surface. We also plan a study on ambiguity; it is common to have highways that also act as avenues when within cities in countries such as Brazil. In this case, should these samples be annotated as highways or high-density roads? We have already pinned several cases of ambiguity, and we plan on tackling these cases before the final version of the dataset to allow for a more robust set of annotations. VWise is also the first step in a research pipeline that is planned within our research group. We plan on using this dataset to develop and deploy vehicular applications that make use of road-type classification.

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2

[3] Bogusław Cyganek and Sławomir Gruszczyński. Hybrid computer vision system for drivers' eye recognition and fatigue monitoring. *Neurocomputing*, 126:78–94, 2014. 2

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[5] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 2

[6] Hiroto Honda and Yusuke Uchida. Clrernet: improving confidence of lane detection with laneiou. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1176–1185, 2024. 1

[7] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 3

[8] Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. The apolloscape dataset for autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 954–960, 2018. 2

[9] Imen Jegham, Anouar Ben Khalifa, Ihsen Alouani, and Mohamed Ali Mahjoub. Mdad: A multimodal and multiview in-vehicle driver action dataset. In *Computer Analysis of Images and Patterns: 18th International Conference, CAIP 2019, Salerno, Italy, September 3–5, 2019, Proceedings, Part I 18*, pages 518–529. Springer, 2019. 2

[10] Wenyu Liu, Gaofeng Ren, Runsheng Yu, Shi Guo, Jianke Zhu, and Lei Zhang. Image-adaptive yolo for object detection in adverse weather conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1792–1800, 2022. 1

[11] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Hanxue Liang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, et al. One million scenes for autonomous driving: Once dataset. *arXiv preprint arXiv:2106.11037*, 2021. 2

[12] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*, pages 1–18. Springer, 2022. 3

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3

[14] Anke Schwarz, Monica Haurilet, Manuel Martinez, and Rainer Stiefelhagen. Driveahead-a large-scale driver head pose dataset. In *Proceedings of the IEEE Conference on

Computer Vision and Pattern Recognition Workshops, pages 1–10, 2017. 2

[15] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 2

[16] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021. 3

[17] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasiu, and Jenq-Neng Hwang. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8797–8806, 2019. 2

[18] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. You only learn one representation: Unified network for multiple tasks. *arXiv preprint arXiv:2105.04206*, 2021. 1

[19] Jing Wang, Wenjing Li, Fang Li, Jun Zhang, Zhongcheng Wu, Zhun Zhong, and Nicu Sebe. 100-driver: a large-scale, diverse dataset for distracted driver classification. *IEEE Transactions on Intelligent Transportation Systems*, 2023. 2

[20] Hang Yin and Christian Berger. When to use what data set for your self-driving car algorithm: An overview of publicly available driving datasets. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–8. IEEE, 2017. 1, 2

[21] Tong Zhao, Junxiang He, Jingcheng Lv, Delei Min, and Yintao Wei. A comprehensive implementation of road surface classification for vehicle driving assistance: Dataset, models, and deployment. *IEEE Transactions on Intelligent Transportation Systems*, 2023. 2