Investigating Vision-Language Model for Point Cloud-based Vehicle Classification

Anonymous CVPR submission

Paper ID *****

Abstract

Heavy-duty trucks pose significant safety challenges due to their large size and limited maneuverability compared to passenger vehicles. A deeper understanding of truck characteristics is essential for enhancing safety perspective of cooperative autonomous driving. Traditional LiDAR-based truck classification methods rely on extensive manual annotations, which makes them labor-intensive and costly. The rapid advancement of large language models (LLMs) trained on massive datasets presents an opportunity to leverage their few-shot learning capabilities for truck classification. However, existing vision-language models (VLMs) are primarily trained on image datasets, which makes it challenging to directly process point cloud data. This study introduces a novel framework that integrates roadside LiDAR point cloud data with VLMs to enable efficient and accurate truck classification, which supports cooperative and safe driving environments. Our approach introduces three key innovations: (1) leveraging real-world LiDAR datasets for model development, (2) designing a preprocessing pipeline to adapt point cloud data for VLM input, including point cloud registration for dense 3D rendering and smoothing techniques to enhance feature representation, and (3) utilizing in-context learning with few-shot prompting to enable vehicle classification with minimal labeled training data. Experimental results demonstrate the effectiveness of this framework, showcasing its potential to reduce annotation efforts while improving classification accuracy.

1. Introduction

Trucks, especially heavy-duty vehicles, present significant challenges in roadway safety due to their large size, weight, and limited maneuverability. Compared to passenger vehicles, trucks require longer stopping distances and have larger blind spots, increasing the likelihood of severe crashes. Understanding truck characteristics, such as body configuration and movement patterns, is crucial for improving traffic safety, optimizing infrastructure design, and enhancing automated vehicle perception systems. Traditional truck classification methods using LiDAR sensors rely on handcrafted feature extraction and extensive manual annotations to build robust datasets [10] [11]. These approaches are time-consuming, expensive, and often lack generalizability across different road environments. Recent advancements in Multi-modal Large Langrage Model, particularly vision-language models (VLMs), have shown remarkable performance in various image-based tasks by leveraging large-scale pretraining and few-shot learning capabilities [7]. However, most VLMs are trained on 2D image datasets, posing a significant challenge when applying them to the point cloud data. This study aims to address the existing gap by leveraging the representational power of VLMs for LiDAR-based heavy-duty truck classification. The key contributions of this work are as follows. First, We utilize roadside LiDAR sensor data to capture detailed point cloud representations of heavy-duty trucks, ensuring the approach's practical applicability in real-world scenarios. Second, we propose a systematic method to adapt point cloud data for VLM input. This includes point cloud registration to generate dense 3D renderings and point cloud smoothing techniques to enhance feature representation, improving the model's ability to process and classify the data. Third, we introduce a few-shot prompting approach that allows VLMs to classify vehicles, particularly heavy-duty trucks, without costly parameter updates. This approach significantly reduces the need for extensive manual annotations, which makes the classification process more efficient and scalable.

2. Preliminary

2.1. Vision Language Model

Vision-language models (VLMs) are a class of multimodal generative models designed to process and understand both visual and textual data. These models take image and text inputs and generate text-based outputs, which en-

ables a wide range of applications. Large VLMs demonstrate strong zero-shot performance and generalize effectively across diverse image types—including documents, web pages, and photographs—and support tasks such as image-based chat, instruction-driven recognition, visual question answering, document understanding, and image captioning [12]. Some advanced VLMs, e.g., DeepSeek-VL, also incorporate spatial reasoning, allowing them to detect, segment, and localize objects within an image [13]. When prompted, they can generate bounding boxes or segmentation masks, identify specific subjects, and answer questions about spatial relationships. The capabilities of VLMs vary significantly based on their training data, image encoding strategies, and architectural design, which leads to diverse strengths across different applications.

2.2. In-Context Learning

In-Context Learning In-context learning (ICL) [8] is a prompt engineering technique in which task demonstrations are embedded within the input prompt in natural language, which allows the model to infer the desired task without explicit parameter updates. This method enables the adoption of pre-trained VLMs for novel tasks without costly fine-tuning.

In-Context Learning with few-shot demonstrations ICL with few-shot demonstrations, also known as few-shot prompting [2] [3], is a prominent approach for multi-class classification using VLM. In the context of VLM-based multi-class classification, this problem can be frame as follows: Given a query input tokenized image x and a set of candidate classes $Y = y_1, ..., y_n$, a pretrained vision language model V predicts the answer with the highest prediction score. This prediction is based on a demonstration set E which consists of an optional task instruction I and k demonstration examples. Therefore, E can be represented in two possible ways: $E = \{I, u(x_1, y_1), ..., u(x_k, y_k)\}$ or $E = \{u'(x_1, y_1, I), ..., u'(x_k, y_k, I)\}$, where $u'(x_1, y_1, I)$ represents an image example tailored to the task. The likelihood of each candidate answer y_i is determined by a scoring function f, which evaluates the entire input sequence. This setup allows the model to choose the most appropriate predictions by considering the input image, a few demonstration examples, and the task instruction.

$$P(y_i \mid x) \triangleq f_V(y_i, F, x) \tag{1}$$

The final prediction can be written as an argument of the maximum of the conditional probability as fellows:

$$\hat{y} = \arg\max_{y_i \in Y} P(y_i|x) \tag{2}$$

3. Vision Language Model (VLM) for Point Cloud-based Truck Classification

3.1. Point Cloud Data Processing

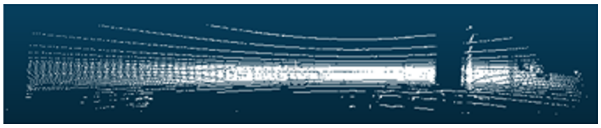
This study adopts the infrastructure-based LiDAR data processing pipeline from [10]. Vehicle point clouds were first segmented using a background subtraction method that divided the LiDAR sensor's conical surface into annular sector-shaped cells, isolating foreground vehicles based on spatial occupancy. DBSCAN clustering then grouped points into distinct vehicle objects [4]. For cross-frame tracking, the SORT algorithm was applied, representing each vehicle by the centroid of its minimum oriented 2D bounding box [1]. Inter-frame displacement was estimated using a Kalman filter, with vehicle assignments optimized via the Hungarian algorithm.

Each individual LiDAR frame is too sparse to accurately capture the configuration of vehicle objects, especially when compared to RGB image-based methods. To create a better 2D rendering of the point cloud for input into the vision-language model (VLM), which is primarily trained on RGB images, this study adopted a vehicle point registration framework to enhance the resolution of the point cloud images [11].

A probabilistic-based pairwise point cloud registration approach was applied to align vehicle objects between consecutive frames [5]. First, vehicle objects from adjacent frames were aligned by minimizing point-to-point distances.This registration was further refined through a point-to-plane strategy, which enhances the precision of vehicle point cloud alignment, particularly when a well-defined surface is established as the vehicle approaches the LiDAR sensors, where the point-to-point method may become less effective. Finally, single-frame vehicle point clouds were reconstructed using the transformation matrices derived from consecutive frames, which improves their resolution and produces a more detailed 3D representation. A visual comparison between a single frame of a tractor-trailer truck and the reconstructed results is presented in Figure 1. The reconstructed truck provides a clearer definition of the vehicle's edges in the point cloud compared to the single frame results, which offers a more precise representation that enhances the VLM model's ability to interpret and perform classification tasks effectively.

3.2. Point Cloud Image Processing

To optimize point cloud images for use with VLMs input, this study applied *statistical outlier removal* [20] and mathematical morphological operators - *opening* (*Erosion* followed by *Dilation*) [14]. These techniques help refine and smooth the contours of foreground objects, which effectively eliminate small noise both from the point cloud and its 2D projects. By mitigating this noise while preserv-



(a) Single Frame

(b) Reconstructed Frame

Figure 1. Comparison of single frame and reconstructed frame.

ing the overall structure of the vehicle, the process results in a cleaner, more continuous 2D representation of the point cloud. This improved representation enhances the suitability of the point cloud for classification task in VLM-based applications.

3.3. Few-shot Prompting

While Large Vision-Language Models show impressive zero-shot abilities for understanding more generalized content, they still struggle with more complex tasks when operating in zero-shot setting, particularly when dealing with point-cloud projected images the haven't seen similar instance during training. To address this issue, few-shot prompting along with ICL was adopted. The VLM model was guided by providing demonstrations within the prompt. These demonstrations conditioning for subsequent examples, which help the model generate more accurate and relevant responses. This study adopted the few-shot prompting strategy proposed by [2] to design the prompt for vehicle point cloud-project image classification. The few-shot prompt design is presented in Figure 2

4. Experimental Results

4.1. Data

The dataset employed to test our approach was collected from the entrance ramp to the San Onofre Truck Scale on the I-5S freeway in Southern California, a major truck corridor between Northern and Southern California (and Mexico). Data collection occurred from July 18 to August 5, 2019, which captures various truck types under both free-flow and congested conditions, with vehicle speeds ranging from 0 to 50 mph.

The site was equipped with a video camera for ground truth data and a Velodyne VLP-32c LiDAR unit for data collection. Both sensors were synchronized and connected to a solid-state field processing unit. The LiDAR sensor, mounted horizontally on a 2 m elevated platform, was aligned parallel to the ground, assuming a level roadway surface. All 12 fine-grained vehicle classes, including 11

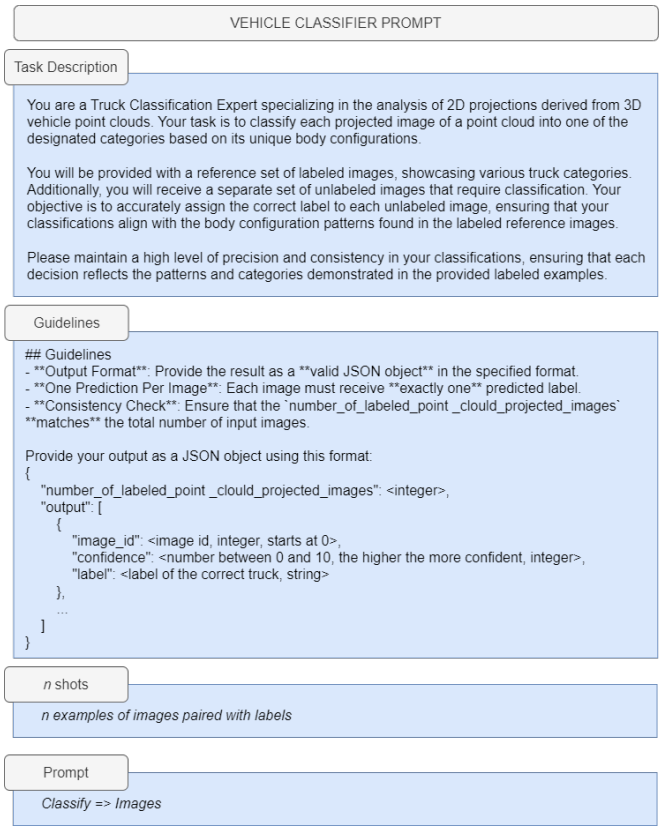


Figure 2. Few-shot Prompt Design for Vehicle Classification

truck categories and 1 passenger vehicle category, were labeled and prepared for the modeling process.

4.2. Experimental Setup

This study adopted Gemini 1.5 [15] VLM to perform the task of few-shot vehicle classification. In order to enhance the efficiency of the prediction process, the tokenized images are divided into five batches . The batching approach not only accelerates processing speed but also prevents errors associated with exceeding the payload size limit of Gemini API. The experiment starts with testing the one-shot capability of Gemini, followed by an evaluation of its few-shot performance as the number of shots is gradually increased. This study compares the model's performance using both original projected 2D point cloud images and the processed image across 1 to 9 shots, with the results presented in Figure 1. Future studes will test and compare various state-of-the art VLM models.

4.3. Results Analysis

In Table I, the classification performance, measured by the F_1 scores, the harmonic mean of *Recall* and *Precision*, is reported over the 12 different types of fine-grained vehicle classes. Between the original images and the proposed

(a) Original Image (b) Opening Image

Figure 3. Illustration of Original and Opening Images

Class name	Processed					Original				
	1 shot	3 shot	5 shot	7 shot	9 shot	1 shot	3 shot	5 shot	7 shot	9 shot
Auto Transporter	0.00	0.45	0.39	0.44	0.51	0.54	0.42	0.52	0.58	0.45
Bobtail	0.83	0.94	0.87	0.88	0.47	0.53	0.89	0.62	0.61	0.83
Platform (SU)	0.34	0.19	0.13	0.00	0.21	0.07	0.12	0.16	0.23	0.10
Tank Tank	0.69	0.81	0.74	0.95	0.83	0.45	0.73	0.74	0.74	0.65
Container	0.00	0.43	0.50	0.22	0.46	0.18	0.12	0.36	0.30	0.31
Dump Tank (Semi)	0.37	0.54	0.68	0.65	0.57	0.42	0.35	0.40	0.40	0.40
Enclosed Van (Semi)	0.29	0.46	0.33	0.48	0.37	0.24	0.26	0.16	0.39	0.23
Enclosed Van (SU)	0.67	0.80	0.69	0.81	0.69	0.64	0.52	0.60	0.81	0.76
Low Boy Platform	0.38	0.29	0.47	0.33	0.51	0.26	0.43	0.35	0.16	0.35
Passenger Vehicle	0.79	0.68	0.72	0.64	0.62	0.90	0.94	0.59	0.70	0.84
Pickup/Utility/Service	0.32	0.41	0.17	0.12	0.49	0.27	0.39	0.20	0.13	0.35
Platform (Semi)	0.15	0.36	0.44	0.29	0.17	0.24	0.35	0.28	0.18	0.34
Avg	**0.40**	**0.53**	**0.51**	**0.48**	**0.49**	**0.39**	**0.46**	**0.41**	**0.44**	**0.47**

Table 1. Performance (F1) comparison between the processed and original images across various shot settings. Note: 'SU' refers to single-unit trucks, 'Semi' denotes semi-trailer trucks, and 'Pickup/Utility/Service' includes a wide range of pickup, utility, and service trucks, both with and without trailers. 'Tank Tank' represents tank truck with tank trailer. Platform trucks encompass both empty and loaded platforms, which exhibit considerable intraclass variation.

processed images, on average our proposed image processing method has better results among all different choices of the number of few-shots, the top performance (0.53) was achieved by 3-shot, beating the no-processing method (0.46) by more the 15% (0.07/0.46). Notably, with just a 3-shot approach and image processing techniques, four vehicle classes were able to achieve an F_1 score greater than 0.60. The vehicle classes "Platform (SU)," "Low Boy Platform," "Pickup/Utility/Service," and "Platform (Semi)" exhibit relatively low F_1 scores due to their high intraclass variability. This variability arises from the diverse range of platform types, which include both empty loads and loaded platforms with various shapes of commodities, making it challenging for few-shot learning techniques to effectively capture the distinctions.

5. Conclusion

In this work, we endeavor to exploit the VLM to classify heavy-duty trucks from point-cloud images via image processing and ICL. To our best knowledge, this is the first such kind of study to transfer the representational power of VLM for LiDAR-based images directly, encouraging results have been observed using our heavy-duty truck data set. This is the preliminary effort to tap into the power of VLM for practical utilities, besides ICL, more work will be conducted to use visual deep nets such as YOLO [16] and few-shot visual learning [18] to be deployed locally as the retrieval-augmented generation (RAG) system [9], together with the Low-Rank Adaption (LoRA) [6] based fine tuning, better classification and segmentation results can be expected. Furthermore, using the Agentic workflow [17] and Chain-of-Thought prompting [19], combining with the image annotation capability and natural language understanding prowess of VLM such as content summary and speech understanding, this line of work can unleash the power of VLM and LLM for more practical use in real-world applications such as traffic safety monitoring.

References

[1] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. Ieee, 2016. 2

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2, 3

[3] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022. 2

[4] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996. 2

[5] Wei Gao and Russ Tedrake. Filterreg: Robust and efficient probabilistic point-set registration using gaussian filter and twist parameterization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11095–11104, 2019. 2

[6] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 4

[7] Hanyao Huang, Ou Zheng, Dongdong Wang, Jiayi Yin, Zijin Wang, Shengxuan Ding, Heng Yin, Chuan Xu, Renjie Yang, Qian Zheng, et al. Chatgpt for shaping the future of dentistry: the potential of multi-modal large language model. *International Journal of Oral Science*, 15(1):29, 2023. 1

[8] Yixing Jiang, Jeremy Irvin, Ji Hun Wang, Muhammad Ahmed Chaudhry, Jonathan H Chen, and Andrew Y Ng. Many-shot in-context learning in multimodal foundation models. *arXiv preprint arXiv:2405.09798*, 2024. 2

[9] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020. 4

[10] Yiqiao Li, Koti Reddy Allu, Zhe Sun, Andre YC Tok, Guoliang Feng, and Stephen G Ritchie. Truck body type classification using a deep representation learning ensemble on 3d point sets. *Transportation Research Part C: Emerging Technologies*, 133:103461, 2021. 1, 2

[11] Yiqiao Li, Andre YC Tok, Zhe Sun, Stephen G Ritchie, and Koti Reddy Allu. Lidar vehicle point cloud reconstruction framework for axle-based classification. *IEEE Sensors Journal*, 23(11):11168–11180, 2023. 1, 2

[12] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26689–26699, 2024. 2

[13] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 2

[14] Laurent Najman and Hugues Talbot. *Mathematical morphology: from theory to applications*. John Wiley & Sons, 2013. 2

[15] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 3

[16] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. Yolov9: Learning what you want to learn using programmable gradient information. In *European conference on computer vision*, pages 1–21. Springer, 2024. 4

[17] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024. 4

[18] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020. 4

[19] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 4

[20] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3d: A modern library for 3d data processing. *arXiv preprint arXiv:1801.09847*, 2018. 2