Deep Learning-Based Multi-Class Semantic Segmentation and Natural Language Scene Description of Multilane Rural Highways Using LiDAR Data

by

Honglin Jiang

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Transportation Engineering

Department of Civil and Environmental Engineering

University of Alberta

© Honglin Jiang, 2024

Abstract

The increasing adoption of light detection and ranging (LiDAR) technology offers a promising avenue for automating the identification of road features. However, due to the complexity and density of the point cloud, most research focuses on extracting single or binary road elements from the LiDAR data. Previous work that attempts to classify more than two categories often suffers from poor accuracy, whether using public datasets like SemanticKITTI and nuScenes or private datasets.

Concurrently, artificial intelligence and natural language processing have emerged as prominent research areas. This thesis demonstrates the effectiveness of leveraging machine learning and large language models to address challenges in the transportation field and propose novel frameworks for asset management, scene understanding, maintenance planning, traffic safety analysis, and intelligent transportation systems.

Recent advancements in Transformer architectures, known for their success in natural language processing, have shown impressive results in handling point cloud data. This thesis aims to identify and enhance the best Transformer-based model capable of simultaneously extracting multiple highway infrastructure elements, thus addressing the current gap in multi-object segmentation. The proposed methods would replace tedious and time-consuming manual processes with advanced deep-learning models that extract valuable features from high-density LiDAR point clouds.

The thesis presents two advanced semantic segmentation approaches that leverage transformer architectures and state-of-the-art natural language models for automating the extraction of rural multilane highway infrastructure elements and generating scene descriptions from LiDAR data. The first approach employs the Point Transformer v2 model, a Transformer-based architecture

tailored for 3D point cloud processing, to process 50-meter highway segments as input along with four additional attributes. This approach leverages the self-attention mechanisms of the Transformer architecture to capture long-range dependencies and contextual information within the point cloud data. The second approach utilizes adaptations of self-attention and cross-attention mechanisms from the Transformer architecture, specifically designed for point cloud data, operating on individual LiDAR points for point-wise classification. This approach leverages the natural language models' ability to process sequential data and applies it to the spatial domain of point clouds, enabling efficient feature extraction and classification.

Experimental results conducted on 2.5 kilometres of highway segments in Alberta, Canada, demonstrate the effectiveness of the proposed approaches. The first approach achieved an overall Mean Intersection over Union (IoU) score of 78.29% and a Mean F1 score of 86.48%, with most individual class accuracies exceeding 95%. The second method achieved an overall Mean IoU score of 86.03% and a Mean F1 score of 92.21%.

This thesis addresses the critical gap between interpreting 3D point cloud data and generating natural language descriptions. The thesis proposes a novel approach that converts semantic segmentation output into multi-view images and integrates the advanced GPT-4o model under specific restrictive conditions. This integration aims to generate accurate and contextually rich textual representations of 3D highway scenes.

This research significantly advances automated infrastructure extraction techniques, providing transportation agencies with a more efficient way to inventory rural highway infrastructure elements. These advancements have direct implications for future autonomous driving, crash environment reproduction for improved highway safety scene understanding, big data analysis, maintenance planning, and asset management, making this study highly relevant and vital.

Preface

The work presented in this thesis has either been published or is under review for publication in various peer-reviewed journals.

Under Review

- **Jiang, H.**, Elmasry, H., Lim, S., El-Basyouny, K. (2024) "Utilizing Deep Learning Models and LiDAR Data for Automated Semantic Segmentation of Infrastructure on Multilane Rural Highways" Canadian Journal of Civil Engineering

*Dedicated to my dear parents, **Jiang Dengxue** and **Zhao Xinjuan**,*

*And my respected high school teacher, **Xia Fei**,*

whose constant support, inspiration, and companionship

have profoundly enriched my life journey

Acknowledgements

I want to express my sincere gratitude to my supervisor, Dr. Karim El-Basyouny, for his invaluable guidance, continuous support, and unwavering encouragement throughout my research journey and the entirety of my Master's program at the University of Alberta. His mentorship and profound expertise have been instrumental in shaping this work.

I extend my heartfelt appreciation to Dr. Tae Kwon, Dr. Wei Liu, and Dr. Qipei Mei, esteemed dissertation committee members, for their valuable participation in my defense. Their constructive feedback significantly enhanced my research, and I am deeply grateful for their guidance and expertise.

My sincere appreciation extends to Dr. Sharon Harper for her editorial recommendations, invaluable advice, and continuous support throughout this endeavour.

I would also like to thank to Gurveer Singh Sohal and Sangwon Lim for their invaluable assistance and support during the model development phase. Additionally, I am grateful to Hesham Elmasry for his assistance and constructive feedback, which kept me consistently on track. I also want to acknowledge Maged Kamal Gouda for sharing his previous work and code related to LiDAR data. Without their guidance, navigating the complexities of my thesis would have been a formidable challenge.

Finally, I wish to express my most profound appreciation to my family, including my parents and cousin, and everyone who has stood by me during this journey. Their unwavering support, unconditional love, and steadfast belief in my abilities have been the cornerstone of my perseverance and achievements. Interestingly, I also find myself grateful to those who doubted me, as their skepticism ultimately fueled my determination to become even stronger. This combination of support and challenge has propelled me towards success.

TABLE OF CONTENTS

List of Tables

List of Figures

Chapter 1. Introduction

1.1 Background

Transportation has experienced considerable progress over the past few decades, and society now approaches a new era in this domain. The future of transportation holds the promise of unprecedented technological advancements, greater sustainability, and improved accessibility. The increasing adoption of light detection and ranging (LiDAR) technology presents a promising avenue for automating the identification of road features. Notably, the interest in utilizing LiDAR technology to map transportation infrastructure has experienced notable growth as LiDAR point clouds deliver precise data of the mapped surroundings.

The future of transportation hinges on automated data collection and big data analytics, with LiDAR technology leading the charge. LiDAR's precision in gathering 3D data of infrastructure data aligns perfectly with big data analytics, offering insight into road systems. Semantic segmentation techniques can accurately identify road features by automating data collection and employing advanced machine learning, such as convolutional neural networks. This synergy not only enhances autonomous vehicle navigation but also enables data-driven decision-making for asset management and maintenance. Integrating LiDAR with automated data analysis promises a future of smarter, sustainable, and accessible mobility solutions.

LiDAR is a remote sensing technology that measures distances by illuminating targets with laser light and analyzing the reflected beam objects (Gargoum & El-Basyouny, 2017). Unlike traditional video or stereo cameras, which depend heavily on lighting conditions, LiDAR offers millimeter-level accuracy, enabling it to scan entire scenes effectively and precisely, regardless of lighting. Additionally, LiDAR provides detailed coordinate information, a feature that vision-based data alone cannot offer. There are various types of LiDAR data, including airborne (gathered via aircraft), spaceborne (obtained via satellites), and terrestrial (sourced from ground level). Additionally, terrestrial LiDAR setups can be either stationary or mobile. In Mobile Laser Scanning (MLS), the scanning equipment is installed on vehicles that travel along specified roads, capturing comprehensive 360-degree images of the surroundings (Williams et al., 2013). MLS integrates laser scanning equipment, Global Positioning Systems (GPS), and inertial navigation technologies into one system that can acquire positional data and intensity information about surrounding objects.

Government agencies and transportation authorities increasingly leverage LiDAR technology to map and monitor transportation infrastructure. This includes roads, highways, bridges, tunnels, railways, and airports. Surveying and geospatial companies increasingly utilize LiDAR technology for its rapid and comprehensive data acquisition capabilities. Moreover, research institutions and academic organizations are conducting studies and experiments to enhance LiDAR technology's applications and effectiveness in transportation mapping. LiDAR point clouds deliver precise data of the mapped surroundings. Nonetheless, prior research has primarily focused on extracting specific road elements such as traffic signs, lane markings, or light poles, neglecting comprehensive examination of the composite road components.

While semantic segmentation and object detection through image processing, typically relying on RGB data, have become common practices, the same cannot be said for LiDAR point clouds. Unlike images, which capture surface appearance, LiDAR provides precise 3D spatial information, enabling more accurate detection and understanding of objects in the environment. LiDAR's advantage lies in its ability to directly measure distances and create detailed 3D models, regardless of lighting conditions or surface texture. LiDAR is particularly valuable in scenarios where visual data alone may be insufficient, such as in low-light conditions or environments with limited texture variation. Additionally, LiDAR data is inherently geometrically accurate, allowing for precise measurements and analysis of transportation infrastructure. Thus, integrating LiDAR technology alongside image processing techniques can significantly enhance the capabilities of transportation mapping and monitoring systems.

Semantic segmentation and object detection are two pivotal techniques in computer vision for analyzing lidar data, particularly point clouds, each offering unique advantages. Semantic segmentation provides a comprehensive understanding of a scene by classifying every pixel or point into predefined categories like road, lane, sign, or vegetation. This approach operates on a per-pixel or per-point basis, leveraging deep learning models such as convolutional neural networks (CNNs) to assign labels based on surrounding context accurately. Unlike object detection, semantic segmentation preserves spatial information, enabling precise localization and detailed scene layout and composition analysis. Segmenting the entire scene into semantic regions offers a holistic view, facilitating robust application decision-making. While object detection offers precise localization of individual objects, semantic segmentation's ability to comprehend the entire scene layout makes it particularly valuable in scenarios where understanding spatial relationships and identifying broader patterns are essential.

Furthermore, semantic segmentation is adept at handling complex scenes with overlapping objects or structures, making it a preferred choice for tasks requiring comprehensive extraction of diverse road and roadside infrastructure components from lidar data. Combining both techniques can enhance the accuracy and completeness of analyses, with semantic segmentation playing a pivotal role in providing rich contextual information for informed decision-making. In addition, recent studies have found that the Transformer structures with their unique self-attention mechanism, which is a natural fit for point clouds because point clouds are essentially sets embedded irregularly in a metric space (Zhao et al., 2021), can potentially offer improved performance for semantic segmentation of lidar data representing road scenes.

In recent years, Natural Language Processing (NLP) and computer vision have emerged as transformative fields within artificial intelligence, each pushing the boundaries of machine understanding in their respective domains. NLP focuses on enabling computers to comprehend, interpret, and generate human language, bridging the gap between human communication and computer understanding. This capability has become increasingly crucial in our digital age, where vast amounts of textual data require efficient processing and analysis. Concurrently, scene understanding has evolved to combine computer vision, linguistics, and AI techniques, allowing machines to interpret and describe visual scenes using natural language.

While significant progress has been made in both NLP and visual scene understanding, a critical gap remains in the realm of 3D point cloud data interpretation and textual description. Point clouds, data points in 3D space, are extensively used in various applications such as autonomous driving, robotics, and architectural modelling. However, transitioning from these complex 3D representations to meaningful textual descriptions poses unique challenges that current research has not fully addressed.

The importance of addressing this gap cannot be overstated. As 3D sensing technologies become more prevalent, the ability to automatically generate human-readable descriptions from point cloud data will significantly enhance human-computer interaction in various fields. This capability could revolutionize applications in autonomous vehicles, improving their ability to communicate scene understanding to passengers or remote operators. In robotics, it could enable more intuitive human-robot collaboration by allowing robots to verbalize their perception of the environment. Additionally, this technology could greatly benefit individuals with visual impairments by providing detailed verbal descriptions of 3D spaces.

This thesis proposes two innovative approaches to the semantic segmentation of rural multilane highways utilizing LiDAR-acquired point cloud data. Furthermore, it proposes an advanced methodology integrating the state-of-the-art GPT-4o model with multi-view images derived from 3D point cloud processing. This integration facilitates the generation of accurate and contextually rich textual representations of 3D scenes.

The proposed method aims to precisely segment various components within rural highway scenes by leveraging extensive point cloud datasets and advanced machine learning models. Furthermore, the proposed semantic segmentation approach leveraging LiDAR data holds the potential to advance autonomous vehicle navigation in rural environments by providing precise mapping of road components. Additionally, the automated identification and segmentation of transportation infrastructure assets could facilitate more data-driven asset management practices, allowing maintenance operations to be prioritized based on empirical insights.

However, the direct implementation of such applications extends beyond the present research scope, which is dedicated to developing an accurate semantic segmentation methodology for rural multilane highways using LiDAR point cloud data.

1.2 Research Motivation

The motivation for this research stems from the pressing need to address the limitations of current methods for managing highway infrastructure elements, which are time-consuming and labour-intensive. As the transportation sector rapidly evolves towards greater automation and data-driven decision-making, there is an urgent demand for efficient and accurate methods to collect and analyze infrastructure data. LiDAR technology, with its ability to provide precise 3D spatial information, offers a valuable tool for detailed infrastructure mapping. However, existing research often focuses on single or binary-object segmentation, leaving a significant gap in the comprehensive extraction of multiple infrastructure elements.

Moreover, there is a critical gap in converting segmented point cloud data directly into meaningful textual descriptions, which could substantially improve highway safety scene understanding, enhance autonomous vehicle safety, streamline maintenance planning, and optimize other decision-making processes.

This research aims to fill these gaps by leveraging advanced deep learning models, particularly Transformer-based architectures and GPT-4o, to significantly enhance the capabilities of automated infrastructure extraction and point cloud-to-text conversion.

The proposed methods aim to streamline data collection processes, reduce operational costs, and improve the accuracy of infrastructure inventories while also providing more accessible and interpretable data outputs. By bridging the gap between raw sensor data and actionable insights, this research has the potential to revolutionize how transportation infrastructure is managed and monitored. Furthermore, the outcomes of this study could have far-reaching implications for the development of autonomous driving systems, potentially enhancing their ability to interpret and navigate complex highway environments.

Ultimately, this research is motivated by the potential to make substantial contributions to the field of transportation engineering, supporting the development of more efficient, safe, and data-driven transportation systems. This work aims to pave the way for smarter, more sustainable mobility solutions by addressing current technological limitations and anticipating future needs in transportation infrastructure management.

1.3 Research Objectives

The primary objective of this thesis is to classify the types of highway infrastructure and facilities located in the province of Alberta and then link the results with natural language processing to describe the scene in a text format. The ultimate goal of this work is to understand and address the challenge of classifying multiple highway infrastructure elements and using natural language processing to identify highway safety concerns in human-readable text, thereby building a framework for using this technology in the transportation field.

In summary, the objectives of the project can be outlined as follows:

1. Conduct a comprehensive review of existing automated extraction methods and natural language models applicable to highway infrastructure analysis
2. Develop and implement automated multi-object semantic segmentation methods for extracting key highway infrastructure elements (e.g., traffic signs, concrete barriers, guardrails, markings).

3. Design and apply a natural language processing model to generate human-readable descriptions of highway environments with important facilities.
4. Evaluate the performance of both semantic segmentation model and natural language processing models using suitable metrics and identify the limitations and future research directions.

1.4 Thesis Structure

The remainder of this thesis is divided into six chapters. Details of the topics covered in each chapter are described below:

Chapter 1 presents an introduction to the thesis, including background information related to the research, the research motivation, and the research objectives. It discusses recent developments in the transportation sector regarding the use of LiDAR data and the difference between semantic segmentation and object detection techniques in the 3D environment. Additionally, based on the background and the research motivation, the chapter summarizes the research objective of this work.

Chapter 2 provides a detailed summary of the literature related to this research project, including an overview of current semantic segmentation techniques, the application of LiDAR in transportation engineering, automated annotation and data collection tools, perviou innovative approaches for infrastructure extraction, advancements in LiDAR technology and data processing, and existing natural language processing techniques for point cloud data.

Chapter 3 outlines the two methodologies adopted for semantic segmentation tasks in this research and any preliminary procedures. This chapter includes a detailed description of both methods, covering the model architecture, data sources, evaluation metrics, and the improvements made.

Chapter 4 specifies the architecture of the pre-trained natural language model and explains the process of converting point cloud data into text descriptions. This chapter elucidates the intricate steps in bridging the gap between 3D spatial data and human-readable textual representations.

Chapter 5 presents a comprehensive analysis of the research outcomes, focusing on three key areas: the semantic segmentation results of both proposed models, a comparative analysis of these models, and the results obtained from the natural language model.

Chapter 6 serves as a comprehensive synthesis of the research, aligning the initial objectives with the findings obtained throughout the study. This chapter critically examines the results and their implications while exploring potential real-world applications of the developed methodologies. Additionally, it addresses the key limitations encountered during the research process, providing a balanced and transparent assessment of the study's scope and constraints.

Chapter 2. Literature Review

Numerous studies in the literature have focused on extracting specific classes of highway infrastructure from LiDAR data due to the significant volume of data points involved. These studies often target single or binary classes such as traffic signs, guardrails, light poles, or lane markings. For instance, research has been conducted on the extraction of traffic signs (S. Gargoum et al., 2017a; Javanmardi et al., 2019; S. Zhang et al., 2019), the detection of guardrails (Gao et al., 2020), urban road facilities such as light poles and bus stations (Yu et al., 2015), and lane markings (Guan et al., 2014; Kumar et al., 2014; Rastiveis et al., 2020; B. Yang et al., 2012). However, few studies have attempted to address the extraction of multiple infrastructure classes simultaneously. This paper seeks to address this void by concentrating on the comprehensive extraction of diverse road and roadside infrastructure components, encompassing lanes, shoulders, markings, traffic signs, light poles, guardrails, concrete barriers, and vegetation.

Computer vision and machine learning have seen significant advancements in recent years, leading to the development of various intelligent systems for object recognition, scene understanding, and semantic segmentation. However, LiDAR-based infrastructure solutions designed for traffic monitoring systems are currently in their early stages of development (A. Wu et al., 2023). LiDAR technology enables data acquisition across diverse environmental scenarios, encompassing low solar elevation, overcast atmospheric conditions, and even nocturnal settings. This leads to an augmented timeframe available for data collection (Veneziano et al., 2002). Electronic Distance Measurement (EDM) and Global Positioning Systems (GPS) are impractical for data collection in vast areas as they are time-consuming and require a lot of equipment movement.

On the other hand, LiDAR demonstrates potential in expediting the acquisition of terrain information compared to current methodologies for data collection. LiDAR represents an integrated technological solution that utilizes laser scanning devices, global navigation satellite systems (GNSS), and inertial measurement units (IMU) within a unified scanning framework. This integration facilitates the acquisition of point cloud data pertaining to the surrounding roadway environment (Gouda et al., 2022). This literature review demonstrates the familiarity of research with LiDAR data processing in different engineering fields and enhancing semantic segmentation for road features.

2.1 Semantic segmentation techniques

Approaches for segmenting 3D point clouds generally fall into two broad categories: voxel-based and point-based networks, each with distinct methodologies and advantages. Considerable prior research has focused on utilizing 3D Convolutional Neural Networks (CNN) to transform point clouds into 2D or 3D volumetric grids, voxels, or analogous representations (Maturana & Scherer, 2015; C. R. Qi et al., 2016; Z. Wu et al., 2015; F. Zhang et al., 2020). Point-based networks, pioneered by PointNet and PointNet++, utilize shared multi-layer perceptions (MLP) to aggregate features from various points in the point cloud data (C. R. Qi et al., 2017). Additionally, specialized networks have been developed for various applications, designed to take raw points as input for sampling directly (J. Yang et al., 2019; Yifan et al., 2019), semantic segmentation (Jiang et al., 2019; Rethage et al., 2018; Wang et al., 2019; Ye et al., 2018; Zhao et al., 2018, 2019), and instance segmentation (Wang et al., 2019; Zhao et al., 2019).

Extracting valuable spatial information from 3D Mobile laser scanning data poses a significant challenge due to the diverse array of 3D objects and the vast volume of data points involved (Lari et al., 2012). As a result, this process is often characterized by its complexity and time-consuming nature. Segmentation emerges as a fundamental step in extracting information from 3D laser point clouds.

In the study by Shen et al., 2020, a new interactive 3D object annotation method was introduced, combining point cloud data and RGB imagery. The methodology involved two stages: initial shape approximation using existing techniques, followed by correction through a user-friendly 2D interface. This interface allowed for easy rectification of 3D errors using scribbles in preferred 2D views, with final adjustments made directly to the object mesh vertices via drag-and-drop. Corrections were integrated into a 3D Graph Convolutional Network, preserving object structure while enhancing local geometry. This interactive ML-driven approach ensures generalizability across object categories and delivers high-quality geometric results.

Wu et al. (2023) found two main hurdles in developing a cost-effective traffic monitoring system with LiDAR. The first was the lack of sufficient LiDAR datasets for infrastructure perception tasks. The second was the time and cost of creating 3D annotations for LiDAR point clouds. They introduced a semi-automated annotation tool to tackle these issues, which used tracking algorithms to annotate LiDAR sequences efficiently. This tool seamlessly integrated multiple-object tracking, single-object tracking, and trajectory post-processing techniques.

2.2 LiDAR applications in transportation engineering

Gargoum & El-Basyouny (2017) highlighted the significant potential of using LiDAR in transportation engineering. However, the body of research in this area remains limited. This could be attributed to researchers not fully recognizing the extensive capabilities of such data or harbouring concerns about the feasibility of processing such voluminous datasets. Previous studies indicate that the utilization of LiDAR in transportation encompasses a broad spectrum of applications, including but not limited to on-road aspects like retrieving road surface details, lane markings, and road boundaries. Additionally, it extends to roadside elements such as identifying traffic signs, light poles, and vegetation and evaluating geometric characteristics pertaining to roadways.

In transportation engineering and planning, gathering highway inventory data is widely acknowledged as crucial for multiple purposes. Comprehensive inventory data are essential for efficient resource allocation in infrastructure management. Additionally, they aid in safety assessment by pinpointing hazardous conditions and prioritizing safety enhancements. Moreover, highway inventory data plays a vital role in supporting transportation planning and design, asset management systems, compliance, reporting requirements, and research in the transportation sector. For instance, the Highway Safety Manual (HSM), published by the American Association of State Highway and Transportation Officials (AASHTO), is crucial for analyzing and predicting roadway safety. However, its accuracy depends on precise input variables, often requiring costly data collection (Jalayer et al., 2014). Various methods are used, but none are comprehensive. Therefore, developing efficient and cost-effective automated data collection methods is essential for enhancing HSM's predictive capabilities and improving roadway safety.

Research shows that the application of LiDAR extends far beyond asset management. For instance, a report by the National Highway Cooperate Research Program (NCHRP) in the United States highlighted several different applications of LiDAR in Transportation (Ai & Tsai, 2016). Moreover, it was found that current and emerging applications for Mobile LiDAR Scanning (MLS) in transportation cover a wide range of topics (S. Gargoum & El-Basyouny, 2018).

Veneziano et al. (2002) introduced a method for enhancing highway location and design processes by integrating of LiDAR and photogrammetric mapping. This approach aimed to expedite these activities while also providing insights into potential reductions in time and costs. The anticipated outcome was notable time efficiency, attributable to the necessity of conducting photogrammetric mapping on a limited scale, as opposed to the extensive area coverage typically demanded during pre-final alignment selection stages.

Gargoum et al. (2018a) proposed a new method for extracting cross-sectional features of roadways that were scanned using LiDAR technology. This algorithm was characterized by its complete automation, enabling efficient evaluation of cross slopes and side slopes across a highway network. Notably, unlike prior research, this algorithm demonstrated the capability to perform such extraction without the necessity of lane marking data, resulting in enhanced ease and efficiency of the extraction process.

Gargoum et al. (2017b) proposed an algorithm that was introduced to facilitate the automated extraction of road signs. To evaluate its performance, the algorithm underwent testing on three distinct highways located within the Alberta province. These specific segments were deliberately selected to encompass a diverse range of geometric characteristics. The resultant algorithms yield a comprehensive compilation of traffic sign data along a designated road stretch. This dataset includes precise positioning of signs, intensity measurements, and elevation attributes for individual points within sign clusters. The gathered insights empower users with the capability to effectively map signs along a given highway, thereby serving the practical purpose of sign maintenance and renewal initiatives. Furthermore, these extracted insights hold considerable value within Intelligent Transportation Systems (ITS), as they can be seamlessly integrated into applications related to connected and autonomous vehicles.

Gargoum et al. (2018c) employed a systematic approach to produce as-built drawings depicting vertical profiles along highways by leveraging LiDAR point cloud information. The outcomes showed the viability of precisely deriving road profiles from LiDAR data. Discrepancies in grade estimations between the suggested technique and GPS data exhibited an average range of 0.023% to 0.061%. Moreover, the proposed methodology exhibited its capacity to capture intricate features within the road profile that remained undetected by GPS data, thereby illustrating the advantageous application of LiDAR for road profile extraction.

Gargoum & El-Basyouny (2019) studied how point density affects the extraction of traffic signs from LiDAR data. They aimed to guide transportation agencies on balancing cost, precision, and coverage when using LiDAR for traffic sign management. Using data from four highway sections in Alberta, Canada, they reduced the LiDAR point cloud density to set proportions and assessed sign detection at these densities. The study found that large panel signs and those on the scanned approach were detectable with as low as 35% of the original point cloud (105 points per square meter). However, detection rates for smaller signs and those on the opposite side of a divided highway decreased when point density dropped below 70% of the original.

2.3 Automated annotation and data collection tools

Li et al. (2020) introduced SUSTech POINTS, an open-source tool for annotating portable 3D point clouds, offering superior visualization and operational capabilities compared to existing tools. It included features like stream playback, object locking, and intelligent box initialization for rapid and precise annotation. SUSTech POINTS utilizes a registration-based inter-frame annotation transfer mechanism, achieving high accuracy and speed. Being web-based enhances its portability and scalability for large-scale annotation projects.

Oliveira & Rocha (2013) classified semantic annotation tools by factors like dynamicity, storage, and scalability. Their survey found manual systems costly and error-prone, while fully automatic ones lacked accuracy. As a result, many researchers now prioritize semi-automated systems, combining human input for better accuracy.

Manikandan & Ganesan (2019) developed an automated video annotation tool for training autonomous driving systems. It employed object detection and classification algorithms to identify vehicles, two-wheelers, and pedestrians, determining their properties like direction and occlusion. Additionally, it recognizes vehicle lanes and pedestrian positions while tracking and detecting object movement. Various algorithms were analyzed for each task, with the best chosen for integration into the tool. The resulting GPU-accelerated system operated 1200 times faster than manual annotation methods.

Gargoum et al. (2018b) developed an algorithm to automatically evaluate sight distance along highway segments using LiDAR point cloud data. The algorithm's application involved the assessment of sight distance along two distinct highway segments in Alberta, Canada. The data obtained from this process was subsequently compared with the theoretical values of Stopping Sight Distance (SSD) and Passing Sight Distance (PSD). The results indicated the algorithm's remarkable efficacy in the automated appraisal of sight distance for highway segments.

2.4 Innovative approaches for infrastructure extraction

Y. Chen et al. (2018) introduced ROAD-Net, a model for segmenting urban environments using synthetic data. It incorporated two main modules: one for adapting style from real images and another for addressing domain distribution differences by integrating layout information. These techniques are vital for adapting the model to various highway environments.

Gouda et al. (2022) enhanced the PointNet++ neural network architecture to improve its efficacy in processing outdoor scenes. They trained several model iterations, incorporating different combinations of geometric attributes and the proposed modifications, using annotated data from seven highway sections in Alberta, Canada. The results showed that the proposed model variations outperformed previous studies' accuracy and computational efficiency for identifying signs using point cloud data. This research demonstrated that when appropriately adjusted, the tailored PointNet++ neural network effectively isolated traffic signs from LiDAR-derived point clouds. Additionally, including local geometric attributes positively impacted the neural network's precision, recall, and F1-score metrics.

Ohgushi et al. (2021) devised a novel road obstacle detection method. They utilized an unsupervised autoencoder with semantic segmentation trained solely on typical road data. This approach did not require prior knowledge of obstacle features and operated with standard in-vehicle camera images. The autoencoder consists of a semantic encoder and a photographic decoder to generate a reconstructed image.

Gouda et al. (2022) proposed an automated methodology to delineate the positions of light poles. This technique evaluated the appropriateness of light pole placement in the clear zones alongside rural highways. The correct positioning of such poles holds significance due to its implications for overall safety, given its association with an escalated collision risk. The proposed algorithm demonstrated promising results across the evaluated segments, yielding favourable performance metrics, including average precision, recall, and F1 scores. However, it is noteworthy to acknowledge a limitation in the study, wherein identifying poles concealed by obstructions during the scanning process might not have been entirely achievable.

2.5 Advancements in LiDAR technology and data processing

Yao et al. (2007) introduced a groundbreaking large-scale ground truth image database for general-purpose use. They developed a unique annotation tool with tailored functional modules capable of executing customized annotation tasks and integrating diverse information forms. Accompanied by a comprehensive database, this tool organized labelled visual knowledge universally. It notably improved object boundary delineation through fine segmentation, which is crucial for small-sized objects. WordNet served as a reference to ensure naming accuracy. Other databases like LabelMe and CalTech 101/256 offer specific features but lack detailed segmentation or real-world context. The Berkeley Segmentation Dataset set a benchmark for ground truth annotation but had scale and content limitations. The UA (Arizona) localized semantics dataset improved semantic annotations, building on Berkeley's information.

Lato et al. (2012) utilized LiDAR technology to assess block size and shape distributions, slope geometry, and ditch profile. This approach aimed to identify potential rockfall hazards along highways and railroads in challenging terrains. The collected data was subjected to a structured hazard rating system to assign a hazard level to the specific location. This technique was proposed to reduce the personal involvement of the evaluating engineer in hazard exposure.

Michele et al. (2021) introduced a novel method for Zero-Shot Learning (ZSL) and Generalized Zero-Shot Learning (GZSL) in 3D data, handling classification and semantic segmentation tasks effectively. ZSL predicted classes unseen in training, while GZSL expanded this to known and unknown classes. This method showed promise for pre-annotating 3D data with applications like autonomous driving. They tested their approach on outdoor 3D GZSL semantic segmentation using the SemanticKITTI dataset, showcasing its performance.

In their recent work, R. Chen et al. (2023) introduced an innovative framework to understand geometric primitives across various object categories. They assimilated shared 3D geometric attributes among known and unknown categories and then used them for new objects. However, the model showed bias towards familiar categories, leading to misclassifications, especially with novel representations resembling recognized classes. To fix this, researchers separated visual representations of unfamiliar categories from semantic representations of familiar ones. The main focus was on addressing misclassification challenges. They evaluated the method using the SemanticKITTI dataset, similar to Michele et al. (2021).

Gao et al. (2021) introduced a fine-grained off-road semantic segmentation and mapping method that utilizes contrastive learning. The paper focused on improving the semantic segmentation accuracy of road scenes and mapping fine-grained categories. By leveraging contrastive learning, the authors effectively learn informative feature representations, which can handle fine-grained variations in off-road scenes and directly align with this paper’s objective of enhancing semantic segmentation for road features.

Zhou et al. (2021) presented an innovative approach for extracting highway alignments and reconstructing 3D models. This was achieved by optimizing an energy function, enabling the precise reconstruction of highway 3D models while adhering to alignment constraints. The methodology exhibited proficiency in delineating the road surface and accurately identifying its boundaries and lane markings, particularly in rugged terrains such as mountainous regions.

Kim et al. (2019) introduced the Highway Driving dataset as a new benchmark for semantic video segmentation. This benchmark serves as a valuable resource for evaluating semantic video segmentation algorithms. The dataset offers meticulously annotated pixel-level information for video frames captured at a rate of 30Hz, making it conducive for training deep neural networks aimed at processing video inputs. Prior to the advent of the Highway Driving dataset, the task of investigating semantic video segmentation encountered challenges due to the absence of annotations densely distributed across temporal frames.

Golparvar-Fard et al. (2015) introduced an algorithm to semantically segment and recognize highway assets by analyzing video frames from a vehicle-mounted camera. The experiment demonstrated strong performance, with segmentation accuracy averaging 76.50% and pixel-level recognition accuracy at 86.75%. These results highlighted the algorithm's effectiveness in identifying 12 categories of assets, showcasing its potential for segmenting and recognizing highway assets from image-derived 3D point clouds.

2.6 Natural language processing techniques for 3D point clouds

Z. Qi et al. (2024) introduced a novel unified framework for point-language understanding and generation, known as GPT4Point. This groundbreaking model introduces the 3D object MLLM, a language model that fully utilizes point clouds to perform various point-text tasks. A key component of GPT4Point is the Bert-based Point-QFormer, which aligns point-text features. These aligned features are then separately input into Large Language Models (LLMs) for text inference tasks and into Diffusion models for 3D object generation tasks. Notably, GPT4Point can generate higher-quality results from low-quality point cloud features while maintaining geometric shapes and colours through controllable text-to-3D generation using point-text aligned features.

Xu et al. (2023) contributed significantly to 3D point cloud understanding and natural language processing by introducing PointLLM. This innovation model tackles three key issues: the lack of training data, the need for an appropriate model architecture, and the absence of comprehensive benchmarks. PointLLM combines a pre-trained point cloud encoder with a powerful language model, employing a two-stage training strategy. First, it aligns the latent spaces between the encoder and the language model, then instruction-tunes the unified model. This approach effectively fuses geometric and appearance information from point clouds with the linguistic capabilities of the language model. The resulting system can process coloured object point clouds alongside human instructions, generating accurate responses demonstrating both 3D data understanding and common-sense knowledge. The authors also provide empirical studies on various design choices, contributing valuable insights to the field of multimodal AI and opening new possibilities for applications in robotics and interactive 3D environments.

2.7 Summary

In conclusion, the literature review demonstrates the growing importance and potential of using LiDAR and semantic segmentation techniques for automated extraction and analysis of road infrastructure features. The implications of this capability extend to multiple crucial aspects of the transportation field and other engineering fields, including highway safety assessment, asset management, and the advancement of autonomous vehicle technology.

In addition, the ability to turn machine language into human-readable text is paramount. Recent advancements like GPT4Point and PointLLM bridge the gap between complex 3D point cloud data and human understanding. This capability allows for improved interpretation of data by non-technical stakeholders, enhanced decision-making by providing clear, actionable information, better communication of complex spatial information, and easier integration of LiDAR-derived insights into existing text-based systems.

The gap in advanced semantic segmentation techniques is particularly evident in their multiclass accuracy, where current methods often fail to deliver precise and reliable results across multiple categories. Additionally, the ability to convert 3D point clouds into human-readable text remains underdeveloped. While significant progress has been made in processing and understanding 3D data at the object level, translating this information into coherent and contextually appropriate natural language descriptions continues to lag.

Chapter 3. Semantic Segmentation Methodology

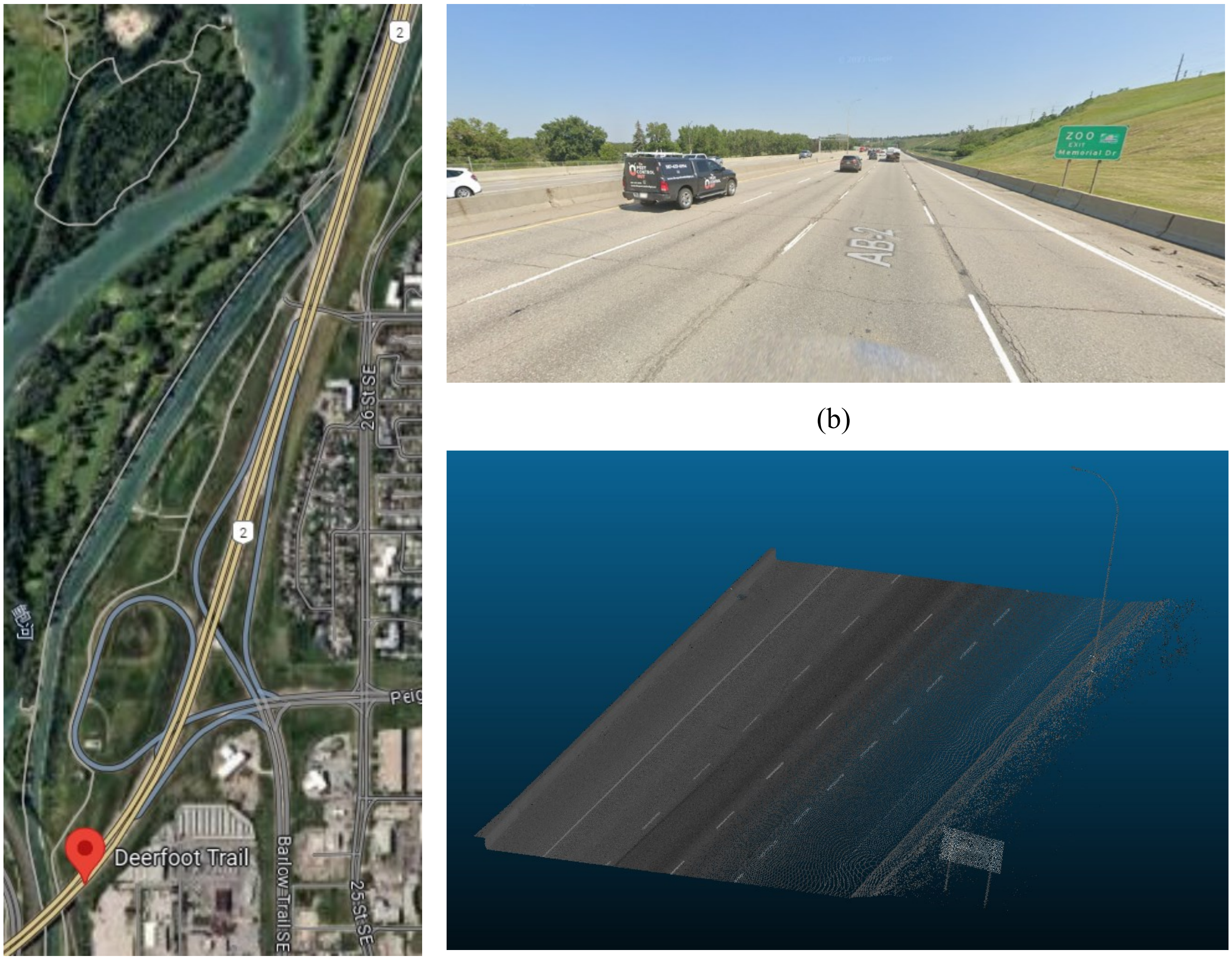
3.1 Data collection & description

Alberta Transportation collected LiDAR point cloud data along rural highways in Alberta, Canada, from 2013 to 2020. This data acquisition utilized a versatile surface profiling vehicle, the Tetra Tech PSP-7000, outfitted with a REIGL VMX-450 system. The vehicle's REIGL VMX-450 system facilitated the capture of 360° LiDAR point clouds along the province's rural highways, with surveys conducted during regular traffic flow at speeds of up to 100 km/h. The resulting data was segmented into individual LAS files, each covering a 2 km section, with point densities ranging from 150 to 1000 points/m^2 at speeds of 90 km/h. These files were approximately 500 MB each in size. Each scanned point within the data set possessed location attributes (X, Y, Z) and point intensity information.

3.2 Dataset development

For the development of the datasets, segments of 2.5 km in both directions (Northbound and Southbound) of Highway 2 (AB-2) were used for training, validation, and testing. The training dataset segments were strategically chosen along the highways to encompass various features while avoiding poor-quality LiDAR data. For instance, segments with elements like signage and guardrails were favoured, given that some highway sections lacked these features. Segments with excessive vehicle occlusion or shallow markings were considered poor-quality LiDAR data and were avoided. Conversely, the test and inference segments were randomly selected to evaluate the model's performance, ensuring no overlap with the training segments. This approach treated the test and inference datasets as entirely new and unseen by the model, providing a robust assessment of its generalization capabilities.

Highway 2 (AB-2), spanning a length of 1273 km, is the longest and busiest highway in the province, featuring a diverse range of characteristics. This paper considered sections with three or more lanes in each direction. The section is separated by a narrow median with either New Jersey barriers or guardrails, as shown in Figure 1. Additionally, segments feature auxiliary lanes for entrances and exits.



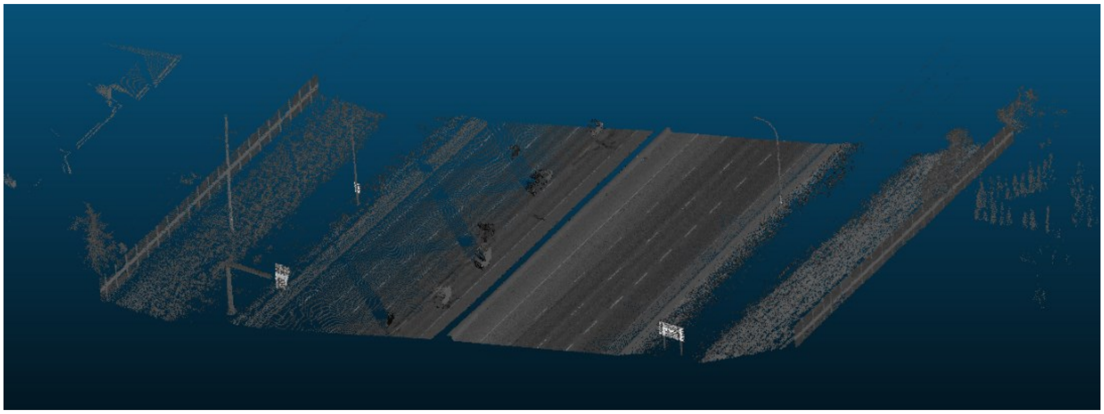
(b)

(a) (c)

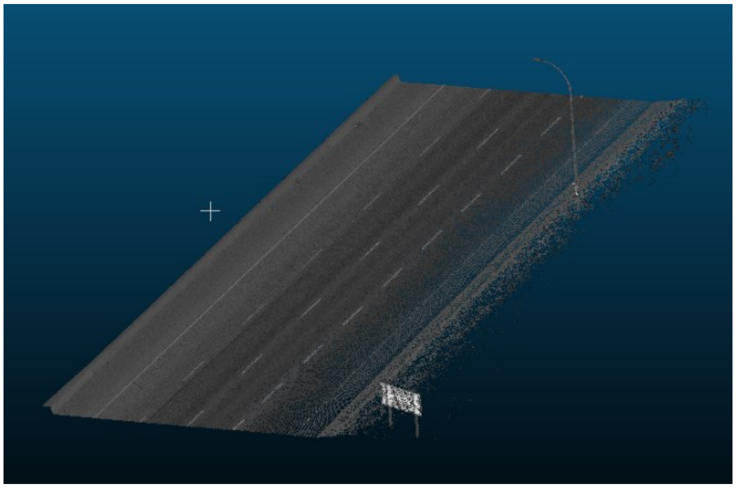
Figure 1. Highway AB-2 (Deerfoot Trail) situated in Alberta, Canada. (a) a satellite image, (b) a street view, and (c) a point cloud sample representing a 50-meter segment

3.3 Data preparation and filtering

The initial data analysis phase involved preparatory procedures to refine the collected dataset. This involved employing CloudCompare, a widely used open-source 3D point cloud editing and processing software primarily for visualizing and filtering point clouds (CloudCompare, 2023). The dataset was segmented into 50-meter-long segments to facilitate subsequent annotation tasks and leverage the computational capabilities of the Graphics Processing Unit (GPU) utilized for model training. Additionally, it is noteworthy that the intensity of points significantly diminished with increasing distance from the LiDAR scanner, necessitating separate annotations for each highway direction. Consequently, points related to the opposite direction were excluded prior to annotation, with annotations for the latter being derived from files capturing scans in the reverse direction. Points corresponding to the median were retained in files for both directions. In addition, data cleaning was performed using the scan angle rank, considering a 35-meter distance from the trajectory on both sides, as shown in Figure 2.



(a)



(b)

Figure 2. A sample of 50-meter segment raw data (a) before and (b) after data cleaning

Additionally, class-specific filters were applied to delineate distinct components, emphasizing splitting lane markings with pavement. Visual inspection of connected components was systematically conducted to validate their classification. Manual adjustments and revisions were diligently undertaken to ensure accuracy and reliability in our analysis, as depicted in Figure 3.

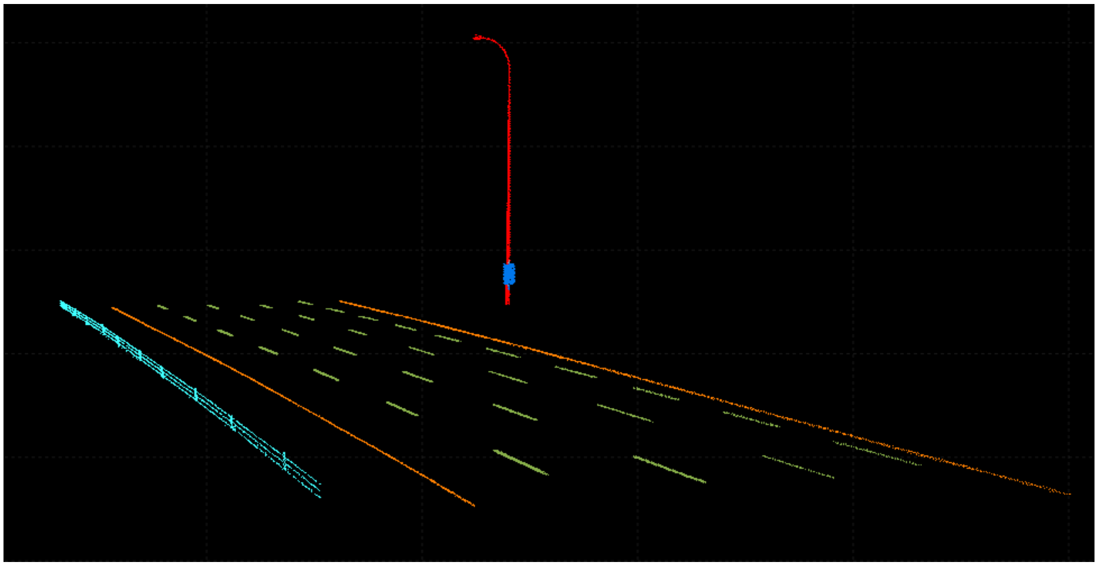


Figure 3. Points filtration using CloudCompare (labelled in different colours)

The following classes were identified as principal components for rural highways, to be annotated in subsequent stages: lane, shoulder, chevron marking, broken line marking, solid line marking, arrow marking, vegetation, traffic sign, guardrail, concrete barrier, and light pole. The primary challenge was distinguishing between points representing road markings and those delineating the road pavement (i.e., lanes and shoulders). We used an intensity filter in CloudCompare to aid in differentiation. Points not falling within the purview of these defined classes, such as moving vehicles, signposts, and electrical lines, were retained without removal; however, they were excluded from the subsequent annotation exercise.

3.4 Data annotation

Data annotation or labelling is crucial for semantic segmentation as it provides the ground truth or reference for training and evaluating machine learning models. In semantic segmentation, the goal is to classify each point in the point cloud into predefined categories. Semantic segmentation is typically performed using supervised learning algorithms, where the model learns from labelled examples. The annotations serve as the target outputs for the model during training, enabling it to learn the relationships between input points and their corresponding semantic classes. Moreover, annotated data allows for the evaluation of model performance. By comparing the model's predictions with the ground truth annotations, metrics such as accuracy, mean IoU (Intersection over Union), or F1 score can be computed to assess the model's accuracy and effectiveness.

SUSTech POINTS was selected as the basis for the annotation tool due to its extensive capabilities, as outlined by Li et al. (2020). Notably, SUSTech POINTS offers a range of features, including the development of visualization modules tailored for fast annotation error localization and seamless annotator-data interactions, the creation of interactive tools enabling annotators to label 3D point clouds and 2D images rapidly, and the implementation of an innovative annotation transfer method facilitating the labelling of identical objects across various data frames. These functionalities significantly contribute to the efficiency and accuracy of the annotation process. Moreover, enhancements were made to the platform to augment its utility further. These improvements included directly utilizing LAS files, implementing a user-friendly interface for dynamic label creation, integrating Google Maps for contextual reference, and optimizations for handling large-scale datasets, as shown in Figure 4.

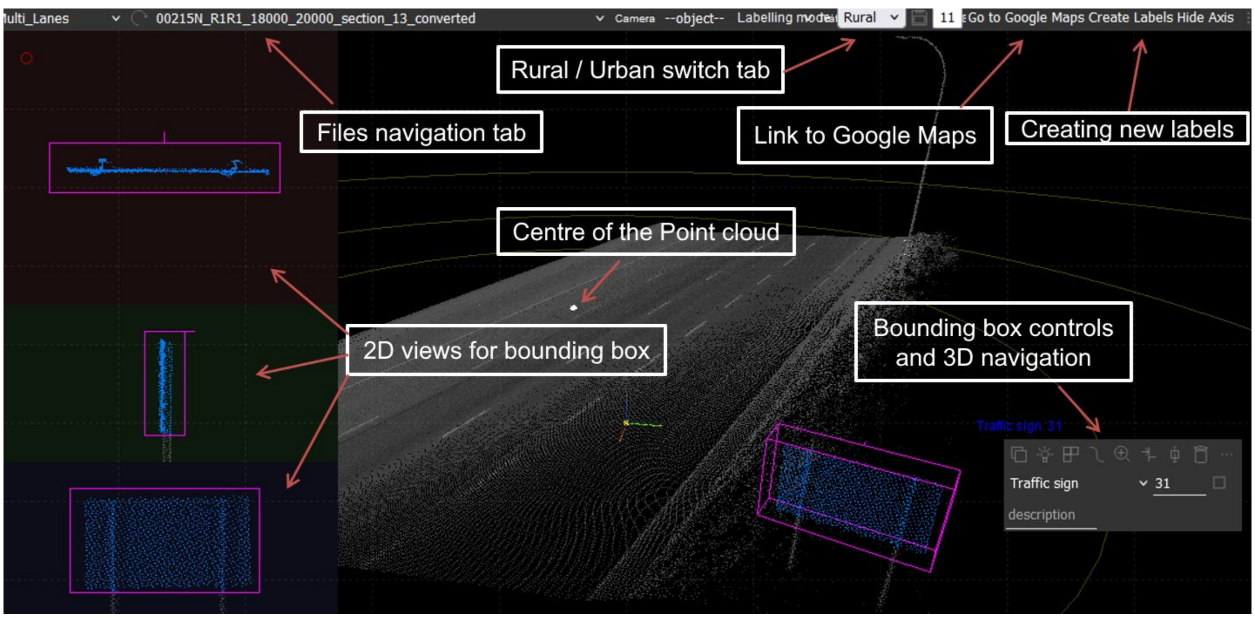


Figure 4. Modified SUSTech POINTS annotation tool

As the figure above shows, each object is delineated using a rectangular bounding box. All points within this box are assigned to the same class, underscoring the critical importance of the filtering and fine-tuning stages. The rectangular nature of the bounding box is fixed, meaning its shape cannot be altered to conform precisely to irregular object boundaries. Consequently, the accuracy of our classification heavily relies on optimizing these preparatory steps to ensure the most representative assignment of points within each bounding box.

The initial tool introduced by Li et al. (2020) exclusively accommodated point clouds formatted in PCD. However, the updated version facilitates automatic conversion from LAS to PCD. Furthermore, a Google Maps tab has been incorporated to provide visualization of annotated segments in street view. Additionally, tabs have been included to toggle between rural and urban modes. Each tab has unique labelling functionality, enabling annotators to supplement additional labels. Moreover, seamless navigation between LAS files depicting different road segments has been implemented. To enhance navigation through each scene (road segment) of the point cloud, the tool automatically calculated the center point of each LAS file, facilitating navigation through the 3D scene, such as zooming and rotating.

By integrating these enhancements, the modified version of SUSTech POINTS provided a comprehensive solution for annotating 3D point clouds, thereby enhancing annotation accuracy and efficiency for diverse applications, as evidenced by experimental evaluations conducted on both public and private datasets (Li et al., 2020).

Furthermore, the annotation tool employed in this study has been tailored to suit our specific requirements by customizating of Python and Docker. In the Python adaptation, the setup procedure entails establishing a virtual environment and installing the essential libraries. This encompasses obtaining the requisite model file and TensorFlow, after which the primary Python script is executed. Conversely, Docker offers an optimized deployment pathway by encapsulating the tool and its adaptations within a container, thereby guaranteeing uniform functionality across diverse computing environments.

The code structure of the annotation tool comprises both backend and frontend components. In the backend, routes are defined within the main Python file (`main.py`) using the CherryPy web framework. These routes facilitate communication between the frontend interface and backend functionalities. Notably, scene-specific information retrieval routes allow users to access scene descriptions and annotation data.

Conversely, the frontend logic resides primarily within JavaScript files in the `/public/js` directory. Key files such as `editor.js`, `view.js`, and `annotation.js` are crucial for user interaction and data visualization. Modifications to these files enable customization and enhancement of the annotation tool's functionalities.

The data structure required by the annotation tool necessitates adherence to a specific format within the `data` directory. This hierarchical structure dictates the organization of calibration, camera, label, lidar, and LAS data folders. Compliance with this structure ensures seamless data access and processing during annotation tasks.

It is worth mentioning that in the domain of computer vision, particularly semantic segmentation or object detection tasks, annotating data to establish ground truth labels is crucial yet prone to imperfections. Despite rigorous data preparation and filtering efforts, achieving 100% accuracy in manual labelling remains challenging due to human error, ambiguity in complex scenes, and the sheer volume of pixel-level annotations required (Cordts et al., 2016). Researchers often incorporate a tolerance threshold in their evaluation metrics to address this inherent uncertainty, typically allowing for up to a 5% discrepancy in label accuracy (Northcutt et al., 2021). This acknowledges the intrinsic subjectivity in pixel-level labelling tasks, particularly in boundary regions between different semantic classes or occlusion cases (Chen et al., 2018).

3.5 Data pre-processing

The initial dataset consisted of point clouds characterized by four attributes: X, Y, Z, and intensity. Labels were assigned to the points outlined in the previous section, utilizing the SUSTech POINTS annotation tool. Points within a bounding box were given the same label as the bounding box itself. Consequently, each point within a bounding box was annotated with a label corresponding to that box's label.

Each annotated road segment underwent meticulous label extraction, discerning various predefined categories: lanes, shoulders, chevron markings, broken and solid line markings, arrow markings, vegetation, traffic signs, guardrails, concrete barriers, and light poles. Unlabelled points, or clutter, were also identified, processed, and segregated into distinct files. The script iterated through each section of the point cloud data, executing analogous procedures to capture the nuanced characteristics of each section. As a result, the script generated a comprehensive set of annotated files that encapsulated the spatial distribution and semantic information of labelled points across the dataset.

The data parsing step involved processing files from the 'las_files', 'labels', and 'centres' directories within the data folder. These files were organized into .txt files according to their corresponding labels. For instance, a section containing three boxes—two traffic signs and one light pole—was organized into individual .txt files (e.g., "traffic-sign_1.txt," "traffic-sign_2.txt," and "light-pole_1.txt") within the parsed folder. This code step was inspired by the nuScenes development kit by Caesar et al. (2020). The modified code defined a function (‘points_in_box’) that checked whether a set of points lie inside a specified 3D box. A reference corner of the bounding box (p1) was chosen for each point within a bounding box, and the vector (v) from this reference point to the target point was computed. This vector was then projected onto the three principal axes of the bounding box. The lengths of these projections were compared to determine the spatial relationship of the point within the bounding box. This method was used to ensure accurate labelling of each point, consistent with its position relative to the bounding box, as initially annotated using the SUSTech POINTS tool.

Subsequently, the pipeline generated a detailed summary outlining the frequency of each label type and the total count of points associated with each class. This systematic approach facilitated the organization and preparation of labelled point cloud data, serving as a crucial step for subsequent analysis and machine learning model training.

3.6 Model development

Two approaches were employed for model training, and the test results from each model were recorded and compared. The first model leveraged the powerful Pointcept codebase, which implements the advanced Point Transformer V2 neural network architecture. In contrast, the second approach adopted the mainstream architecture in natural language processing, known as Transformer, for point-wise classification, leveraging the self-attention and cross-attention modules of Point Transformer (Zhao et al., 2021) and PU-Ray (Lim et al., 2024). The Point transformer v2 was trained using an NVIDIA® GeForce RTX™ 4070Ti GPU with 12 GB VRAM. Due to the smaller model sizes of the self-attention and cross-attention modules in Point Transformer and PU-Ray, the second approach was feasible on a smaller GPU, the NVIDIA® GeForce RTX™ 3080 Lite Hash Rate (LHR), with a 10 GB memory size.

3.6.1 Point Transformer v2

Point Transformer v2 Architecture

The proposed semantic segmentation model adopts the Point Transformer v2 architecture in Figure 5, which follows an encoder-decoder structure inspired by the U-Net design. The encoder branch is responsible for extracting features and progressively downsampling the input point cloud through four stages with the block depths [2, 2, 6, 2]. At each stage, a grid size multiplier is applied for downsampling, with values [x3.0, x2.5, x2.5, x2.5], resulting in progressively smaller point sets. The initial feature dimension is set to 48, which is doubled at each stage, leading to dimensions [96, 192, 384, 384]. Correspondingly, the number of attention groups is set to [12, 24, 48, 64] with a k-nearest-neighbor (k-NN) of 16 (X. Wu et al., 2022).

The decoder branch mirrors the encoder, consisting of four stages with a block depth of [1, 1, 1, 1]. It utilizes grid up modules to reconstruct the point cloud, ultimately restoring it to the original input size while incorporating the enriched feature information from the encoder branch.

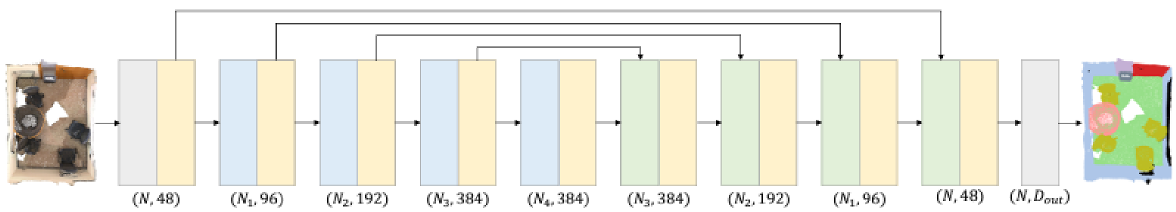


Figure 5 Point Transformer v2 (X. Wu et al., 2022) network for semantic segmentation

Point Transformer v2 introduces several key modifications to improve computational efficiency and performance for point cloud processing tasks compared to the original Point Transformer model, as shown in Figure 6.

Firstly, it employs a novel grouped vector attention mechanism instead of calculating individual attention weights for each point, which can be computationally inefficient and limit generalization. it divides value vector $v = \mathbb{R}^c$ evenly into g groups $(1 \leq g \leq c)$, and points within the same group share a single scalar attention weight from the corresponding group's attention vector (X. Wu et al., 2022). The detailed formulas are shown in Eq. 1. This approach reduces computational complexity while maintaining the ability to capture contextual information.

$$\omega_{ij} = \omega(\gamma(q_i, k_j)), \qquad f_i^{attn} = \sum_{x_j}^{M(pi)} \sum_{l=1}^{g} \sum_{m=1}^{c/g} \text{Softmax}(W_i)_{jl} v_j^{lc/g+m}$$

Eq. 1

Secondly, a position encoding multiplier is incorporated to address the potential capacity limitations of the grouped attention mechanism, enhancing the model's ability to capture positional information within the point cloud data.

Thirdly, Point Transformer v2 introduces partition-based pooling to address non-spatially aligned query sets and uncontrollable information density in traditional sampling-based pooling methods. The point cloud is divided into non-overlapping subsets, where each subset undergoes max pooling for feature aggregation and mean pooling for position aggregation, resulting in a fused point set for the next encoding stage. During unpooling, the fused set is mapped back to the original point set using recorded locations, ensuring efficient and accurate reconstruction. This partition-based pooling approach maintains spatial alignment and controlled information density within the subsets, leading to improved performance and generalizable representations while addressing the limitations of traditional pooling methods.

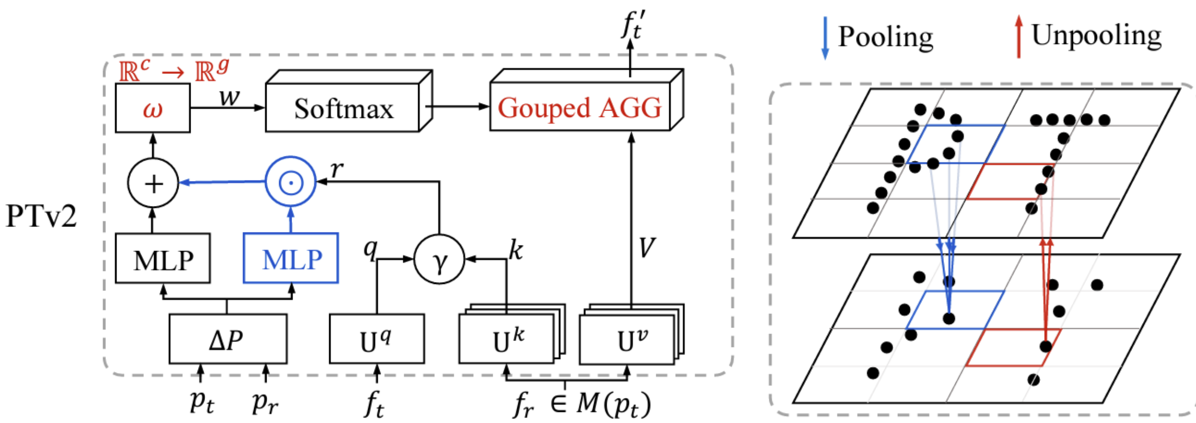


Figure 6 Left: group vector attention (denoted by red), improved position encoding (denoted by blue), and Right: partition-based pooling and unpooling in PTv2 (X. Wu et al., 2022)

Point Transformer v2 Implementation and Improvement

The local features for each point (X, Y, Z, intensity), along with the assigned labels from the annotation exercise, were insufficient to differentiate between points from various classes. Consequently, additional geometric features were computed for each point using the CloudComPy library. These features included the Z-gradient, roughness, intensity gradient, and neighbour density. The gradient refers to the rate of change in any feature value at a point (e.g., elevation or intensity) compared to its neighbours. Definitions of these additional features are detailed in Table 1 (*CloudCompare*, 2023).

The geometric features were computed prior to down-sampling to maintain higher accuracy while simultaneously reducing computational costs. These features were derived for each point based on neighbouring points within a spherical neighbourhood of radius R. Although the down-sampling process inevitably resulted in the loss of fine-shape details, the calculated geometric features retained essential information about the local point distribution, thus providing critical data to the network.

Selecting features and determining the optimal sphere radius for each feature were conducted through an iterative trial-and-error process. This approach allowed for fine-tuning the feature set, balancing the trade-off between information richness and computational efficiency.

Table 1 Additional features computed for each point

Additional feature	Definition
Z-gradient	the gradient of the elevation of a point relative to its neighbours within a sphere of radius R
Roughness	the distance between a point and the best fitting plane to its neighbours within a sphere of radius R
Intensity gradient	the gradient of the intensity of a point relative to its neighbors within a sphere of radius R
Neighbour density	counts the number of neighbour points within a sphere of radius R

The additional features provided valuable information about the surrounding 3D environment of each point. For example, points belonging to broken and solid line markings exhibited similar local features. However, by calculating the intensity gradient for each point, the training model was expected to better distinguish between the classes of these points.

The 50-meter segments still contained too many points to fit the necessary Point Transformer calculations within the memory of an economical GPU. Therefore, the dataset was down-sampled before being fed into the neural network using voxelization. Voxelization down-sampling is a technique used in computer graphics and 3D data processing to reduce the complexity of a point cloud by grouping points into a regular 3D grid called voxels. This process facilitates managing and processing large point cloud datasets by converting them into a more structured form that is easier to handle. Voxelization divides the 3D space into a grid of small, cube-shaped cells known as voxels. Each voxel represents a small volume in 3D space, analogous to a pixel in a 2D image but in three dimensions. Points within a voxel are replaced by their centroid, the average position of all points within the voxel. The voxel size, which is the edge length of each cubic voxel, is a critical parameter. Smaller voxel sizes result in a finer resolution, with more voxels covering the same space, retaining more detail with a decrease in the reduction of the number of points. Larger voxel sizes yield a coarser resolution with fewer voxels, significantly reducing the number of points and thus the data size but also losing more detail. For each 50-meter segment, the dataset initially comprised approximately 1,000,000 points, efficiently down-sampled to around 250,000 points without losing too much spatial relationship. In addition, the three positions value and five features x, y, z, intensity, roughness, density, z-gradient, and intensity gradient, are normalized so that the values are in the range of [0, 1].

In the model development, various loss functions were employed to optimize the training process. These included the Cross-entropy loss, Focal loss, Lovasz loss, and Dice loss. Each of these loss functions serves a distinct purpose in guiding the training of the neural network. The Cross-entropy loss is a standard choice for multi-class classification tasks, providing a measure of dissimilarity between predicted and actual class distributions. Focal loss, on the other hand, addresses the issue of class imbalance by down-weighting the contributions of well-classified examples. Lovasz loss is particularly effective for optimizing intersection-over-union (IoU) based metrics while being robust to class imbalance. The Dice loss is designed to address class imbalance in segmentation tasks. It measures the overlap between predicted and ground truth segmentation masks. By minimizing the Dice loss during training, models learn to produce accurate segmentation masks, handling scenarios with imbalanced foreground and background classes. Additionally, learning rates and epochs were explored to optimize performance. This iterative process involved testing various combinations of learning rates, determining the step size for updating model parameters, and epochs, representing the number of complete passes through the training dataset. By systematically varying these hyper parameters, the aim was to identify the configuration that yielded the most effective convergence and generalization of the neural network. The optimal combination of learning rates and epochs was determined to enhance the model's ability to improve predictive accuracy through experimentation and evaluation of model performance metrics, such as accuracy, F1 score and intersection over union.

Two models were evaluated: the full model and the reduced model. The full model incorporated all 11 classes (plus clutter), whereas the reduced model consolidated similar classes.

Specifically, it merged the lane and shoulder into a single 'pavement' class, combined chevron, broken line, solid line, and arrow markings into a single 'marking' class, and left the remaining classes unchanged. This resulted in 7 classes (plus clutter) for the reduced model.

The model's performance was assessed using mean Intersection over Union (mIoU), prediction precision, and F1 score for each class. The mIoU offered a comprehensive measure of the model's performance across all classes by averaging the IoU values, as illustrated in Eq. 2. Meanwhile, the precision and recall for each class provided insight into how accurately the model predicted each class, considering both correct and incorrect predictions, as calculated in Eq. 3.

$$IoU_i = \frac{Intersection}{Union} = \frac{TP_i}{TP_i + FP_i + FN_i}$$

$$mIoU = \frac{1}{N}\sum_{i=1}^{N} IoU_i$$

Eq. 2

where:

TP_i = the number of true positive predictions for the class i.

FP_i = the number of false positive predictions for the class i.

FN_i = the number of false negative predictions for the class i.

N = the number of classes.

IoU_i = the IoU for class i.

$mIoU$ = the mean Intersection over Union.

$$Precision = \frac{TP_i}{TP_i + FP_i}$$

$$Recall = \frac{TP_i}{TP_i + FN_i}$$

$$F1\ score_i = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall}\right)$$

$$mean\ F1\ score = \frac{1}{N}\sum_{i=1}^{N} F1\ score_i$$

Eq. 3

where:

TP_i = the number of true positive predictions for the class i.

FP_i = the number of false positive predictions for the class i.

FN_i = the number of false negative predictions for the class i.

F1 score = the $\mathrm{F1}\ score$ for each class i.

Mean F1 score = the mean F1 for all classes

A confusion matrix showing the accuracy of the predicted class compared to the ground truth class for the test files has been constructed. The confusion matrix is a key tool for evaluating the performance of a classification model. For a multi-class classification problem, the confusion matrix provided a comprehensive view of the model's predictive capabilities by displaying the true positives, false positives, true negatives, and false negatives for each class. This allowed for a detailed analysis of where the model performed well and where it struggled, enabling targeted improvements to enhance overall accuracy. This evaluation identified specific classes where the model exhibited high accuracy and others where further refinement was necessary, thereby guiding future model development and optimization efforts.

The complete data processing and implementation procedure is visually demonstrated in Figure 7.

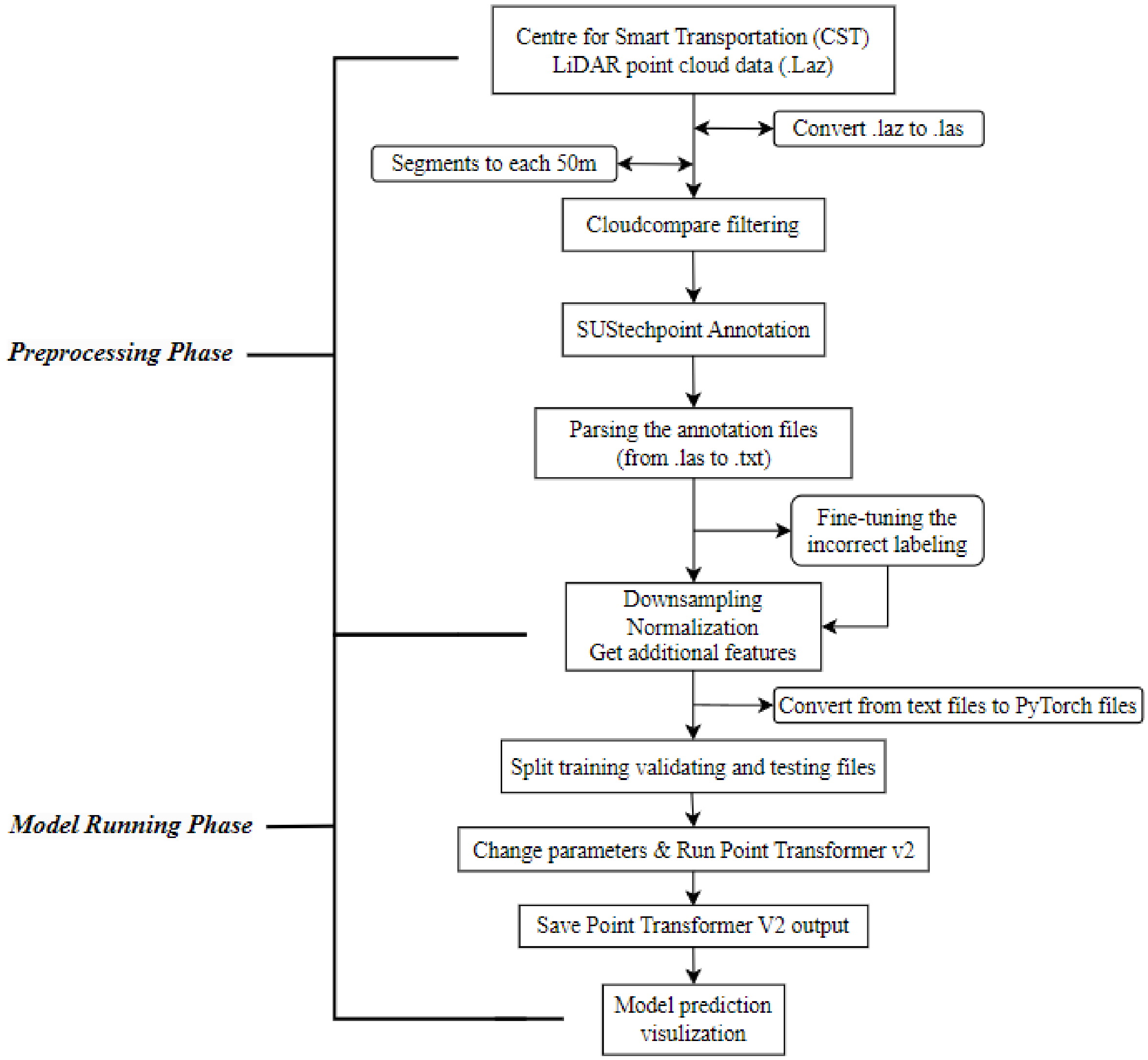


Figure 7 Point Transformer v2 flowchart diagram

3.6.2 Transformer-Base point classification

A point cloud is defined as $P \subset \mathbb{R}^{n\times 5}$. A point, $p \in P$, consists of $[x \;\; y \;\; z \;\; i \;\; d]^T$, the 3D coordinates (x, y, z), the reflection intensity i and the density feature d. An approach for point cloud segmentation is by point-wise classification by a neural network function $f(\cdot)$, which outputs the one-hot vector representation of a p's label class $c \in \mathbb{N}^{|c|}$, where $\sum_i^{|c|} c_i = 1$. The function $f(\cdot)$ takes in the input point p and its neighbouring patch $N \subset \mathbb{R}^{k\times 5}$:

$$c = f(p, N).$$

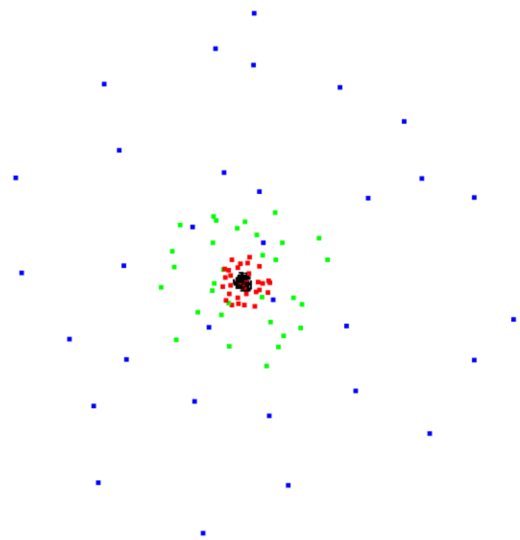


Figure 8 A sample patch on a 2D point cloud generated with a uniform distribution. Black, red, green and blue points are sampled using voxel sizes of 0.1 metre, 0.3 metre, 1.0 metre and 3.0 metre.

Patch points are selected by k-nearest-neighbor (k-NN) sampling on the downsampled point clouds of P with voxelization using different voxel sizes. In other words, 32 nearest points from p are selected in each downsampled point cloud using voxel sizes of 0.1 metre, 0.3 metres, 1 metre and 3 metres, resulting in a patch size $|N| = 128$. For voxelization downsampling, one point in a voxel is selected by max-pooling using the i value (i.e. one point with the highest intensity in a voxel is selected), which is based on the assumption that points with a higher intensity carry more information. The multi-resolution point sampling with different voxel sizes imitates human vision, where the spatial resolution is higher in the central vision near the focusing point and gets lower in the peripheral vision farther away from the focusing point. The voxel sizes are determined to minimize the intersections between k-NNs sampled with different voxel sizes (Figure 8). For example, in uniformly distributed points on a surface, 32 k-NN points that are evenly spaced with 0.3-meter intervals are clustered in a circular area with approximately a 1-meter radius (e.g. $32 \times 0.3^2 \approx \pi \times 1.0^2$). Thus, the points with a finer resolution are less likely to intersect with those with a coarser resolution.

All 3D coordinates (x, y, z) of p and N are translated for relative position encoding (Zhao et al., 2021) so that the points are centred at the average point of P'. The first four features x, y, z

and i are normalized so that most values are in the range [-1, 1]. In addition, each patch point $p' \in N$ has an extra fifth feature d, which represents the density of the voxel that p' belongs in. The equation for calculating the density is as follows:

$$d = \frac{\#\ of\ points\ within\ voxel}{(voxel\ size * coefficient)^2}$$

Where the coefficient, determined though empirical studies, is set to 50.

For feature vector formatting, the d of p is filled with a dummy feature 0. Point densities in LiDAR point clouds vary in different regions depending on the distance from this sensor. Therefore, the density feature compensates for the information loss in the regions that may suffer from sparser densities. The information is obtained by counting the input points and normalizing the number differently for each voxel size. Random translation, rotation, scaling, and jittering are applied for data augmentation, as suggested by (Zhu et al., 2024).

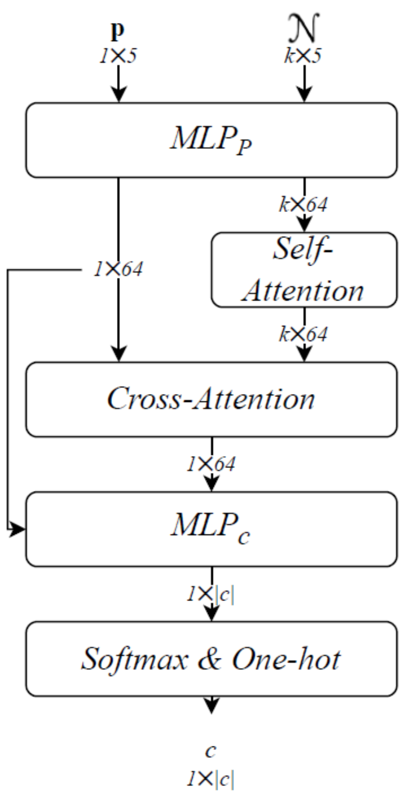


Figure 9 Network architecture for point cloud classification

Point clouds are permutation-invariant data that follow the characteristics of the set data structure. The unorderedness hinders the utilization of early convolutional neural network approaches in computer vision, which led to the adoption of the transformer architecture from the natural language processing domain (Vaswani et al., 2017) for point cloud processing (Lim et al., 2024; Zhao et al., 2021). The method adopts a variation of the architecture proposed by Lim et al. (Lim et al., 2024) to utilize the transformer network architecture in this study (Figure 9). Firstly, all points, including p and N, are fed into a multi-layer perception (MLP_{p}) to extract point-wise features in a latent space. In addition, the self-attention layer of Zhao et al. (Zhao et al., 2021) encodes each point feature of N to extract inter-point relationships within the patch. After obtaining the future vectors of all points, a cross-attention layer (Lim et al., 2024) naturally accumulates all the features of N relating to the latent features of p. The resulting cross-attention feature vector is concatenated with a skip layer of p's features resulting from MLP_{p}. The one-hot vector representation results from MLP_c and a softmax function given the concatenation.

Chapter 4. Natural Language Scene Description Methodology

A significant gap persists in directly converting 3D point cloud data to textual descriptions without the intermediary step of 2D image conversion. While existing models, such as PointLLM (Xu et al., 2023b), have attempted to address this challenge, their efficacy is primarily limited to scenarios on which they were previously trained. Consequently, the results obtained using this model for more complex or multi-object point cloud data are suboptimal, underscoring the need for further research and development in this domain.

Given these limitations, converting point cloud data to 2D images is necessary. However, to mitigate information loss during this process, we propose utilizing multiple views of an image. These multi-view representations can then be fed into a pre-trained model to generate the desired textual output, potentially preserving more original 3D information and improving the quality of the resulting descriptions.

4.1 Large language model (LLM) procedure

GPT-4 is a Transformer-based model pre-trained to predict the next token in a document (Achiam et al., 2023). For this natural language scene description task, we leverage GPT-4o by OpenAI, an advanced iteration of GPT-4 and GPT-4 Turbo. This model can accept both text and image inputs while outputting text, boasting twice the processing speed and 50% improved cost-effectiveness compared to GPT-4 Turbo. However, it's important to note that GPT-4o is not open-source, and a purchased API key is required for its utilization. The figure below illustrates the full flowchart of the large language model (LLM) procedure, demonstrating how GPT-4o is integrated into our methodology.

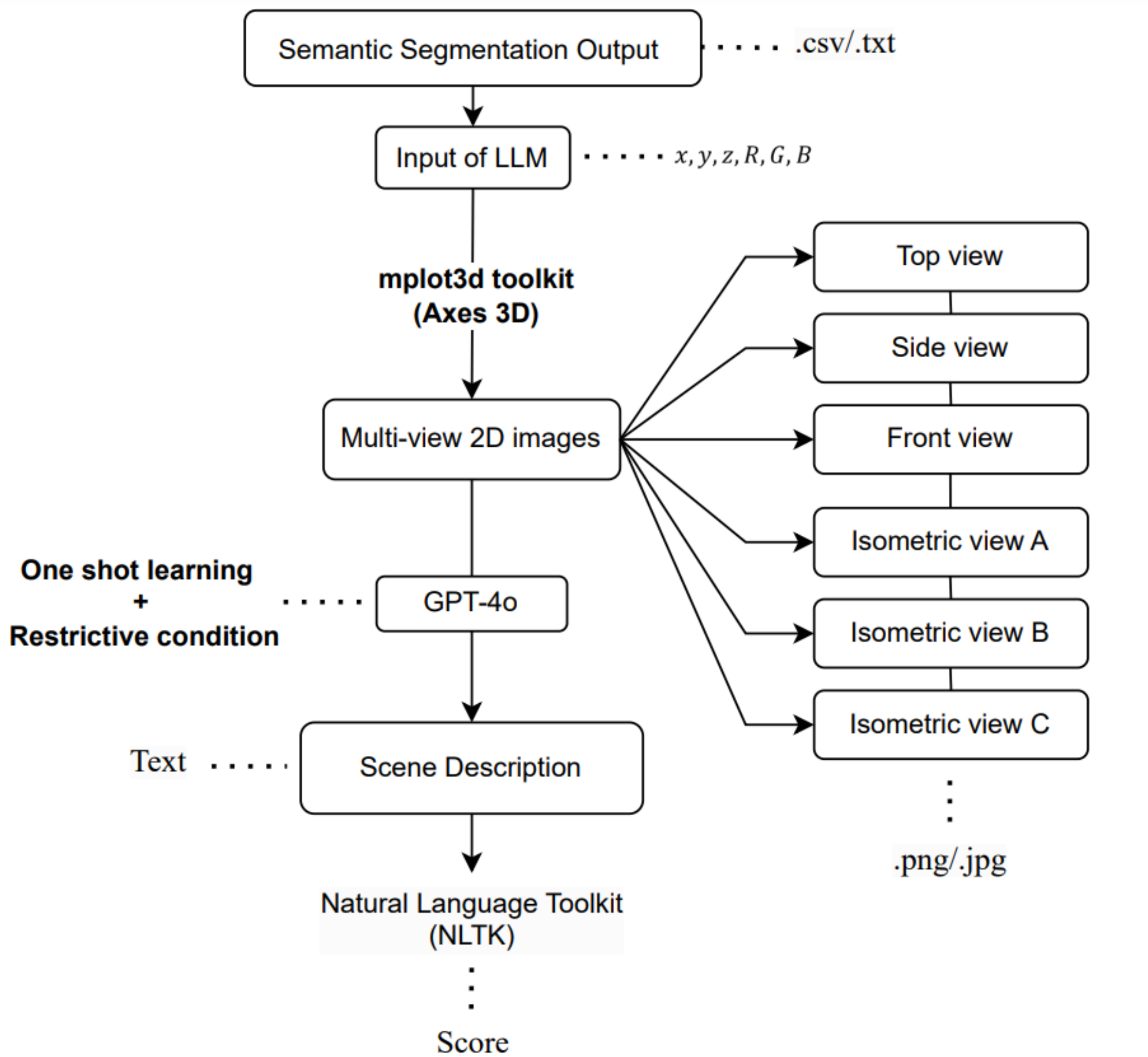


Figure 10 Natural language scene description procedure

The methodology for understanding the scene description is straightforward. We begin with the output from the previous semantic segmentation, stored in .csv or .txt files. Each row in these files contains the local or global coordinates [x, y, z, R, G, B]. Using the Axes3D toolkit from mplot3d, we generate multi-view 2D images. Six views are selected for each scene to input into the GPT-4o model. This number of views is chosen because increasing the number of views does not significantly improve the results, while fewer views can lead to inaccuracies in describing the scene.

GPT-4o, as a Large Language Model (LLM) trained on nearly 2 trillion parameters, is designed to handle many real-life scenarios. To optimize its performance for this specific task, a one-shot learning method was used and provided a desired description for the model:

"The highway point cloud data represents a multi-lane highway consisting of four lanes. These lanes are marked by three broken lines and two solid lines. A concrete barrier median separates the lanes for opposing traffic flows. Alongside the road, there is a light pole, and an overhead traffic sign provides guidance for drivers."

Additionally, the model was given contextual information to frame its analysis:

- Role: "You are an advanced AI assistant installed on the autonomous vehicle, equipped with conversational analysis capabilities for discussing autonomous driving scenarios. The perspective presented is from the point-of-view of the autonomous vehicle, where the camera is mounted. It's important to note that the autonomous vehicle itself is not visible in the images provided."
- Scenario Overview: "The images uploaded represent scenes captured from the viewpoint of the autonomous vehicle's camera. The images are taken from the coloured point cloud data. The colours of point denote different objects."
- Clarification: "Your role as the AI assistant is to analyze these scenes, considering the position, number of lanes or markings, quantify the number of the object, if possible, potential interactions of the autonomous vehicle with other objects in the environment."
- Additional Information: "Feel free to ask questions or seek clarification regarding the scene or any specific details you need to accurately analyze and discuss the autonomous driving scenario presented in the images."

Utilizing these restrictive conditions in the GPT-4o model, along with multi-view images, enables the generation of desired descriptive results that align closely with the specified requirements, ensuring a high level of confidence. Once the output was generated by the GPT-4o model, three evaluation metrics were used to assess the numerical performance of the generated text against the reference text. The reference text was derived from manually checked ground truth files. The evaluation focused on three key metrics:

1. METEOR (Metric for Evaluation of Translation with Explicit ORdering): Assesses the quality of the generated text by comparing it to a reference, focusing on alignment, paraphrasing, and linguistic variations. It focuses on precision and recall, using exact word matches, as well as matching synonyms and stems to assess the alignment between the generated and reference text (Banerjee & Lavie, 2005; Saadany & Orasan, 2021).

2. Semantic Similarity: Measures how closely the meanings of the generated text and the reference text align. It goes beyond word-level matches to assess whether the generated text conveys the same message or information as the reference (Chandrasekaran & Mago, 2021).
3. Entity Score: Evaluates the performance of a system in recognizing, extracting, and classifying named entities in text. It measures how accurately the system identifies these entities compared to a reference or ground truth (Daiber et al., 2013; Zhong et al., 2015).

Each score was calculated using the Natural Language Toolkit (NLTK), with the final total score based on a weighted system: 0.1 for METEOR, 0.8 for Semantic Similarity, and 0.1 for the Entity Score. This weighting ensures that the evaluation emphasizes semantic meaning, reducing the bias of human evaluation and providing a more objective assessment of the model's overall performance.

Chapter 5. Results & Discussion

5.1 Semantic Segmentation Results

The dataset distribution for each class is illustrated in the Figure below. Since this is a semantic segmentation task rather than object detection, it is more important to focus on the total number of points per class rather than the number of individual objects. The chart shows the number of points for each class in the full model.

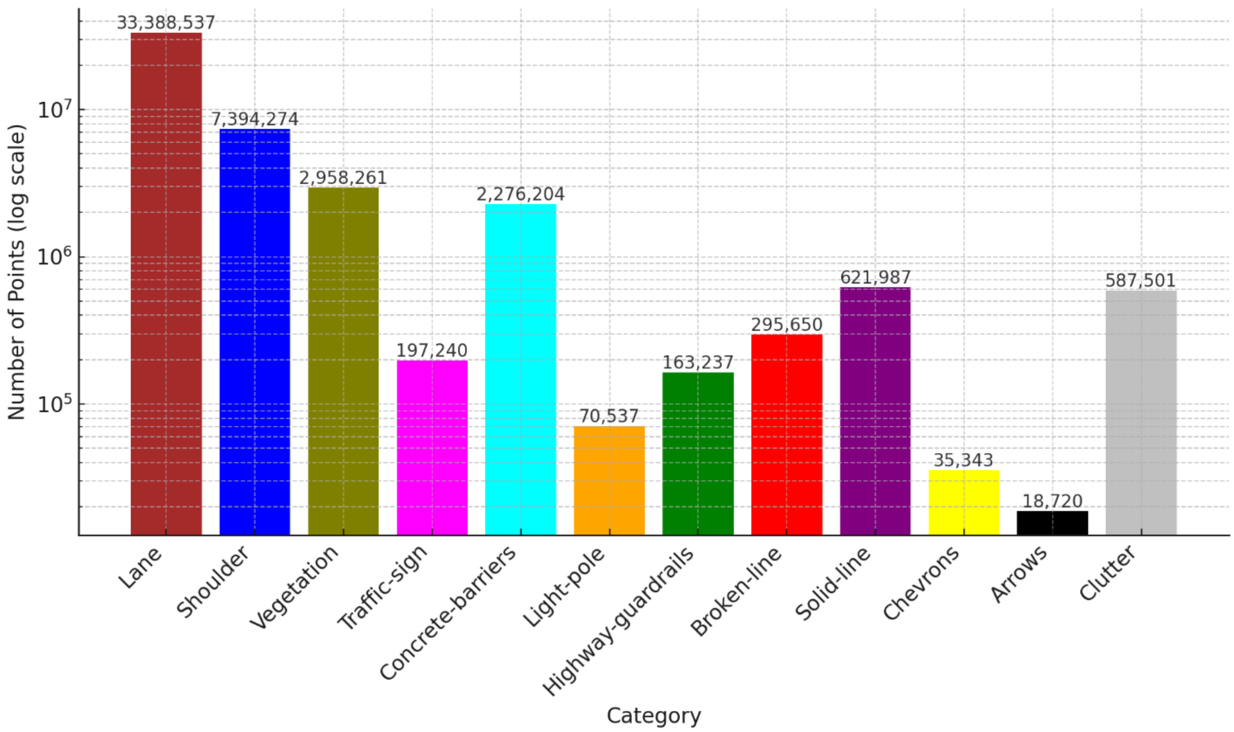


Figure 11 Dataset point-wise distribution Full model

In the reduced model, the “Lane” and “Shoulder” classes are combined into a single category called “Pavement.” Additionally, all marking-related classes, such as “Broken-line,” “Solid-line,” “Chevrons,” and “Arrows,” are grouped together under the category “Marking.”

5.1.1 Point Transformer v2 result:

The experiments evaluated a 2.5km highway segment in Alberta, Canada, using the proposed model with default parameters: roughness radius 1m, density radius 1m, z-gradient radius 0.2m, and intensity-gradient radius 1.5m. The model was trained for 200 epochs with a learning rate of 0.001 and a decay rate of 0.05. Four additional attributes were incorporated into the model (as per

Section 3.6.1) to improve accuracy. The evaluation considered 400m divided into eight randomly selected segments. The key metrics were intersection over union (IoU) and F1 score for well-rounded evaluations.

Table 2 presents the detailed accuracy results for the model before and after incorporating these additional features.

Table 2 Model output for both with and without four additional features

Experiment Results for the four additional attributes						
Model type	Roughness	Density	Z-gradient	Intensity-gradient	Mean IoU	Mean F1 score
Full model	Yes	Yes	Yes	Yes	64.30	71.91
	N/A	N/A	N/A	N/A	49.24	59.02
Reduce model	Yes	Yes	Yes	Yes	78.29	86.48
	N/A	N/A	N/A	N/A	60.66	72.95

A sensitivity analysis of the four additional attributes was performed to show the proposed strategy's necessity and effectiveness. The results show a significant improvement in additional attributes for both models. The local features cannot give enough information for the model to learn the shape and pattern. Most researchers use IoU for their evaluation in semantic segmentation tasks since it quantifies how well the model can distinguish objects from their backgrounds. The full model mIoU improved from 49.24% to 64.30%, and the mean F1 score increased from 59.02% to 71.91%. On the other hand, the reduced model mIoU improved from 60.66% to 78.29%, and the mean F1 score increased from 72.95% to 86.48%. The results demonstrate the improved performance by including the new attributes in the model.

Table 3 Semantic segmentation results on Alberta Highway 2 datasets (reduced model)

Class	F1 score	IoU
Clutter	57.73	40.58
Concrete barriers	95.78	91.90
Highway guardrails	85.00	73.91
Light pole	90.43	82.53
Marking	74.19	58.97
Pavement	97.91	95.91
Traffic sign	97.32	94.78

Vegetation	93.48	87.76
Mean	86.48	78.29
Total Accuracy	95.03	

Table 4 Semantic segmentation results on Alberta Highway 2 datasets (full model)

Class	F1 score	IoU
Arrows	0.00	0.00
Broken Line	45.80	29.70
Chevrons	14.38	7.74
Clutter	90.22	82.18
Concrete Barrier	95.61	91.59
Highway Guardrail	83.13	71.14
Lane	97.15	94.45
Light Pole	89.24	80.57
Shoulder	88.56	79.46
Solid Line	67.04	50.42
Traffic Sign	95.42	91.24
Vegetation	96.42	93.09
Mean	71.91	64.30
Total Accuracy	94.57	

Figure 12 and Figure 13 present the optimized models, highlighting areas of misclassification. A control variable approach was employed during optimization to ensure optimal model performance. For each trial, only one parameter was altered, allowing for precise identification of its impact on the model.

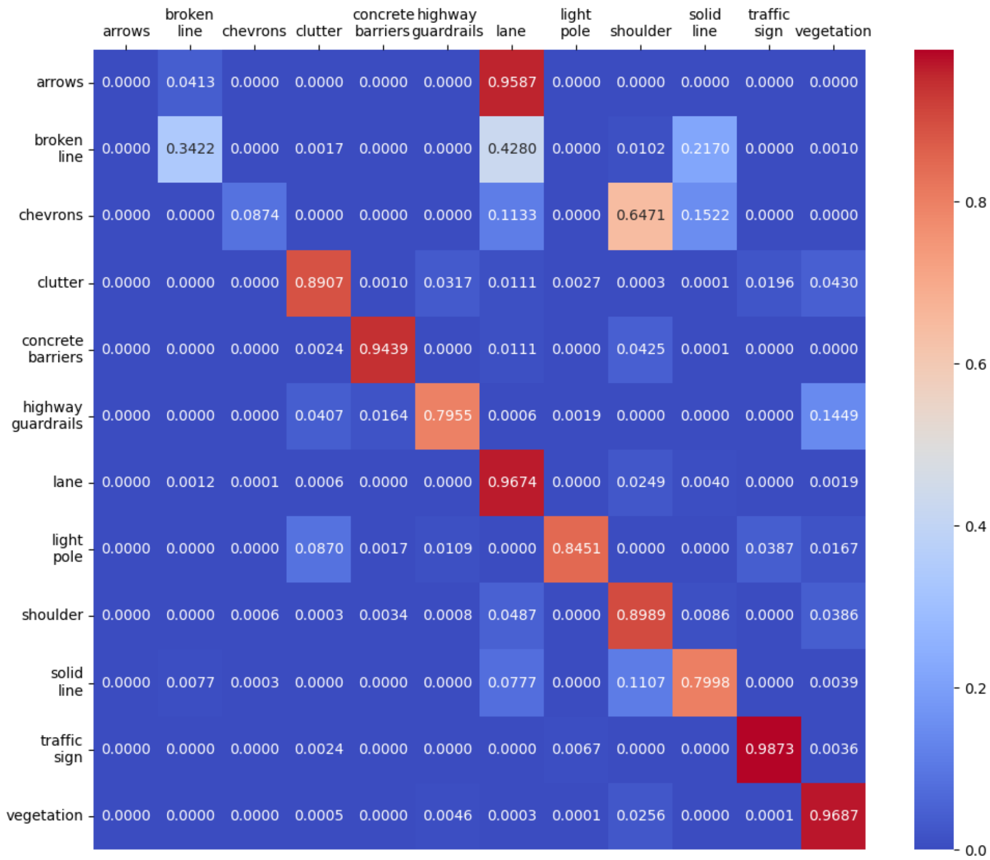


Figure 12 Confusion matrix for Point Transformer v2 full model

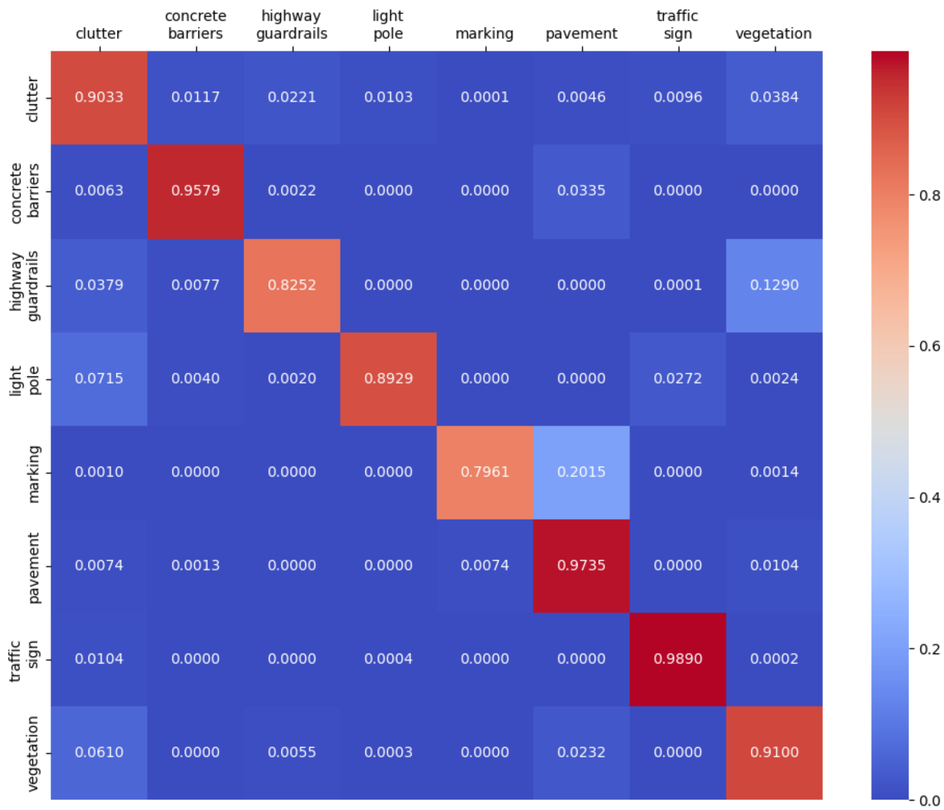


Figure 13 Confusion matrix for Point Transformer v2 reduced model

Both models achieved their optimal performance when trained for 200 epochs. Due to GPU memory constraints, each epoch consisted of only one batch. The AdamW optimizer was utilized with an initial learning rate of 0.001 and a weight decay of 0.05 to prevent overfitting. A combination of focal loss, dice loss, Lovász loss, and cross-entropy loss functions were employed to address data imbalance.

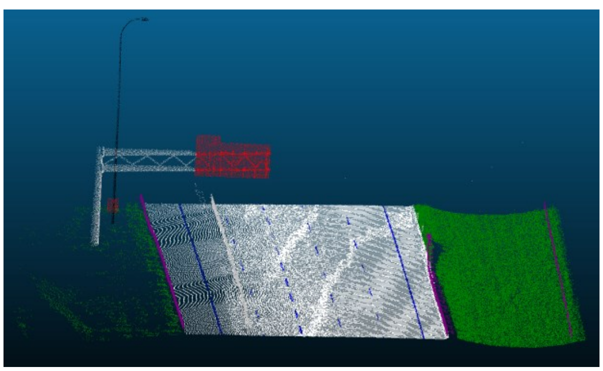
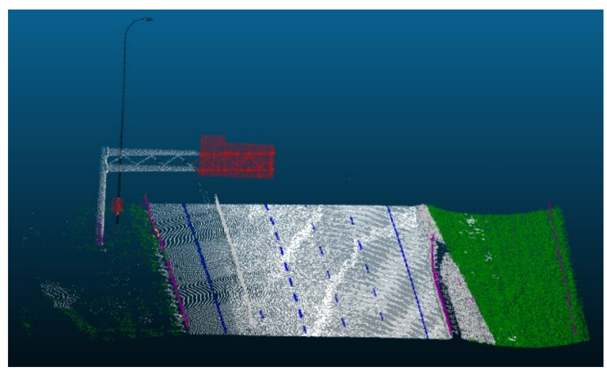
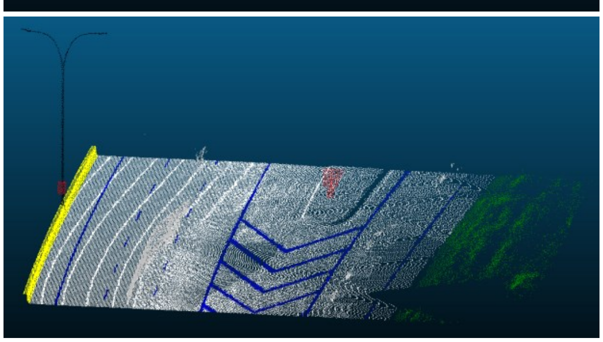
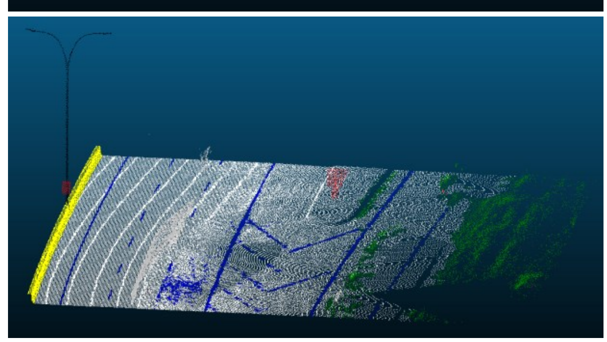
After experimenting with multiple trials, the full model achieved the best performance using a combination of focal loss and cross-entropy loss with the following parameters: a density radius of 1 meter, a roughness radius of 1 meter, a Z-gradient radius of 0.2 meters, and an intensity gradient of 1.5 meters. Conversely, cross-entropy loss yielded the best performance for the reduced model with a roughness radius of 1 meter, a density radius of 1.5 meters, a Z-gradient of 1 meter, and an intensity gradient of 0.875 meters.

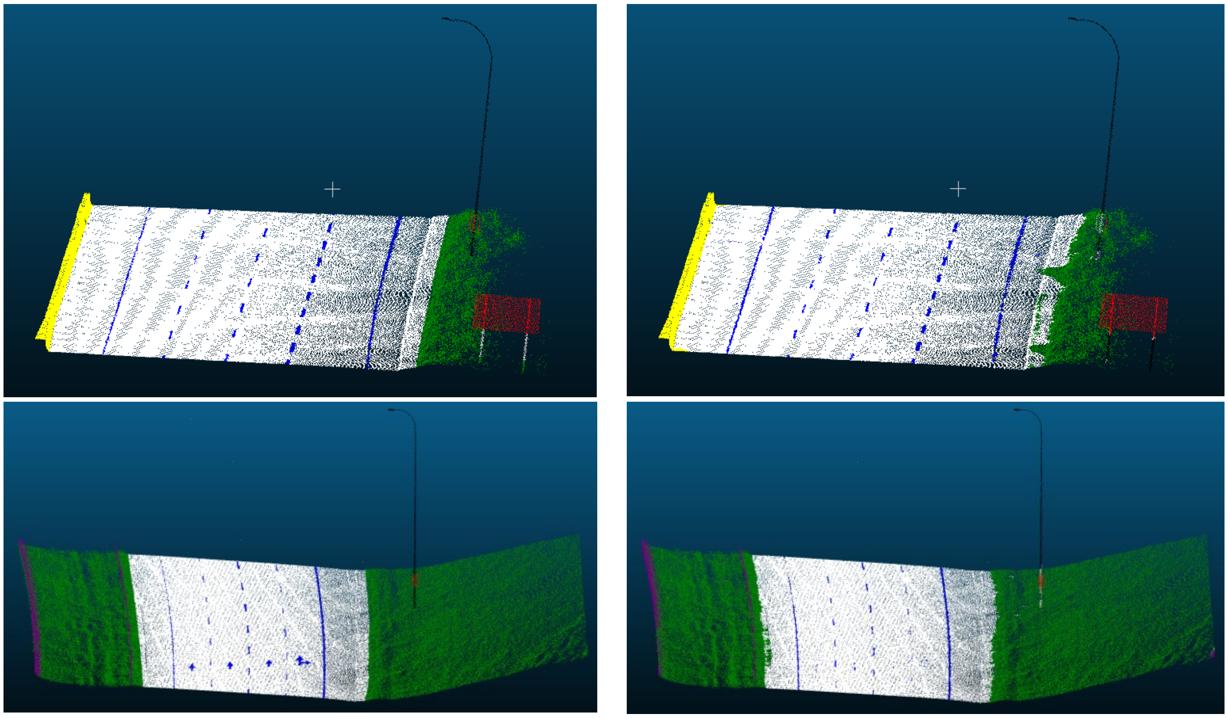
The diagonal values in both confusion matrices represent the correct label percentages. In the full model, Chevron, broken line, and arrow marking categories performed the worst. This poor performance is attributed to the relatively low number of training examples for these categories in the overall training set. Additionally, some markings were faded and unclear, further contributing to misclassification. The similarity among some of these markings also led to confusion in the model's predictions.

While the F1 score provides a balanced measure of both precision and recall across all classes, in this case, prioritizing recall is more critical due to the safety risks posed by false negatives. Although precision, which relates to false positives, is important, its impact is less severe, primarily causing traffic congestion or disruption. For example, a false negative would miss an actual traffic sign, leading to potential safety hazards, while a false positive would detect a non-existent sign, resulting in minor inconveniences. Therefore, the model should prioritize recall over precision to minimize safety risks.

The visualization below compares the best-performing models with their ground truth values. The ground truth colours are assigned as follows: White represents Pavement, Blue represents Marking, Green represents Vegetation, Red represents Traffic-sign, Magenta represents Highway-guardrails, Yellow represents Concrete-barriers, Black represents light poles, and Silver represents Clutter.

Table 5 Point Transformer v2 Reduced model visualization

Ground truth	Prediction
	
	



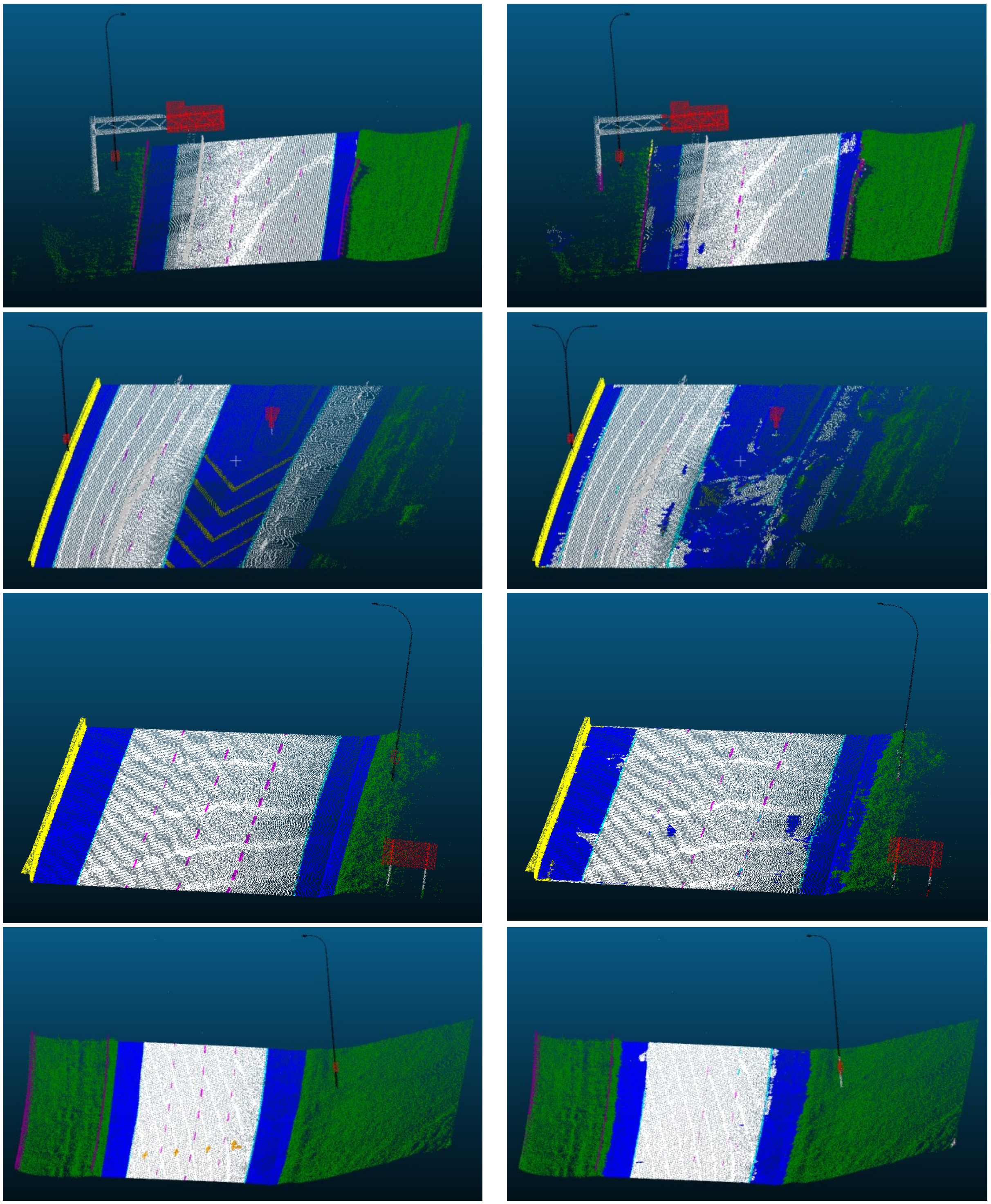
A subset of four files from the original eight was carefully selected to visualize model predictions based on their representation of diverse environmental conditions. The selection process aimed to comprehensively evaluate the model's performance across varied scenarios, providing a more robust assessment of its capabilities and potential limitations in different contexts. The remaining test files can be found in Appendix A.

The visualizations reveal that the model generally extracted markings correctly, although there were some instances where arrow markings were incorrectly labelled as pavement. Additionally, the model showed some confusion with traffic signs attached to light poles. Despite these minor issues, the model demonstrated good overall performance in detecting necessary infrastructure. This success across various environments highlights the model's potential for future transportation-related applications.

The ground truth colours for this full model were more extensive due to the inclusion of additional classes. The colour assignments are as follows: White represents Lane, Blue represents Shoulder, Olive represents Chevrons, Purple represents Broken line, Cyan represents Solid line, and Orange represents arrows. These additional classes provide a more detailed categorization of road markings. The remaining colours maintain their previous assignments: Green for Vegetation, Red for Traffic signs, Magenta for Highway guardrails, Yellow for Concrete barriers, Black for light poles, and Silver for Clutter. This expanded colour scheme allows a more nuanced visualization of the various elements in transportation infrastructure scenes.

Table 6 Point Transformer v2 Full model visualization

Ground truth	Prediction



For the full model, as the class specificity increases, its weaknesses become more apparent. In addition to the previously mentioned issues with lane markings, the model also experienced confusion between lanes and shoulders, especially in cases where a shoulder is present in the middle, leading to misclassification. Despite these challenges, the extraction of traffic signs and vegetation was excellent. However, the model still requires further modifications and improvements to handle the intricacies of all 12 classes effectively.

5.1.2 Point classification result:

The point classification semantic segmentation results and the corresponding confusion matrix are below. The point classification results were obtained in the same format as the Point Transformer v2 for comparative purposes.

Table 7 Semantic segmentation results on Alberta Highway 2 datasets (reduced model)

Class	F1 score	IoU
Clutter	94.12	88.90
Concrete Barrier	97.28	94.70
Highway Guardrail	90.37	82.43
Light Pole	82.73	70.55
Marking	83.71	71.99
Pavement	97.84	95.78
Traffic Sign	96.36	92.97
Vegetation	95.25	90.94
Mean	92.21	86.03
Total Accuracy	96.69	

Table 8 Semantic segmentation results on Alberta Highway 2 datasets (full model)

Class	F1 score	IoU
Arrows	82.65	70.43
Broken Line	77.81	63.68
Chevrons	52.48	35.57
Clutter	90.76	83.09
Concrete Barrier	97.27	94.69
Highway Guardrail	88.17	78.84
Lane	97.44	95.01
Light Pole	82.70	70.51
Shoulder	87.56	77.87

Solid Line	82.18	69.76
Traffic Sign	96.95	94.07
Vegetation	95.36	91.13
Mean	85.94	77.05
Total Accuracy	94.52	

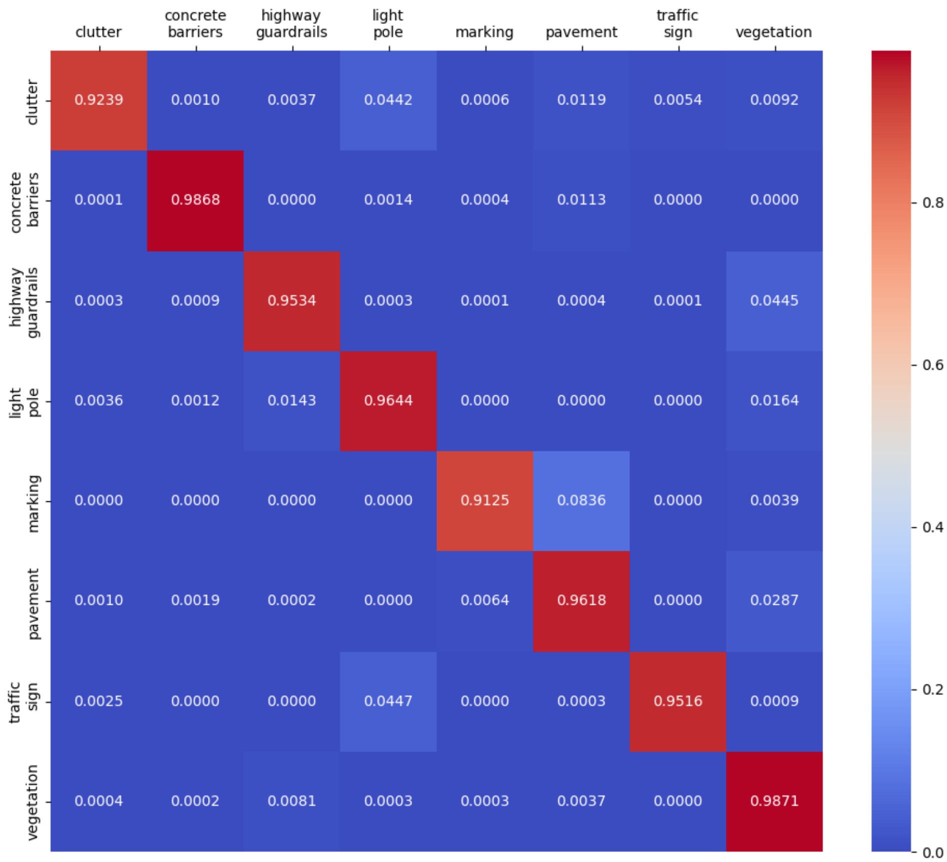


Figure 14 Point classification confusion matrix reduced model

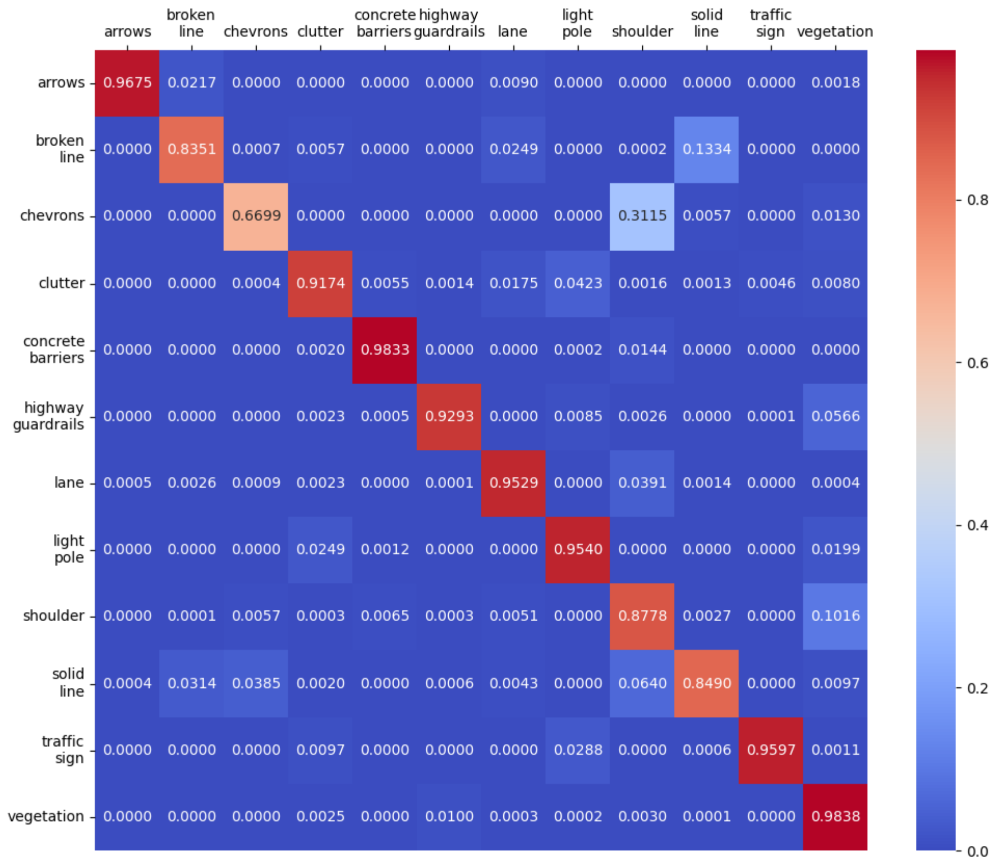


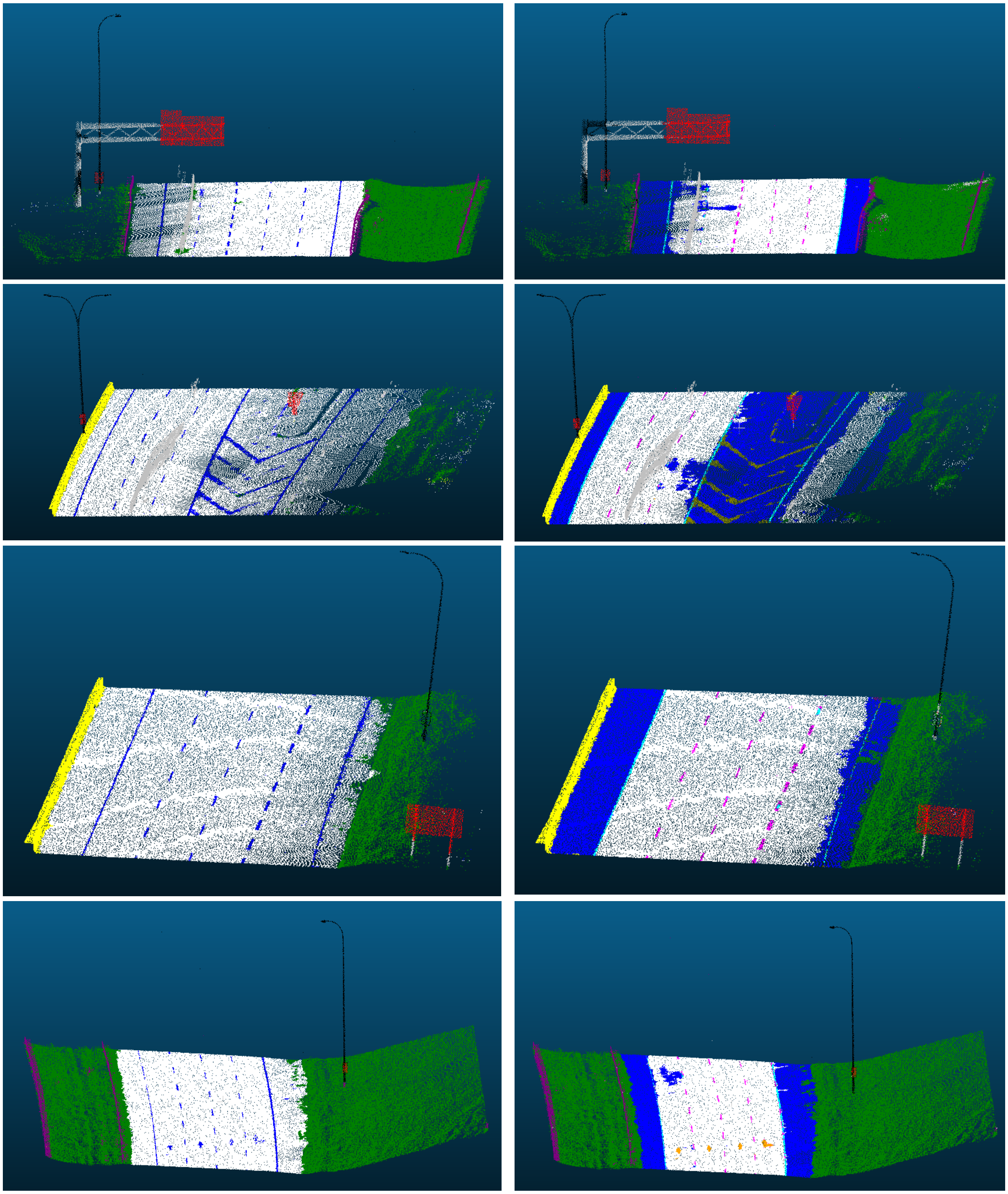
Figure 15 Point classification confusion matrix full model

The point classification models were trained using different hyperparameter configurations to optimize their performance. Both the full and reduced models employed focal loss as the objective loss function, with a modulating factor of $\alpha = 0.25, \gamma = 2$. The full model was trained for 300 epochs, while the reduced model was trained for 200 epochs. A learning rate of 0.001 was used, with a decay rate of 1% applied after each epoch. The models were trained with a batch size of 32, as detailed in the methodology section. The evaluation results demonstrated the superiority of the reduced model over the full model in terms of both the mean F1 score and the mean Intersection over Union (mIoU) metrics. Notably, the reduced model achieved an impressive mean F1 score of 92.21% and a mIoU of 86.03%, outperforming the full model, which obtained a mean F1 score of 85.94% and a mIoU of 77.05%.

Table 9 shows the reduced and full model prediction visualization using the point classification approach.

Table 9 Transformer-based point classification visualization

Reduce model prediction | Full model prediction



5.2 Semantic segmentation result comparison

A comprehensive evaluation was conducted on both the full and reduced semantic segmentation models, with training and testing on the same dataset to ensure a consistent and fair comparison.

Notably, the same voxel size of 0.1m was employed across both models, further reinforcing the validity of the comparative analysis. A detailed examination of the results revealed distinct performance characteristics between the two approaches.

Both the Point Transformer v2 and the Transformer-based point classification model demonstrated strong performance. However, the Transformer-based point classification model yielded superior results when applied to the full model with 12 classes. This model exhibited an enhanced capability to differentiate between various markings, leading to more precise classifications, as evidenced by higher accuracy metrics.

In terms of visualization, both models produced high-quality results, effectively demonstrating misclassifications and laying a strong foundation for subsequent natural language scene descriptions. However, when considering processing time, the Point Transformer v2 was faster in both training and inference. Specifically, it required approximately 2.5 hours to train for 200 epochs and 3 minutes for inference on a single segment. In contrast, the Transformer-based point classification model required roughly twice as much time for both training and inference, primarily due to the use of group vector attention in the Point Transformer v2 model and the 128-per-patch configuration in the Transformer-based point classification model.

These contrasting strengths highlight an inherent trade-off between classification accuracy and processing time:

1. Segmentation Performance: The Transformer-based point classification model excelled in overall accuracy for both full and reduced datasets.
2. Visualization Quality: Both models demonstrated excellent visualization quality, effectively representing the highway environment.
3. Processing time: The Point Transformer v2 exhibited significantly faster processing times, making it more efficient for both training and inference compared to the Transformer-based point classification model.

This dichotomy in performance characteristics highlights the importance of aligning model selection with the specific requirements of the application. Tasks that prioritize precise classification may benefit more from the Transformer-based point classification model, while applications where processing time is critical may find greater value in the Point Transformer v2, even if it results in some compromise on classification accuracy.

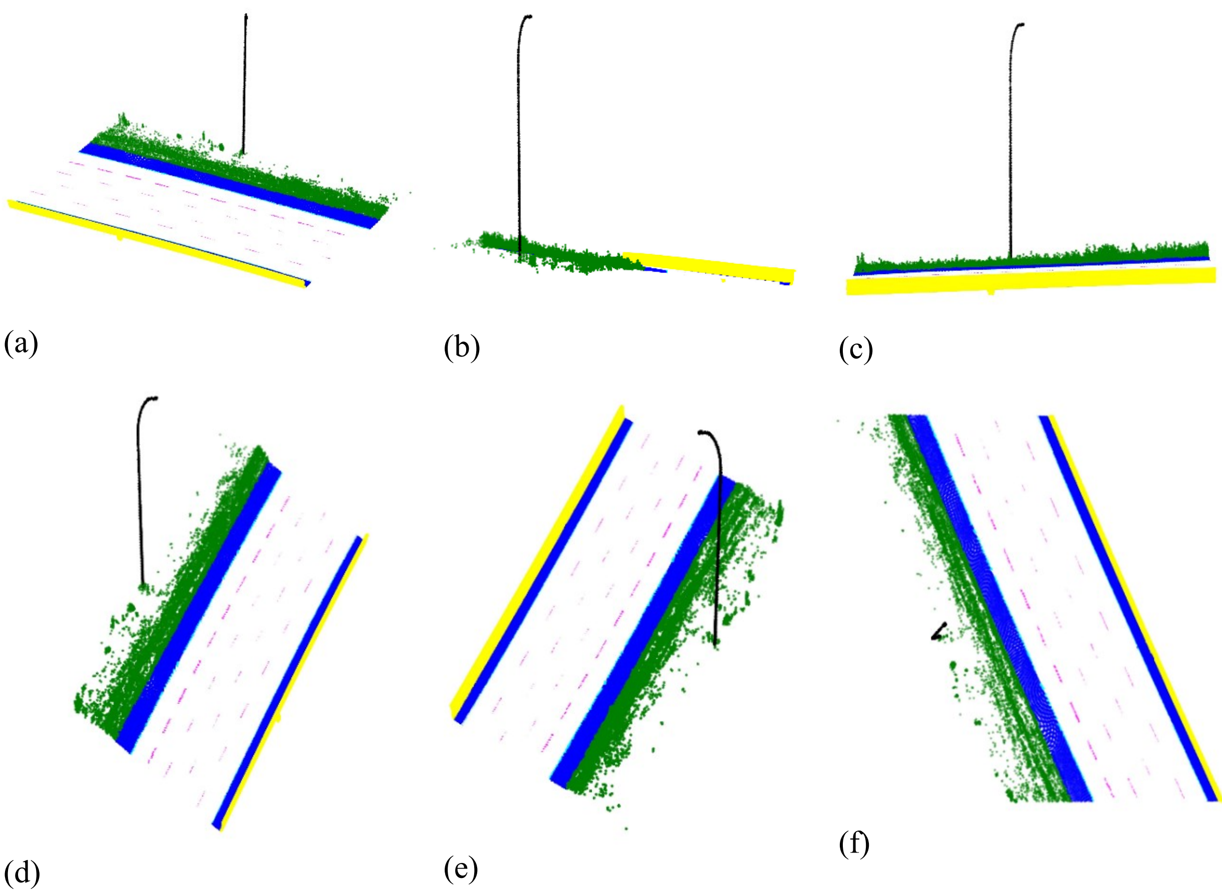
Ultimately, selecting the most appropriate model requires careful consideration of the trade-off between classification accuracy and processing time, based on the intended use case. The decision should be driven by the specific needs and priorities of the task, ensuring that the chosen model aligns optimally with the project's objectives.

5.3 Natural language scene description results

A comparative study was conducted at the first stage to evaluate the description change regarding different views, as illustrated in the Figure below. This analysis employed the identical restrictive conditions outlined in the methodology section. The resultant outputs from this comparative analysis are presented below. The generated description results can be split into two categories: Ground truth results and Predicted results.

<u>Scene description using Ground truth files:</u>

Full model scenarios (12 classes):



(a) (b) (c)

(d) (e) (f)

Figure 16 Ground truth of a sample (section 16 - Full model) multi-view images (a) Isometric View A (b) Front View (c) Side View (d) Isometric View B (e) Isometric View C (f) Top View

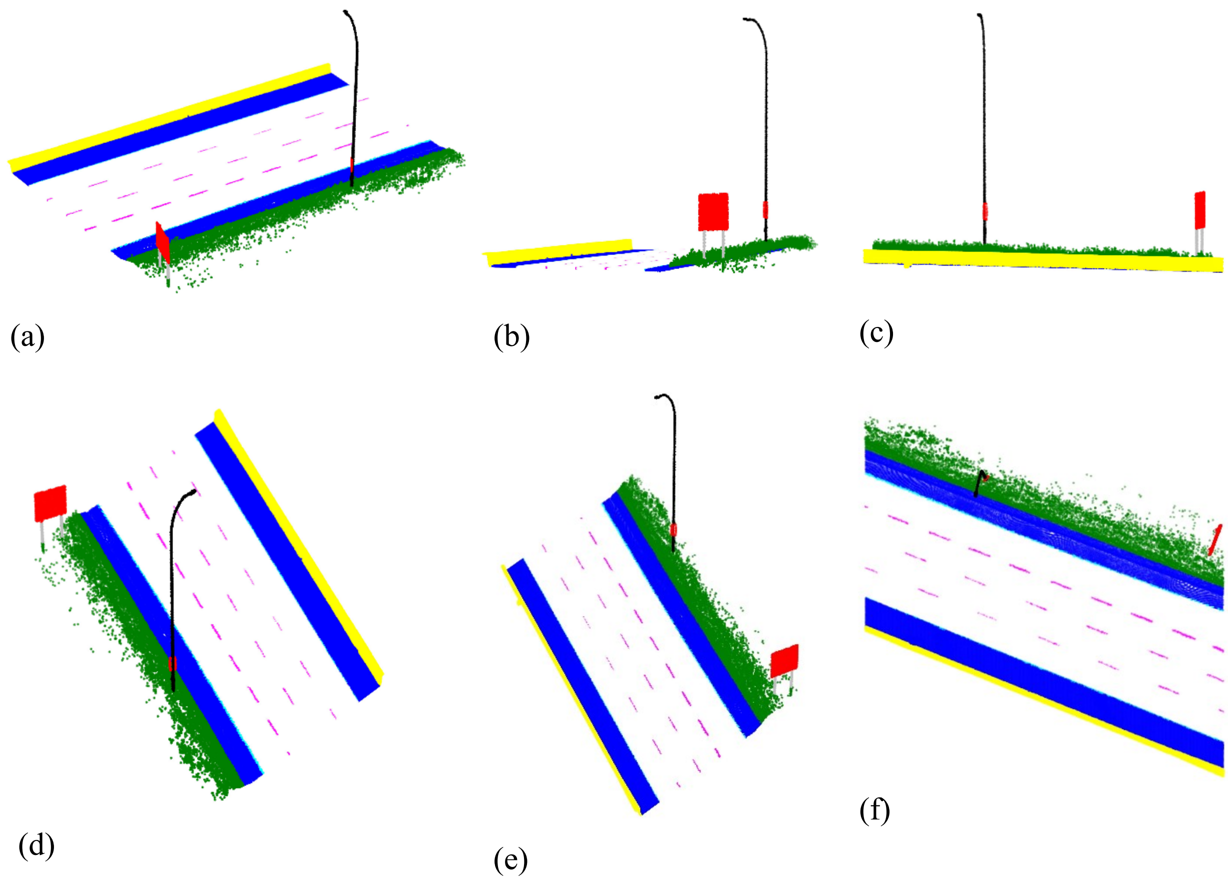
Table 10 Comparison of different views in generating descriptions (section 16) - Full model

Multi-views	Generated Description
Front, Side, Top, Isometric A	The point cloud data depicts a divided highway with either three or four lanes, distinguished by broken lines and one solid line marking the lanes. The highway features concrete barriers on both sides, indicating a separated corridor for traffic directions. Along the side of the road, there is a light pole and a stretch of vegetation, providing a clear demarcation between the roadway and the non-motoring area. This scenario highlights the infrastructure's design for organized traffic flow and safety measures with physical separations and illumination.
Top, Isometric A, Isometric B, Isometric C	The point cloud data reveals a highway environment with three lanes, as indicated by four broken lines and one solid line. A concrete barrier divides the lanes from what appears to be a shoulder area. Vegetation is present adjacent to the shoulder on one side of the highway. Additionally, a light pole is visible near the vegetation, providing illumination to the area. The scene suggests a well-structured highway with clear demarcations and safety elements in place.
Front, Side, Top, Isometric A, Isometric B, Isometric C	The point cloud data illustrates a divided highway scene with a total of five lanes, delineated by multiple broken lines and solid lines. The highway includes a concrete barrier along one edge, ensuring clear separation from the adjacent area. Vegetation can be seen along the divided portion. Additionally, a light pole is clearly present, providing illumination for the roadway. This setup facilitates clear guidance and ensures safety for autonomous navigation by defining proper lane boundaries and emphasizing the divided nature of the highway.

From the Table above, all generated outputs effectively describe the scene, albeit with varying word choices. A critical factor in evaluating these descriptions is their ability to quantify features, thereby extracting more detailed information.

The first description exhibits uncertainty, questioning whether the road contains three or four lanes. This ambiguity reduces its utility for precise feature extraction. The second description incorrectly states the presence of three lanes. However, it accurately identifies four broken lines, demonstrating partial accuracy in feature quantification. The confusion likely stems from the proximity of solid lines to the shoulder, which appears to have misled the GPT-4o model in its lane count. The final description is the most accurate, correctly identifying five lanes in total. Although it does not separately quantify broken and solid lines, its overall lane count is correct. This accuracy in total lane identification makes it the most valuable description for feature extraction purposes despite the lack of granular line-type differentiation.

While analyzing a single section provides valuable insights, it is insufficient to draw definitive conclusions about the model's performance. Additional sections are presented and analyzed below to establish a more robust and representative assessment.



(a) (b) (c) (d) (e) (f)

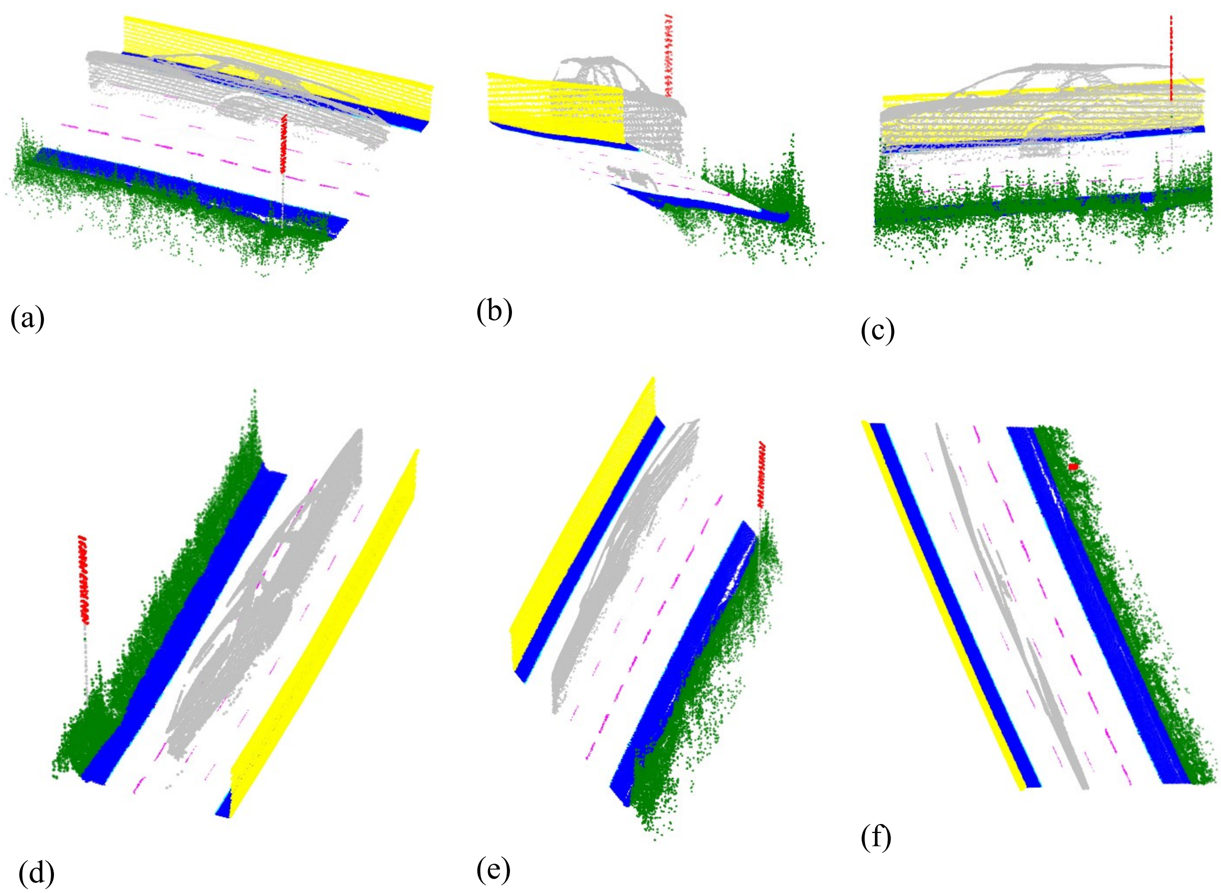
Figure 17 Ground truth of a sample (section 13- Full model) multi-view images (a) Isometric View A (b) Front View (c) Side View (d) Isometric View B (e) Isometric View C (f) Top View

Table 11 Comparison of different views in generating descriptions (Section 13) - Full model

Multi-views	Generated Description
Front, Side, Top, Isometric A	The scene depicts a multi-lane highway consisting of three lanes. The lanes are delineated by two broken lines and are flanked by a solid shoulder line

	on the right side. The highway is divided from opposite traffic by concrete barriers positioned on both sides. There is vegetation on the right side of the highway, along with a light pole and two traffic signs. The signage and markings offer essential guidance for drivers navigating this section of the road.
Top, Isometric A, Isometric B, Isometric C	The point cloud data depicts a highway with clearly marked lanes, including broken lines marking the lanes. There is a substantial shoulder on one side and vegetation alongside the road. The highway is bordered by vegetation and infrastructure items. Notable features include a light pole and traffic signs situated along the roadside. The configuration suggests a safe and organized environment designed to facilitate smooth vehicular movement.
Front, Side, Top, Isometric A, Isometric B, Isometric C	The scene depicts a multi-lane divided highway featuring several broken-line lane markings. The highway is segmented by concrete barriers separating the lanes from a shoulder area. Vegetation lines the side of the shoulder, and the road is supported by multiple traffic signs. Additionally, a light pole is present on the roadside near the shoulder, contributing to the overall infrastructure and navigational aids visible in the scene.

This section demonstrates the model's varying performance in quantifying and describing scene elements. In the first description, the model accurately identified the number of traffic signs and light poles, showcasing its ability to quantify certain features correctly. However, it incorrectly stated the highway had three lanes instead of four, indicating some limitations in quantifying lane markings. While unable to provide precise quantification of features, the second and third descriptions offered overall descriptions that were sensible and readily comprehensible to humans. This variation in performance makes it challenging to determine which description is superior definitively. The model's ability to provide useful information even when exact quantification is impossible highlights its flexibility.



(a) (b) (c) (d) (e) (f)

Figure 18 Ground truth of a sample (section 18 - Full model) multi-view images (a) Isometric View A (b) Front View (c) Side View (d) Isometric View B (e) Isometric View C (f) Top View

Table 12 Comparison of different views in generating descriptions (Section 18) - Full model

Multi-views	Generated Description
Front, Side, Top, Isometric A	The point cloud data captures a section of a divided highway with a light shoulder and multiple features. The highway consists of three lanes discerned by two broken lines and a solid line. A yellow concrete barrier separates the traffic flow on opposite sides of the highway. Vegetation lines one side of the highway, which is bounded by a shoulder area in blue. A red traffic sign is visible along the side of the highway, providing guidance or information to drivers. Each lane is clearly demarcated to facilitate the autonomous vehicle's navigation and lane-keeping abilities.
Top, Isometric A, Isometric B, Isometric C	The scene depicts a divided multi-lane highway with a concrete barrier separating two directions of traffic. The highway has three lanes delineated by broken lines, indicating the presence of multiple lanes for traffic flow.

	A shoulder runs along the right side of the highway, with vegetation situated by the shoulder. Additionally, a traffic sign is present along the side of the road, presumably for providing essential information to drivers. The top and multiple isometric views provide clarity on the layout and different sections of road markings, shoulders, and boundaries.
Front, Side, Top, Isometric A, Isometric B, Isometric C	The scene captures a section of a multi-lane highway. This highway consists of four lanes, indicated by three broken lines and two solid lines, accompanied by a solid lane marking on the outer edges. A concrete barrier is present on one side, making it a divided highway, ensuring separation from the opposing traffic. On the other side, there is a clear shoulder followed by a stretch of vegetation. A traffic sign is evident by the roadside. Additionally, a vehicle is detected with a detailed depiction from multiple viewing angles. The presence of clear lane markings, a separated highway, and a vehicle suggests an organized traffic environment well-suited for autonomous navigation.

This scene consists of a vehicle on the road which may affect model performance. However, in the six-view trial, GPT-4o could still distinguish four lanes with three broken lines and two solid lines, along with other key infrastructure elements.

Overall, the model performs best when six views are input, demonstrating a high probability of correctly quantifying the number of features. This suggests that multiple viewpoints give the model more comprehensive information, leading to more accurate results. Interestingly, in cases where the model cannot confidently confirm the exact number of features, it resorts to broader descriptions instead of producing an incorrect output. This approach makes sense allowing the model to convey useful information even when precise quantification is challenging. Overall, these findings indicate that the model effectively utilizes multiple perspectives to enhance its understanding and description of objects or scenes, with six views seemingly offering an optimal balance of information for accurate feature quantification.

It is important to note that GPT-4o, a large language model (LLM), exhibits some stochastic behaviour in its outputs. The generated descriptions may vary slightly even when presented with identical input images. This variability is an inherent characteristic of the model, designed to handle diverse scenarios while maintaining natural language fluency. Consequently, the output descriptions possess a certain level of randomness, which should be considered when interpreting this comparative study's results.

Reduced model scenarios (8 classes):

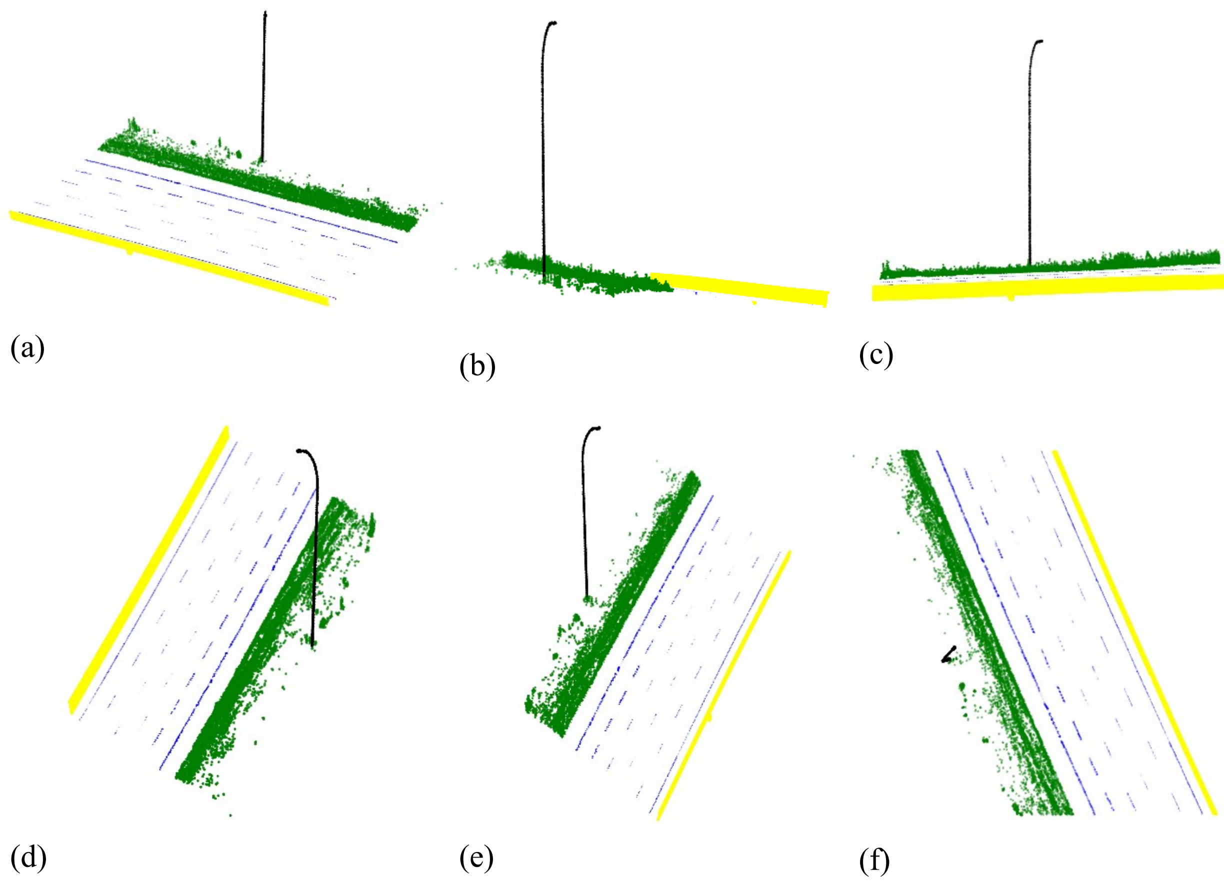


Figure 19 Ground truth of a sample (section 16 - Reduced model) multi-view images (a) Isometric View A (b) Front View (c) Side View (d) Isometric View B (e) Isometric View C (f) Top View

Table 13 Multi-views generated descriptions (Section 16) - Reduced model

Multi-views	Generated Description
Front, Side, Top, Isometric A, Isometric B, Isometric C	The point cloud data depicts a highway scene encompassing a divided highway with a total of five lanes, demarcated by four broken lines and two solid lines. A concrete barrier is present, indicating division from opposing traffic. Vegetation is visible along the side of the road, and there is a light pole prominently located near the vegetation. The arrangement suggests a clear and organized highway structure suitable for vehicular navigation.

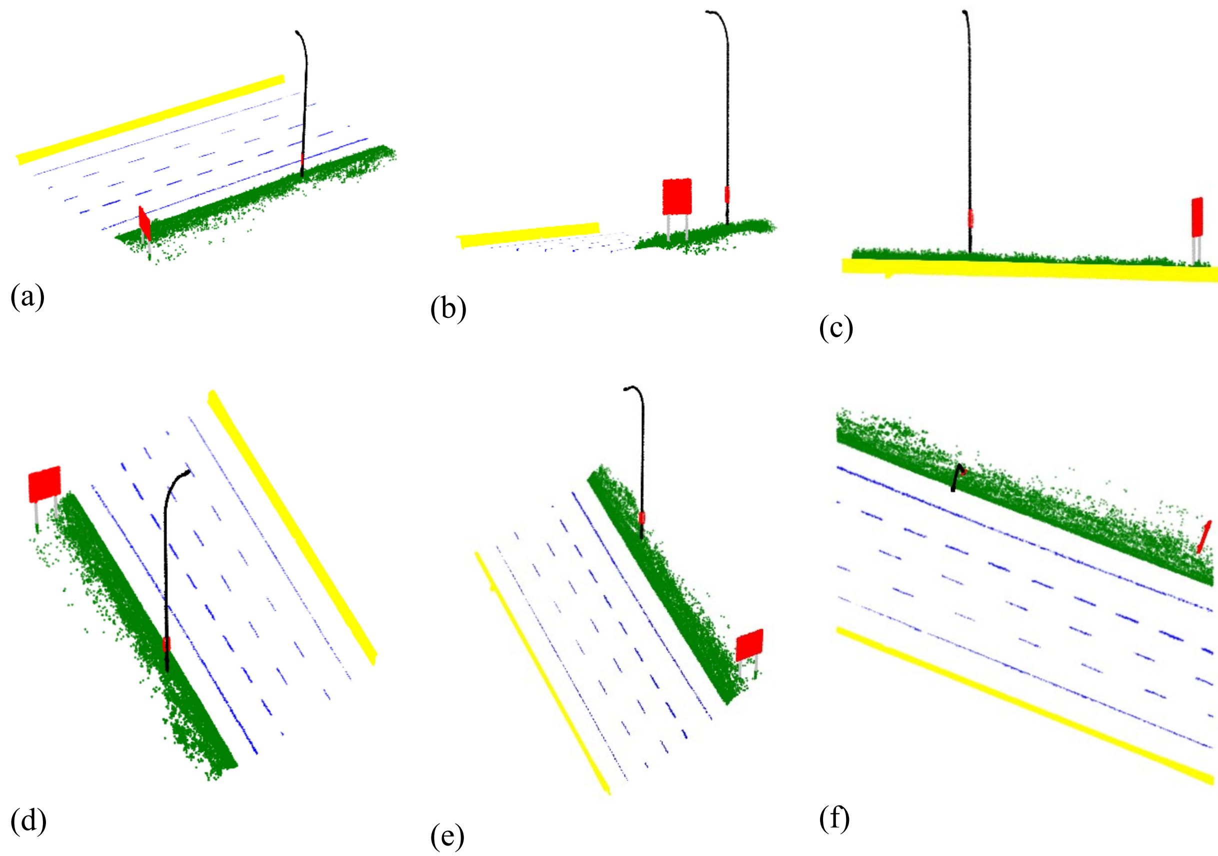


Figure 20 Ground truth of a sample (section 13 - Reduced model) multi-view images (a) Isometric View A (b) Front View (c) Side View (d) Isometric View B (e) Isometric View C (f) Top View

Table 14 Multi-views generated descriptions (Section 13) - Reduced model

Multi-views	Generated Description
Front, Side, Top, Isometric A, Isometric B, Isometric C	The point cloud data depicts a multi-lane highway with a total of three lanes, marked by broken and solid lines. The highway is bordered by a concrete barrier on one side, indicating it is a divided highway. There is also a light pole and a traffic sign located on the side of the road with vegetation present in the area. The overall scene suggests a well-defined roadway suitable for autonomous driving with clear lane demarcations, a physical divider, and navigational aids such as the traffic signs and light pole.

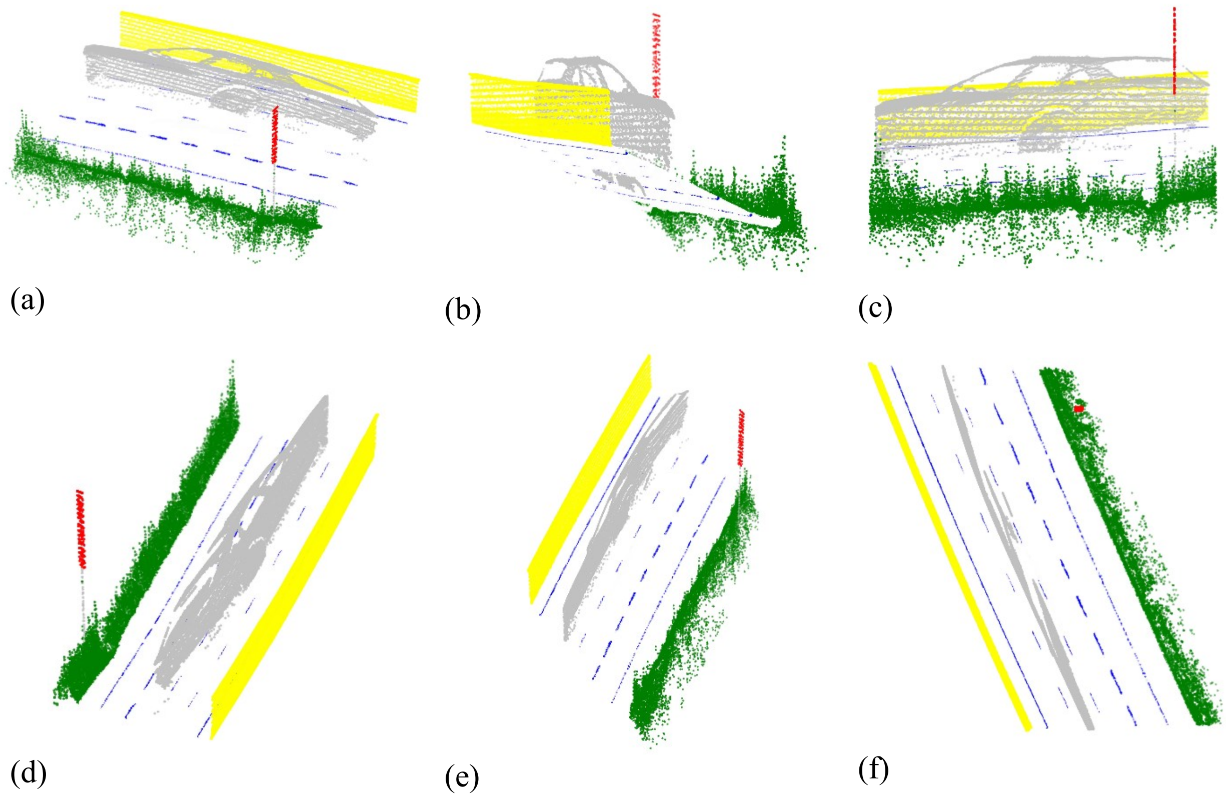


Figure 21 Ground truth of a sample (section 18 - Reduced model) multi-view images (a) Isometric View A (b) Front View (c) Side View (d) Isometric View B (e) Isometric View C (f) Top View

Table 15 Multi-views generated descriptions (Section 18) - Reduced model

Multi-views	Generated Description
Front, Side, Top, Isometric A, Isometric B, Isometric C	The point cloud data depicts a divided highway with a total of four lanes, featuring two sections with broken lines and two sections with solid lines. A concrete barrier separates the opposing traffic flows. Adjacent to the highway on the right side is a stretch of vegetation, and a traffic signpost is visible near this area. On the left side, there seems to be another lane or shoulder lane. Additionally, a vehicle is present in the scene, traveling parallel to the highlighted traffic lanes.

Sections 16 and 18 were described correctly. However, Section 13 contains some mistakes in describing the total number of lanes. Even though the six-view approach appears to be the best input method, there are still some errors in quantifying certain features. This may be due to the inherent randomness of GPT-4o or any Large Language Model (LLM).

Scene description using predicted files:

The purpose of using both ground truth and predicted files in the model is to evaluate the performance when noise or misclassification is involved. This approach aims to determine whether such factors will affect the overall description generated by the model. Additionally, if the predicted files perform well, the intention is to automate the entire process.

Full model scenarios (12 classes):

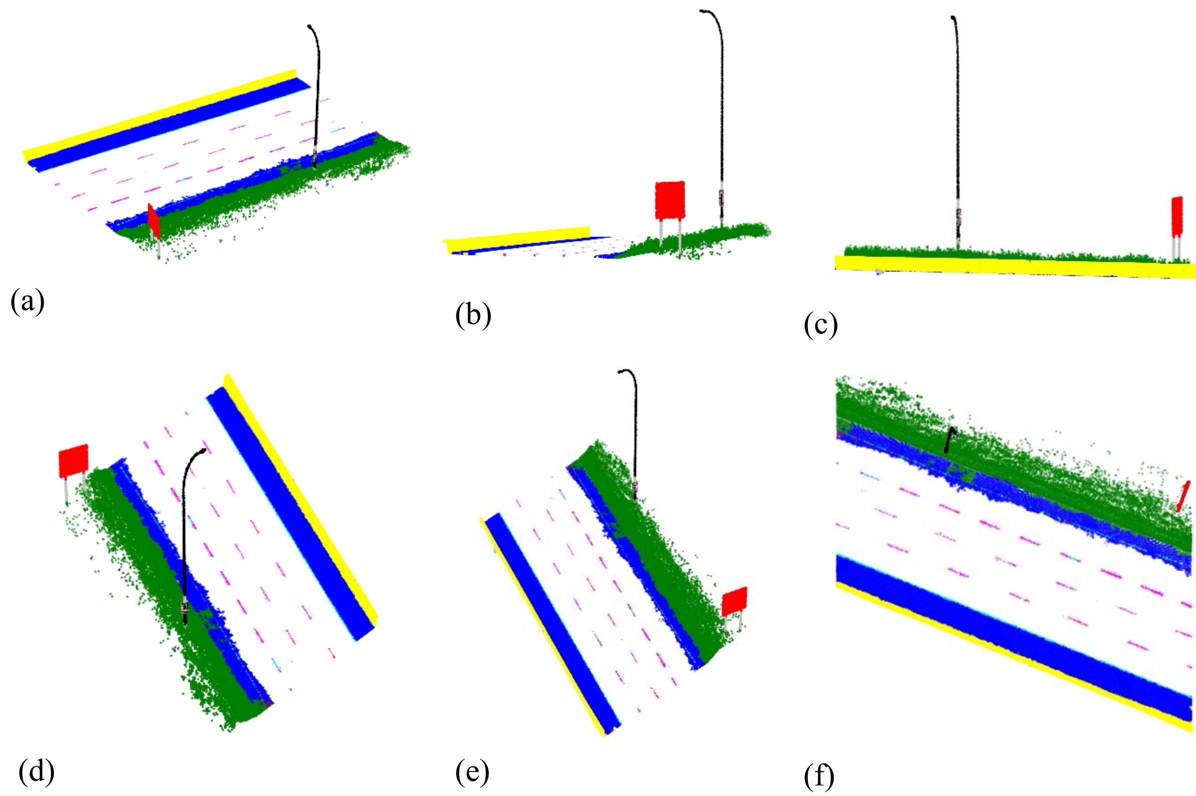


Figure 22 Prediction of a sample (section 13- Full model) multi-view images (a) Isometric View A (b) Front View (c) Side View (d) Isometric View B (e) Isometric View C (f) Top View

Table 16 Multi-views generated descriptions (Section 13) - Full model

Multi-views	Generated Description
Front, Side, Top, Isometric A, Isometric B, Isometric C	The point cloud data depicts a multi-lane divided highway. The road consists of four lanes separated by three broken lines. There is a wide shoulder on one side of the road, marked by a single solid line. Alongside the road, there is vegetation and a light pole. A traffic sign is also visible at the side of the road, providing guidance for drivers. The clear demarcation

	of lanes, shoulder, and road objects suggests a well-structured roadway environment suitable for autonomous driving.

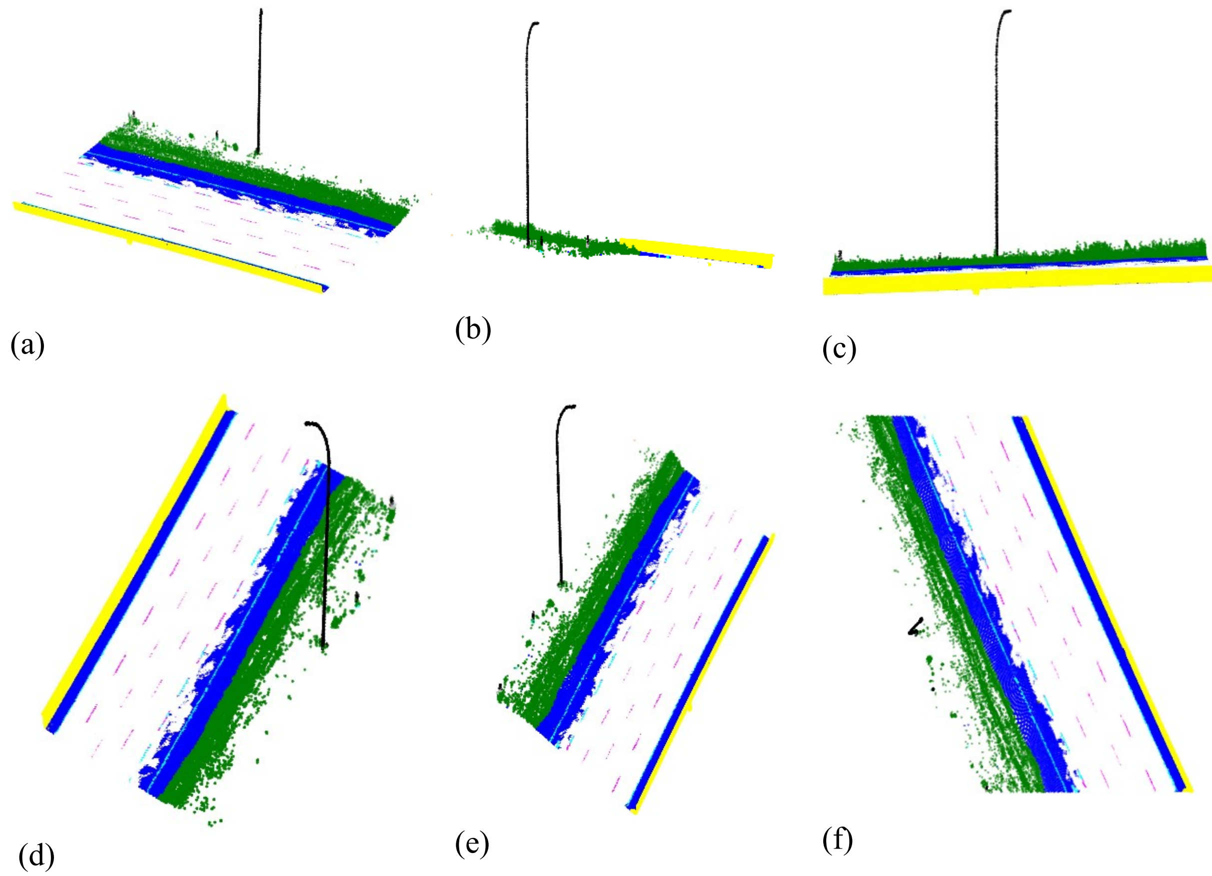
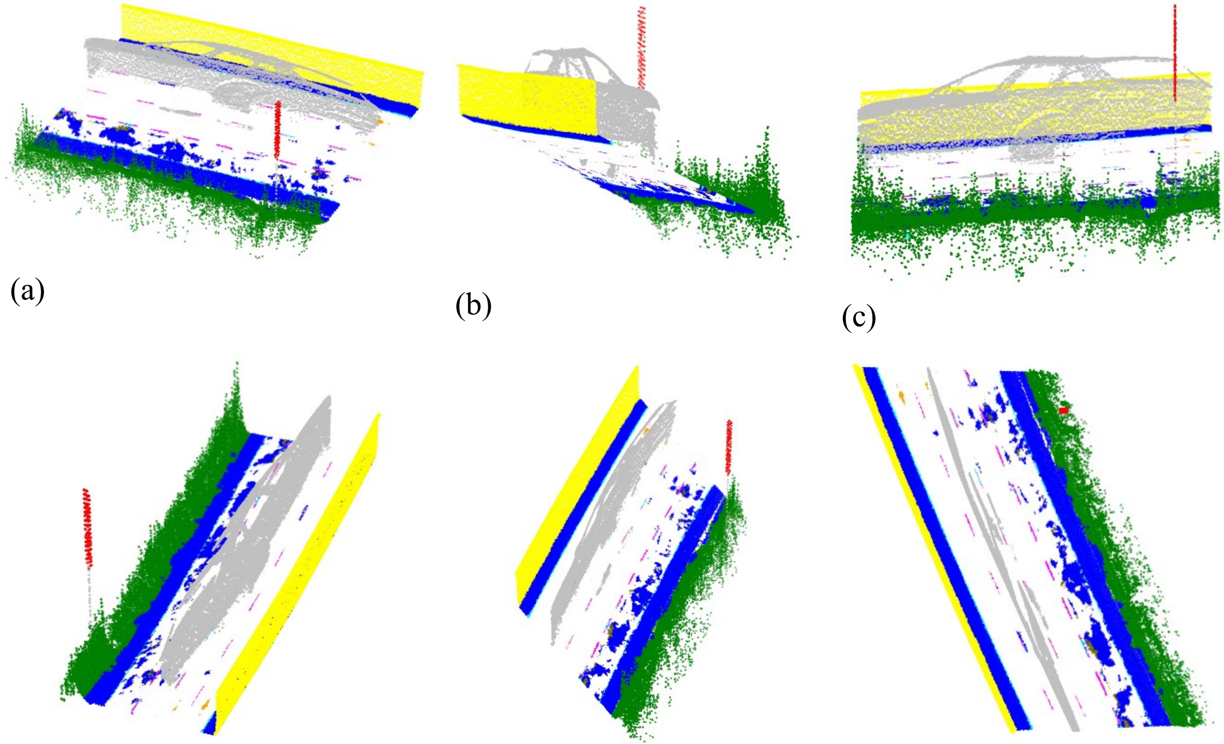


Figure 23 Prediction of a sample (section 16 - Full model) multi-view images (a) Isometric View A (b) Front View (c) Side View (d) Isometric View B (e) Isometric View C (f) Top View

Table 17 Multi-views generated descriptions (Section 16) - Full model

Multi-views	Generated Description
Front, Side, Top, Isometric A, Isometric B, Isometric C	The scene depicts a section of a multi-lane divided highway comprising three lanes marked by broken lines. The highway is bordered by a shoulder on one side and vegetation on the other. A concrete barrier separates the lanes in opposing traffic flows, ensuring safety. Additionally, a light pole (street light) is visible on the side of the road, which would provide illumination during nighttime or low-visibility conditions. This setup indicates a standard highway structure with appropriate safety features and road markings for lane guidance.



(a) (b) (c) (d) (e) (f)

Figure 24 Prediction of a sample (section 18 - Full model) multi-view images (a) Isometric View A (b) Front View (c) Side View (d) Isometric View B (e) Isometric View C (f) Top View

Table 18 Multi-views generated descriptions (Section 18) - Full model

Multi-views	Generated Description
Front, Side, Top, Isometric A, Isometric B, Isometric C	The point cloud data represents a segment of a highway with three lanes, distinguished by solid and broken lane markings. A concrete barrier separates the lanes from the shoulder on one side, indicating this is a divided highway. The shoulder includes blue markings and is lined with vegetation on one side. Additionally, there is a traffic sign alongside the road, and a vehicle is present in the scene, situated near the concrete barrier. There are also scattered pieces of clutter throughout the scene.

As some misclassification came into play, the model's performance was slightly affected. Sections 16 and 18 have misclassifications between solid and broken lines, making the number of lanes challenging to distinguish. However, the model showed slightly better performance when using the ground truth files.

Reduced model scenarios (8 classes):

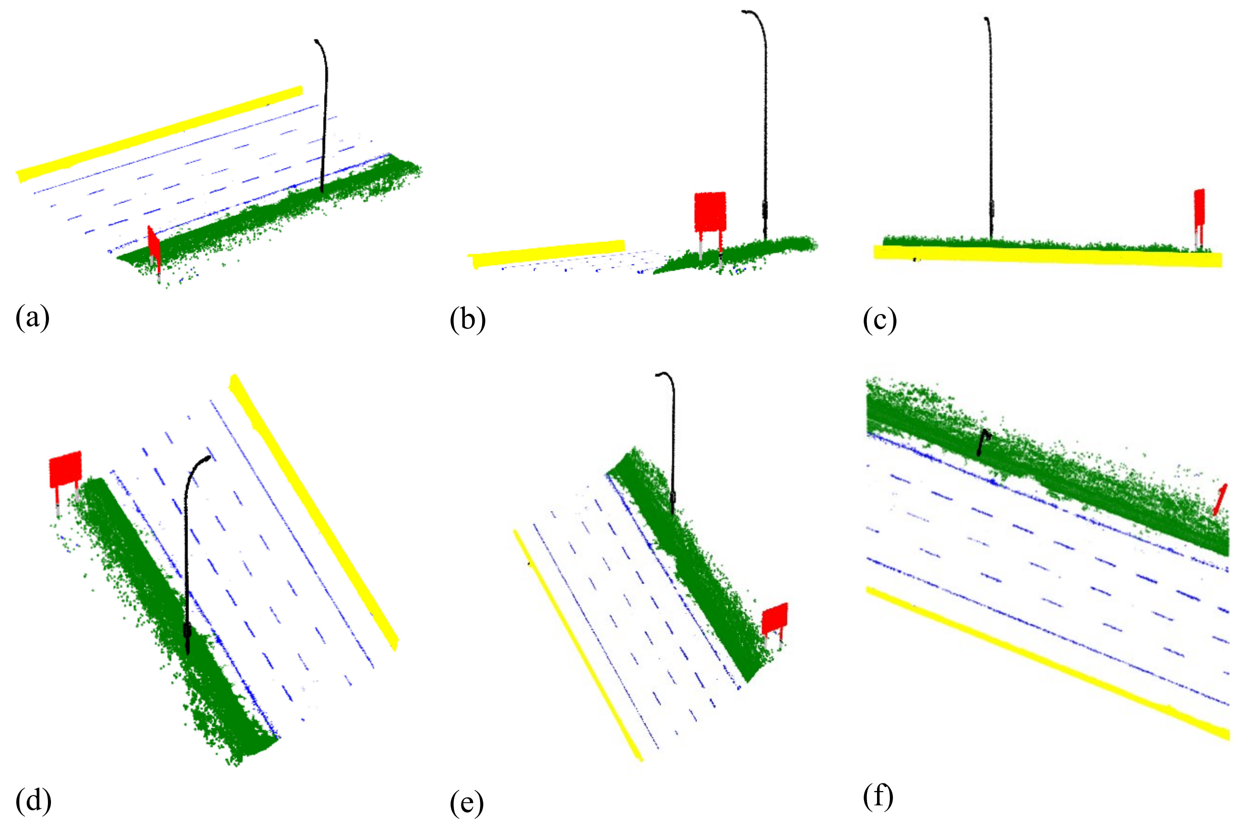


Figure 25 Prediction of a sample (section 13 - Reduced model) multi-view images (a) Isometric View A (b) Front View (c) Side View (d) Isometric View B (e) Isometric View C (f) Top View

Table 19 Multi-views generated descriptions (Section 13) - Reduced model

Multi-views	Generated Description
Front, Side, Top, Isometric A, Isometric B, Isometric C	The point cloud data of the scene depicts a three-lane road with a single lane marking on each side and two lane markings separating the lanes. There is a concrete barrier on the left side of the road, indicating a divided highway. The road is lined with vegetation to the right, interspersed with a light pole and a traffic sign. The breakdown lane adjacent to the concrete barrier offers additional space. The lanes are clearly marked with dashed lines, ensuring proper lane guidance.

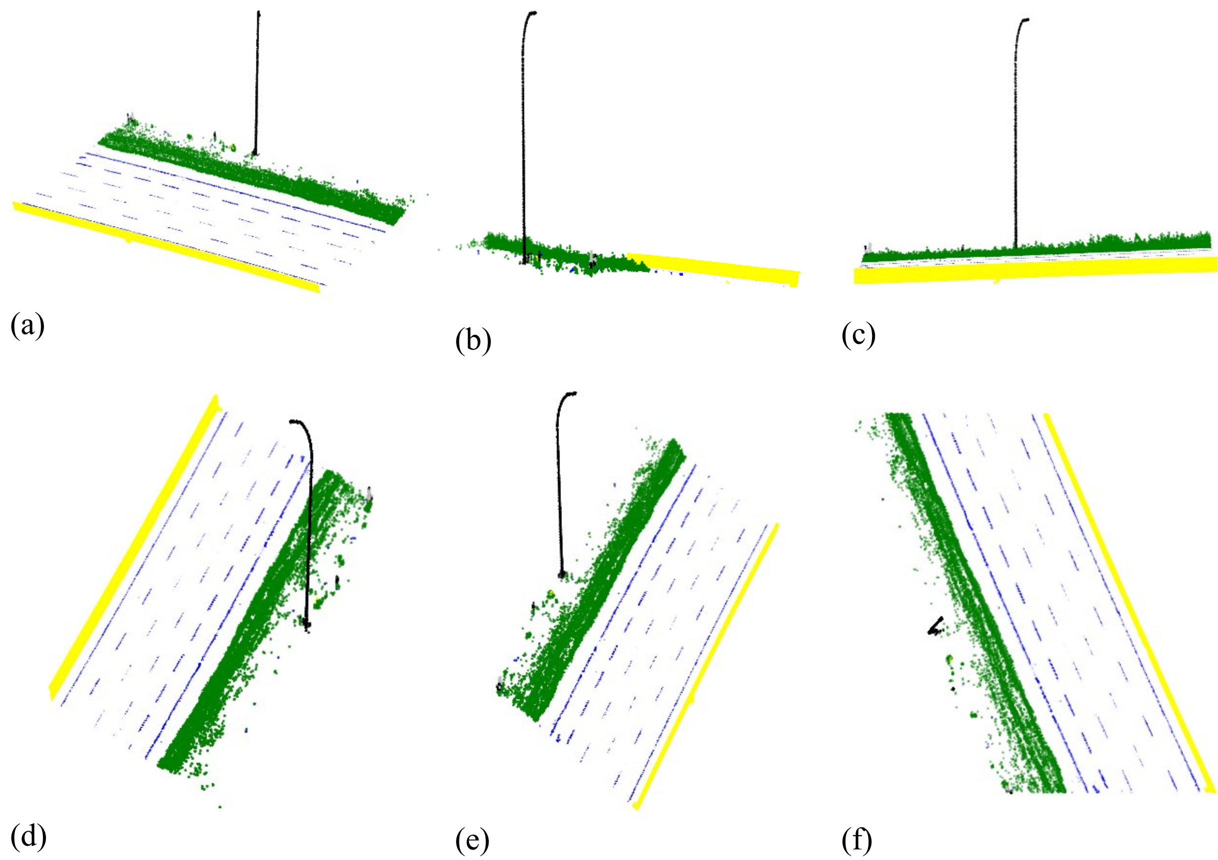


Figure 26 Prediction of a sample (section 16 - Reduced model) multi-view images (a) Isometric View A (b) Front View (c) Side View (d) Isometric View B (e) Isometric View C (f) Top View

Table 20 Multi-views generated descriptions (Section 16) - Reduced model

Multi-views	Generated Description
Front, Side, Top, Isometric A, Isometric B, Isometric C	The scene depicts a stretch of a multi-lane highway with several lanes marked by broken lines, indicating multiple driving lanes. The highway is divided by concrete barriers, ensuring separation between the opposing traffic flows. There is visible vegetation along the side of the road. A light pole is also present on the roadside, indicating proper illumination for nighttime driving. The presence of these elements signifies that the environment is well-structured for autonomous driving, providing clear demarcations and necessary infrastructure for safe navigation.

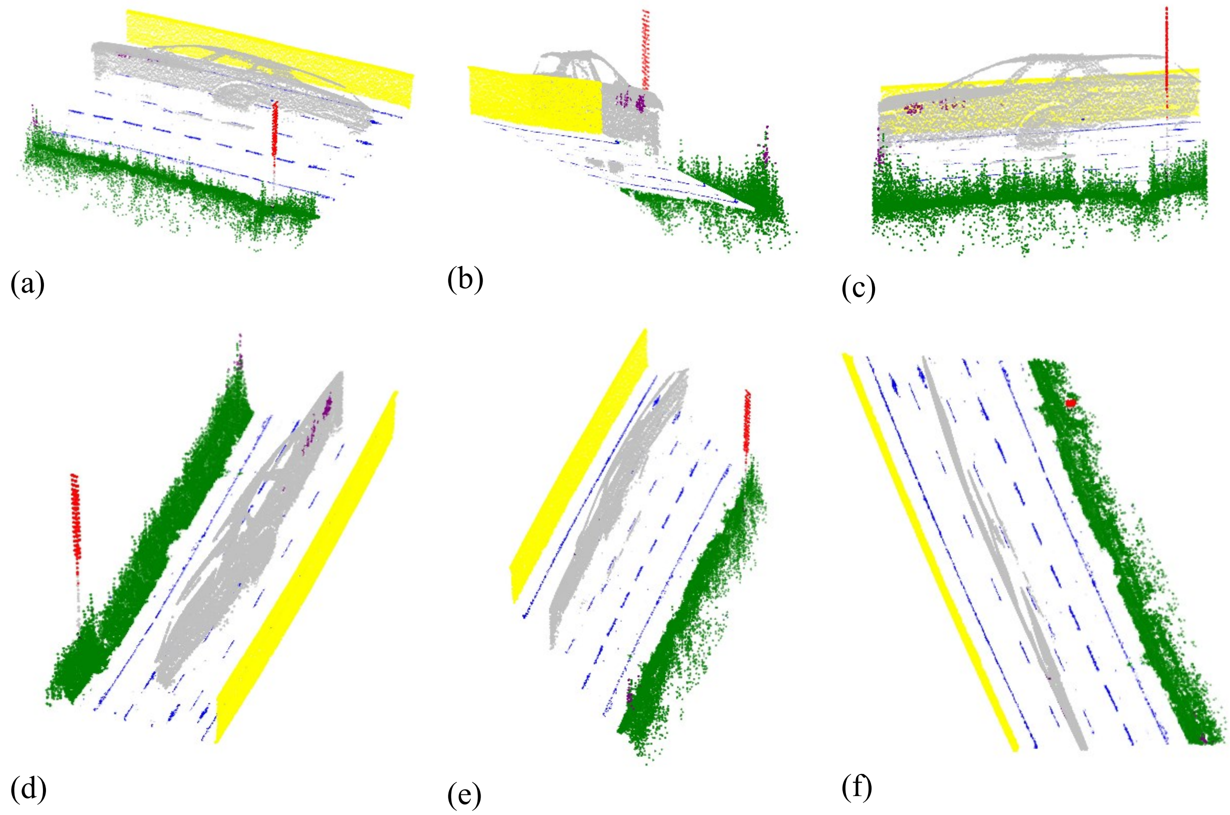


Figure 27 Prediction of a sample (section 18 - Reduced model) multi-view images (a) Isometric View A (b) Front View (c) Side View (d) Isometric View B (e) Isometric View C (f) Top View

Table 21 Multi-views generated descriptions (Section 18) - Reduced model

Multi-views	Generated Description
Front, Side, Top, Isometric A, Isometric B, Isometric C	The point cloud data shows a scene of a multi-lane highway consisting of four lanes. The lanes are demarcated by a mix of solid and broken lines. There is a concrete barrier on one side, indicating a divided highway, while the other side is bordered by vegetation. In addition, an overhead traffic sign is present to provide guidance. The highway shows the presence of another vehicle and some miscellaneous objects, potentially roadside clutter. The information suggests the autonomous vehicle's immediate environment, including crucial elements for navigation and safety, like lane markings, barriers, and potential obstacles.

In summary, the reduced model provides less information than the full model due to combining marking and pavement into two classes rather than more split classes. However, the accuracy of the predicted files did not significantly affect the GPT-4o results. All the generated descriptions are reasonable and make sense to humans. The outputs for all other files are provided in Appendix B.

Chapter 6. Conclusion, Contribution, and Future Works

6.1 Conclusions

LiDAR technology holds many potential applications that have yet to be fully developed, and the use of Natural Language Processing (NLP) in the transportation sector is even less explored. This research addresses this gap by utilizing existing highway LiDAR data, without RGB and image inputs, to extract highway infrastructure elements using two advanced Point Transformer models. This approach transcends traditional single or binary segmentation tasks and explores the conversion of segmented point cloud data into textual descriptions, laying the groundwork for future applications.

The study conducted semantic segmentation tasks on a 2.5km highway segment in Alberta, Canada. The Point Transformer v2 achieved an overall Mean Intersection over Union (IoU) score of 78.29% and a Mean F1 score of 86.48%. The Transformer-based point classification achieves an overall Mean IoU score of 86.03%, a Mean F1 score of 92.21%, and demonstrated strong visualization capabilities. These results underscore the efficacy of Transformer structures in processing point cloud data.

The natural language processing tasks integrated multi-view imagery derived from semantic segmentation outputs with the GPT-4o model, subject to specific constraints. When utilizing ground truth data, the model effectively quantified the number of lanes and other infrastructure elements in most cases. As expected, the full class model provides more comprehensive information than the reduced class model. However, occasional inaccuracies arose due to scene complexity, such as unavoidable vehicle trajectories or indistinct road markings, as well as the inherent stochasticity of large language models. Furthermore, discrepancies between the predicted segmentation output and the ground truth led to descriptions with diminished accuracy compared to those generated from ground truth data.

This research not only advances state-of-the-art point cloud processing and interpretation but also provides a foundation for future work in various domains. Potential applications include autonomous vehicle navigation, infrastructure maintenance and asset management, infrastructure planning and design, environmental monitoring, accessibility assessment, more accurate accident scene reconstruction, and climate change impact assessment.

The interdisciplinary nature of this work underscores its potential for wide-ranging impact and further research opportunities. This study contributes to developing more natural and effective interfaces between humans and 3D-aware AI systems by leveraging advanced machine learning models and extensive point cloud datasets. This lays the groundwork for innovative applications across multiple fields, emphasizing this research's broad applicability and significance.

6.2 Contributions

This research presents a comprehensive pipeline for processing raw point cloud data into human-readable descriptions, leveraging two Transformer-based architectures and the GPT-4o model. The contributions of this work are as follows:

- Enhancement of Annotation Tools: improved upon the SUSTech POINT system, developing a more user-friendly online annotation format. This enhancement facilitates more efficient and accurate labelling of point cloud data, potentially accelerating future research in this domain.
- Novel Application of Transformer-based Models for Highway Infrastructure Segmentation: Proposes and implements two Transformer-based semantic segmentation models tailored explicitly for extracting multiple highway infrastructure elements in Alberta's rural highway environments. This application demonstrates the adaptability of Transformer architectures to complex 3D point cloud data in real-world scenarios.
- Integration of large language model (LLM) for Automated Scene Understanding: Introduce a framework of using multi-view images and GPT-4o model to convert segmented point cloud data into human-readable text descriptions. This novel integration bridges the gap between raw sensor data and easily interpretable scene understanding, opening new possibilities for applications in autonomous driving, infrastructure management, and beyond.
- End-to-End Pipeline for LiDAR Data Interpretation: By combining these elements, we present a complete process flow from raw point cloud data to semantic segmentation and finally to natural language descriptions. This end-to-end approach represents a significant step in automating complex 3D environmental data interpretation.
- Expansion of LiDAR Applications: Our work demonstrates the potential for extending LiDAR applications across various fields. We facilitate broader adoption and utilization of LiDAR technology in research and practical applications by providing a framework for converting point cloud data into human-readable formats.

6.3 Research Limitations & Future Works

6.3A Semantic segmentation limitation and future works:

While the method demonstrates promising accuracy and visualization capabilities, there remains room for improvement, particularly in the accuracy of solid and broken markings in full-class (12 classes) segmentation tasks. Future work will focus on refining these aspects to further enhance the reliability and applicability of LiDAR data in infrastructure management and planning.

Several limitations of this study must be considered. Firstly, the model was trained, tested, and optimized based on 2.5 km road segments specific to the Alberta highway environment. Applying this model to other highway environments may not achieve the desired accuracy. Additionally, LiDAR sensors can only capture data from surfaces directly visible to the sensor so that occluded objects can affect segmentation accuracy.

Despite addressing class imbalances with different loss functions, the natural characteristics of the LiDAR data from Alberta highways resulted in highly imbalanced classes. The large volume of highway point cloud data presents significant computational challenges, and limited GPU memory restricts the ability to process multiple segments simultaneously. Furthermore, the limited training data could lead to overfitting, highlighting the need for an expanded dataset.

Achieving real-time semantic segmentation is challenging due to high computational demands and the need to process continuous data streams rapidly. The lack of colour information in LiDAR data makes segmentation more difficult. Clear markings are more likely to be correctly identified than faded ones, indicating a limitation of the current approach.

Transformer models, while powerful for semantic segmentation tasks, face challenges due to their substantial computational resource requirements compared to traditional CNNs. This can limit their deployment in real-time or resource-constrained environments.

Future work in semantic segmentation should focus on three key areas. First, expanding the training dataset beyond the 2.5 km currently used for both models can enhance performance, improve generalization, and reduce the risk of overfitting. Second, while the current accuracy is strong, optimizing processing time and achieving multi-segment capabilities are critical for improving efficiency. Lastly, exploring alternative deep learning architectures beyond transformers is essential, as newer models continue to demonstrate high accuracy. However, this exploration requires careful experimentation, as the optimal solution varies depending on the specific segmentation task, available computational resources, and the desired balance between accuracy and efficiency.

6.3B Natural language scene description limitation and future works:

The scene descriptions were generated using a pre-trained GPT-4o model, which is a large language model. While powerful, this approach has several limitations. First, since the model is not open-source, it's impossible to fine-tune the parameters to best suit specific use cases. Additionally, large language models inherently produce outputs with considerable randomness. The model also exhibits instability in quantifying certain features in particular environments, with complex scenes proving particularly challenging to describe accurately. These limitations highlight the trade-offs between the model's flexibility and consistency, especially when dealing with intricate visual information. The inability to adjust the model's parameters and its variable performance across different scenarios underscores the challenges of using general-purpose language models for specialized tasks like detailed scene description and feature quantification. Future work can focus on developing a custom LLM model to address these limitations using domain-specific datasets. Additionally, exploring direct point cloud-to-text conversion methods using clustering techniques could improve scene descriptions' accuracy and consistency, bypassing the need for intermediate image generation.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., & Anadkat, S. (2023). Gpt-4 technical report. *ArXiv Preprint ArXiv:2303.08774*.

Ai, C., & Tsai, Y. J. (2016). An automated sign retroreflectivity condition evaluation methodology using mobile LIDAR and computer vision. *Transportation Research Part C: Emerging Technologies*, *63*, 96–113.

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72.

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

Chandrasekaran, D., & Mago, V. (2021). Evolution of semantic similarity—a survey. *ACM Computing Surveys (CSUR)*, *54*(2), 1–37.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European Conference on Computer Vision (ECCV)*, 801–818.

Chen, R., Zhu, X., Chen, N., Li, W., Ma, Y., Yang, R., & Wang, W. (2023). *Zero-shot Point Cloud Segmentation by Transferring Geometric Primitives*.

Chen, Y., Li, W., & Gool, L. Van. (2018). ROAD: Reality Oriented Adaptation for Semantic Segmentation of Urban Scenes. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 7892–7901. https://doi.org/10.1109/CVPR.2018.00823

CloudCompare. (2023). Https://Www.Cloudcompare.Org/Doc/Wiki/Index.Php/Main_Page.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3213–3223.

Daiber, J., Jakob, M., Hokamp, C., & Mendes, P. N. (2013). Improving efficiency and accuracy in multilingual entity extraction. *Proceedings of the 9th International Conference on Semantic Systems*, 121–124.

Gao, B., Hu, S., Zhao, X., & Zhao, H. (2021). Fine-Grained Off-Road Semantic Segmentation and Mapping via Contrastive Learning. *IEEE International Conference on Intelligent Robots and Systems*, 5950–5957. https://doi.org/10.1109/IROS51168.2021.9636033

Gao, J., Chen, Y., Junior, J. M., Wang, C., & Li, J. (2020). Rapid extraction of urban road guardrails from mobile LiDAR point clouds. *IEEE Transactions on Intelligent Transportation Systems*, *23*(2), 1572–1577.

Gargoum, S. A., El-Basyouny, K., Froese, K., & Gadowski, A. (2018). A fully automated approach to extract and assess road cross sections from mobile LiDAR data. *IEEE Transactions on Intelligent Transportation Systems*, *19*(11), 3507–3516. https://doi.org/10.1109/TITS.2017.2784623

Gargoum, S. A., El-Basyouny, K., & Sabbagh, J. (2018). Assessing Stopping and Passing Sight Distance on Highways Using Mobile LiDAR Data. *Journal of Computing in Civil Engineering*, *32*(4), 1–13. https://doi.org/10.1061/(asce)cp.1943-5487.0000753

Gargoum, S. A., El-Basyouny, K., Shalkamy, A., & Gouda, M. (2018). Feasibility of extracting highway vertical profiles from LiDAR data. *Canadian Journal of Civil Engineering*, *45*(5), 418–421. https://doi.org/10.1139/cjce-2017-0620

Gargoum, S., & El-Basyouny, K. (2017). Automated extraction of road features using LiDAR data: A review of LiDAR applications in transportation. *2017 4th International Conference on Transportation Information and Safety, ICTIS 2017 - Proceedings*, 563–574. https://doi.org/10.1109/ICTIS.2017.8047822

Gargoum, S., & El-Basyouny, K. (2018). Transportation Infrastructure Asset Management using LiDAR Remote Sensing Technology. *Int. J. Trend Res. Dev*, *143*, 143–152.

Gargoum, S., & El-Basyouny, K. (2019). Effects of LiDAR Point Density on Extraction of Traffic Signs: A Sensitivity Study. *Transportation Research Record*, *2673*(1), 41–51. https://doi.org/10.1177/0361198118822295

Gargoum, S., El-Basyouny, K., Sabbagh, J., & Froese, K. (2017a). Automated highway sign extraction using lidar data. *Transportation Research Record*, *2643*(1), 1–8.

Gargoum, S., El-Basyouny, K., Sabbagh, J., & Froese, K. (2017b). Automated highway sign extraction using lidar data. *Transportation Research Record*, *2643*, 1–8. https://doi.org/10.3141/2643-01

Golparvar-Fard, M., Balali, V., & de la Garza, J. M. (2015). Segmentation and Recognition of Highway Assets Using Image-Based 3D Point Clouds and Semantic Texton Forests. *Journal of Computing in Civil Engineering*, *29*(1), 1–14. https://doi.org/10.1061/(asce)cp.1943-5487.0000283

Gouda, M., Epp, A., Tilroe, R., & El-Basyouny, K. (2022). Traffic sign extraction using deep hierarchical feature learning and mobile light detection and ranging (LiDAR) data on rural highways. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, *0*(0), 1–22. https://doi.org/10.1080/15472450.2022.2074792

Gouda, M., Shalkamy, A., Li, X., & El-Basyouny, K. (2022). Fully Automated Algorithm for Light Pole Detection and Mapping in Rural Highway Environment Using Mobile Light Detection and Ranging Point Clouds. *Transportation Research Record*, *2676*(7), 617–629. https://doi.org/10.1177/03611981221082531

Guan, H., Li, J., Yu, Y., Wang, C., Chapman, M., & Yang, B. (2014). Using mobile laser scanning data for automated extraction of road markings. *ISPRS Journal of Photogrammetry and Remote Sensing*, *87*, 93–107.

Jalayer, M., Zhou, H., Gong, J., Hu, S., & Grinter, M. (2014). A comprehensive assessment of highway inventory data collection methods. *Journal of the Transportation Research Forum*, *53*(2), 73–92.

Javanmardi, M., Song, Z., & Qi, X. (2019). Automated traffic sign and light pole detection in mobile LiDAR scanning data. *IET Intelligent Transport Systems*, *13*(5), 803–815.

Jiang, L., Zhao, H., Liu, S., Shen, X., Fu, C.-W., & Jia, J. (2019). Hierarchical point-edge interaction network for point cloud semantic segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10433–10441.

Kim, B., Yim, J., & Kim, J. (2019). Highway driving dataset for semantic video segmentation. *British Machine Vision Conference 2018, BMVC 2018*.

Kumar, P., McElhinney, C. P., Lewis, P., & McCarthy, T. (2014). Automated road markings extraction from mobile laser scanning data. *International Journal of Applied Earth Observation and Geoinformation*, *32*, 125–137.

Lari, Z., Habib, A., & Kwak, E. (2012). An adaptive approach for segmentation of 3D laser point cloud. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *38*, 103–108.

Lato, M. J., Diederichs, M. S., Hutchinson, D. J., & Harrap, R. (2012). Evaluating roadside rockmasses for rockfall hazards using LiDAR data: Optimizing data collection and processing protocols. *Natural Hazards*, *60*(3), 831–864. https://doi.org/10.1007/s11069-011-9872-y

Li, E., Wang, S., Li, C., Li, D., Wu, X., & Hao, Q. (2020). SUSTech POINTS: A Portable 3D Point Cloud Interactive Annotation Platform System. *IEEE Intelligent Vehicles*

Symposium, Proceedings, *April 2022*, 1108–1115. https://doi.org/10.1109/IV47402.2020.9304562

Lim, S., El-Basyouny, K., & Yang, Y. H. (2024). PU-Ray: Domain-Independent Point Cloud Upsampling via Ray Marching on Neural Implicit Surface. *IEEE Transactions on Intelligent Transportation Systems*.

Manikandan, N. S., & Ganesan, K. (2019). *Deep Learning Based Automatic Video Annotation Tool for Self-Driving Car*.

Maturana, D., & Scherer, S. (2015). Voxnet: A 3d convolutional neural network for real-time object recognition. *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 922–928.

Michele, B., Boulch, A., Puy, G., Bucher, M., & Marlet, R. (2021). Generative Zero-Shot Learning for Semantic Segmentation of 3D Point Clouds. *Proceedings - 2021 International Conference on 3D Vision, 3DV 2021*, *2*, 992–1002. https://doi.org/10.1109/3DV53792.2021.00107

Northcutt, C., Jiang, L., & Chuang, I. (2021). Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, *70*, 1373–1411.

Ohgushi, T., Horiguchi, K., & Yamanaka, M. (2021). Road Obstacle Detection Method Based on an Autoencoder with Semantic Segmentation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12627 LNCS*, 223–238. https://doi.org/10.1007/978-3-030-69544-6_14

Oliveira, P., & Rocha, J. (2013). Semantic annotation tools survey. *Proceedings of the 2013 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2013 - 2013 IEEE Symposium Series on Computational Intelligence, SSCI 2013*, 301–307. https://doi.org/10.1109/CIDM.2013.6597251

Qi, C. R., Su, H., Mo, K., & Guibas, L. J. (2017). Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 652–660.

Qi, C. R., Su, H., Nießner, M., Dai, A., Yan, M., & Guibas, L. J. (2016). Volumetric and multi-view cnns for object classification on 3d data. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5648–5656.

Qi, Z., Fang, Y., Sun, Z., Wu, X., Wu, T., Wang, J., Lin, D., & Zhao, H. (2024). Gpt4point: A unified framework for point-language understanding and generation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26417–26427.

Rastiveis, H., Shams, A., Sarasua, W. A., & Li, J. (2020). Automated extraction of lane markings from mobile LiDAR point clouds based on fuzzy inference. *ISPRS Journal of Photogrammetry and Remote Sensing*, *160*, 149–166.

Rethage, D., Wald, J., Sturm, J., Navab, N., & Tombari, F. (2018). Fully-convolutional point networks for large-scale point clouds. *Proceedings of the European Conference on Computer Vision (ECCV)*, 596–611.

Saadany, H., & Orasan, C. (2021). BLEU, METEOR, BERTScore: evaluation of metrics performance in assessing critical translation errors in sentiment-oriented text. *ArXiv Preprint ArXiv:2109.14250*.

Shen, T., Gao, J., Kar, A., & Fidler, S. (2020). Interactive Annotation of 3D Object Geometry Using 2D Scribbles. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12362 LNCS*, 751–767. https://doi.org/10.1007/978-3-030-58520-4_44

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Veneziano, D., Souleyrette, R., & Hallmark, S. L. (2002). Evaluation of LIDAR for Highway Planning, Location and Design. *Pecora 15/Land Satellite Information IV/ISPRS Commission I/FIEOS*.

Wang, X., He, J., & Ma, L. (2019). Exploiting local and global structure for point cloud semantic segmentation with contextual point representations. *Advances in Neural Information Processing Systems*, *32*.

Williams, K., Olsen, M. J., Roe, G. V, & Glennie, C. (2013). Synthesis of transportation applications of mobile LiDAR. *Remote Sensing*, *5*(9), 4652–4692.

Wu, A., He, P., Li, X., Chen, K., Ranka, S., & Rangarajan, A. (2023). *An Efficient Semi-Automated Scheme for Infrastructure LiDAR Annotation*. 1–10.

Wu, X., Lao, Y., Jiang, L., Liu, X., & Zhao, H. (2022). Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in Neural Information Processing Systems*, *35*, 33330–33342.

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., & Xiao, J. (2015). 3d shapenets: A deep representation for volumetric shapes. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1912–1920.

Xu, R., Wang, X., Wang, T., Chen, Y., Pang, J., & Lin, D. (2023a). Pointllm: Empowering large language models to understand point clouds. *ArXiv Preprint ArXiv:2308.16911*.

Xu, R., Wang, X., Wang, T., Chen, Y., Pang, J., & Lin, D. (2023b). Pointllm: Empowering large language models to understand point clouds. *ArXiv Preprint ArXiv:2308.16911*.

Yang, B., Fang, L., Li, Q., & Li, J. (2012). Automated extraction of road markings from mobile LiDAR point clouds. *Photogrammetric Engineering & Remote Sensing*, *78*(4), 331–338.

Yang, J., Zhang, Q., Ni, B., Li, L., Liu, J., Zhou, M., & Tian, Q. (2019). Modeling point clouds with self-attention and gumbel subset sampling. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3323–3332.

Yao, B., Yang, X., & Zhu, S. C. (2007). Introduction to a large-scale general purpose ground truth database: Methodology, annotation tool and benchmarks. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *4679 LNCS*, 169–183. https://doi.org/10.1007/978-3-540-74198-5_14

Ye, X., Li, J., Huang, H., Du, L., & Zhang, X. (2018). 3d recurrent neural networks with context fusion for point cloud semantic segmentation. *Proceedings of the European Conference on Computer Vision (ECCV)*, 403–417.

Yifan, W., Wu, S., Huang, H., Cohen-Or, D., & Sorkine-Hornung, O. (2019). Patch-based progressive 3d point set upsampling. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5958–5967.

Yu, Y., Li, J., Guan, H., & Wang, C. (2015). Automated extraction of urban road facilities using mobile laser scanning data. *IEEE Transactions on Intelligent Transportation Systems*, *16*(4), 2167–2181.

Yue, Y., Gouda, M., & El-Basyouny, K. (2021). Automatic detection and mapping of highway guardrails from mobile lidar point clouds. *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2520–2523.

Zhang, F., Guan, C., Fang, J., Bai, S., Yang, R., Torr, P. H. S., & Prisacariu, V. (2020). Instance segmentation of lidar point clouds. *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 9448–9455.

Zhang, S., Wang, C., Lin, L., Wen, C., Yang, C., Zhang, Z., & Li, J. (2019). Automated visual recognizability evaluation of traffic sign based on 3D LiDAR point clouds. *Remote Sensing*, *11*(12), 1453.

Zhao, H., Jiang, L., Fu, C.-W., & Jia, J. (2019). Pointweb: Enhancing local neighborhood features for point cloud processing. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5565–5573.

Zhao, H., Jiang, L., Jia, J., Torr, P. H. S., & Koltun, V. (2021). Point transformer. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16259–16268.

Zhao, H., Zhang, Y., Liu, S., Shi, J., Loy, C. C., Lin, D., & Jia, J. (2018). Psanet: Point-wise spatial attention network for scene parsing. *Proceedings of the European Conference on Computer Vision (ECCV)*, 267–283.

Zhong, H., Zhang, J., Wang, Z., Wan, H., & Chen, Z. (2015). Aligning knowledge and text embeddings by entity descriptions. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 267–272.

Zhou, Y., Huang, R., Jiang, T., Dong, Z., & Yang, B. (2021). Highway alignments extraction and 3D modeling from airborne laser scanning point clouds. *International Journal of Applied Earth Observation and Geoinformation*, *102*, 102429. https://doi.org/10.1016/j.jag.2021.102429

Zhu, Q., Fan, L., & Weng, N. (2024). Advancements in point cloud data augmentation for deep learning: A survey. *Pattern Recognition*, 110532.

Appendix A

Table 22 Reduced and full model ground truth visualization

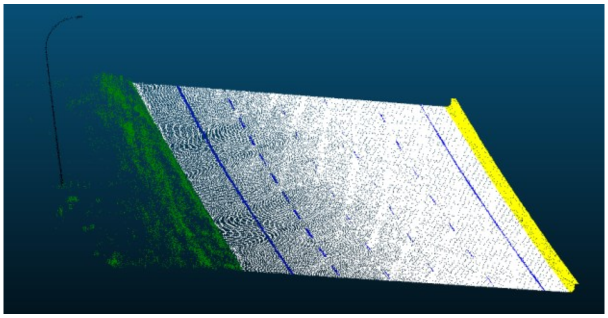
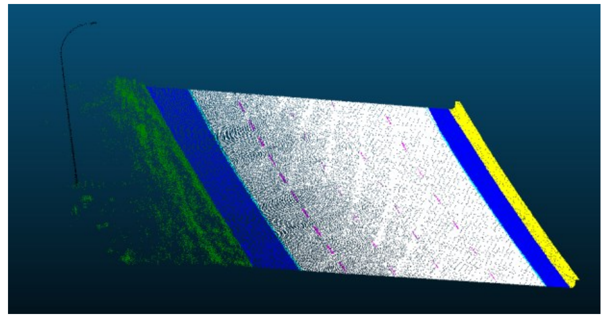
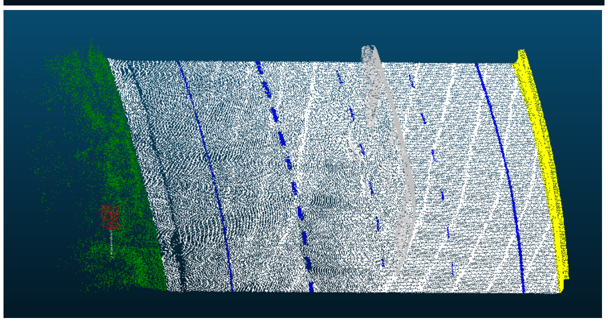
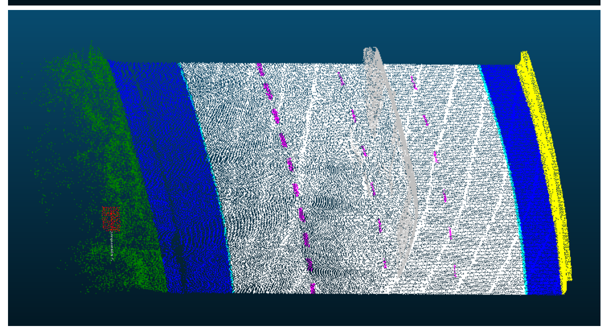
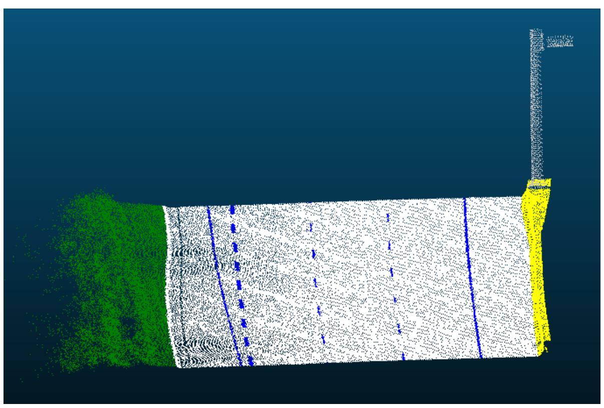
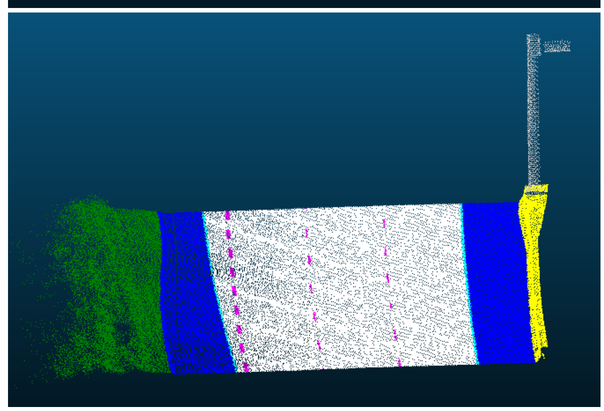
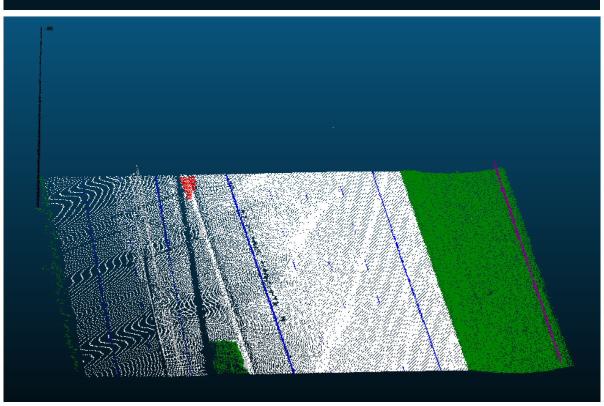
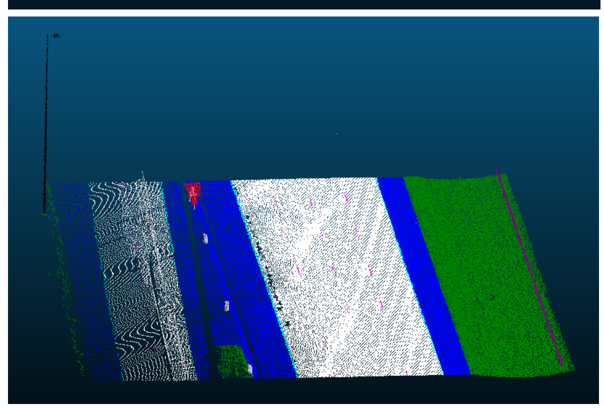
8 classes (reduced model)	12 classes (full model)
	
	
	
	

Table 23 The rest prediction visualization (4 out of 8) for Full model

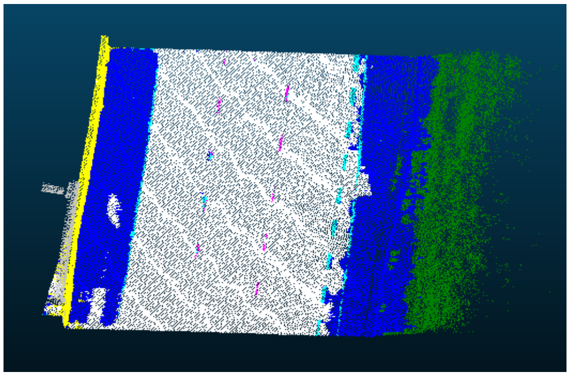
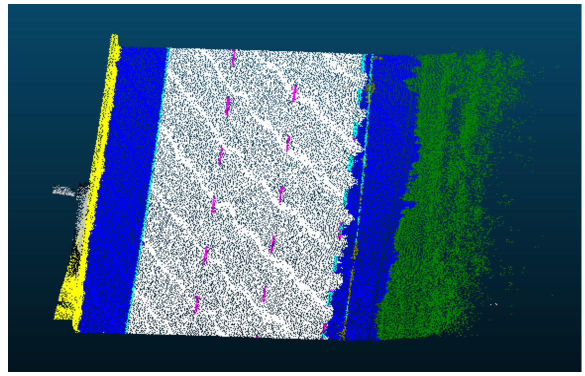
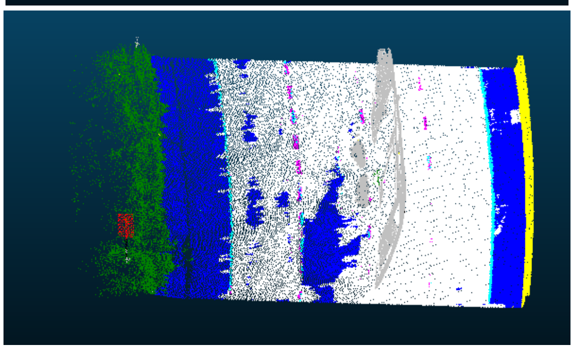
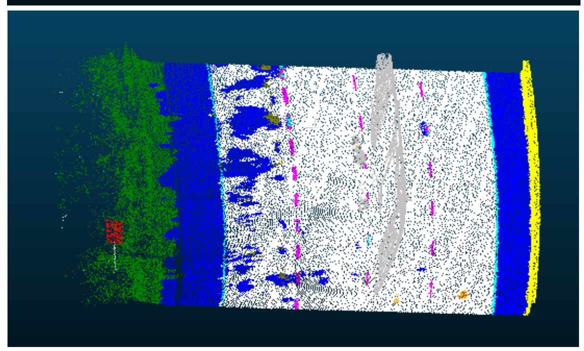
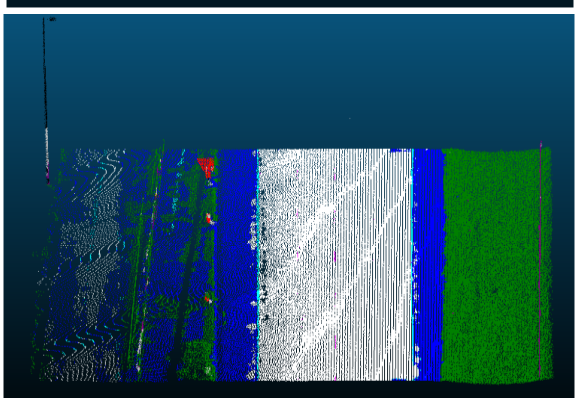
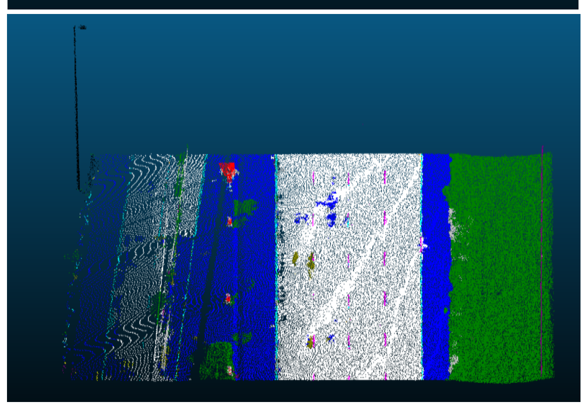
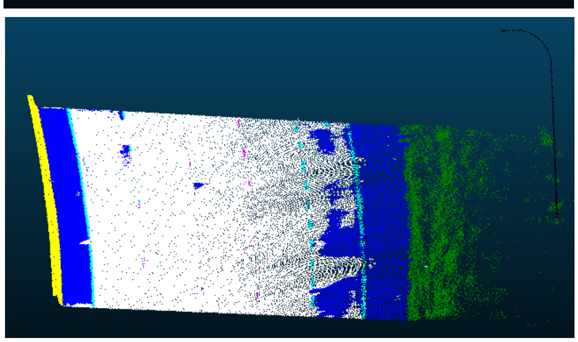
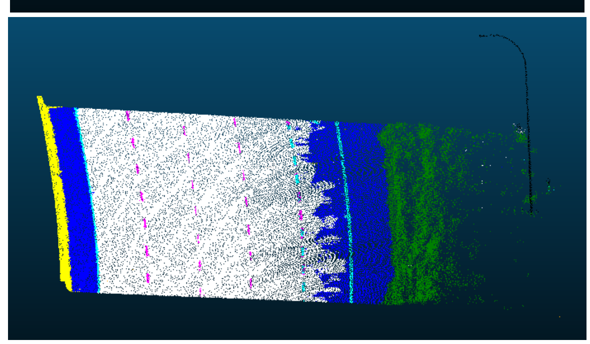
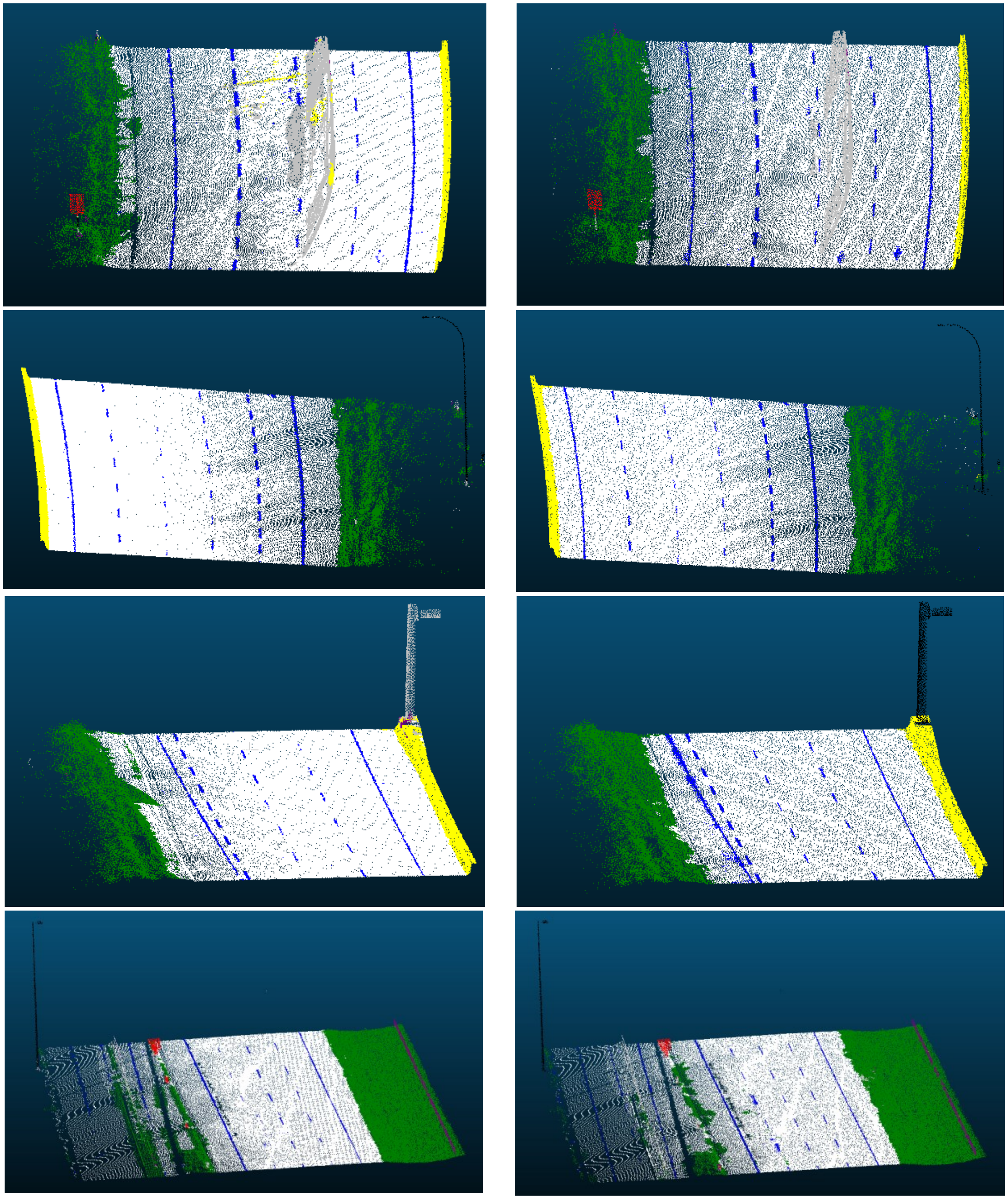
Point Transformer v2	Transformer-based Point Classification
	
	
	
	

Table 24 The rest prediction visualization (4 out of 8) for Reduced model

Point Transformer v2	Transformer-based Point Classification



Appendix B

Full model:

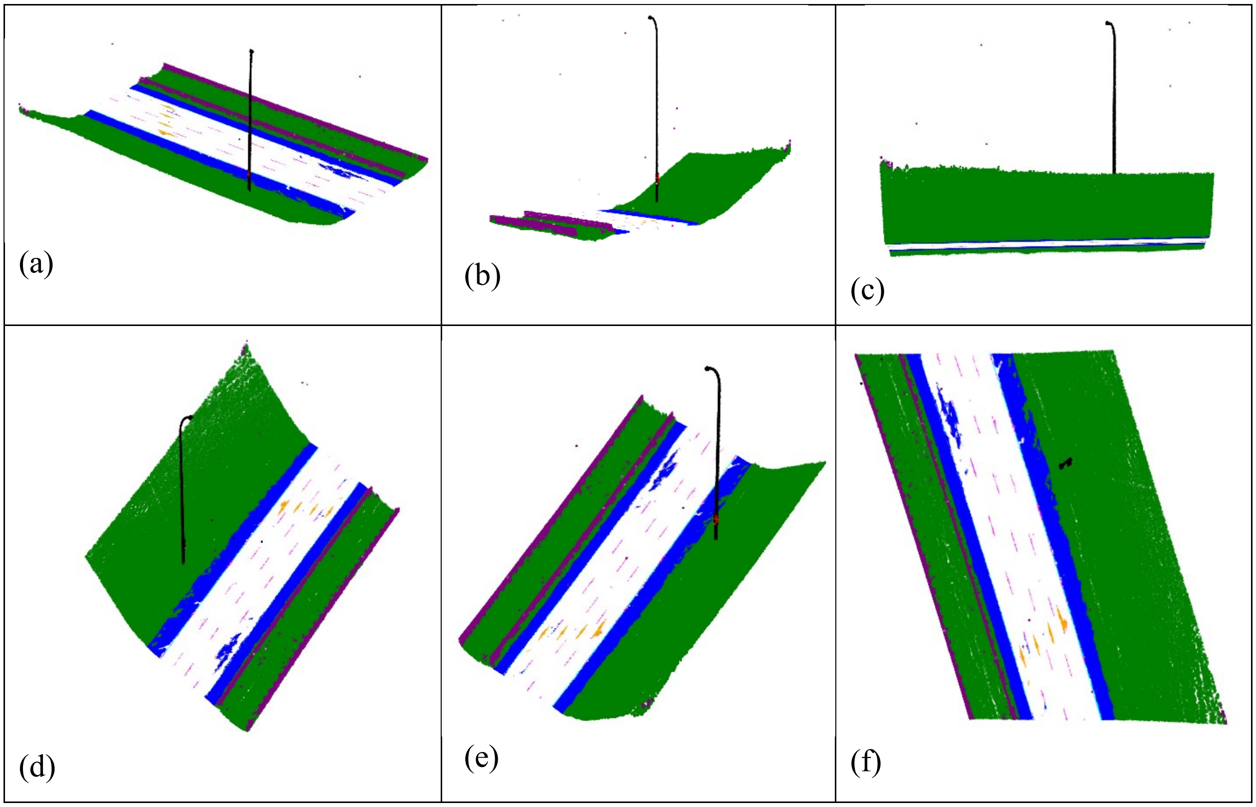


Figure 28 Prediction of a sample (section 7 - Full model) multi-view images (a) Isometric View A (b) Front View (c) Side View (d) Isometric View B (e) Isometric View C (f) Top View

Table 25 Multi-views generated descriptions (Section 7) - Full model

Multi-views	Generated Description
Front, Side, Top, Isometric A, Isometric B, Isometric C	The point cloud data reveals a scene of a multi-lane highway with vegetation alongside. The highway is divided by highway guardrails, indicating it is a divided highway. There are three lanes, distinguished by broken lines and solid lines, ensuring structured traffic flow. Additionally, the presence of a light pole suggests sufficient illumination for nighttime driving. The scene provides adequate space for vehicles to navigate and is bordered by vegetation, likely offering a buffer between the roadway and surrounding areas.

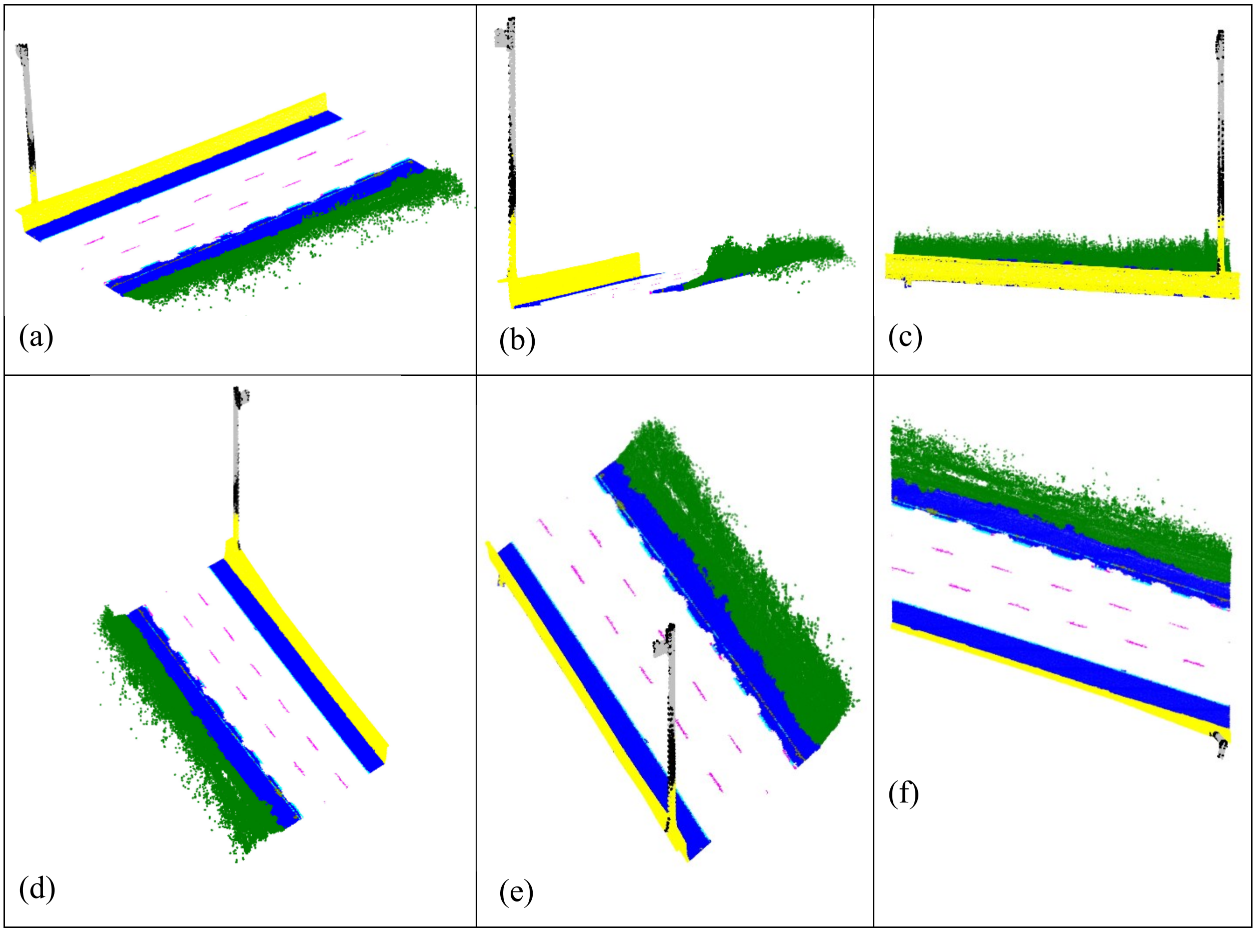


Figure 29 Prediction of a sample (section 8 - Full model) multi-view images (a) Isometric View A (b) Front View (c) Side View (d) Isometric View B (e) Isometric View C (f) Top View

Table 26 Multi-views generated descriptions (Section 8) - Full model

Multi-views	Generated Description
Front, Side, Top, Isometric A, Isometric B, Isometric C	The point cloud data captured represents a multi-lane road with three lanes demarcated by two solid lines and two broken lines. The roadway is a divided highway, evident by the presence of a concrete barrier that separates the direction of travel. To the right of the roadway, there is vegetation noted. Additionally, a light pole is present near the edge of the road, providing illumination for the area. The scene suggests an environment designed for organized and clearly-separated lane usage, optimal for autonomous navigation.

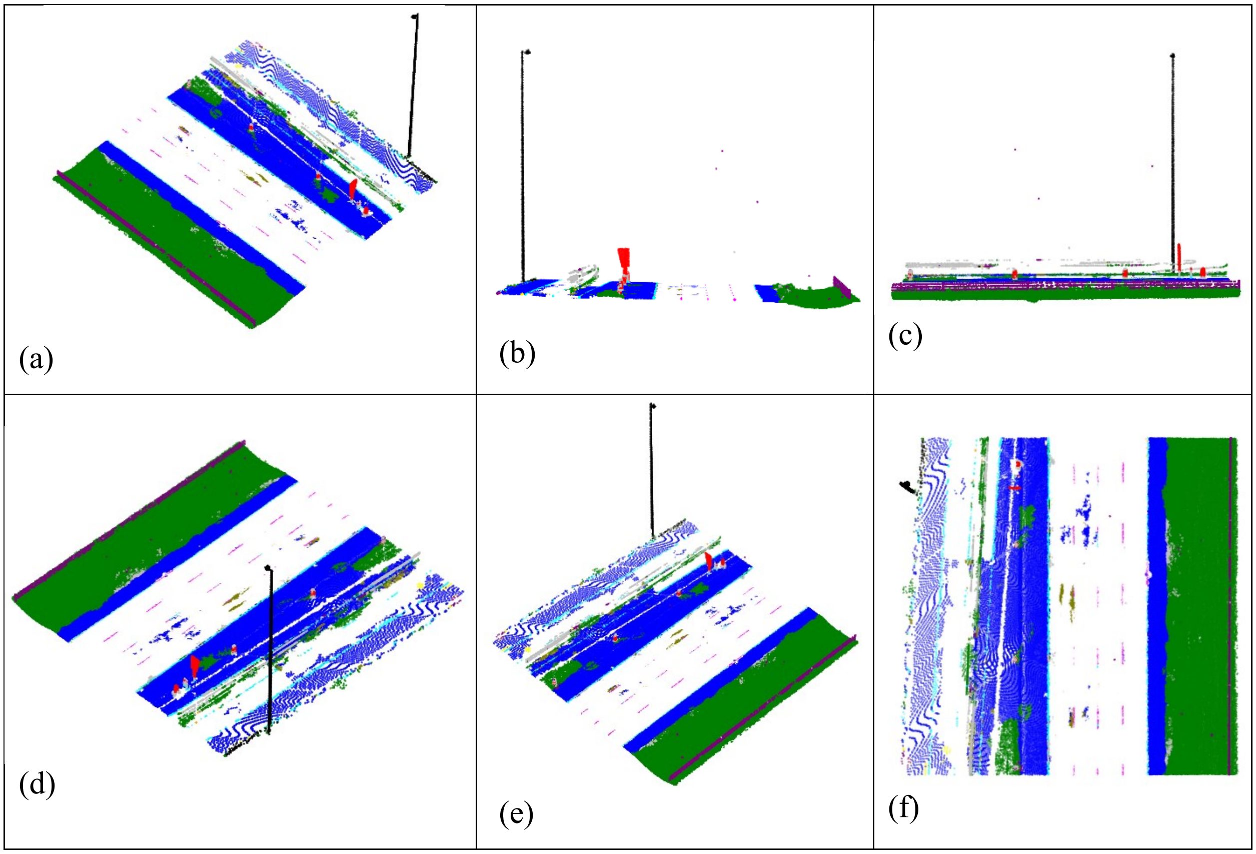


Figure 30 Prediction of a sample (section 21 - Full model) multi-view images (a) Isometric View A (b) Front View (c) Side View (d) Isometric View B (e) Isometric View C (f) Top View

Table 27 Multi-views generated descriptions (Section 21) - Full model

Multi-views	Generated Description
Front, Side, Top, Isometric A, Isometric B, Isometric C	The point cloud data shows a divided highway with multiple lanes, consisting of solid and broken lines that mark four lanes in total. The highway is divided by concrete barriers and highway guardrails, providing separation for opposing traffic flows. Along the road, there are multiple traffic signs and a light pole for guidance and illumination. Vegetation is visible on the side of the road, and shoulders are clearly marked. The scene reflects a typical highway setup with necessary road signage and infrastructure elements for traffic management and safety.

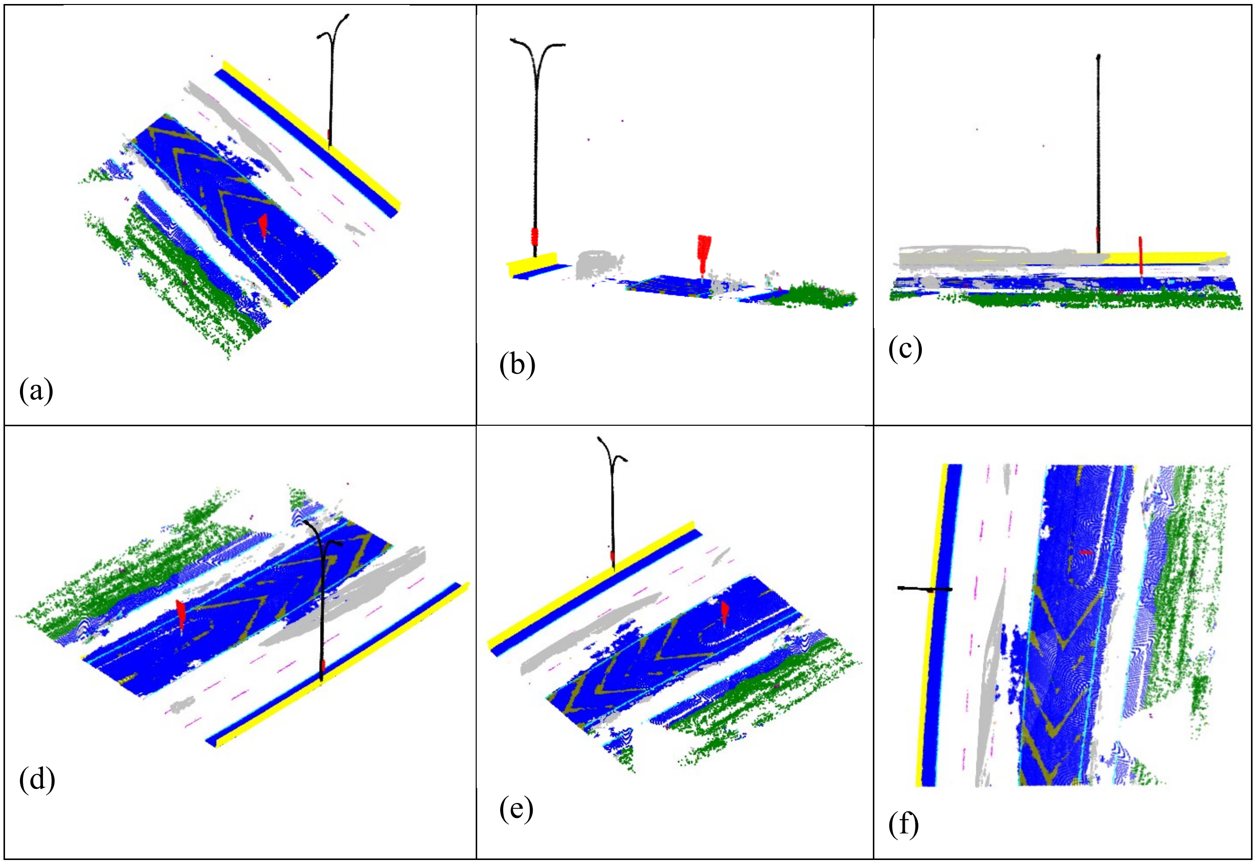


Figure 31 Prediction of a sample (section 26 - Full model) multi-view images (a) Isometric View A (b) Front View (c) Side View (d) Isometric View B (e) Isometric View C (f) Top View

Table 28 Multi-views generated descriptions (Section 26) - Full model

Multi-views	Generated Description
Front, Side, Top, Isometric A, Isometric B, Isometric C	The provided point cloud data represents a scene of a roadway featuring two lanes marked by broken lines and solid lines, indicating a divided highway. There are chevron markings in the lanes, suggesting a merge or a start of an offramp or onramp. On the left side of the roadway, there is a concrete barrier, highlighting the separation between traffic directions. Near the roadway, a light pole is present, providing illumination for the area. Additionally, there is a visible traffic sign, potentially giving important traffic information or warnings to the drivers. Vegetation borders the right side of the roadway, further indicating the roadside environment.

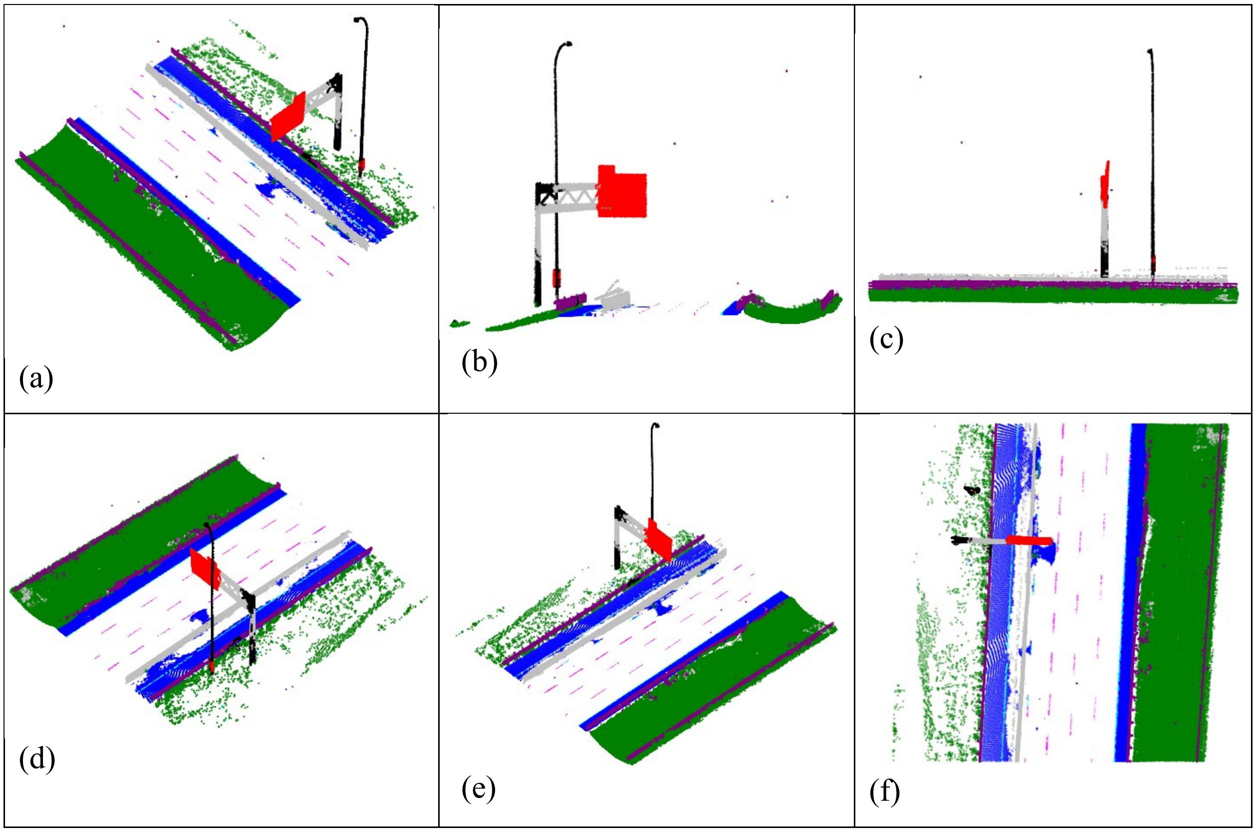


Figure 32 Prediction of a sample (section 34 - Full model) multi-view images (a) Isometric View A (b) Front View (c) Side View (d) Isometric View B (e) Isometric View C (f) Top View

Table 29 Multi-views generated descriptions (Section 34) - Full model

Multi-views	Generated Description
Front, Side, Top, Isometric A, Isometric B, Isometric C	The point cloud data captures a scene on a divided highway with multiple lanes separated by guardrails. There are three lanes, each marked with broken lines. On one side of the road, there is a shoulder, and alongside it, vegetation is present. On the other side, there's clearly defined vegetation as well. Additionally, an overhead traffic sign is visible, which is mounted on a pole, providing guidance for drivers. The scene also contains light poles, enhancing visibility in this highway environment.

Reduced model:

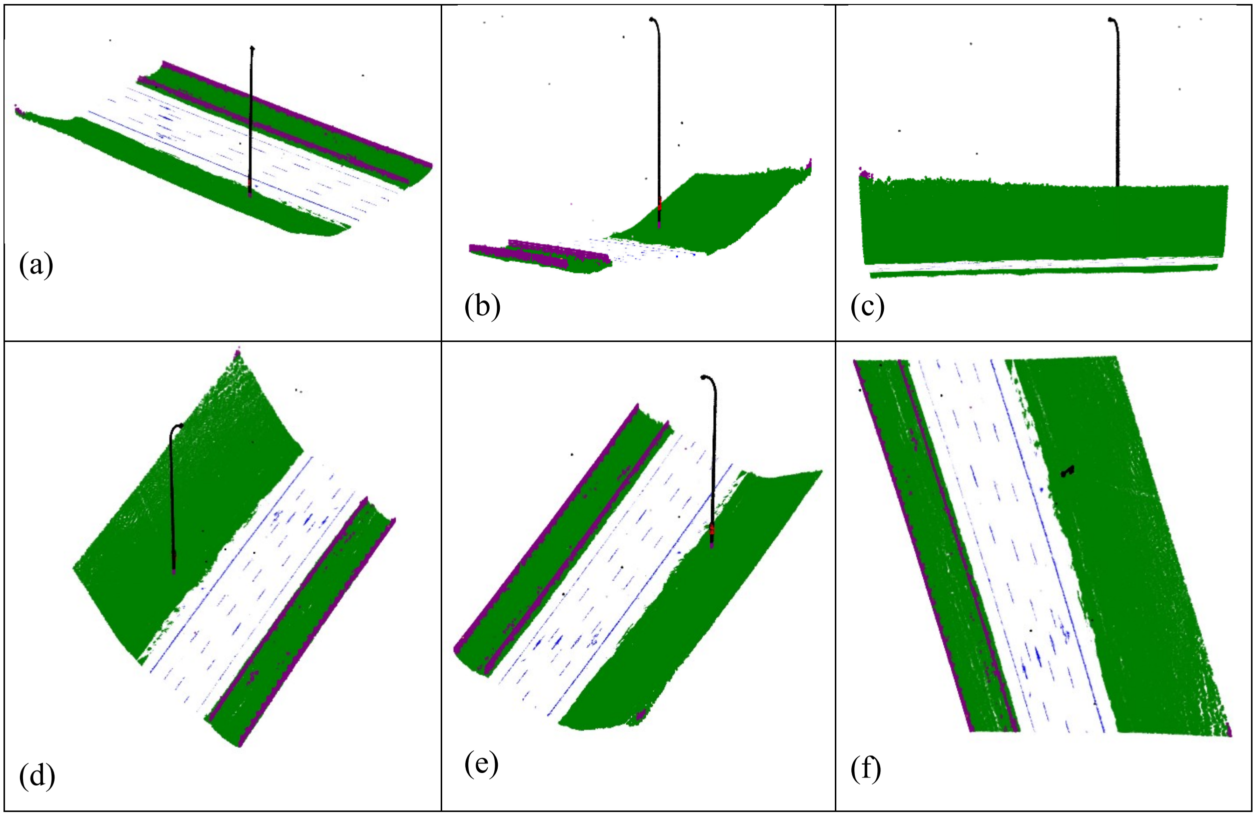


Figure 33 Prediction of a sample (section 7 - Reduced model) multi-view images (a) Isometric View A (b) Front View (c) Side View (d) Isometric View B (e) Isometric View C (f) Top View

Table 30 Multi-views generated descriptions (Section 7) – Reduced model

Multi-views	Generated Description
Front, Side, Top, Isometric A, Isometric B, Isometric C	The point cloud data depicts a highway scene with three lanes. The lanes are delineated by broken line markings. Dividing the highway from the surroundings is a guardrail, indicating it is a divided highway. Vegetation is observed on both sides of the road, providing a clear boundary. Additionally, a tall light pole is present on the right side of the scene, contributing to roadway illumination.

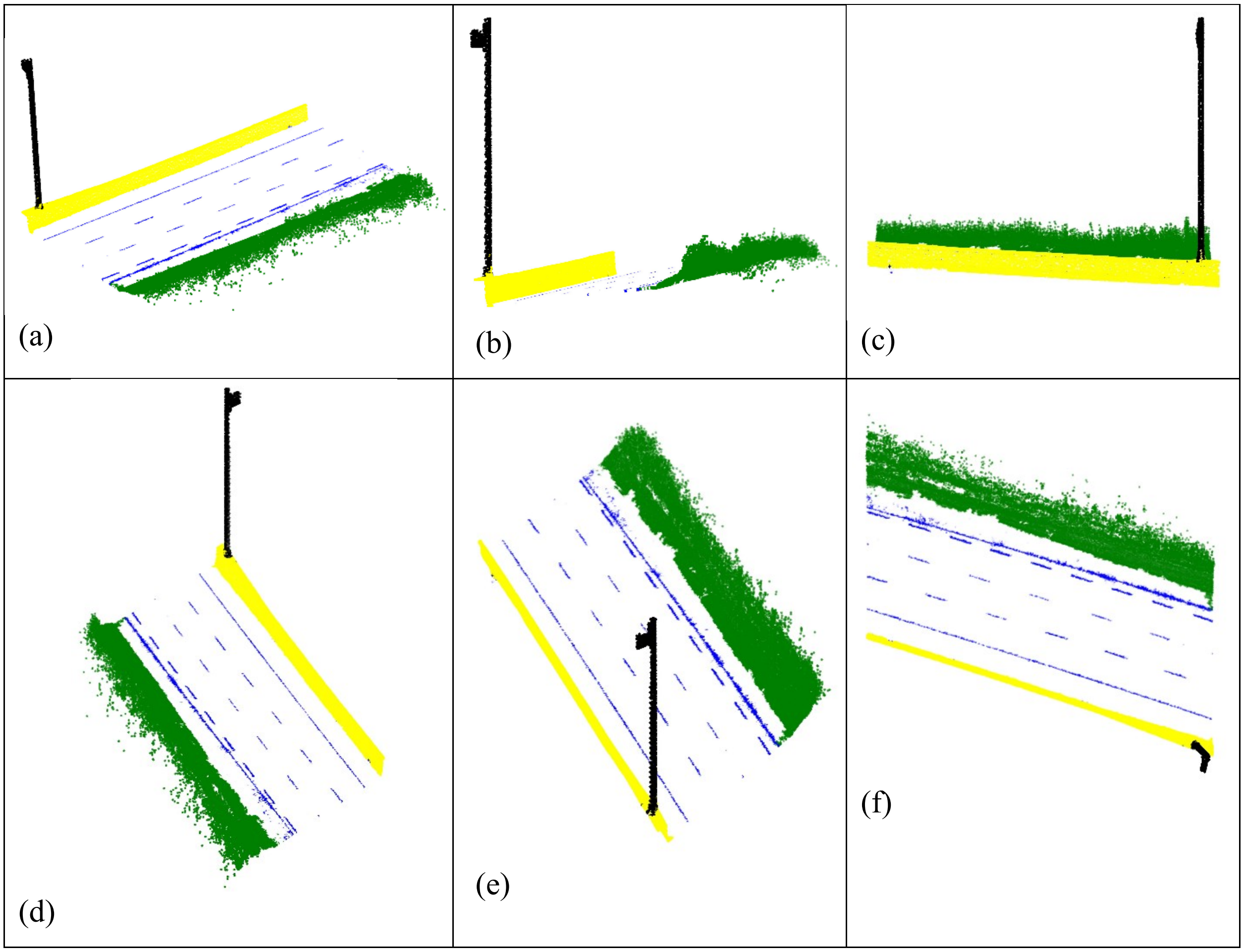


Figure 34 Prediction of a sample (section 8 - Reduced model) multi-view images (a) Isometric View A (b) Front View (c) Side View (d) Isometric View B (e) Isometric View C (f) Top View

Table 31 Multi-views generated descriptions (Section 8) - Reduced model

Multi-views	Generated Description
Front, Side, Top, Isometric A, Isometric B, Isometric C	The scene depicts a multi-lane divided highway from several perspectives. The highway features three lanes, each marked by broken lines. The lanes are bordered by vegetation along the right side and a concrete barrier median on the left, indicating it is a divided highway. A light pole is visible near the concrete barrier, providing illumination. The presence of vegetation alongside the road further adds to the highway's environmental context. This setup provides a clear and orderly driving environment for the autonomous vehicle.

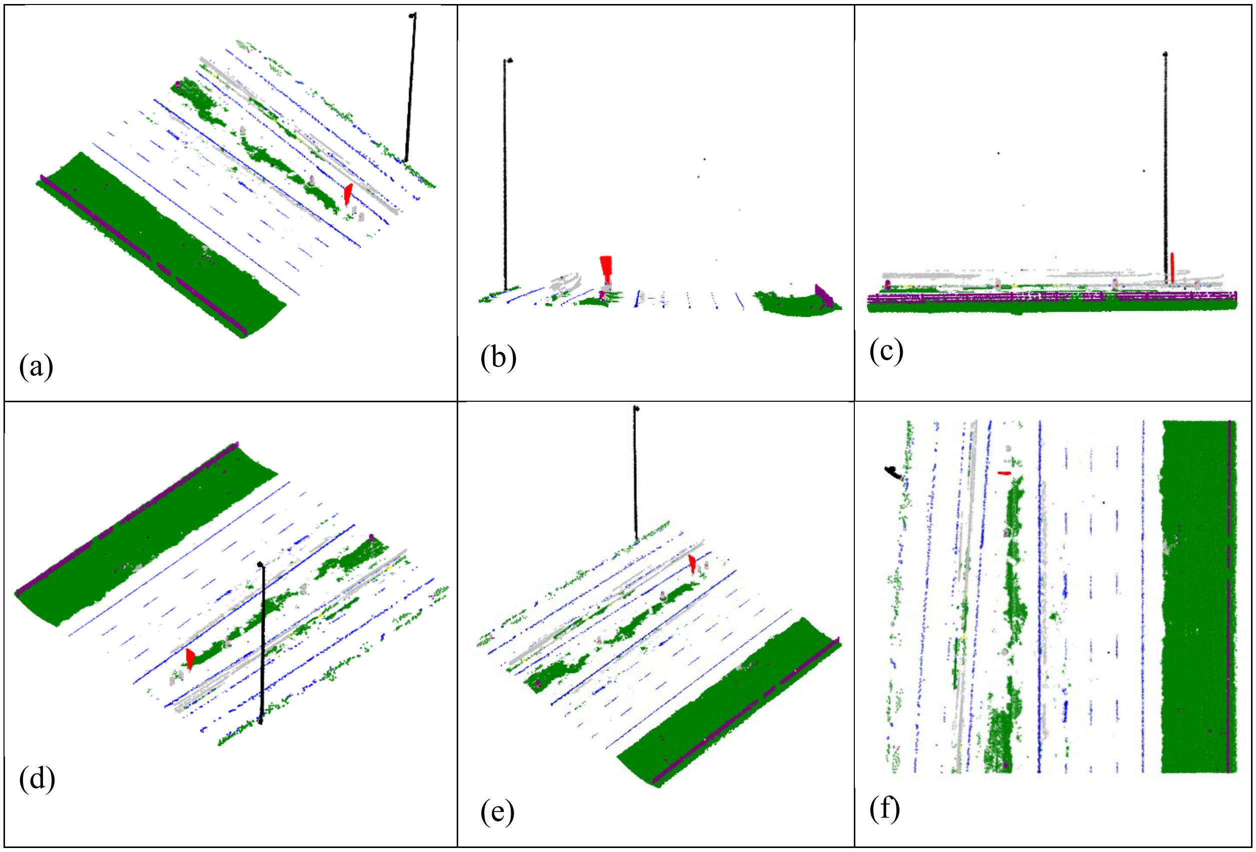


Figure 35 Prediction of a sample (section 21 - Reduced model) multi-view images (a) Isometric View A (b) Front View (c) Side View (d) Isometric View B (e) Isometric View C (f) Top View

Table 32 Multi-views generated descriptions (Section 21) - Reduced model

Multi-views	Generated Description
Front, Side, Top, Isometric A, Isometric B, Isometric C	The provided colored point clouds depict a highway scene with multiple lanes marked by dotted lines. The highway is divided by concrete barriers, indicating opposing traffic flows. Vegetation runs along the highway's edges and the median, while traffic signs are present along the route. A light pole is visible on the left side of the highway, contributing to the road's illumination. This scene includes all elements necessary for safe navigation and lane-keeping in autonomous driving scenarios.

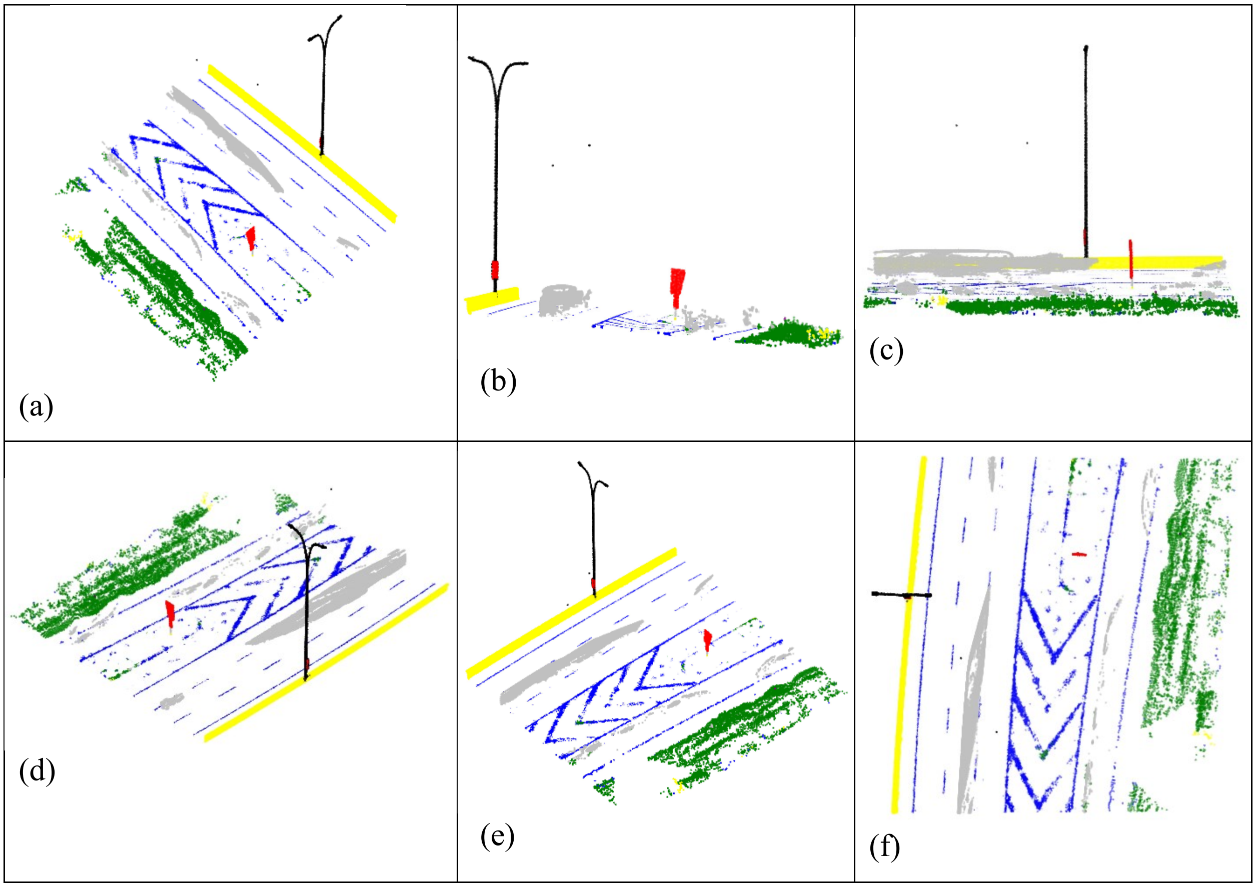


Figure 36 Prediction of a sample (section 26 - Reduced model) multi-view images (a) Isometric View A (b) Front View (c) Side View (d) Isometric View B (e) Isometric View C (f) Top View

Table 33 Multi-views generated descriptions (Section 26) - Reduced model

Multi-views	Generated Description
Front, Side, Top, Isometric A, Isometric B, Isometric C	The point cloud data represents a multi-lane road with clear lane markings and directional arrows on the pavement, indicating structured traffic flow. There are four visible lanes, with two solid outer lines and multiple internal broken lines. A light pole is positioned adjacent to the road. The scene includes a concrete barrier on one side, indicating a divided highway. Vegetation is present to the side, suggesting a roadside or median green space. Additionally, a traffic sign is visible near one of the lanes, providing important guidance for drivers.

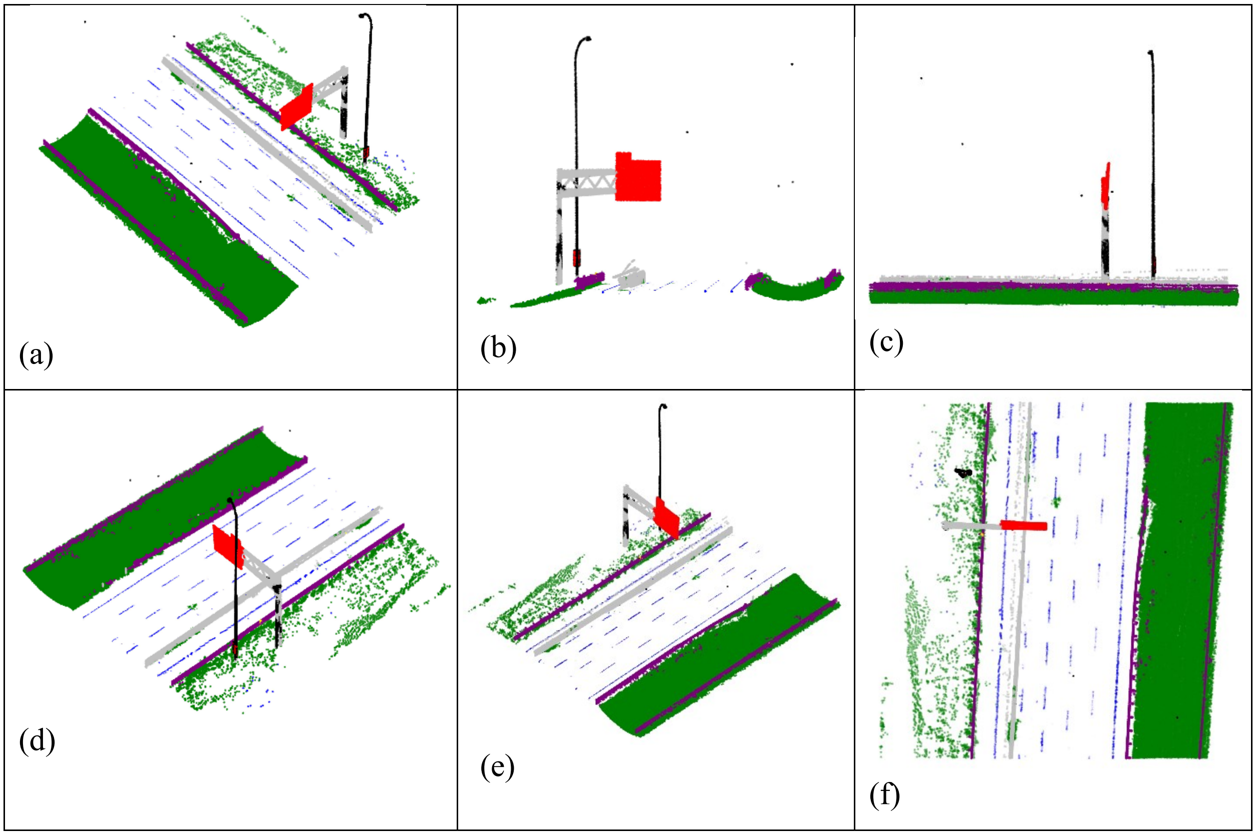


Figure 37 Prediction of a sample (section 34 - Reduced model) multi-view images (a) Isometric View A (b) Front View (c) Side View (d) Isometric View B (e) Isometric View C (f) Top View

Table 34 Multi-views generated descriptions (Section 34) - Reduced model

Multi-views	Generated Description
Front, Side, Top, Isometric A, Isometric B, Isometric C	The point cloud data represents a multi-lane highway scenario with six lanes, delineated by five broken lane markings and two solid markings on either side. The highway is divided by guardrails separating the opposite traffic flows. Vegetation can be seen on both sides of the highway. There is an overhead traffic sign supported by a structure extending over the roadway and a light pole nearby. The traffic sign is positioned above the roadway, likely for visibility and guidance purposes, while the light pole contributes to road illumination.