applied sciences

MDPI

Article

Urban Road Anomaly Monitoring Using Vision–Language Models for Enhanced Safety Management

Hanyu Ding [1,2,*], Yawei Du [3] and Zhengyu Xia [1,2,*]

[1] State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China
[2] Institute of Media Science, Communication University of China, Beijing 100024, China
[3] Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong 999077, China; duyawei@whu.edu.cn
* Correspondence: day9011@gmail.com (H.D.); xia_zhengyu@126.com (Z.X.)

Abstract: Abnormal phenomena on urban roads, including uneven surfaces, garbage, traffic congestion, floods, fallen trees, fires, and traffic accidents, present significant risks to public safety and infrastructure, necessitating real-time monitoring and early warning systems. This study develops Urban Road Anomaly Visual Large Language Models (URA-VLMs), a generative AI-based framework designed for the monitoring of diverse urban road anomalies. The InternVL was selected as a foundational model due to its adaptability for this monitoring purpose. The URA-VLMs framework features dedicated modules for anomaly detection, flood depth estimation, and safety level assessment, utilizing multi-step prompting and retrieval-augmented generation (RAG) for precise and adaptive analysis. A comprehensive dataset of 3034 annotated images depicting various urban road scenarios was developed to evaluate the models. Experimental results demonstrate the system's effectiveness, achieving an overall anomaly detection accuracy of 93.20%, outperforming state-of-the-art models such as InternVL2.5 and ResNet34. By facilitating early detection and real-time decision-making, this generative AI approach offers a scalable and robust solution that contributes to a smarter, safer road environment.

Keywords: urban road safety; vision large language model; road anomaly; real-time monitoring; safety management

check for updates

Academic Editor: Pedro Couto

Received: 22 January 2025
Revised: 20 February 2025
Accepted: 24 February 2025
Published: 26 February 2025

Citation: Ding, H.; Du, Y.; Xia, Z. Urban Road Anomaly Monitoring Using Vision–Language Models for Enhanced Safety Management. *Appl. Sci.* **2025**, *15*, 2517. https://doi.org/10.3390/app15052517

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Urban road anomalies, such as cracks and holes in road surfaces, garbage accumulation, traffic congestion, floods, fallen trees, fires, and traffic accidents, pose significant risks to public safety and urban infrastructure worldwide.

According to the Global Status Report on Road Safety 2023, approximately 1.19 million people worldwide lost their lives due to road traffic accidents in 2021 [1]. In China, road traffic accidents continue to be a major public safety concern. In 2023, 255,000 road traffic accidents were recorded, resulting in direct economic losses amounting to 1.18 billion yuan [2]. Road anomalies are significant contributors to accidents, substantially affecting the safety and comfort of both drivers and passengers. This situation has prompted governments to invest heavily in maintenance [2,3].

These statistics underscore the urgent need for enhanced road safety measures to mitigate the risks associated with urban road anomalies [4,5]. Addressing these challenges necessitates timely road anomalies monitoring [6] and interventions [7]. The real-time system would provide early warnings [8], enabling timely interventions to prevent accidents

and reduce economic costs [9], thereby creating safer, more resilient urban environments and optimizing infrastructure management [10].

Monitoring road anomalies in real-time employs a variety of foundational techniques, each presenting its own characteristics. Conventional monitoring approaches frequently struggle to effectively evaluate risks and can be influenced by weather factors, leading to possible underestimations and insufficient responses [11,12]. While ground-based monitoring systems are capable of collecting high-resolution, instantaneous data, they are associated with significant installation and maintenance costs and have limited coverage [13]. Remote sensing technologies enable efficient information gathering about Earth's surface features and infrastructure without the necessity of direct contact, supporting disaster evaluation and prompt relief efforts; however, their effectiveness may be compromised by adverse weather conditions [14–16]. Social media serves as a real-time information source but requires careful screening and processing of the data available therein [17,18].

Recent studies have shown that advancements in artificial intelligence (AI) have greatly enhanced the capabilities of monitoring road anomalies [19]. Machine learning techniques, including K-Nearest Neighbors (KNNs) and Decision Trees (DTs), have been utilized in multiple investigations to minimize the need for human input and reduce the time necessary for mapping road anomalies [20–22]. An asphalt pavement pothole detection and segmentation method employs a wavelet energy field to emphasize pothole regions and achieves an overall accuracy of nearly 87%. However, the overall process was time-consuming and involved complex processing [23]. A monitoring system for road conditions was created in two stages to identify distressed areas. In the second stage, a single-layer artificial neural network (ANN) was trained [24]. For road waterlogging, an investigation designed an algorithm for flood mapping that integrated the Normalized Difference Water Index (NDWI) with a supervised classifier, achieving an overall accuracy of above 70% for different contexts [25]. Another research study illustrates a deep neural network methodology combined with Canny edge detection to identify submerged stop signs and assess floodwater depth [26]. These analyses indicate that hybrid or ensemble models tend to produce higher accuracy compared to standalone models, suggesting that the potential for developing highly accurate automated road conditions mapping tools is yet to be thoroughly investigated [27,28].

Current studies have demonstrated feasible methods; however, there are notable gaps in data volume, accuracy, and real-time performance. A primary concern is that many machine learning models require substantial volumes of high-quality historical data for effective training [29], which may not always be accessible [30]. Second, traditional AI methods struggle to integrate external prior knowledge, whereas large models employ techniques such as in-context learning (ICL) and retrieval-augmented generation (RAG) to provide real-time prior knowledge, thereby facilitating adaptation to new tasks and improving accuracy [6,31]. Moreover, these models often lack robustness; changes in camera angles can significantly degrade performance due to variations in the captured images, which limited their abilities in real-time application. This limitation requires the collection of new data in innovative scenarios to fulfill performance metrics.

The emergence of next-generation sensors [32] and data-driven technologies [33,34] has transformed the ways data are collected and interpreted, leading to significant progress in road anomaly monitoring. There is an increasing necessity to investigate more advanced AI models, including visual large language models (VLMs) and various generative AI structures. The rise of generative artificial intelligence represents a significant milestone [35,36]. These models display outstanding abilities in comprehending human natural language, which allows them to effectively execute tasks across a variety of domains. Trained on progressively larger datasets, large language models reveal exceptional skill in deriving

insightful information from images, even in the absence of specialized training data. Nevertheless, their performance remains constrained in some practical applications [37,38], underscoring the necessity for additional refinement and customization for specific use cases.

Recent research has employed pre-trained generative AI models for the categorization of road anomalies [39]. A vision large language model has been developed for adaptive object detection through interactions in natural language. Evaluations suggest that the VLM framework performs comparably to leading models in road anomaly detection [40]. A GPT-4 model's Python (3.11.10) API was utilized to measure road anomalies and enhance estimation results by referencing specialized materials [41]. A prototype of an autonomous GIS (Geographic Information System) was developed, based on the GPT-4 API, aiming to accept tasks through natural language and autonomously solve spatial problems [42]. These models can offer intelligent solutions that closely resemble human reasoning, allowing us to leverage general artificial intelligence for addressing challenges in a variety of applications.

However, there is a need for further improvement in the monitoring models of road anomalies. Given the capabilities of large language models, their potential influence extends across multiple domains, including the real-time safety management of road anomalies that can be explored.

This study uses China as a case study and investigates the automatic real-time detection of urban anomalies through vision large language models. It presents three key contributions to the realm of anomaly monitoring and management. Firstly, it employs advanced AI technologies, specifically focusing on tailored vision large language models, to enhance real-time surveillance of anomaly events, thus closing existing methodological gaps. Secondly, by applying innovative modeling approaches, it allows for more accurate assessments of real-time anomalies compared to traditional deep learning models. Lastly, this research aims to educate drivers, policymakers, and stakeholders on effective strategies for managing road surface risks and promoting environmental sustainability. It aids in the development of more resilient urban infrastructure that can respond to evolving climate changes while enhancing overall resilience.

In this paper, we propose anomalies monitoring vision large language models, referred to as Urban Road Anomaly Visual Large Language Models (URA-VLMs). First, we developed a comprehensive database by collecting road anomaly images and video feeds, leveraging both open-source data and a self-collected database, along with real-time access to road video streams. This effort resulted in a dataset of road anomalies, consisting of 15,223 images, with 80% allocated for training and 20% for testing and validation. Second, we compared the AI models for adaptability for this monitoring purpose and selected InternVL as the primary model, with ResNet serving as the comparison model. Third, we designed a Chain of Thought framework to evaluate the pre-trained visual language model, focusing on the identification of anomalies and assessing the safety level. This model was optimized through the use of customized prompts and retrieval-augmented generation, which targeted specific monitoring indicators. We emphasized the selection and assignment of reference objects, facilitating the continuous optimization of the monitoring workflow. This step was essential for conducting analyses of video streams, thereby ensuring accurate detection results. Lastly, we evaluated the models to assess their performances.

2. Methods

This study presents an innovative and fully automated framework, URA-VLMs designed to achieve lightweight operational efficiency and rapid response times. It selects a base model that aligns with the monitoring objectives. Utilizing the Chain of Thought strategy, we emphasize the model's lightweight capabilities and the accuracy of Urban

Road Anomaly evaluations, further enhancing it through a systematic refinement and testing methodology.

The research process is as follows (Figure 1): (1) data collection and annotation, (2) model selection, (3) training and optimization, and (4) performance evaluation.

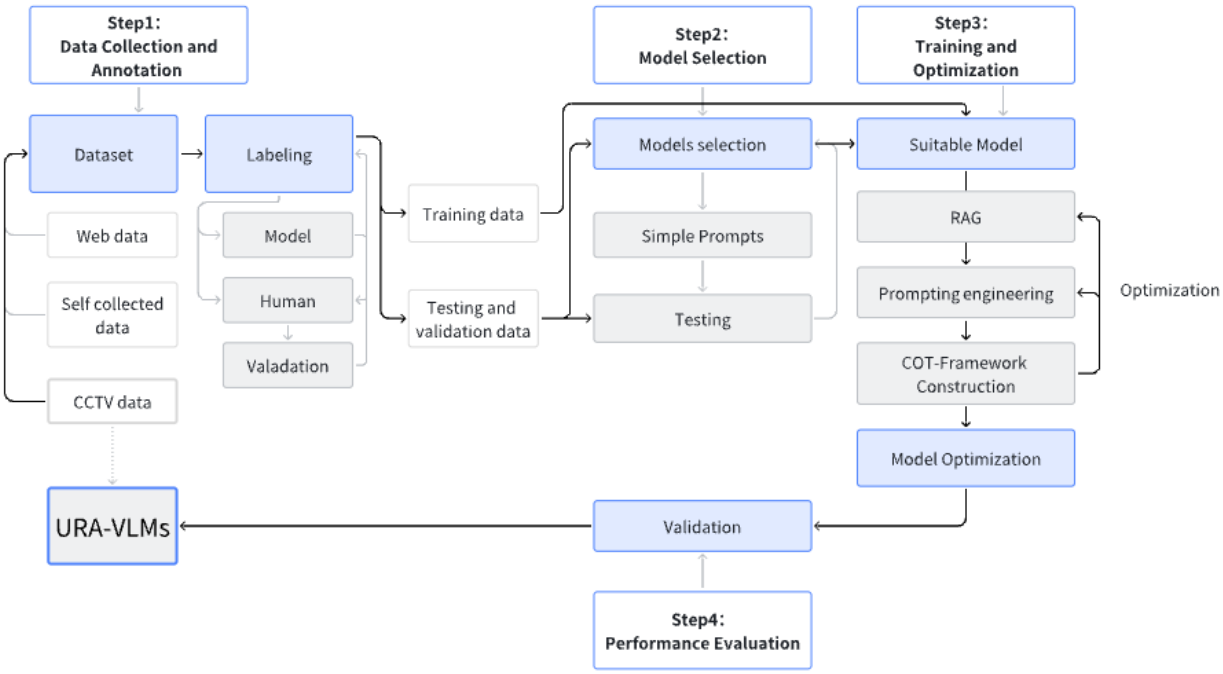


Figure 1. The research process.

The proposed system employs a cloud-based framework (Figure 2) to facilitate the monitoring and analysis of roads in real time. Video streams captured by CCTV cameras are transmitted over a network to an online cloud service platform. Within this platform, the Urban Road Anomaly VLMs carry out functions such as classifying road debris and estimating the depth of floodwaters. The cloud infrastructure makes use of distributed computing capabilities to guarantee efficient and scalable data processing. Results from the analysis are transmitted instantly to mobile devices, including smartphones, laptops, and computers, through secure communication protocols. This architecture offers an effective, centralized, and accessible approach for monitoring road conditions in real time, rendering it suitable for both urban roadway networks.

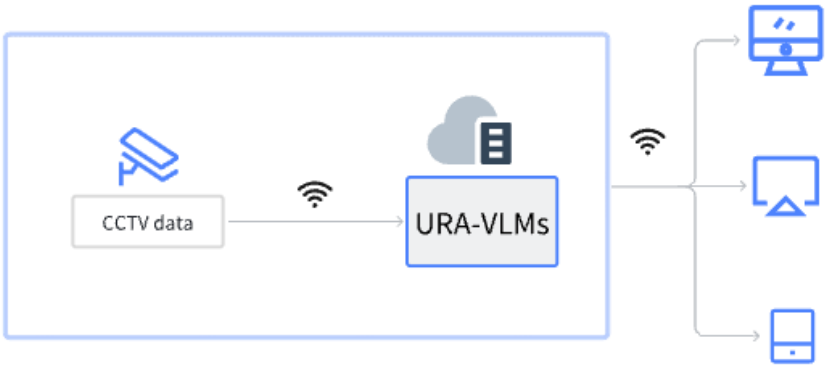


Figure 2. The cloud-based system framework.

To closely simulate real-world scenarios, we conducted primarily tests on the entire process. First, we selected a lightweight, high-definition Wi-Fi camera (360 D1PRO, Chengdu Panorama Intelligent Technology Co., Ltd., Chengdu, China) that can be powered

by either a battery or an electrical source, offering a resolution of 1080P. This camera is positioned at the front of the car for monitoring purposes. The API interface facilitates the connection of the video to the VLMs. We employed both (1) cloud, as well as (2) edge and cloud methods to evaluate the response time under varying weather conditions, specifically sunny and rainy over two separate days in February 2025.

2.1. Data Collection and Annotation

The initial stage involves gathering and filtering visual data suitable for training and testing. We sourced over 18,000 Urban Road Anomaly images from open databases, search engines, and video screenshots, and self-shoot photos, ensuring that they met the necessary size and quality requirements for training. During the image search process, we focused on images depicting various road anomalies in urban areas, mainly the various road anomalies—such as bad roads, fallen trees, fires, floods, garbage, traffic accidents, and jams—that threaten road safety and traffic efficiency. Additionally, we considered different environmental conditions, including daytime and nighttime, as well as sunny and rainy weather, to enhance the urban road dataset. After data cleaning, 15,223 images remained as the original dataset. Among them, approximately 82% were labeled as daytime, while about 18% were labeled as nighttime. Around 77% of the data were labeled under sunny or cloudy conditions, approximately 18% during rainy or windy weather, and 5% during freezing or snowy conditions.

The annotation process utilized a dual verification approach that combined human and machine collaboration. The InternVL2.5-MPO [43] model and a human annotated the images, with a second human annotator reviewed for final confirmation. The dataset was divided into 80% for training and 20% for testing and validation.

We annotated the image data according to the following three steps: first, the eight types of road data were classified (Table 1). Initially, rules to indicate urban road anomalies were established. Our focus was exclusively on road safety; non-road safety events was not included in our analysis.

Table 1. The urban road anomaly annotations.

No.	Types	Description
1	Bad road	The road exhibits significant cracks and potholes, posing potential hazards to vehicle safety and traffic flow.
2	Fallen tree	There are fallen tree obstructs the roadway, creating a safety hazard for vehicles.
3	Fire	A fire is present on the road, and poses a risk to nearby traffic.
4	Flooding	The road is subject to varying levels of flooding, which may impede vehicle traffic flow.
5	Garbage	Accumulated garbage is evident on the road, presenting obstacles and safety risks to traffic.
6	Normal	The road appears in good condition with no noticeable hazards, facilitating normal traffic flow.
7	Traffic accident	A traffic accident has occurred on the road, potentially disrupting traffic and posing safety concerns.
8	Traffic jam	A traffic accident has occurred on the road, potentially disrupting traffic and posing safety concerns.

Next, for the road flooding data, we delve deeper into the assessment of waterlogging depths on roadways. The depth of floodwater plays a critical role in flood mapping [9,10]. Accurately understanding water depth is essential for assessing flood impacts and developing strategies to mitigate flood risks [44]. To classify the levels of road flooding based on depth, we referenced the Beijing Urban Waterlogging Risk Map Report published by

the Beijing Water Authority [45]. Specifically, water depths measuring less than 10 mm are deemed to have an almost negligible impact. Such shallow depths present minimal risks to both traffic flow and infrastructure. Depths ranging from 10 mm to 150 mm are classified as low-risk, whereas those equal to or exceeding 150 mm are classified as waterlogging, which presents considerable safety hazards [46,47]. This classification provides a practical framework for evaluating flood severity and road safety.

To further evaluate the impact of these road issues on traffic, we classified them according to their safety levels (Table 2). The road safety classification employs three categories—GREEN, LIGHT GREEN, and YELLOW—to effectively monitor and manage road conditions in real-time. GREEN signifies optimal safety with no anomalies, allowing for no more than 10 mm of standing water. LIGHT GREEN indicates minor issues such as traffic congestion, garbage accumulation, or slight flooding, advising drivers to exercise caution and necessitating minor interventions like waste removal or drainage improvements. YELLOW denotes severe problems, including road damage, fires, deep flooding, or accidents, advising drivers to avoid the area and requiring urgent actions such as road repairs or traffic rerouting. This classification provides a clear framework for road authorities and drivers to respond appropriately to varying road conditions, thereby enhancing safety and efficiency in urban environments [48].

Table 2. The urban road safety level annotations.

No.	Safety Levels	Types	Description
1	GREEN	Normal Flooding (depth $\leq$ 10)	The road is in good condition, exhibiting no noticeable hazards, and has waterlogging measuring less than 10 mm, which facilitates normal traffic flow.
2	LIGHT GREEN	Garbage Traffic jam Flooding (10 mm < depth $\leq$ 150 mm)	The presence of traffic congestion, garbage accumulation, and minor flooding (10 mm < depth $\leq$ 150 mm) has a negligible impact on vehicular traffic.
3	YELLOW	Bad road Fallen tree Fire Traffic accident Flooding (depth > 150 mm)	Road damages, fires, substantial flooding (water depth > 150 mm), or traffic accidents pose a significant threat to vehicular traffic flow.

To conduct the models' tests, we randomly selected 20% of the images from each type to form the test set after completing the dataset annotation. The distribution of images in the test set is as follows (Tables 3 and 4).

Table 3. The training and testing datasets.

No.	Types	Number of Data Points (Total)	Number of Data Points (Training Dataset)	Number of Data Points (Testing Dataset)
1	Bad road	1068	854	214
2	Fallen tree	1480	1184	296
3	Fire	1523	1218	305
4	Flooding	1598	1288	310
5	Garbage	1819	1455	364
6	Normal	3375	2700	675
7	Traffic accident	2163	1730	433
8	Traffic jam	2185	1748	437
	Total	15,211	12,177	3034

Table 4. The testing dataset.

No.	Level	Types	Number of Data Points (Testing Dataset)	
1	GREEN	Normal Flooding (depth $\leq$ 0)	675 39	714
2	LIGHT GREEN	Garbage Traffic jam Flooding (10 < depth $\leq$ 150)	364 437 43	844
3	YELLOW	Bad road Fallen tree Fire Traffic accident (depth > 150 mm)	214 296 305 433 228	1476
	Total		3034	3034

2.2. Model Selection

Our approach commenced with an initial screening of object detection models, Vision–Language Models, and image classification models. This preliminary analysis highlighted the inherent features of these models, their effectiveness for real-time assessments, and their adaptability for monitoring road anomalies and describing road waterlogging.

Based on this evaluation, we selected the InternVL model due to its high applicability. Additionally, we chose ResNet34 as a reference model to facilitate result testing and comparison. This decision enhances the analytical depth of our research scenario and establishes a solid foundation for subsequent optimization and experimentation (Table 5).

Table 5. The comparison of popular vision models.

	Model	Main Features	Suitability for Real-Time Monitoring	Suitability for Urban Road Anomaly Classifications	Suitability for Urban Road Waterlogging Monitoring
1	ResNet34 (Image classification model)	A fast, simple convolutional neural network that performs classification.	Medium It focuses on image classification rather than real-time processing.	High It focuses on image classification.	Medium-Low It is limited in detecting flood depth. It should be used in conjunction with other models.
2	YOLO (You Only Look Once) (Object detection model)	Real-time processing with rapid inference speeds, well-suited for detecting multiple objects for specific classes [49].	High It is fast in inference speed and efficient in object detection capabilities.	High It is suitable for detecting and classifying	Medium-Low It should be used in conjunction with other models for optimal results.
3	Faster R-CNN (Object detection model)	High detection accuracy with Region Proposal Networks (RPN), ideal for detailed segmentation tasks [50].	Low Less suitable due to slower inference speeds compared to single-stage detectors.	High Highly suitable for accurate anomaly detection and classification.	Medium-Low The model can be applied to this scenario; however, it shares the same limitations as YOLO in capturing water depth.

Table 5. *Cont.*

	Model	Main Features	Suitability for Real-Time Monitoring	Suitability for Urban Road Anomaly Classifications	Suitability for Urban Road Waterlogging Monitoring
4	Mask R-CNN (Object detection model)	Capable of instance segmentation with pixel-level accuracy [51].	Medium-Low Moderate suitability due to its additional segmentation overhead, which slows down processing.	High Highly suitable for detailed anomaly detection and pixel-level segmentation.	Low It requires integration with other models or methods for depth estimation or event detection.
5	Efficient Det (Object detection model)	Balances accuracy and speed using a compound scaling method [52].	High Highly suitable due to its balance between speed and accuracy.	High Suitable for detecting and classifying road anomalies.	Medium-Low It requires integration with other models or machine learning methods for depth estimation.
6	InternVL (Vision–Language Model)	Integrates visual and language understanding for contextual and vision monitoring [53].	Medium-High Highly suitable due to its balance between speed and accuracy.	High Suitable for detecting and classifying.	Medium-High The model can be applied to this scenario and can be trained and optimized based on the context.

2.3. Training and Optimization

In this section, we focus on optimization using InternVL-2.5 MPO and the training of ResNet34 for monitoring road anomalies. Notably, the optimization process for InternVL does not include fine-tuning, which results in a training set of zero. The training of ResNet34 was conducted using three approaches to enhance performance. The specific steps undertaken are outlined below.

2.3.1. URA VLMs

Based on the InternVL selected as relatively suitable for real-time monitoring of road anomalies, we proceeded with the URA VLMs training and optimization using prompt engineering, a retrieval-augmented generation (RAG) database, and chain-of-thought (COT) techniques. This phase focused on enhancing the model's capability to accurately classify the road and interpret and describe flood-related data; the key feature of RAG is presented in Appendix A Table A1.

The inference establishes a comprehensive judgment framework consisting of the following steps: (1) First, identify road anomalies, which refer to the occurrence of seven types of road anomalies. (2) For flooding cases, proceed to the reference detection VLM, which identifies a reference object within the image that can be utilized to describe depth. The RAG provides parameters for the reference object to ascertain depth in relation to the flooding situation. (3) These parameters are then combined to generate prompts. Next, the VLM evaluates the depth of water accumulation and determines the severity based on this depth. (4) After concluding the types of road anomalies and assessing the waterlogging depth, merge the results to generate the final safety level for the dashboard (Figure 3).

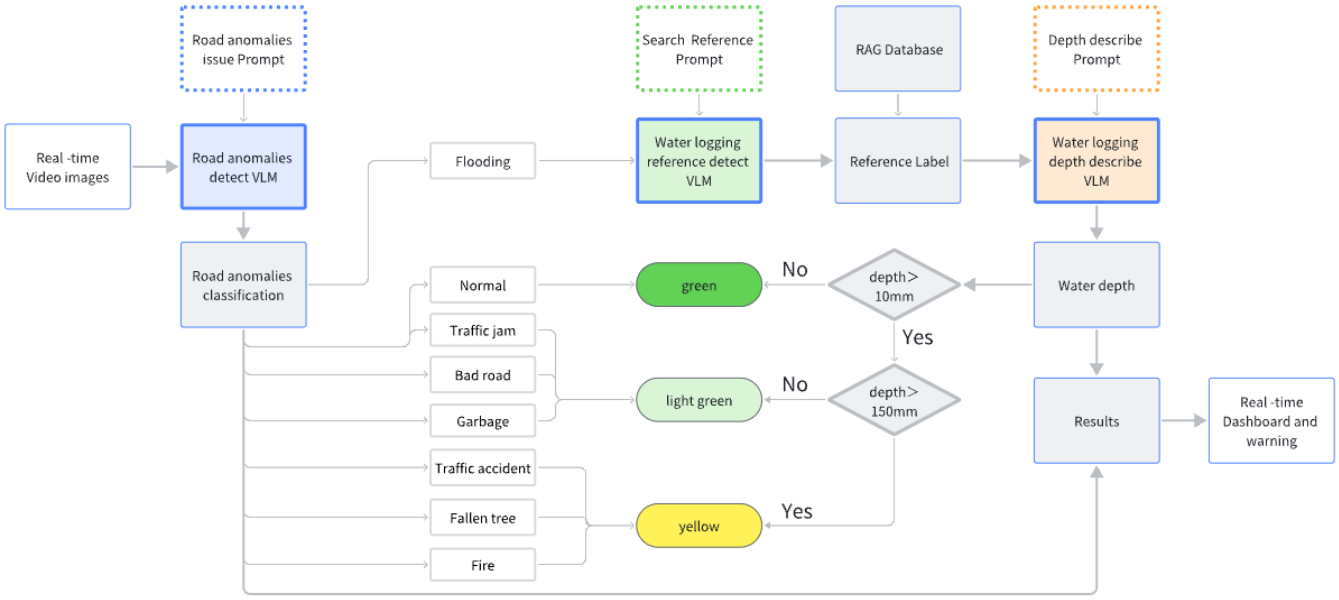


Figure 3. The urban road flooding VLM inference process.

We constructed three specialized expert models—Road Anomalies Detect VLM, Waterlogging Reference Object Detect VLM, and Waterlogging Depth Describe VLM—for specific tasks using a COT design that employs multi-step prompts. In this framework, each step requires the URA-VLMs to perform distinct tasks.

This approach not only enhances the accuracy of road anomaly detection but also contributes to the efficient allocation of resources for real-time monitoring. The purpose of this research method is to decompose an end-to-end (E2E) task into multiple steps, thereby improving both accuracy and efficiency.

Our model aims to classify types of road anomalies and further assess the depth of flooding to categorize the severity of road abnormalities. By decomposing the task into several subtasks, we significantly enhance accuracy. To optimize resource utilization, the first step involves classifying road anomalies and determining the presence of a waterlogging issue, as such problems are often absent. This strategy effectively reduces the number of VLM calls in subsequent steps, alleviating server load. Additionally, the independent optimization of sub-task modules enables parallel optimization of multiple tasks, promoting collaboration and division of labor. This innovative approach not only streamlines the workflow but also enhances the overall performance of the model in real-time flood monitoring applications.

2.3.2. ResNet34

We utilize 80% of the dataset, which consists of a total of 12,177 images, to train the ResNet34 model, a widely recognized model in image classification, for detecting road abnormal events. Three approaches are employed to train the ResNet34.

In the first approach, inspired by the linear probing [54], we freeze the pre-trained model and do not update its parameters; only the parameters of the final linear layer are adjusted. ResNet34 is a model that has been pre-trained on the ImageNet dataset. Given that the resolution of the training set data is suboptimal, our objective is to leverage the benefits of the pre-trained model's weights for image processing tasks. In this approach, only the affine layer is fine-tuned during training.

In the second approach, after loading the model, we utilize the xavier_init() function to initialize all weights, followed by updating all weights throughout the training process.

In the third approach, we load the pre-trained model, increase the learning rate (lr) of the last layer by a factor of ten, initialize the weights of the last layer using xavier_init(), and then proceed to train the entire model.

The results of the three approaches for the ResNet34 model are as follows (Table 6). It can be seen that the overall accuracy of the first approach is expected to be higher; therefore, it will be selected for subsequent comparisons.

Table 6. Testing ResNet34 of three approaches.

No.	Types	Number of Data Points (Testing Dataset)	Model Accuracy (%)		
			ResNet34 (First Approach: Linear Probing)	ResNet34 (Second Approach: Not Finetuned, Training)	ResNet34 (Third Approach: Finetuned, Other Layer Weights Are 1 Times LR, and the Last FC Layer Is 10 Times)
1	Bad road	214	79.44	78.04	79.91
2	Fallen tree	296	81.08	93.58	90.54
3	Fire	305	91.8	84.92	82.3
4	Flooding	310	62.42	53.73	64.91
5	Garbage	364	67.03	70.88	73.08
6	Normal	675	89.93	75.7	86.81
7	Traffic accident	433	92.15	66.74	63.97
8	Traffic jam	437	79.44	78.04	79.91
	Overall accuracy (%)	3034	81.48	75.31	77.38

2.4. Performance Evaluation

This part involved testing the model's accuracy with the remaining 20% of the images, as well as evaluating response times in real-time situations. First, we assessed the accuracy of the road anomalies classification, water depth estimation, and the safety level accuracy to ensure the model's robustness and reliability. This methodological approach, grounded in recent advancements and practices from the latest scientific literature, ensures a rigorous and effective development of a visual language model specifically tailored for real-time road anomalies monitoring applications [9,11,41]. Second, real-time tests were made to evaluate the VLM's response time and robustness.

2.4.1. Model Performance

- Road anomaly classification

To assess the performance of the URA-VLMs, the presence of road anomalies, and the accuracy rate of the sub-type, and the overall accuracy were defined as follows:

$$Acc = \frac{\sum_{i=1}^{n} F(i) = \begin{cases} 1 \; if \; right \\ 0 \; if \; error \end{cases}}{N} \quad (1)$$

- Road waterlogging depth estimation

Additionally, the relative error rate was computed to quantitatively evaluate the accuracy of the URA-VLMs flood depths estimations against the manually determined estimations. The relative error was expressed as a percentage.

$$Error = \frac{\sum_{i=1}^{n} \min\left(\frac{|a_i - b_i|}{b_i}, 1\right)}{N} \quad (2)$$

Furthermore, the Mean Absolute Error (MAE) was calculated to measure the accuracy of the URA-VLMs estimations in comparison to the manual estimations (refer to Formula (3)). In this context, 'a' represents the described value, while 'b' denotes the labeled value, thereby expressing the average value of the relative error.

$$MAE = \sum_{i=1}^{n} \frac{|a_i - b_i|}{n} \quad (3)$$

- Road safety level

To assess the performance of the URA-VLMs for the green, light green, and yellow road safety levels, the level accuracy rate was defined as follows:

$$Safety\ Level\ Acc = \frac{\sum_{i=1}^{n} F(i) = \begin{cases} 1\ if\ right \\ 0\ if\ error \end{cases}}{N} \quad (4)$$

2.4.2. Real-Time System Architecture

To test real-time performance, (1) cloud server processing and (2) edge cloud server processing were tested based on the A800 Server and NVIDIA Jetson (NVIDIA Corporation, Santa Clara, CA, USA). First, the time for the full-process inference using the cloud server and the edge was tested. After optimization, (2) edge cloud server processing was proposed, where the edge makes a preliminary inference. Once an event on the road is detected, the cloud server is then used for further inference (Figure 4). This approach enables more efficient collaboration between the edge and the cloud.

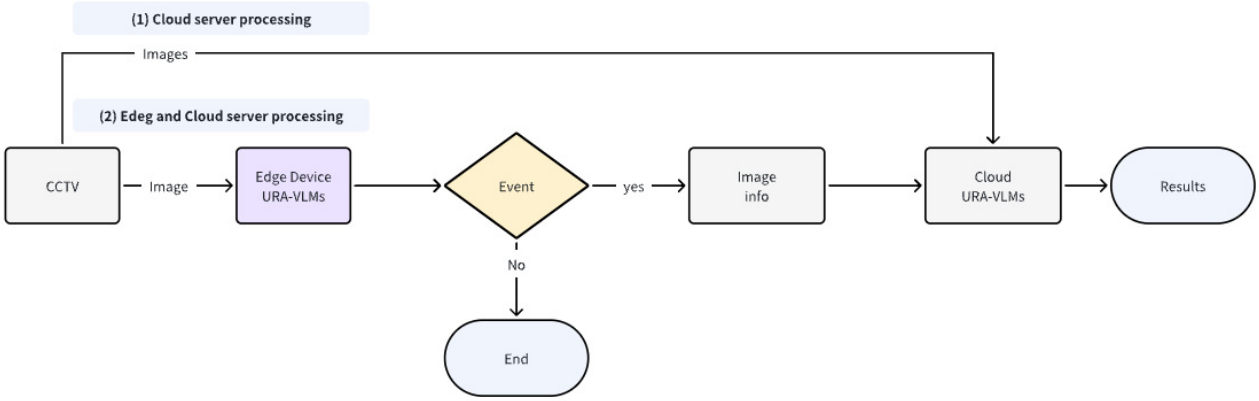


Figure 4. The two paths for real-time testing.

Under our current testing conditions, the downstream speed is 26 Mbps, which fully meets the requirements. According to the specifications of the Vision–Language Model and actual test results, images with a resolution of 448 dpi × 448 dpi or higher are sufficient.

The basic parameters of the cloud and the edge are shown in Table 7.

Table 7. The basic parameters of the cloud and the edge servers.

Servers	A800	NVIDIA Jetson
GPU	80 GB × 8	64 GB
CPU		58 G
Memory		
GPU Memory		61.4 G

- Response time

To further evaluate the model's response time in real-world scenarios, we conducted real-time tests on the roads of a community in Shenzhen over two separate days in February 2025. These tests occurred during both daytime and nighttime, encompassing various weather conditions, including sunny, rainy, and windy. The camera was mounted at the front of the car for monitoring, and various road conditions were selected for video capture. The real-time video stream from the camera was transmitted to both the cloud service and the edge computer via APIs for testing. We assessed the response time of each system.

- Robustness under diverse environmental conditions

Regarding the robustness under diverse environmental conditions, we carried out basic qualitative tests. We collected and tested real-time data at night, during winds, and in rainy weather. Then, we made judgments on the collected data to analyze the possible situations.

3. Results

This section addresses the following three key components of the results: the annotation of road anomalies data, a comparison of models in the classification of road anomalies, and the performance evaluation of URA-VLMs.

3.1. Data Annotation

First, we classified the data into seven categories of road anomalies and one category of normal road conditions based on the data descriptions. Subsequently, we further annotated the flooding-related road data. We observed that classifying water severity directly could lead to ambiguity. Therefore, we first annotated the depth values and then categorized the water severity levels.

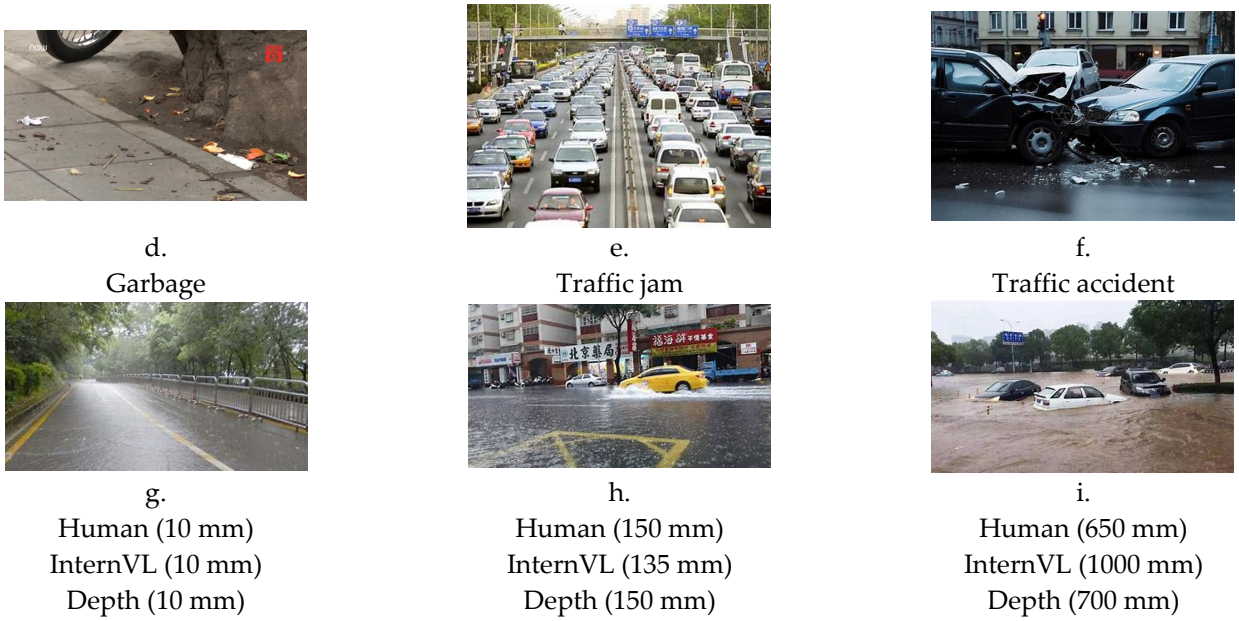
Annotating flood depth presents a significant challenge. Initially, we excluded images with overly complex content, where annotators struggled to reach a consensus. In this process, we initially used InternVL and manual annotation separately, followed by a second annotator to finalize the labels. This approach ensures higher accuracy and consistency in the annotation process. (Figure 5).

a.
Bad road

b.
Fallen tree

c.
Fire

Figure 5. *Cont.*



d. Garbage
e. Traffic jam
f. Traffic accident
g. Human (10 mm) InternVL (10 mm) Depth (10 mm)
h. Human (150 mm) InternVL (135 mm) Depth (150 mm)
i. Human (650 mm) InternVL (1000 mm) Depth (700 mm)

Figure 5. Examples of annotation.

3.2. Models Classification Performance Comparison

During the model selection phase, we conducted comparative experiments between the vision large language model InternVL-2.5 MPO and the deep learning model ResNet-34. The InternVL-2.5 MPO was evaluated using classification prompts without fine-tuning, while ResNet-34 was fine-tuned by optimizing the final fully connected layer with the following hyperparameters: 25 epochs and a batch size of 8, using a training dataset of 12,177 samples.

The results demonstrate that (with inference design and preliminary classification prompts) the InternVL-2.5 MPO model outperformed the fine-tuned ResNet-34 in terms of accuracy across multiple evaluation metrics. The overall accuracy of the InternVL-2.5 MPO (90.35%) surpasses that of the fine-tuned ResNet-34 (81.48%) by near 10%. In every sub-type except the traffic jam, the accuracy of the InternVL-2.5 MPO model exceeds that of ResNet-34. Notably, the most significant difference is observed in the recognition of flooding, where the accuracy is 31.06% higher than that of ResNet-34. Conversely, the smallest gap is found in the recognition of traffic jam, which is only 0.26% lower than ResNet-34. This highlights the superior generalization capability and robustness of the InternVL-2.5 MPO model for the given task (Table 8).

Table 8. Preliminary testing of InternVL-2.5 MPO with ResNet34.

No.	Types	Number of Data Points (Testing Dataset)	Model Accuracy (%)	
			InternVL-2.5 MPO (Preliminary Classification Prompt)	ResNet34 (First Approach: Linear Probing)
1	Bad road	214	90.19	79.44
2	Fallen tree	296	96.62	81.08
3	Fire	305	96.72	91.80
4	Flooding	310	93.48	62.42
5	Garbage	364	84.62	67.03
6	Normal	675	92.00	89.93
7	Traffic accident	433	92.84	92.15
8	Traffic jam	437	79.18	79.44
	Overall accuracy (%)	3034	90.35	81.48

3.3. Performance Evaluation of URA-VLMs

Through the training and optimization of the URA-VLMs, we achieved significant improvements based on the optimization of the inference process and prompt engineering. Next, we further evaluated the model's performance from the following three perspectives: road anomaly classification, road waterlogging estimation, and road safety levels.

3.3.1. Model Performance

- Road anomalies classification

After the model optimization, the overall accuracy of the URA-VLMs in road anomalies classification rose from 90.35% to 93.20%. Except for the identification and classification of road garbage (87.64%), the accuracy rates for classifying other road anomaly issues all reached 90% or higher (Table 9).

Table 9. Performance evaluation of URA-VLM before and after optimization.

No.	Types	Number of Data Points (Testing Dataset)	Model Accuracy (%)	
			InternVL-2.5 MPO (Preliminary Classification Prompt)	InternVL-2.5 MPO (Optimized Prompt)
1	Bad road	214	90.19	95.33
2	Fallen tree	296	96.62	96.62
3	Fire	305	96.72	97.05
4	Flooding	310	93.48	92.86
5	Garbage	364	84.62	87.64
6	Normal	675	92.00	92.00
7	Traffic accident	433	92.84	93.53
8	Traffic jam	437	79.18	93.59
	Overall accuracy (%)	3034	90.35	93.20

Identifying road garbage poses greater challenges than detecting other anomalies, leading to lower accuracy rates. The low accuracy in road debris identification is attributed to its diversity, complex backgrounds, dataset limitations, and environmental factors [55]. These challenges are consistent across models, underscoring the need for further research and optimization in this area.

Compared to state-of-the-art road anomaly AI models, URA-VLMs have achieved relatively high accuracy. A YOLOv3 was also employed for pothole detection, achieving a mean Average Precision of 65.65% [56]. A deep learning (DL) framework employing a convolutional neural network (CNN) classified four types of pavement cracks: longitudinal, transverse, alligator, and potholes, achieving an accuracy of 76%, which is considered reasonable compared to binary classification methods [57]. A proposed Mask R-CNN [58,59] was utilized for pixel-wise segmentation of pavement regions to detect potholes and calculate their areas, achieving 90% accuracy in pothole area computation using a manually annotated dataset. The related research exhibits accuracies ranging from 65.65% to 90%, depending on the specific task and dataset.

URA-VLMs have reached a notably high level of accuracy, showcasing better performance in comparison to alternative AI techniques. These large models utilize lightweight architectures and require minimal training data.

- Road waterlogging depth estimation

In assessing water depth within the validation dataset of 310 images, the URA-VLMs recorded a depth error of 33.57%, with a Mean Absolute Error (MAE) of 97.21 mm. The most

significant discrepancy between the URA-VLMs and the annotated values was 470 mm (Table 10, Figure 6).

Table 10. Performance evaluation of URA-VLMs depth estimation before and after optimization.

Number of Data Points (Testing Dataset)	Depth Error (%)		MAE (mm)	
	InternVL-2.5 MPO (Preliminary Classification Prompt)	InternVL-2.5 MPO (Optimized Prompt)	InternVL-2.5 MPO (Preliminary Classification Prompt)	InternVL-2.5 MPO (Optimized Prompt)
310	45.99	33.57	238.79	97.21

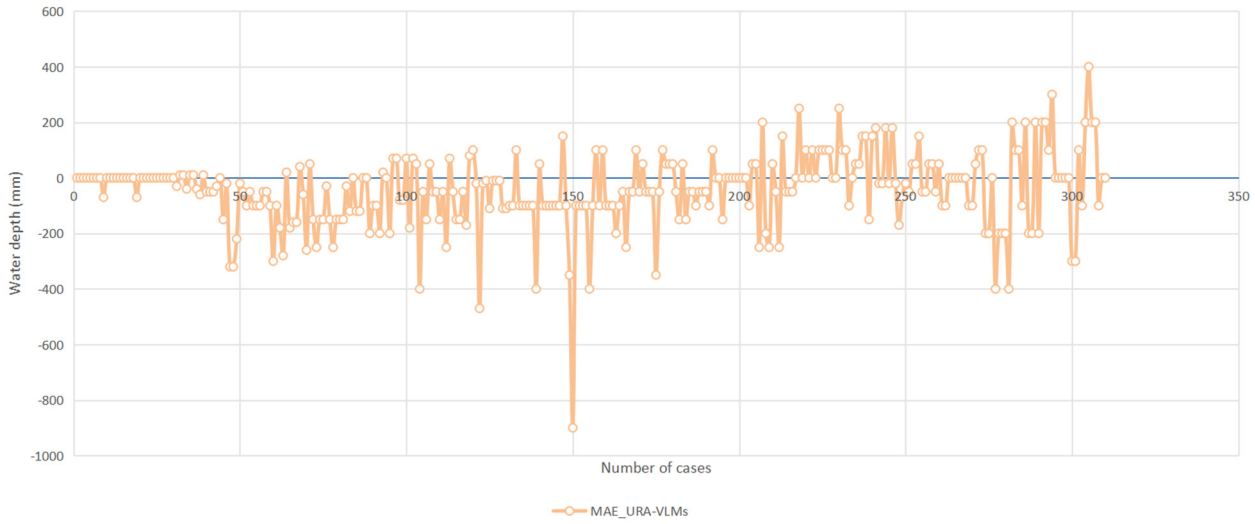


Figure 6. Performance evaluation of URA-VLMs depth Mean Absolute Error estimation after optimization.

Although the MAE indicates reasonable accuracy in depth estimation, it also reveals the potential for significant deviations. Notably, the maximum discrepancy between URA-VLMs and the annotated, highlighting the need for model refinement in certain extreme cases.

In comparison, this finding aligns with previous research and enhances the MAE. The flood-GPT estimations yielded an MAE of 270 mm when compared to manual assessments [41]. Previous studies reported MAE values ranging from 70 mm to 310 mm [13,26], which also employed deep learning methods to estimate floodwater depth from images sourced from online open dataset. The consistency in these results underscores the reliability of AI-driven approaches in flood depth estimation, with our model's performance demonstrating competitive accuracy within the established range of prior studies.

- Road Safety level

In the assessment of model safety levels, it is evident that the URA-VLM demonstrates strong performance, particularly the optimized version. The classification accuracy for the categories GREEN (92.42%), LIGHT GREEN (88.19%), and YELLOW (96.45%) approaches around 90% (Table 11).

Table 11. Performance evaluation of URA-VLMs safety level before and after optimization.

No.	Safety Level	Number of Data Points (Testing Dataset)	Safety Level Accuracy (%)	
			InternVL-2.5 MPO (Preliminary Classification Prompt)	InternVL-2.5 MPO (Optimized Prompt)
1	GREEN	714	91.81	92.42
2	LIGHT GREEN	844	79.86	88.19
3	YELLOW	1476	95.7	96.45

Although the model's performance at the light green level is slightly lower, it has improved by nearly 10 percentage points after optimization compared to its performance before optimization. This improvement can be attributed to the fact that both Garbage and Flooding (10 mm < depth $\leq$ 150 mm) fall within the light green type, which is relatively more challenging. Consequently, this category exhibits the lowest expressiveness.

3.3.2. Real-Time Performance

The real-time tests aim to conduct a basic assessment of the situations that the model may encounter in practical applications under conditions that closely mimic real-life scenarios. Here, we evaluate the response time and the performance of the model in complex environments.

- Response time

Firstly, we conducted tests using the InterVL2.5 27B model on a cloud server and the InterVL2.5 7B model on the edge. The time taken for each step is presented in the following table. On cloud servers, the average time for the entire inference process was 9.3 s, whereas on the edge, it was 22.6 s. Additionally, for edge cloud computing, the average time for the entire inference process was recorded as 12.1 s (Table 12).

Table 12. The response times of the servers.

No.	Tasks	(1) Cloud Computing Response Time (ms) InterVL2.5 26B	Edge Computing Response Time (ms) InterVL2.5 7B	(2) Edge Cloud Computing Response Time (ms)
1	Road anomaly classification	2656.59	5450.09	5450.09
2	Road waterlogging depth estimation	2884.02	8319.64	2884.02
3	Road safety level	3763.04	8903.92	3763.04
	Total	9303.65	22,673.66	12,097.15

- Robustness under diverse environmental conditions

The road anomalies monitoring methods are highly susceptible to the influence of extreme weather [60,61]. To assess robustness under diverse environmental conditions, we conducted basic qualitative tests. Our findings indicate that the time of day—daytime versus nighttime—does not significantly affect the model's judgment, provided that the human annotator can accurately assess the situation.

However, adverse weather conditions such as strong winds and rain do impact the model's judgment. In some cases, despite extreme weather conditions, the collection of video and images remains unaffected, allowing for the acquisition of clear images for inference, thereby not influencing the model's judgment. Conversely, heavy storms or rains

can hinder data collection, as car window glasses may become contaminated by water stains, resulting in compromised data quality that adversely affects the model's judgment.

In conclusion, due to extreme weather conditions, it becomes difficult to identify images. For example, large areas of flare and light pollution not only make it challenging for the model to make judgments but also for humans. This affects the results and tends to cause illusions in the model, making it hard to accurately describe the images. However, changes in lens angles and small areas of image blurring do not significantly impact the model's judgment, and it can still perform normal reasoning.

4. Discussion

This research proposed the Urban Road Anomaly Visual Large Language Models (URA-VLMs), marking substantial progress in the real-time monitoring of urban road anomalies. Based on a pre-trained generative AI model, InternVL 2.5 MPO, the URA-VLMs exhibit remarkable accuracy and flexibility in identifying and evaluating different types of road anomalies, achieving a comprehensive detection accuracy rate of 93.20% in classifying road anomalies. Specifically, for instances of road flooding, the URA-VLMs noted a depth error of 33.57%, alongside a Mean Absolute Error of 97.21 mm. Regarding road safety levels, the URA-VLMs achieved approximately 90% accuracy across the categories GREEN (92.42%), light green (88.19%), and YELLOW (96.45%). This performance surpasses that of leading existing models, such as InternVL 2.5 and ResNet 34, highlighting the potential of generative AI models to enhance urban safety management.

The URA-VLMs framework facilitates real-time identification and classification of urban road anomalies by leveraging a cloud-based system that integrates CCTV cameras with vision large language models. This innovative approach enables continuous monitoring and analysis of road conditions, delivering immediate results to mobile devices. The capabilities are essential for establishing effective early warning systems.

The integration of generative AI marks a significant advancement in road anomaly monitoring, enhancing the system's capacity to promptly detect anomalous events and assist in decision-making processes. In comparison with the other AI models for road anomalies monitoring, the URA-VLMs based on InternVL 2.5 MPO outperformed the fine-tuned ResNet-34 in terms of accuracy across multiple evaluation metrics. Compared to state-of-the-art road anomaly AI models, URA-VLMs have achieved relatively high accuracy. For instance, in the related similar roads anomalies motoring tasks, the YOLOv3 attained a mean Average Precision of 65.65% [56], while the CNN classified four types of pavement cracks, achieving an accuracy of 76% [57]. A proposed Mask R-CNN [58,59] for pixel-wise segmentation of pavement regions achieving 90% accuracy.

The significance of data annotation is critical, particularly in the context of road flood monitoring. Given the scarcity of real depth measurements, data annotation frequently depends on human judgment. To address this challenge, we advocate for a multi-source data evaluation approach that integrates diverse observations to achieve a comprehensive understanding of flood conditions. Establishing consistent standards for calibrating depth estimates using reference objects [41], along with a combination of human and machine assessments, improves reliability [18,41].

The Urban Road Anomaly VLMs have undergone numerous optimizations in the inference processes, employing multi-step reasoning to enhance both efficiency and accuracy. The VLMs consist of three specialized expert models: the Road Anomalies Detect VLM, the Waterlogging Reference Object Detect VLM, and the Waterlogging Depth Describe VLM. These models employ a Chain of Thought (COT) design with multi-step prompts. This methodology facilitates targeted training on specific datasets, thereby enhancing performance and effectively addressing the challenges associated with mixed data training. By

systematically adjusting prompts and optimizing the retrieval process, we can achieve a balance between the accuracy and efficiency of the large language models. This ensures that the models not only produce precise outputs but also operate within a reasonable timeframe, making them suitable for real-time applications.

It is worth noting that URA-VLMs exhibit relatively poor accuracy in recognizing garbage, which needs further discussion. Constructing a large-scale and content-rich garbage dataset requires substantial resources and incurs significant costs. Consequently, zero-shot learning facilitates the transfer of attributes learned from visible classes to invisible classes, enabling the classification of invisible classes without the need for labeled samples. This approach warrants exploration for the long-term usability of garbage classification [62]. Moreover, a distributed framework based on federated learning [63] and a novel three-stage method utilizing Ensemble ELM [64] have demonstrated accuracy exceeding 90% on targeted datasets. The accuracy of image recognition can be significantly improved for specific types of waste. However, in scenarios involving mixed types of waste, further investigation is still required.

These URA-VLMs would be applied in road monitoring or autonomous driving. We conducted tests using lightweight Wi-Fi cameras. In the tests on the A800 cloud server, the response time for one image was an average of 9.3 s. On the edge cloud server, the response time was 12.1 s. These tests verified the feasibility of the system and model design. However, in practical applications, the complexity of the scenarios will still affect the accuracy. Further design and optimization in terms of response time, privacy, and ethics are required.

One crucial aspect demanding particular emphasis is privacy. While AI brings certain privacy advantages, it also presents challenges [65]. In highly recognizable scenarios, AI can assist us in swiftly identifying risks and desensitizing data. However, in numerous areas that remain underexplored, AI harbors long-term latent risks yet to be identified. At least three key issues, like vehicle authorization, privacy protection, and report source validation, should be highlighted in the road anomalies monitoring [66]. AI methods, such as the FDMME, have been utilized for privacy preservation in the detection of road anomaly conditions [67]. The hybrid approach combining federated learning, differential privacy, and homomorphic encryption has been employed to maximize user data privacy without compromising the overall accuracy of the AI model [68].

Moreover, there are also non-AI methods for privacy protection. One such approach is a traffic route management scheme that employs homomorphic encryption and blockchain technology, achieving security objectives and demonstrating good performance [69]. A novel Vehicular Flexible Anonymous System (VFAS) has been developed, enabling vehicles to autonomously generate pseudonyms and keys for enhanced privacy and memory efficiency [70]. AI methodologies provide numerous effective solutions for privacy protection. At the same time, non-AI techniques protect data from various perspectives, including cross-validation and secure data transmission. A wide array of strategies is available.

In light of the ethical risks, it is essential to ensure that algorithm designs adhere to ethical standards. However, a unified ethical standard encompassing transparency, fairness, and privacy has yet to be established [71]. A thorough assessment of the possible dangers that might affect both individuals and society needs to be conducted, along with the formulation of proactive strategies. A complete framework should be established to monitor the full progression of AI development, assessment, and implementation. This includes training staff and enhancing privacy safeguards and ethical evaluation processes. For generative AI, further institutional advancements are essential [72,73].

This study presents a rapid, convenient, and innovative generative model for real-time road anomaly monitoring; however, several limitations should be addressed. First,

while we focused on the InternVL model, there is significant potential to explore other models in related monitoring tasks with different data sources that could further enhance performance [40,74]. Second, our research utilized a small sample size without fine-tuning, indicating that additional fine-tuning with larger datasets may lead to improvements in model accuracy. Although our model demonstrated marked improvements compared to Internals and ResNet34 [74], there remains room for enhancement in garbage detection, Mean Absolute Error and depth accuracy. Third, in the absence of ground truth data, the standards for data annotation need to be established and refined more thoroughly. Finally, the evaluation framework for assessing model accuracy is an area that requires further exploration by experts in the relevant fields. Addressing these limitations will be crucial for advancing road anomaly monitoring technologies and ensuring their effectiveness in practical applications.

As urban road anomaly challenges intensify due to climate change and urbanization, future research in flood monitoring and management should focus on several critical areas: (1) Future studies should consider the combination of diverse data sources, including satellite imagery, IoT sensor data, social media reports, and historical records. This multimodal approach would contribute to enhancing models' performance and provide deeper insights into road anomalies and flood dynamics [15,60]. (2) Research should concentrate on developing explainable AI models, or the combination of models of generative AI and mathematics models that clarify the decision-making processes behind generation [75]. A transparent and logical model inference process will enable stakeholders to better understand and rely on these systems during emergencies. (3) Adaptive learning techniques should be established to allow models to continuously update and improve as new data become available [76]. This capability is essential in a rapidly changing climate, as it enhances the accuracy and robustness of predictions over time [77]. (4) Future systems should incorporate IoT devices and edge computing technologies to strengthen real-time capabilities [19,22].

5. Conclusions

The proposed cloud-based framework, powered by Urban Road Anomaly Visual Large Language Models (URA-VLMs), represents a significant leap forward in real-time urban road monitoring and safety management. This study emphasizes the critical importance of real-time road anomalies for effective road safety management, highlighting its significance for informed disaster response and strategic planning. By leveraging the advanced capabilities of large vision language models, we introduced a novel approach for automatically determining road anomalies, floodwater depths, and safety levels from photographs and videos, utilizing recognized reference objects and tailored instructions.

Our findings demonstrate that optimized vision large language models provide significant advantages over traditional approaches by offering a universal pipeline for urban road analyses, eliminating the need for multiple individual models. This versatility not only streamlines the process but also enhances cost-effectiveness, thereby enabling rapid situational awareness for decision-makers and improving safety within the road environment.

Author Contributions: Conceptualization, H.D. and Y.D.; methodology, H.D. and Y.D.; software, H.D.; validation, H.D., Y.D. and Z.X.; formal analysis, H.D.; data curation, Y.D.; writing—original draft preparation, H.D. and Y.D.; writing—review and editing, Z.X.; visualization, Y.D.; supervision, Z.X.; project administration, Z.X.; funding acquisition, Z.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are unavailable due to privacy or ethical restrictions.

Acknowledgments: During the preparation of this manuscript/study, the authors used DeepSeek-V3 for the purposes of Chinese and English translation. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. The retrieval-augmented generation.

Category	Component/Feature	Height (mm)	Description
Buses, coaches, and large trucks	Overall Height	3000–3500	The total height of the bus
	Wheel Height	900–1000	Height of the bus wheels
Cars	Overall Height	1400–1600	The total height of the car
	Wheel Height	600–800	Height of the car wheels
Bicycle	Bicycle Overall Height	1000	The total height of a bicycle
	Bicycle Pedal Height	120–150	Height of the pedals from the ground,
	Bicycle Tire Height	650–700	Height of the tires
	Bicycle Tire Thickness	30	The thickness of the tire
People (Adult measurement)	Ankle Height	70–90	The height of an adult's ankle
	Knee Height	400–500	Height of an adult's knee
	Hip Height	800–1000	Height of an adult's hip
Other Facilities	Traffic Cone Height	750	Height of a traffic cone used for directing vehicle flow and safety
	Railing Top Rail Height	900–1200	Height of the top rail
	Footpath (Curb/Sidewalk) Height	50–150	Height of curbs and sidewalks
	Base Height of Sign Post	800	Base height of sign post

References

1. WHO. *Global Status Report on Road Safety 2023*; World Health Organization: Geneva, Switzerland, 2023.
2. Silva, N.; Soares, J.; Shah, V.; Santos, M.Y.; Rodrigues, H. Anomaly Detection in Roads with a Data Mining Approach. *Procedia Comput. Sci.* **2017**, *121*, 415–422. [CrossRef]
3. Xing, H.; Yang, F.; Qiao, X.; Li, F.; Huang, X. Enhanced End-to-End Regression Algorithm for Autonomous Road Damage Detection. *J. Supercomput.* **2025**, *81*, 380. [CrossRef]
4. *UNDRR Annual Report (2021)*; United Nations Office for Disaster Risk Reduction: Geneva, Switzerland, 2021.
5. Devitt, L.; Neal, J.; Coxon, G.; Savage, J.; Wagener, T. Flood Hazard Potential Reveals Global Floodplain Settlement Patterns. *Nat. Commun.* **2023**, *14*, 2801. [CrossRef]
6. Lin, Z.; Guan, S.; Zhang, W.; Zhang, H.; Li, Y.; Zhang, H. Towards Trustworthy LLMs: A Review on Debiasing and Dehallucinating in Large Language Models. *Artif. Intell. Rev.* **2024**, *57*, 243. [CrossRef]
7. The Office of the National Disaster Prevention, Reduction and Relief Committee. *Natural Disasters in China in the First Half of 2024*; The Office of the National Disaster Prevention, Reduction and Relief Committee: Beijing, China, 2024.
8. Garcia, V.M.; Granados, R.P.; Medina, M.E.; Ochoa, L.; Mondragon, O.A.; Cheu, R.L.; Villanueva-Rosales, N.; Rosillo, V.M.L. Management of Real-Time Data for a Smart Flooding Alert System. In Proceedings of the 2020 IEEE International Smart Cities Conference (ISC2), Piscataway, NJ, USA, 28 September–1 October 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–8.
9. Li, Z.; Wang, C.; Emrich, C.T.; Guo, D. A Novel Approach to Leveraging Social Media for Rapid Flood Mapping: A Case Study of the 2015 South Carolina Floods. *Cart. Geogr. Inf. Sci.* **2018**, *45*, 97–110. [CrossRef]
10. Matini, N.; Qiao, Y.; Sias, J.E. Development of Time–Depth–Damage Functions for Flooded Flexible Pavements. *J. Transp. Eng. Part B Pavements* **2022**, *148*, 04022011. [CrossRef]

11. Henao Salgado, M.J.; Zambrano Nájera, J. Assessing Flood Early Warning Systems for Flash Floods. *Front. Clim.* **2022**, *4*, 787042. [CrossRef]
12. Kuller, M.; Schoenholzer, K.; Lienert, J. Creating Effective Flood Warnings: A Framework from a Critical Review. *J. Hydrol.* **2021**, *602*, 126708. [CrossRef]
13. Chaudhary, P.; D'Aronco, S.; Leitão, J.P.; Schindler, K.; Wegner, J.D. Water Level Prediction from Social Media Images with a Multi-Task Ranking Approach. *ISPRS J. Photogramm. Remote Sens.* **2020**, *167*, 252–262. [CrossRef]
14. Shah, S.A.; Seker, D.Z.; Hameed, S.; Draheim, D. The Rising Role of Big Data Analytics and IoT in Disaster Management: Recent Advances, Taxonomy and Prospects. *IEEE Access* **2019**, *7*, 54595–54614. [CrossRef]
15. Lopez, T.; Al Bitar, A.; Biancamaria, S.; Güntner, A.; Jäggi, A. On the Use of Satellite Remote Sensing to Detect Floods and Droughts at Large Scales. *Surv. Geophys.* **2020**, *41*, 1461–1487. [CrossRef]
16. Munawar, H.S.; Hammad, A.W.A.; Waller, S.T. Remote Sensing Methods for Flood Prediction: A Review. *Sensors* **2022**, *22*, 960. [CrossRef] [PubMed]
17. Soomro, S.; Boota, M.W.; Zwain, H.M.; Soomro, G.-Z.; Shi, X.; Guo, J.; Li, Y.; Tayyab, M.; Aamir Soomro, M.H.A.; Hu, C.; et al. How Effective Is Twitter (X) Social Media Data for Urban Flood Management? *J. Hydrol.* **2024**, *634*, 131129. [CrossRef]
18. Li, J.; Cai, R.; Tan, Y.; Zhou, H.; Sadick, A.-M.; Shou, W.; Wang, X. Automatic Detection of Actual Water Depth of Urban Floods from Social Media Images. *Measurement* **2023**, *216*, 112891. [CrossRef]
19. Ajim, I. Pathan An IoT and AI Based Flood Monitoring and Rescue System. *Int. J. Eng. Res.* **2020**, *9*, 564–567. [CrossRef]
20. Benoudjit, A.; Guida, R. A Novel Fully Automated Mapping of the Flood Extent on SAR Images Using a Supervised Classifier. *Remote Sens.* **2019**, *11*, 779. [CrossRef]
21. Campolo, M.; Andreussi, P.; Soldati, A. River Flood Forecasting with a Neural Network Model. *Water Resour. Res.* **1999**, *35*, 1191–1197. [CrossRef]
22. Rani, D.S.; Jayalakshmi, G.N.; Baligar, V.P. Low Cost IoT Based Flood Monitoring System Using Machine Learning and Neural Networks: Flood Alerting and Rainfall Prediction. In Proceedings of the 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bangalore, India, 5–7 March 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 261–267.
23. Wang, P.; Hu, Y.; Dai, Y.; Tian, M. Asphalt Pavement Pothole Detection and Segmentation Based on Wavelet Energy Field. *Math. Probl. Eng.* **2017**, *2017*, 1604130. [CrossRef]
24. Banharnsakun, A. Hybrid ABC-ANN for Pavement Surface Distress Detection and Classification. *Int. J. Mach. Learn. Cybern.* **2017**, *8*, 699–710. [CrossRef]
25. Shahabi, H.; Shirzadi, A.; Ghaderi, K.; Omidvar, E.; Al-Ansari, N.; Clague, J.J.; Geertsema, M.; Khosravi, K.; Amini, A.; Bahrami, S.; et al. Flood Detection and Susceptibility Mapping Using Sentinel-1 Remote Sensing Data and a Machine Learning Approach: Hybrid Intelligence of Bagging Ensemble Based on K-Nearest Neighbor Classifier. *Remote Sens.* **2020**, *12*, 266. [CrossRef]
26. Alizadeh Kharazi, B.; Behzadan, A.H. Flood Depth Mapping in Street Photos with Image Processing and Deep Neural Networks. *Comput. Environ. Urban Syst.* **2021**, *88*, 101628. [CrossRef]
27. Prakash, G.; Gupta, P.K.; Rao, G.V.; Pratap, D. Flood Inundation Mapping and Depth Modelling Using Machine Learning Algorithms and Microwave Data. *J. Geomat.* **2021**, *15*, 221–229.
28. Hocini, N.; Payrastre, O.; Bourgin, F.; Gaume, E.; Davy, P.; Lague, D.; Poinsignon, L.; Pons, F. Performance of Automated Methods for Flash Flood Inundation Mapping: A Comparison of a Digital Terrain Model (DTM) Filling and Two Hydrodynamic Methods. *Hydrol. Earth Syst. Sci.* **2021**, *25*, 2979–2995. [CrossRef]
29. Song, S.; Zhang, C.; Zhang, P.; Li, P.; Song, F.; Zhang, L. Urban Waterlogging Detection: A Challenging Benchmark and Large-Small Model Co-Adapter. In Proceedings of the ECCV 2024, Milan, Italy, 29 September–4 October 2024.
30. Munappy, A.R.; Bosch, J.; Olsson, H.H.; Arpteg, A.; Brinne, B. Data Management for Production Quality Deep Learning Models: Challenges and Solutions. *J. Syst. Softw.* **2022**, *191*, 111359. [CrossRef]
31. Iivanainen, S.; Lagus, J.; Viertolahti, H.; Sippola, L.; Koivunen, J. Investigating Large Language Model (LLM) Performance Using in-Context Learning (ICL) for Interpretation of ESMO and NCCN Guidelines for Lung Cancer. *J. Clin. Oncol.* **2024**, *42*, e13637. [CrossRef]
32. Sony, S.; Laventure, S.; Sadhu, A. A Literature Review of Next-Generation Smart Sensing Technology in Structural Health Monitoring. *Struct. Control Health Monit.* **2019**, *26*, e2321. [CrossRef]
33. Sony, S.; Dunphy, K.; Sadhu, A.; Capretz, M. A Systematic Review of Convolutional Neural Network-Based Structural Condition Assessment Techniques. *Eng. Struct.* **2021**, *226*, 111347. [CrossRef]
34. Haldar, A.; Al-Hussein, A. *Recent Developments in Structural Health Monitoring and Assessment—Opportunities and Challenges*; World Scientific: Singapore, 2022; ISBN 978-981-12-4300-4.
35. Shahriar, S.; Lund, B.D.; Mannuru, N.R.; Arshad, M.A.; Hayawi, K.; Bevara, R.V.K.; Mannuru, A.; Batool, L. Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency. *Appl. Sci.* **2024**, *14*, 7782. [CrossRef]
36. Floridi, L.; Chiriatti, M. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds Mach.* **2020**, *30*, 681–694. [CrossRef]

37. Friha, O.; Amine Ferrag, M.; Kantarci, B.; Cakmak, B.; Ozgun, A.; Ghoualmi-Zine, N. LLM-Based Edge Intelligence: A Comprehensive Survey on Architectures, Applications, Security and Trustworthiness. *IEEE Open J. Commun. Soc.* **2024**, *5*, 5799–5856. [CrossRef]
38. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. A Survey of Large Language Models. *arXiv* **2024**, arXiv:2303.18223.
39. Feng, Y.; Brenner, C.; Sester, M. Flood Severity Mapping from Volunteered Geographic Information by Interpreting Water Level from Images Containing People: A Case Study of Hurricane Harvey. *ISPRS J. Photogramm. Remote Sens.* **2020**, *169*, 301–319. [CrossRef]
40. Shafiq, S.; Awan, H.M.; Khan, A.A.; Amin, W. Driving Like Humans: Leveraging Vision Large Language Models for Road Anomaly Detection. In Proceedings of the 2024 3rd International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering (ETECTE), Lahore, Pakistan, 26–27 November 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1–6.
41. Akinboyewa, T.; Ning, H.; Lessani, M.N.; Li, Z. Automated Floodwater Depth Estimation Using Large Multimodal Model for Rapid Flood Mapping. *Comput. Urban Sci.* **2024**, *4*, 12. [CrossRef]
42. Li, Z.; Ning, H. Autonomous GIS: The next-Generation AI-Powered GIS. *Int. J. Digit. Earth* **2023**, *16*, 4668–4686. [CrossRef]
43. Wang, W.; Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Zhu, J.; Zhu, X.; Lu, L.; Qiao, Y.; et al. Enhancing the Reasoning Ability of Multimodal Large Language Models via Mixed Preference Optimization. *arXiv* **2024**, arXiv:2411.10442.
44. Cian, F.; Marconcini, M.; Ceccato, P.; Giupponi, C. Flood Depth Estimation by Means of High-Resolution SAR Images and Lidar Data. *Nat. Hazards Earth Syst. Sci.* **2018**, *18*, 3063–3084. [CrossRef]
45. Beijing Water Authority. *Beijing Urban Waterlogging Risk Map Report*; Beijing Water Authority: Beijing, China, 2022.
46. *GB 51222-2017*; Technical Specification for Urban Waterlogging Prevention and Control. Ministry of Housing and Urban-Rural Development of the People's Republic of China: Beijing, China, 2017.
47. *GB 50014-2021*; Outdoor Drainage Design Standards. Ministry of Housing and Urban-Rural Development of the People's Republic of China: Beijing, China, 2021.
48. Tian, Y.; Wang, Q.; Guo, Z.; Zhao, H.; Khan, S.; Mao, W.; Yasir, M.; Zhao, J. A Hybrid Deep Learning and Ensemble Learning Mechanism for Damaged Power Line Detection in Smart Grids. *Soft Comput.* **2022**, *26*, 10553–10561. [CrossRef]
49. Sarda, A.; Dixit, S.; Bhan, A. Object Detection for Autonomous Driving Using YOLO [You Only Look Once] Algorithm. In Proceedings of the 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, 4–6 February 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1370–1374.
50. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
51. Bharati, P.; Pramanik, A. Deep Learning Techniques—R-CNN to Mask R-CNN: A Survey. In *Computational Intelligence in Pattern Recognition*; Springer: Singapore, 2020; pp. 657–668.
52. Huang, Y.; Chen, J.; Huang, D. UFPMP-Det:Toward Accurate and Efficient Object Detection on Drone Imagery. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 1026–1033. [CrossRef]
53. Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024.
54. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked Autoencoders Are Scalable Vision Learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
55. Pytorch Team. Deep Residual Networks Pre-Trained on ImageNet. Available online: https://pytorch.org/hub/pytorch_vision_resnet/ (accessed on 19 February 2025).
56. Yik, Y.K.; Alias, N.E.; Yusof, Y.; Isaak, S. A Real-Time Pothole Detection Based on Deep Learning Approach. *J. Phys. Conf. Ser.* **2021**, *1828*, 012001. [CrossRef]
57. Aslan, O.D.; Gultepe, E.; Ramaji, I.J.; Kermanshachi, S. Using Artifical Intelligence for Automating Pavement Condition Assessment. In Proceedings of the International Conference on Smart Infrastructure and Construction 2019 (ICSIC), Cambridge, UK, 8–10 July 2019; ICE Publishing: London, UK, 2019; pp. 337–341.
58. Arjapure, S.; Kalbande, D.R. Deep Learning Model for Pothole Detection and Area Computation. In Proceedings of the 2021 International Conference on Communication Information and Computing Technology (ICCICT), Mumbai, India, 25–27 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–6.
59. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
60. Bai, Z.; Wang, Y.; Zhang, A.; Wei, H.; Pan, G. Road Surface Condition Monitoring in Extreme Weather Using a Feature-Learning Enhanced Mask–RCNN. *J. Transp. Eng. Part B Pavements* **2024**, *150*, 04024030. [CrossRef]

61. Jiang, J.; Xu, G.; Wang, H.; Yang, Z.; Sun, B.; Guan, C.; Feng, J.; Ma, Y.; Chen, X. High-Accuracy Road Surface Condition Detection through Multi-Sensor Information Fusion Based on WOA-BP Neural Network. *Sens. Actuators A Phys.* **2024**, *378*, 115829. [CrossRef]
62. Li, J.; Chen, X.; Zhang, A.; Li, C.; Lin, D. Survey of Garbage Classification Methods Based on Deep Learning. *Comput. Eng.* **2022**, *48*, 4–9.
63. Ahmed Khan, H.; Naqvi, S.S.; Alharbi, A.A.K.; Alotaibi, S.; Alkhathami, M. Enhancing Trash Classification in Smart Cities Using Federated Deep Learning. *Sci. Rep.* **2024**, *14*, 11816. [CrossRef] [PubMed]
64. Nahiduzzaman, M.; Ahamed, M.F.; Naznine, M.; Karim, M.J.; Kibria, H.B.; Ayari, M.A.; Khandakar, A.; Ashraf, A.; Ahsan, M.; Haider, J. An Automated Waste Classification System Using Deep Learning Techniques: Toward Efficient Waste Recycling and Environmental Sustainability. *Knowl.-Based Syst.* **2025**, *310*, 113028. [CrossRef]
65. Martin, K.D.; Zimmermann, J. Artificial Intelligence and Its Implications for Data Privacy. *Curr. Opin. Psychol.* **2024**, *58*, 101829. [CrossRef]
66. Wang, Y.; Ding, Y.; Wu, Q.; Wei, Y.; Qin, B.; Wang, H. Privacy-Preserving Cloud-Based Road Condition Monitoring With Source Authentication in VANETs. *IEEE Trans. Inf. Forensics Secur.* **2019**, *14*, 1779–1790. [CrossRef]
67. Narayanan, K.L.; Naresh, R. Privacy-Preserving Dual Interactive Wasserstein Generative Adversarial Network for Cloud-Based Road Condition Monitoring in VANETs. *Appl. Soft Comput.* **2024**, *154*, 111367. [CrossRef]
68. Krishna, S.; Siri, S.; Kamalsha, S.; Amruth, S.; Jadon, S. PRIVATE-AI: A Hybrid Approach to Privacy-Preserving AI. In Proceedings of the 2023 IEEE/ACIS 8th International Conference on Big Data, Cloud Computing, and Data Science (BCD), Ho Chi Minh City, Vietnam, 14–16 December 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 170–175.
69. Zhang, J.; Fang, H.; Zhong, H.; Cui, J.; He, D. Blockchain-Assisted Privacy-Preserving Traffic Route Management Scheme for Fog-Based Vehicular Ad-Hoc Networks. *IEEE Trans. Netw. Serv. Manag.* **2023**, *20*, 2854–2868. [CrossRef]
70. Cheng, H.; Yang, J.; Shojafar, M.; Cao, J.; Jiang, N.; Liu, Y. VFAS: Reliable and Privacy-Preserving V2F Authentication Scheme for Road Condition Monitoring System in IoV. *IEEE Trans. Veh. Technol.* **2023**, *72*, 7958–7972. [CrossRef]
71. Radanliev, P. AI Ethics: Integrating Transparency, Fairness, and Privacy in AI Development. *Appl. Artif. Intell.* **2025**, *39*, 2463722. [CrossRef]
72. Ye, X.; Yan, Y.; Li, J.; Jiang, B. Privacy and Personal Data Risk Governance for Generative Artificial Intelligence: A Chinese Perspective. *Telecommun. Policy* **2024**, *48*, 102851. [CrossRef]
73. Saeed, M.M.; Alsharidah, M. Security, Privacy, and Robustness for Trustworthy AI Systems: A Review. *Comput. Electr. Eng.* **2024**, *119*, 109643. [CrossRef]
74. Aravindkumar, S.; Varalakshmi, P.; Alagappan, C. Automatic Road Surface Crack Detection Using Deep Learning Techniques. In *Artificial Intelligence and Technologies*; Springer: Singapore, 2022; pp. 37–44.
75. Ali, S.; Abuhmed, T.; El-Sappagh, S.; Muhammad, K.; Alonso-Moral, J.M.; Confalonieri, R.; Guidotti, R.; Del Ser, J.; Díaz-Rodríguez, N.; Herrera, F. Explainable Artificial Intelligence (XAI): What We Know and What Is Left to Attain Trustworthy Artificial Intelligence. *Inf. Fusion* **2023**, *99*, 101805. [CrossRef]
76. Kabudi, T.; Pappas, I.; Olsen, D.H. AI-Enabled Adaptive Learning Systems: A Systematic Mapping of the Literature. *Comput. Educ. Artif. Intell.* **2021**, *2*, 100017. [CrossRef]
77. Yan, F.; Zhang, H.; Li, Y.; Yang, Y.; Liu, Y. End-to-End: A Simple Template for the Long-Tailed-Recognition of Transmission Line Clamps via a Vision-Language Model. *Appl. Sci.* **2023**, *13*, 3287. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.