scientific reports

OPEN

Ensemble deep learning with image captioning for visual pollution detection, classification, and reporting

Haya Almalki[1] & Nahlah Algethami[2]

With the rapid urban development and initiatives such as Saudi Vision 2030, efforts have been directed toward improving services and quality of life in Saudi cities. As a result, multiple environmental challenges have emerged, including visual pollution (VP), which significantly impacts the quality of life. Current approaches to these challenges rely on reporting through an online application managed by the Ministry of Municipalities and Housing, which is prone to errors due to manual data entry. This study proposes an AI-driven framework that integrates deep learning models (YOLOv5 and EfficientDet), along with ensemble techniques. Additionally, the study proposes using Bootstrapping Language-Image Pre-training (BLIP) to automatically generate text descriptions based on the content of images in reports. This framework was developed using the public dataset "Saudi Arabia Public Roads Visual Pollution Dataset" from Mendeley. This study is the first to combine the results of the YOLOv5 and EfficientDet models to detect VP and automatically generate descriptions using BLIP-2, thereby facilitating the production of citizen-monitored reports. The proposed system aims to improve decision-making, reduce errors, and enhance urban management by automating the detection, classification, and reporting of VP. This ensemble approach achieved a Mean Average Precision (mAP) of 0.95, a recall of 0.95, a precision of 0.91, and an F1 score of 0.93, surpassing the performance of the individual models. In image captioning, the "BLIP2-Flan-T5-XL" model achieved an accuracy 80% based on human evaluation, demonstrating the effectiveness of AI-generated text in urban reporting. This suggests that the system could help automate VP reporting and improve reporting accuracy, thereby contributing to more sustainable cities.

Keywords Visual pollution, Ensemble, EfficientDet, YOLOv5, Image captioning

In recent years, the world has witnessed unprecedented transformations in the transition toward sustainable and smart urban cities. This transformation has led to several advancements, including the Quality of Life Program, one of the key initiatives under Saudi Vision 2030, which aims to enhance the urban landscape of Saudi cities and improve the quality of services provided[1]. To achieve this, it is essential to tackle the challenges of environmental pollution that endangers both the environment and human health[2]. Previous studies have identified environmental pollution, including noise, air, and water pollution, as one of the most serious dangers to human health, with significant impacts on the quality of life (Gao et al.[3]). While water and air pollution represent significant environmental challenges, noise pollution—caused by transportation, urban expansion, and industrial activities—is often harder to control, particularly amid rapid urban growth. In addition to these traditional forms of pollution, a new concept called visual pollution (VP) has recently emerged, which often comes from the same sources as noise pollution, such as uncontrolled urban growth, chaotic construction, and excessive advertising[4].

VP refers to the distortion of a city's appearance, which may degrade its aesthetics, hinder concentration, and obstruct visibility and mobility in public roads. In other words, VP is a form of sensory pollution that interferes with the senses and has been relatively ignored despite its significant impact[5]. In addition, VP can also affect mental health by causing negative feelings, leading to modes, including anxiety, fear, and insecurity (Gao et al. 2024). It includes many types like graffiti, potholes, bad streetlights, broken signage, faded signage,

[1]Department of Computer Science, Saudi Electronic University, Riyadh, Saudi Arabia. [2]Computer Science Department, College of Computing and Informatics, Saudi Electronic University, 11673 Riyadh, Saudi Arabia. email: n.algethami@seu.edu.sa

and construction roads. Addressing VP is essential for improving the quality of life in urban areas and achieving citizen satisfaction[2].

In 2024, the statistics from the Ministry of Municipalities and Housing in the Kingdom of Saudi Arabia indicate that approximately 1.2 million VP-related reports had been addressed[6]. Currently, the process of identifying and reporting VP relies on citizens submitting manual reports via a mobile application operated by the ministry. This submission involves uploading an image, selecting a category, and writing a description. However, this manual process is inefficient, fails to utilize modern technology, and does not meet the evolving needs of urban management. Additionally, submitting reports while driving poses safety risks.

In the digital era, new technologies have become essential in the decision-making process. The integration of artificial intelligence (AI) and new technologies has become essential to efficiently manage smart and sustainable cities. Recent studies have used AI to address VP in Saudi Arabia. For example, (AlElaiwi, Al-antari et al.[7]) proposed a framework based on deep learning techniques to detect and classify VP. Also, (Mazroa et al.[4]) developed a computer vision model with explainable AI to detect VP. However, existing research has not addressed automatically generating a description of VP from the image. Automating image description with VP detection is essential for providing a comprehensive and accurate report without human intervention.

In this study, we propose to develop a framework that integrates deep learning, ensemble techniques, and image captioning. With deep learning, advanced models such as YOLO (You Only Look Once), and EfficientDet can be created to detect and classify types of VP from images, including potholes, excavation barriers, and dilapidated sidewalks. The ensemble technique is used to enhance the model's performance. Finally, image captioning using Bootstrapping Language-Image Pre-training (BLIP) will use the image and the model's output to generate a description of the detected type of pollution. The integration of these techniques facilitates a comprehensive and more accurate framework for managing VP in urban environments.

To support government organizations in addressing VP more efficiently, this work proposes an AI-based framework that not only detects and classifies pollution but also generates automatic descriptions based on the image. Unlike previous studies that focused on detection alone, our approach integrates deep learning and image captioning into a single framework, offering a fully automated solution that reduces manual effort and improves decision-making.

The key contributions of this study are:

- Development of a framework that integrates deep learning, ensemble techniques, and image captioning to detect, classify, and describe VP from images.
- Implementation of an ensemble approach by combining YOLO and EfficientDet models to improve detection accuracy and robustness.
- Integration of BLIP for automated description generation of detected pollution based on model results and images.
- Development of an automated reporting system that fills out reports based on images to reduce errors in the reporting process and improve decision-making.

The rest of this paper is structured as follows. Section "Literature review" elaborates on related studies and previous work. Section "Methodology" explains the technical aspects in detail. Section "Results and discussion" provides the analysis of findings and sub-conclusions. Finally, Sect. "Conclusion" concludes the study and suggests future work.

Literature review

Overview

VP is a modern environmental phenomenon that has emerged as a result of rapid urban expansion and industrialization, leading to the introduction of inconsistent elements that distort the visual landscape of cities. This pollution includes excessive outdoor advertising, disorganized structures, poor infrastructure planning, and random modifications to buildings that lack architectural harmony. Its impact also extends to affect mental health, as it contributes to increased levels of stress and visual fatigue, thereby reducing the quality of life. Urban areas and commercial centers are among the environments most affected by this problem. In addition, neglect and mismanagement of public and natural spaces contribute to the exacerbation of the problem, making VP an increasing environmental challenge that requires regulatory strategies and advanced technologies to reduce its impact and improve the quality of the urban landscape[3].

In the next section, we will review existing studies related to our study, divided into three main parts. The first part covers studies on the detection and classification of VP, focusing on different approaches, methodologies, and their results. The second part highlights conducted in Saudi Arabia. The third part covers studies that apply image captioning techniques and their role in generating meaningful descriptions.

Visual pollution detection and classification techniques

VP is classified as a global problem that adversely affects the visual appeal and quality of life in urban environments. Research in recent years has focused on the detection and classification of VP, exploring innovative techniques for identifying visual distortions in urban areas.

Earlier studies used manual methods like crowdsourcing or 3D modeling to analyze VP. Although helpful, these methods faced limitations, including high cost, limited scalability, and insufficient attention to personal or cultural differences. The study byKucharikova & Simko used crowdsourcing techniques to detect VP from outdoor advertising by collecting user-submitted photos and reports[8]. These were analyzed with clustering algorithms and image processing techniques, with Mean Squared Error (MSE) and Mean Error (ME) to evaluate accuracy. The results indicated that the combination of crowdsourcing and visual analysis improves the accuracy of VP identification.

A different approach was presented in another study, which applied geometric analysis using 3D modeling techniques such as 3D Isovist and Voxel to assess the impact of outdoor billboards on urban scenery[9]. The study assessed how billboards influence visual openness in urban areas using 3D simulation models and environmental data. Results showed that large billboards significantly reduce visibility and negatively affect the aesthetic experience for pedestrians and drivers. While 3D modeling provides accurate measurements, it requires costly spatial data and overlooks psychological and cultural differences in how VP is perceived.

Recently, with the advancement of technologies, many studies have focused on addressing VP by using deep learning models to detect and classify it from images. In Bangladesh, Ahmed et al.[10] used a Convolutional Neural Network (CNN) to detect and classify four types of VP: billboards, wires, towers, and street waste. They trained the model on a dataset of 800 images collected from Google Image Search, applying manual filtering and augmentation to enhance the dataset. The model achieved 95% training accuracy and 85% validation accuracy. The study highlights the importance of expanding datasets to include diverse geographic regions, thereby enhancing the model's generalizability and applicability across various environments.

Another study by Hossain et al.[11] proposed an automated approach to detect VP in Dhaka using Google Street View images and the YOLOv5 model. Unlike studies using general online images, they created a custom dataset of 1400 images covering six pollution types. The data, labeled with CVAT, was split into an 80/20 ratio for training and testing. The model achieved 78% for both precision and recall, and 82% Mean Average Precision (mAP). However, the performance was limited due to the limited dataset size and lack of diversity, highlighting the importance of increasing the size of the dataset and diversifying its sources.

In addition, many research papers have discussed road damage—considered a type of VP—with the objective of enhancing road safety and maintenance, thereby reducing VP. For example, Pham et al.[12] used the YOLOv7 model in detecting and classifying road damage. Their study aimed to improve detection accuracy by the integration of Coordinate Attention layers and ensemble learning. Three versions of YOLOv7 were developed: (1) a baseline model with updated image augmentation, (2) an enhanced model incorporating three Coordinate Attention layers in the detection head, and (3) an optimized model with additional Coordinate Attention layers in the backbone. The researchers improved detection accuracy by using an ensemble method that merged the predictions from the three top models. The study used the Crowdsensing-based Road Damage Detection Challenge (CRDDC2022) dataset[12] which includes road images from different countries like the United States, India, Japan, and Norway. It aimed to identify four types of road damage: longitudinal cracks, transverse cracks, alligator cracks, and potholes. The model was assessed using the F1-score, achieving 81.7% on U.S. Road damage data and 74.1% on all test images. However, challenges like differences in the dataset and varying image quality impacted the overall results[13]. Another study proposed by Pramanik et al.[14] used a custom dataset containing 1490 images with two classes (Plain Road and Pothole Road) collected from various cities in Bangladesh and applied classification using pre-trained models: VGG16, and ResNet50. The models were evaluated using accuracy, precision, recall, F1-score, and specificity. ResNet50 model was the best with an accuracy of 98.66%. It outperformed VGG16 due to its deeper architecture with residual connections, which enhance feature extraction for better accuracy in pothole detection. The study suggested increasing the dataset size and utilizing real-time models like YOLO to enhance results.

A pothole detection method in real-time using the YOLOv8 model was developed by Mirajkar et al.[15]. The dataset in the study was obtained from the Kaggle website and contained 200 pothole images, divided into two classes: potholes and non-potholes. It was split into 128 images for training, 44 images for validation, and 28 images for testing. The results were a 78.27% mAP, 80.55% precision, 73.17% recall, and 76.68% F1 score. The research gaps found are expanding the dataset and evaluating real-time deployment in different urban environments.

Similarly, the study (Kumari et al.[16]) proposed a pothole detection system utilizing YOLOv8. The study used the "Pothole Detection Image Dataset" from Roboflow (Alahmari n.d.), which contains of 8000 images. The dataset was split into 70% for training, 15% for validation, and 15% for testing. Multiple YOLOv8 versions (YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, YOLOv8x) were trained and evaluated using metrics like precision, recall, mAP@0.5, and mAP@0.5:0.95. The results indicated that YOLOv8l and YOLOv8x outperformed the others, with mAP@0.5 scores of 78.7% and 78.5%, respectively.

On the other hand, some studies have focused on solutions for textile-based VP. In (Tasnim et al.), an automated system to classify textile visual pollutants in Bangladesh using deep learning[17]. The dataset included 1709 images of cloth garbage, billboards, and dyeing waste, collected from search engines and local sources, and annotated using LabelImg and Roboflow. Three models were evaluated: EfficientDet, Faster R-CNN, and YOLOv5. EfficientDet achieved the highest accuracy with 97% in training and 93% in testing, using fewer parameters, while YOLOv5 showed competitive results with faster training, making it suitable for real-time applications. The study recommended improving performance by applying advanced preprocessing and expanding the dataset.

Titu et al.[18] proposed a deep learning-based system for real-time VP detection in Dhaka using Google Street View and robotic vision. The system targeted nine types of urban and textile pollutants and evaluated four models: YOLOv5, YOLOv7, Faster SegFormer, and EfficientDet. YOLOv5 performed best with 98% accuracy, 71% mAP, 50% Intersection over Union (IoU), 70% precision, and 67% recall. It was deployed on a Raspberry Pi 4B for real-time use. The study recommended enhancing dataset diversity and optimizing models for edge devices.

Table 1 shows a summary of studies that applied deep learning to the detection and classification of visual.

Year	Author	Dataset	Number of classes	Model used	Result
2019	Ahmed et al.[10]	800 images from the Google image Search engine	4 (billboards, electrical and communication wires, communication towers, and street waste)	CNN	95% training accuracy and 85% validation accuracy
2021	Hossain et al.[11]	1400 images from Google street view	6 (billboards, street litters, construction materials, wires, bricks, and towers)	Yolov5	82% mAP, 78% recall, and 78% precision
2022	Pham et al.[13]	26,336 from from Google street view	4 (longitudinal cracks, transverse cracks, alligator cracks, and potholes)	YOLOv7	74.1% F1-score
2021	Pramanik et al.[14]	1490 images from a custom dataset	2 (potholes and plain roads)	VGG16, ResNet50	ResNet50 with 98.66% accuracy
2024	Mirajkar et al.[15]	200 images from Kaggle	2 (potholes and non-potholes)	Yolov8	78.27% mAP. 80.55% precision, 73.17% recall, and 76.68% f1 score
2023	Kumari et al.[16]	8000 images from Roboflow	2 (potholes and non-potholes)	Yolov8	79% mAP50, 50% mAP50-95, 83% Precision, and 73% Recall
2023	Tasnim et al.[17]	1709 images were collected from search engines and local textile factories and streets	3 (cloth garbage, textile billboards and textile dye)	Faster R-CNN, YOLOv5, and EfficientDet	The EfficientDet with 97% and 93% training and test accuracies
2024	Titu et al.[18]	5653 images collected from web scraping	9 (billboards, bricks, construction materials, street litters, towers, wires, industrial clothes dump, industrial textile billboard, and industrial textile dye)	Faster SegFormer, YOLOv5, YOLOv7, and EfficientDet	YOLOv5 with 98% Accuracy

Table 1. Summary of studies for VP detection and classification.

Visual pollution detection and classification techniques in Saudi Arabia

Despite the growing focus on urban sustainability and smart city initiatives, VP remains insufficiently explored. Due to the subjective nature of VP and the variations in its perception between cultures and regions, this section sheds light on recent studies that use AI techniques to detect and classify VP in Saudi Arabia.

AlElaiwi, Al-antari et al.[19] in their study, proposed a framework for detecting and classifying visual pollutants using a deep active learning approach. They applied their framework using a dataset consisting of 34,460 images with three classes: excavation barriers, potholes, and dilapidated sidewalks, collected from the Ministry of Municipal and Rural Affairs and Housing (MOMRAH) in Saudi Arabia[5]. In their approach, they implemented multiple data preprocessing steps to enhance model performance, including the manual removal of irrelevant images by experts, normalization of pixel intensity values to a standard range of [0, 255], resizing all images to 460 × 600 pixels, and splitting the dataset into 70% for training, 10% for validation, and 20% for testing. They compared five models for object detection: MobileNetSSDv2, EfficientDet, Faster R-CNN, Detectron2, and YOLO. The models were evaluated using training loss, validation loss, recall, precision, and mAP. YOLO achieved the best results, with 89% precision, 88% recall, 89% F1-score, and 93% mAP. Due to the limited number of classes in the dataset, the researchers suggest expanding VP classes to improve model performance. They also suggested integrating Explainable AI (XAI) to better interpret model outputs.

Mazroa et al.[4] proposed the Modeling of Computer Vision with Explainable Artificial Intelligence for Visual Pollution Detection (MCVXAI-VPD) framework, which applies deep learning techniques for detecting and classifying VP in Saudi Arabia. They used eight classes from a Kaggle dataset and applied preprocessing techniques such as mean filter-based noise removal and Contrast Limited Adaptive Histogram Equalization (CLAHE) for contrast enhancement. The framework integrated YOLOv5 for object detection, Bidirectional Long Short-Term Memory (BiLSTM) for classification, PDO (Prairie Dog Optimization) for hyperparameter tuning, and LIME (Local Interpretable Model-agnostic Explanations) for explainability. The model achieved 98.20% accuracy, 90.74% precision, 82.48% recall, and 85.09% F1-score. While the study focused on the detection and classification of VP, the current study extends this by integrating image captioning to enhance report automation and assist in minimizing human errors.

On the other hand, Rafe Alamri et al.[20] developed an automated system to detect abandoned vehicles in Riyadh using deep learning and remote sensing images. They used WorldView-3 satellite imagery collected over three months in 2020 and applied a Mask R-CNN model for vehicle detection. The process included data collection, preprocessing, training, and temporal analysis to identify abandoned vehicles based on a 50% overlap across time. The model achieved 95% accuracy, 91% recall, 93% F1-score, and 87% mAP.

Table 2 shows a summary of related work on the detection and classification of VP in Saudi Arabia.

Image captioning techniques

As technology continues to evolve, several studies have been conducted to develop techniques and AI models to generating image captions that mimics human writing. Image captioning is the process of generating a description of an image. It is one of the most innovative models in the neural networks field.

In past years, traditional approaches relied on using two models, one to extract features from images and another model to generate sequences of words. For example, a deep learning-based image captioning model was developed using Convolutional Neural Networks (CNN) as an encoder to extract image features and Long Short-Term Memory (LSTM) as a decoder to generate text descriptions[21]. The model was trained on the Flickr8k dataset[22], where CNN focused on capturing the visual aspects of images, while LSTM processed these features to generate meaningful captions. The model's performance was evaluated using the Bilingual Evaluation

Year	Author	Dataset	Number of classes	Model Used	Result
2022	AlElaiwi et al.[5]	34,460 images from MOMRAH in Saudi Arabia	3 (excavation barriers, potholes, and dilapidated sidewalks)	MobileNetSSDv2, EfficientDet, Faster R-CNN, Detectron2, and YOLO	Yolov5 with 89% precision, 88% recall, 89% F1-score, and 93% mAP
2024	Mazroa et al.[4]	Open dataset "the Urban visual pollution dataset" from Kaggle	8 (graffiti, faded signage, garbage, construction road, bad streetlight, sand on road, clutter sidewalk, unkept facade)	YOLOv5, BiLSTM, and the XAI LIME model	98.20% accuracy, 90.74% precision, 82.48% recall, and 85.09% F1-score
2024	Rafe Alamri et al.[19]	WorldView-3 satellite images	2 (abandoned and non-abandoned)	Mask-RCNN	95% Accuracy, 91% Recall, 93% F1-score, and 87% mAP

Table 2. Summary of studies for VP detection and classification in Saudi arabia.

Understudy (BLEU) metric. Regarding architectural complexity, future improvements could involve adopting advanced models like ResNet or GoogleNet and integrating bidirectional LSTMs to enhance caption accuracy.

Chen et al.[23] introduced an attribute-driven attention model for image captioning using a CNN-RNN framework. The study used the Microsoft COCO dataset[24] and integrated a visual attention mechanism to enhance feature extraction and maintain attribute dependencies. The model consisted of three main components: (1) an inference module for attribute prediction, (2) an attention module for feature selection, and (3) a generation module for sentence construction. It was evaluated using BLEU, ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation – Longest Common Subsequence (LCS)), and CIDEr (Consensus-based Image Description Evaluation) scores. To improve caption generation, need to expand the dataset and improve attribute inference[23].

Recently, vision and language models (VLMs) have advanced significantly. VLM integrates two models: a deep learning model to extract features from images and a natural language model to generate text to enhance image understanding and captioning. Ghosh et al.[25] classified VLMs into three categories based on their input and output capabilities. The first category, Vision-Language Understanding (VLU) models, such as Contrastive Language-Image Pre-Training (CLIP), process only visual input for comprehension tasks. The second category, Text Generation with Multimodal Input models, like Large Language and Vision Assistant (LLaVA) and BLIP-2, generates textual descriptions from both visual and textual inputs. The final category, Multimodal Output with Multimodal Input models, such as Compositional Diffusion for Multimodal Generation (CoDi), produces outputs in multiple formats, including text and images, by processing multimodal inputs.

According toWu et al.[26] who conducted a survey on Multimodal Large Language Models, multiple models were compared based on their functionality and technical components. The study discussed the strengths and limitations of different models. Among them, BLIP-2 was found to be very effective for generating text from images without prior training. This is because it combines pre-trained image encoders, a querying transformer, and frozen large language models. These features make BLIP-2 a great option for this study to generate text descriptions of VP images and support report automation.

Anagnostopoulou et al.[27] explored the role of large language models (LLMs) and large multimodal models (LMMs) in enhancing journalism by generating captions for news images. Two main approaches were used: the first, Contextualized Image Captioning (CIC), is a two-step process where an image description is first generated using an image captioning model like BLIP-2, then refined by a language model such as GPT-3 (Generative Pre-trained Transformer-3) or LLaMA (Large Language Model Meta AI) to incorporate contextual information. The second approach relies on LMMs like BLIP-2, GPT-4v, and LLaVA, which analyze images and text together to generate an integrated caption. The results indicated that using named entities, like people's names and locations, was more effective than providing the full article as context, leading to improved caption quality. Furthermore, BLIP-2 with GPT-3 achieved the highest accuracy with Bidirectional Encoder Representations from Transformers Score (BERTScore) of 0.898.

The study by Zhoul et al.[28] investigates the use of pre-trained large language-image models for medical image captioning, focusing on generating accurate and contextually relevant descriptions for medical images. The researchers utilized BLIP-2 with a vision transformer - giant (ViT-g) and the Open Pre-trained Transformer (OPT2.7B) as the foundation model. The model performed well in the ImageCLEFmedical Caption 2023 competition[29] ranking second overall with a BERTScore of 0.6281. It also achieved top results in Bilingual Evaluation Understudy with Representations from Transformers (BLEURT) of 0.3209 and Consensus-based Image Description Evaluation (CIDEr) of 0.2377, which indicates strong and accurate performance in medical image descriptions.

Jeon et al.[30] in their study, explore the use of image captioning to improve risk assessment in surveillance systems. The study presents a novel approach where BLIP-2 is tuned to generate descriptive captioning detailing objects, behaviors, and spatial context within surveillance footage. These captions are then analyzed using BERT to classify scenes into seven different risk levels. This approach improves surveillance monitoring by providing clear descriptions of activities and assisting in risk evaluation. The study needs to incorporate more data sources and enable real-time processing to improve accuracy.

Conclusion

Based on these reviews, deep learning techniques, especially object detection models such as YOLOv5 and EfficientDet, have demonstrated high accuracy and speed in detecting and classifying VP. Therefore, this study proposes a framework that combines these models to improve detection accuracy and processing speed[31]. In addition, to enhance the generalizability of the model and ensure greater data diversity, the study will use

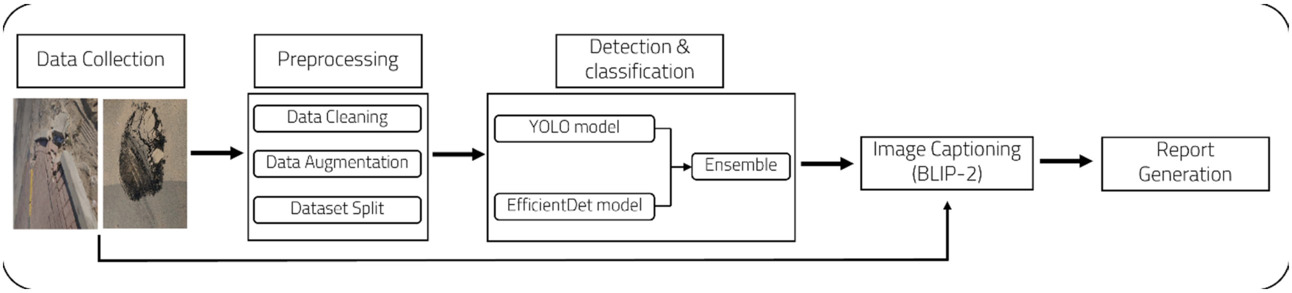


Fig. 1. Conceptual diagram of the proposed AI-based framework.

Class name	Number of train labels	Number of validation labels	Total
Barriers	6579	1838	**8417**
Sidewalks	5663	1749	**7412**
Pothole	19,960	6015	**25,975**
Total	**32,202**	**9602**	**41,804**

Table 3. The distribution of labels in the first dataset[5]. Bold values indicate the totals for each row and column.

two datasets containing VP images collected from Saudi Arabia, with the aim of supporting government organizations[5,32].

Moreover, unlike most existing studies that focus only on detection and classification, this study presents an important extension by integrating image translation techniques using BLIP-2. BLIP-2 is the best choice for this study due to its efficient use of pre-trained image encoders and frozen large language models (LLMs), which reduces the demand for intensive computational resources and large amounts of training data[29]. By automatically generating text descriptions from detected images, the system aims to enhance reporting automation, improve data interpretability, and support better decision-making while reducing human error. This integration thus provides a more comprehensive solution for VP management.

Methodology

The proposed framework automates the detection, classification, and description of VP in cities, addressing the challenges of manual reporting, which is slow and can lead to road hazards, by providing an intelligent solution to improve the accuracy of submitted reports. It uses YOLOv5 and EfficientDet to detect and classify VP, then applies an ensemble approach to improve accuracy. In addition, BLIP-2 generates detailed descriptions of the detected pollution to make the reports more informative. As shown in Fig. 1, the methodology follows multiple processes, starting with data collection and preprocessing, followed by VP detection and classification, then the interpretation of model results, and ending with the generation of captioning.

Data collection

This study aims to support government organizations by utilizing AI-based technology to automatically detect, classify, and generate descriptions of VP from images. Therefore, the focus was on datasets containing images of VP challenges in Saudi Arabia. The first dataset, titled "Visual Pollution of Public Roads in Saudi Arabia" from Mendeley[5] was created through a government initiative, where Saudi citizens and expatriates were asked to take photos of potential visual pollutants and upload them to the Balady mobile app, managed by the Ministry of Municipal and Rural Affairs and Housing (MOMRAH)[5]. These reports were then compiled into a structured dataset containing 31,795 images of potholes, excavation barriers, and dilapidated sidewalks, where some images contain more than one object class, resulting in a total of 41,804 labels. Table 3 presents the distribution of labels for this dataset, while Fig. 2 shows sample images.

The second dataset, titled "Urban Visual Pollution Dataset" from Kaggle[28], consists of 9966 images collected from Saudi Arabia and includes eleven classes of pollution, namely garbage, road construction, potholes, cluttered sidewalk, bad billboard, graffiti, sand on roads, unkept facade, faded signage, broken signage, and bad streetlight, where some images contain more than one annotated object, resulting in a total of 19,950 labels. These images were collected from moving vehicles, providing a real-world view of VP. The distribution of labels for this dataset is detailed in Table 4, and Fig. 3 provides sample images.

By combining these two datasets, this study aims to train the model with diverse and real data, which helps improve its ability to detect and classify VP.

Data preprocessing

Preprocessing is the second step in this framework to ensure that the dataset is clean, structured, and optimized for deep learning models. Since the two datasets will be merged, differences in image size and class distribution require preprocessing to create a unified dataset and ensure that the model is trained on properly structured data.

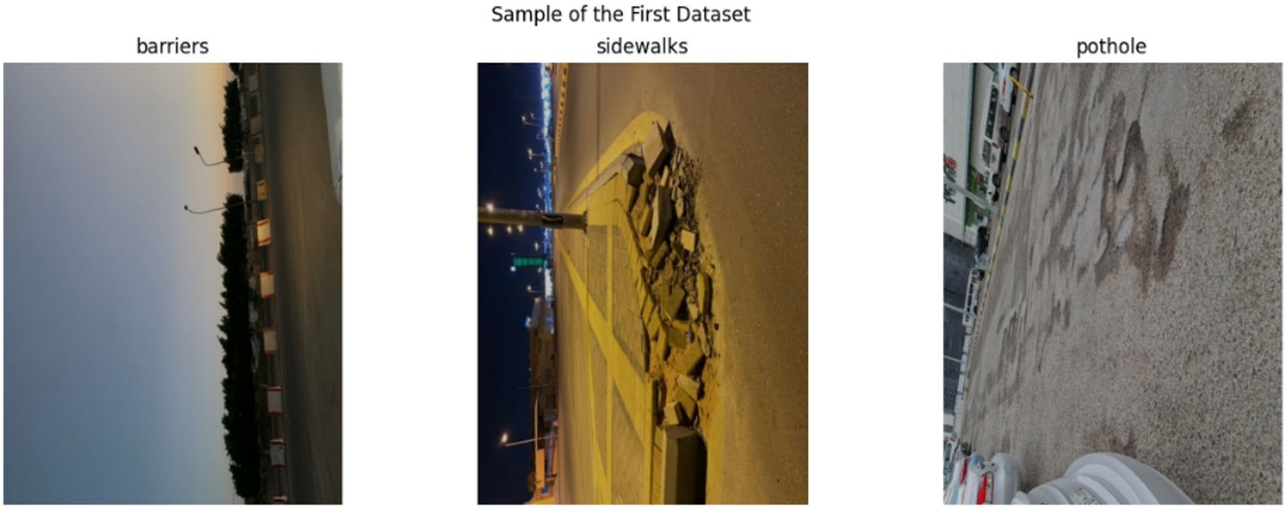


Fig. 2. A sample of the first dataset (AlElaiwi, Ahmad et al.[5]).

Class name	Number of labels
GARBAGE	8597
CONSTRUCTION_ROAD	2730
POTHOLES	2625
CLUTTER_SIDEWALK	2253
BAD_BILLBOARD	1555
GRAFFITI	1124
SAND_ON_ROAD	748
UNKEPT_FACADE	127
FADED_SIGNAGE	107
BROKEN_SIGNAGE	83
BAD_STREETLIGHT	1
Total	**19,950**

Table 4. The distribution of labels in the second dataset (Panigrahi[28]). Bold value indicates the totals for each row and column.

The preprocessing steps included: (1) identifying relevant classes, (2) handling class imbalance, (3) checking bounding boxes, (4) resizing images, (5) converting dataset files into appropriate formats—such as *.yaml*, *.txt*, and *.json*—based on model requirements, and (6) splitting the dataset into training, validation, and test sets.

The first step was to identify the relevant classes. For the first dataset, which contains classes relevant to Saudi Arabia with sufficient samples in each category, all classes were retained. By contrast, the second dataset contained eleven classes, some of which had very few samples. These underrepresented classes were removed, retaining only the meaningful ones: "GARBAGE," "CONSTRUCTION_ROAD," "POTHOLES," "CLUTTER_SIDEWALK," "BAD_BILLBOARD," and "GRAFFITI."

Class imbalance occurs when there is a significant disparity in the number of images across classes, which may lead to biased predictions as the model favors majority classes while underperforming on minority ones. In the first dataset, the potholes class contained substantially more samples than the others, while in the second dataset several classes were underrepresented (e.g., BAD_STREETLIGHT, BROKEN_SIGNAGE, FADED_SIGNAGE, UNKEPT_FACADE, and SAND_ON_ROAD). To address this, data augmentation (flipping, rotation, brightness adjustment, and noise addition) was applied to minority classes, while Random Under-Sampling (RUS) was used for the potholes class to reduce bias.

Since object detection models rely on accurate bounding boxes, all annotations were reviewed and corrected using Roboflow, which also enabled conversion into compatible formats (e.g., YOLOv5, EfficientDet) and automatic dataset structuring[17]. To standardize input sizes, all images were resized to 640 × 640 pixels (the first dataset originally ranged 450–460 × 600 pixels, while the second was 1920 × 1080). Finally, the processed dataset was split into training (70%), validation (15%), and testing (15%) sets to ensure effective training, fine-tuning, and evaluation.

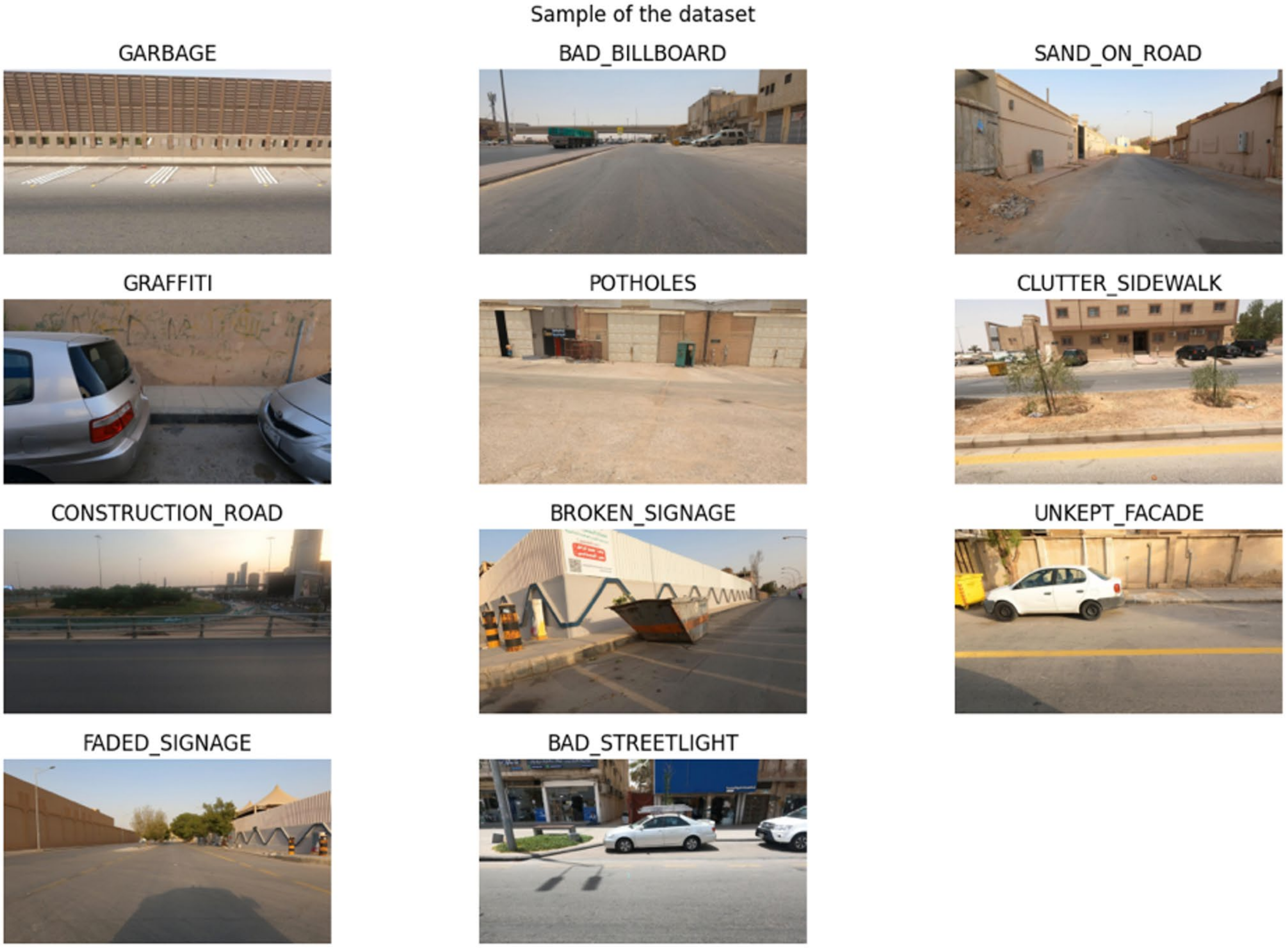


Fig. 3. A sample of the second dataset (Panigrahi[28]).

Detection and classification models

The proposed framework uses an ensemble approach that combines two models (EfficientDet and YOLOv5) to improve detection accuracy and reliability. Ensemble learning using multiple models together to get better results. This approach helps reduce errors and makes the model more robust in detecting VP.

EfficientDet model

EfficientDet is an accurate object detection model developed by Google. It uses EfficientNet as its backbone and incorporates a Bi-directional Feature Pyramid Network (BiFPN) to improve detection across different object sizes. BiFPN allows data to flow in both directions, improving the model's ability to detect both small and large objects[34]. In addition, EfficientDet employs compound scaling, which means it adjusts its layers and image size together to optimize speed and accuracy. The model predicts object locations using anchor boxes at multiple scales[31]. Figure 4 illustrates the architecture of the EfficientDet model.

YOLOv5 model

In this study, YOLOv5 was selected because prior research using the same dataset achieved excellent results with this version, confirming its effectiveness and reliability for detecting visual pollution in Saudi Arabia. In addition to its proven performance, YOLOv5 is a fast and lightweight model for real-time object detection that improves learning efficiency and reduces processing time[4]. It employs Spatial Pyramid Pooling (SPP) to capture multi-scale features and a Path Aggregation Network (PANet) to enhance feature fusion and detection accuracy. The architecture is composed of three main components: the Backbone, which extracts features using BottleNeckCSP layers and the SPP module; the PANet, which refines features through concatenation, upsampling, and convolution operations; and the Output layer, which generates the final object predictions[18]. Figure 5 shows the architecture of the YOLOv5 model.

Ensemble model

The two models (EfficientDet and YOLOv5) were first trained individually and then combined using an ensemble approach. Both models generate bounding boxes, class labels, and confidence scores for the detected VP. To merge their results and reduce redundancy, the Non-Maximum Suppression (NMS) algorithm was applied as a

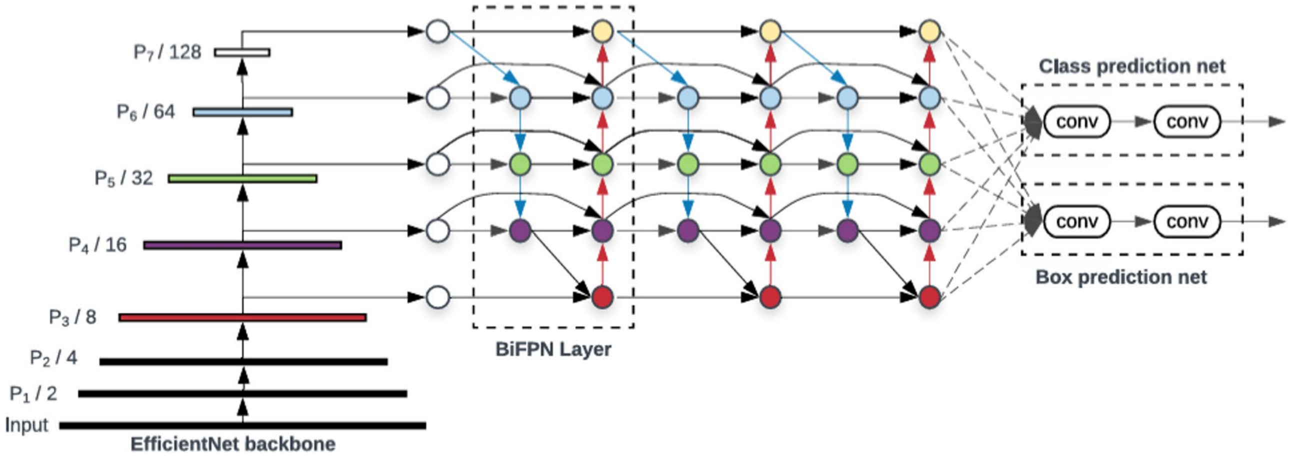


Fig. 4. The architecture of EfficientDet[30].

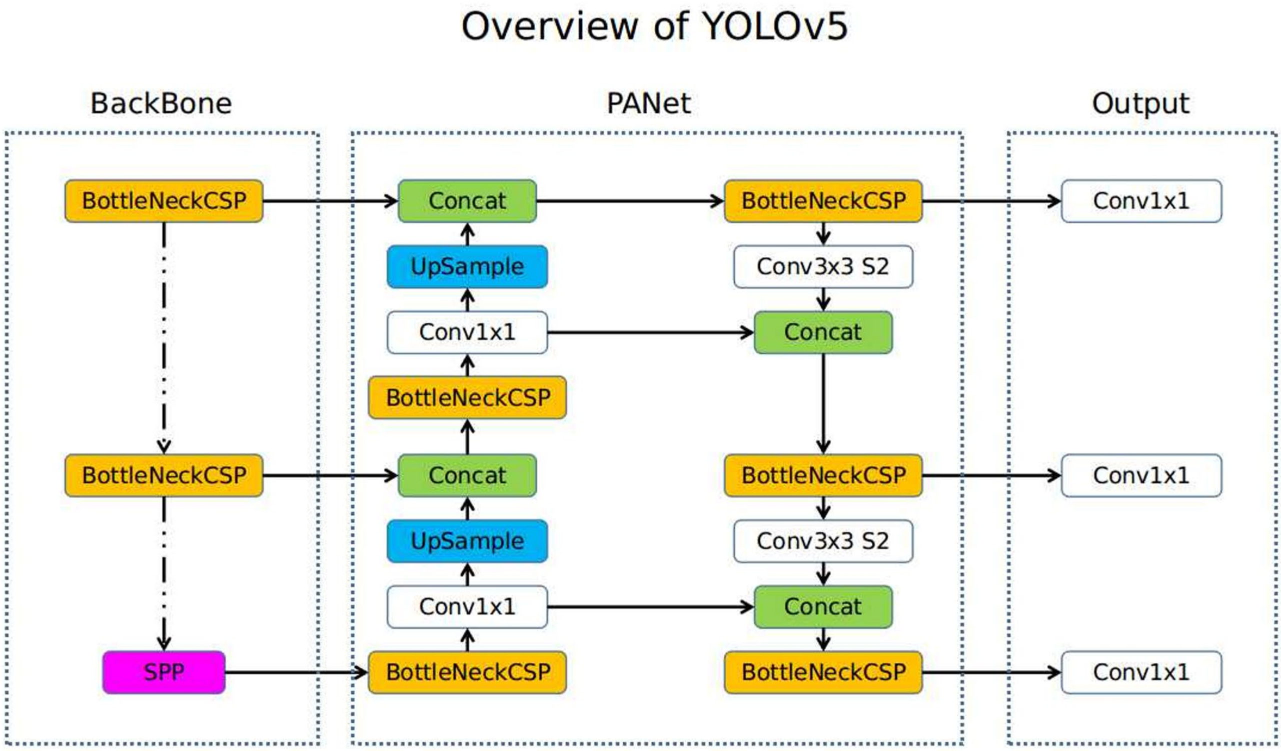


Fig. 5. The architecture of YOLOv5[4].

post-processing step. NMS removes redundant and overlapping bounding boxes by retaining only the one with the highest confidence score, thereby improving detection accuracy and reducing false positives[31,35].

Image captioning using BLIP-2

To improve VP reporting, this study uses BLIP-2 for automatic image captioning. By generating textual descriptions of detected VP, BLIP-2 reduces the need for manual input and supports the creation of structured and informative reports. The model takes two inputs: the image and the detection result (class name), as shown in the proposed framework in Fig. 1.

BLIP-2 is a VLM that bridges the gap between computer vision and natural language processing. It adopts a two-stage pretraining approach, linking a frozen image encoder such as a Vision Transformer (ViT), with a frozen language model such as BERT. This architecture enables BLIP-2 to produce highly relevant and contextually appropriate descriptions of given objects, by effectively combining the capabilities of both vision and language models to interpret and express images in a way that closely matches the textual context[33].

In this study, BLIP-2 uses bounding box outputs from the ensemble model to generate captions that describe the type of pollution. These captions are automatically incorporated into structured reports, facilitating reporting automation, reducing the need for manual text entry, and minimizing the potential for human error. By integrating these automatically generated descriptions with detection and classification results, the system enables the generation of more accurate reports.

Performance metrics

To evaluate the performance of the proposed framework, various metrics were used to assess both object detection (YOLOv5 and EfficientDet) and image captioning (BLIP-2). These metrics ensure that the system provides accurate detections, relevant captions, and clear explanations.

$$mAP = \frac{1}{n}\sum{}_{K=1}^{K=n} APK$$

For object detection, the primary metric is mAP, which measures the model's ability to identify objects accurately by analyzing the precision-recall curve at different IoU thresholds[35].

where n represents the total number of classes, and APK is the average precision for class K.Also, precision and recall are key concepts. Precision looks at how many pollution cases were correctly identified, focusing on reducing false positives (Hossain et al. 2021):[11].

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

On the other hand, recall measures how many real pollution cases were found, with the goal of lowering false negatives (Hossain et al. 2021):[11].

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

The F1-score, which combines precision and recall, is also used for a comprehensive evaluation of detection accuracy (Pramanik et al. 2021):[14].

$$F1\ Score = 2\ \times \frac{Precision\ \times\ Recall}{Precision + Recall}$$

For image captioning, two important metrics are used to evaluate the quality of the generated descriptions. BLEU (Bilingual Evaluation Understudy) checks how closely the generated captions match reference descriptions by comparing word overlaps. CIDEr is also used, as it looks at the importance of certain words and compares the captions to multiple human-created references, making sure the output is relevant and descriptive. Furthermore, human evaluation was applied by sharing images with two captions to evaluate the quality of the captions generated by two different model versions. Three independent evaluators rated each caption on a scale from 1 (poor) to 5 (excellent) based on relevance and clarity. The caption with the higher average score was chosen as the better option. This method helped identify which model version generated more effective descriptions[36].

Results and discussion

Data preprocessing

This section presents the preprocessing results used to create a clean and balanced dataset for training deep learning models in VP detection and captioning. Two datasets were employed: one from Mendeley and another from Kaggle, as described in the methodology section.

Three datasets were prepared to compare training results and evaluate model performance across different data sources. The dataset that achieved the highest evaluation scores was ultimately selected for this study, while the others remain valuable for future improvements. Both datasets were first cleaned to ensure label quality. To avoid redundancy, overlapping classes such as potholes, sidewalks, and barriers were retained only from the first dataset, which contained a larger number of annotated samples. Some bounding boxes were incorrectly placed and did not align with the true object locations, as shown in Fig. 6. These were corrected by adjusting coordinates, removing misaligned annotations, and converting all labels into YOLO-compatible .txt format using custom Python scripts.

The class imbalance across both datasets was addressed using RUS for majority classes and data augmentation for minority ones. A fixed target of 8000 instances per class was set, with augmentation. This balancing step was applied to the merged dataset, creating a more uniform distribution for fair and stable model training (Fig. 7).

The processed datasets were merged into a single dataset containing six main classes: barriers, sidewalks, garbage, potholes, graffiti, and bad billboards. The final combined dataset includes 42,547 images with well-defined bounding box annotations (Fig. 8).

The merged dataset was uploaded to Roboflow for further refinement, including visual inspection of bounding boxes, removal of irrelevant or mislabeled instances, standardization of class labels, and resizing all images to 640 × 640. The class "sidewalk" was renamed to "fracture" to better describe both broken sidewalks and asphalt cracks not classified as potholes. The dataset was then split into 70% training, 15% validation, and 15% testing, and exported in YOLO and COCO formats for compatibility with the selected architectures.

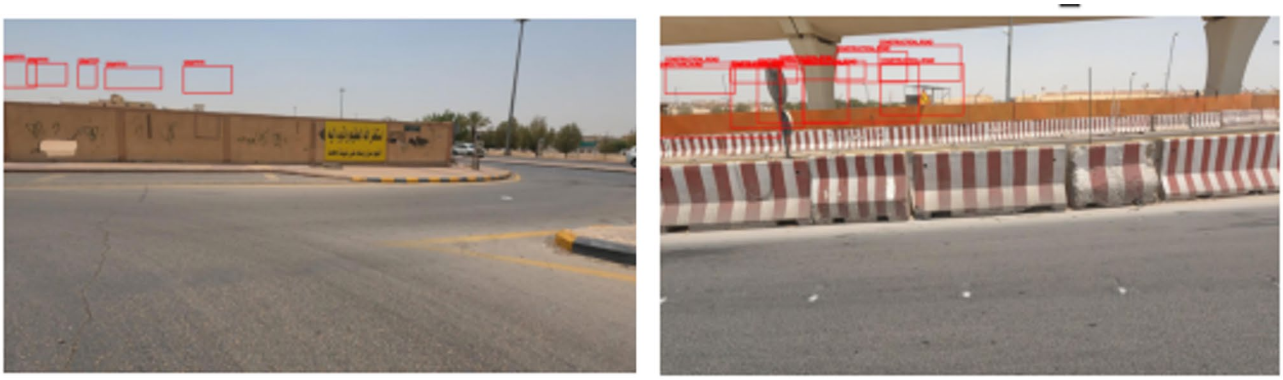


Fig. 6. Examples of an incorrect bounding box annotation from the Kaggle dataset.

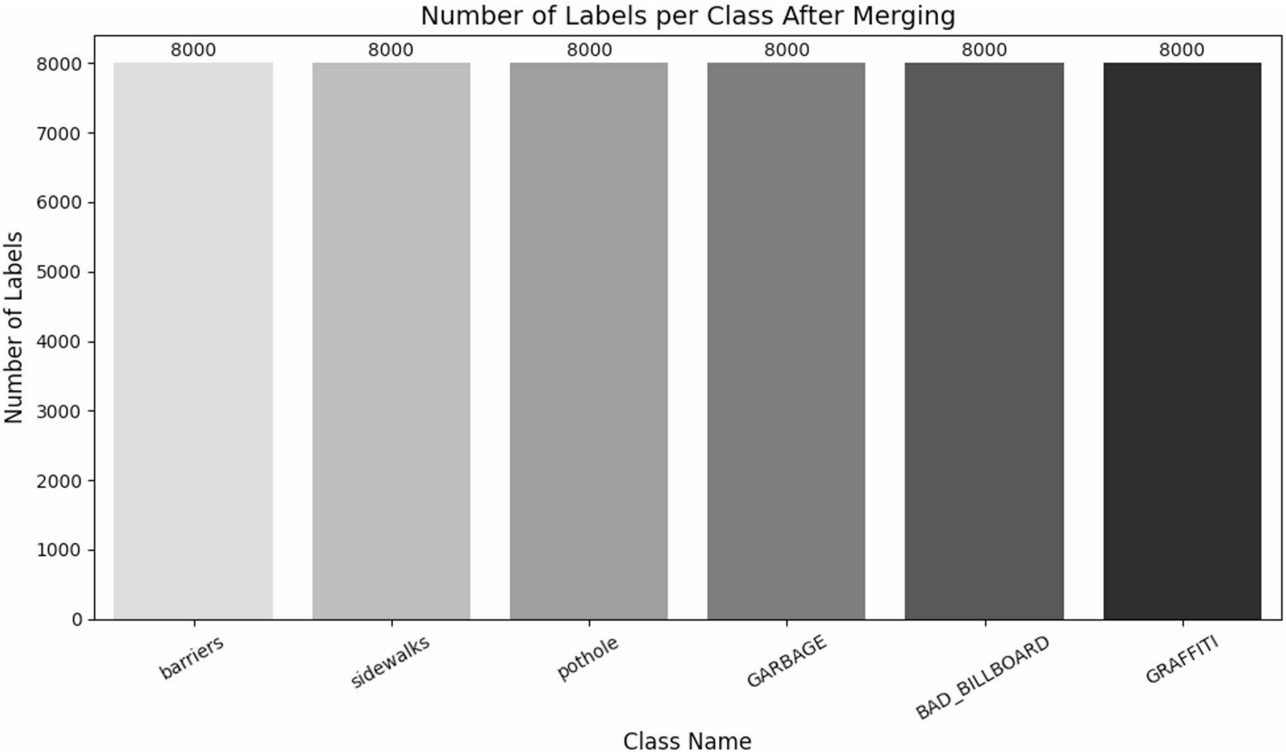


Fig. 7. Class distribution of the merged dataset.

Although merged, each dataset was also balanced individually to ensure fairness and enable separate evaluation. This balancing was used only RUS based on the smallest class size in each dataset. In the first dataset, barriers and potholes were reduced to 5663 labels to match sidewalks (Fig. 9). In the second dataset, each of the six classes was balanced to 1124 samples (Fig. 10). These steps ensured consistency before splitting and training.

Finally, Finally, three dataset variations were used for training and evaluation: (1) the cleaned and balanced version of the first dataset, (2) the cleaned and balanced version of the second dataset, and (3) the merged version combining both datasets. This strategy enabled comparative experiments and provided a thorough evaluation of each model's robustness and generalizability.

For the captioning task, 3000 images were randomly selected from the final dataset with equal representation across all VP classes. Initial captions were generated using BLIP-2 and then refined with ChatGPT prompts to make them more natural and human-like. These refined captions were subsequently used to fine-tune BLIP-2, enhancing its ability to capture visual-language associations specific to urban VP, as shown in Fig. 11.

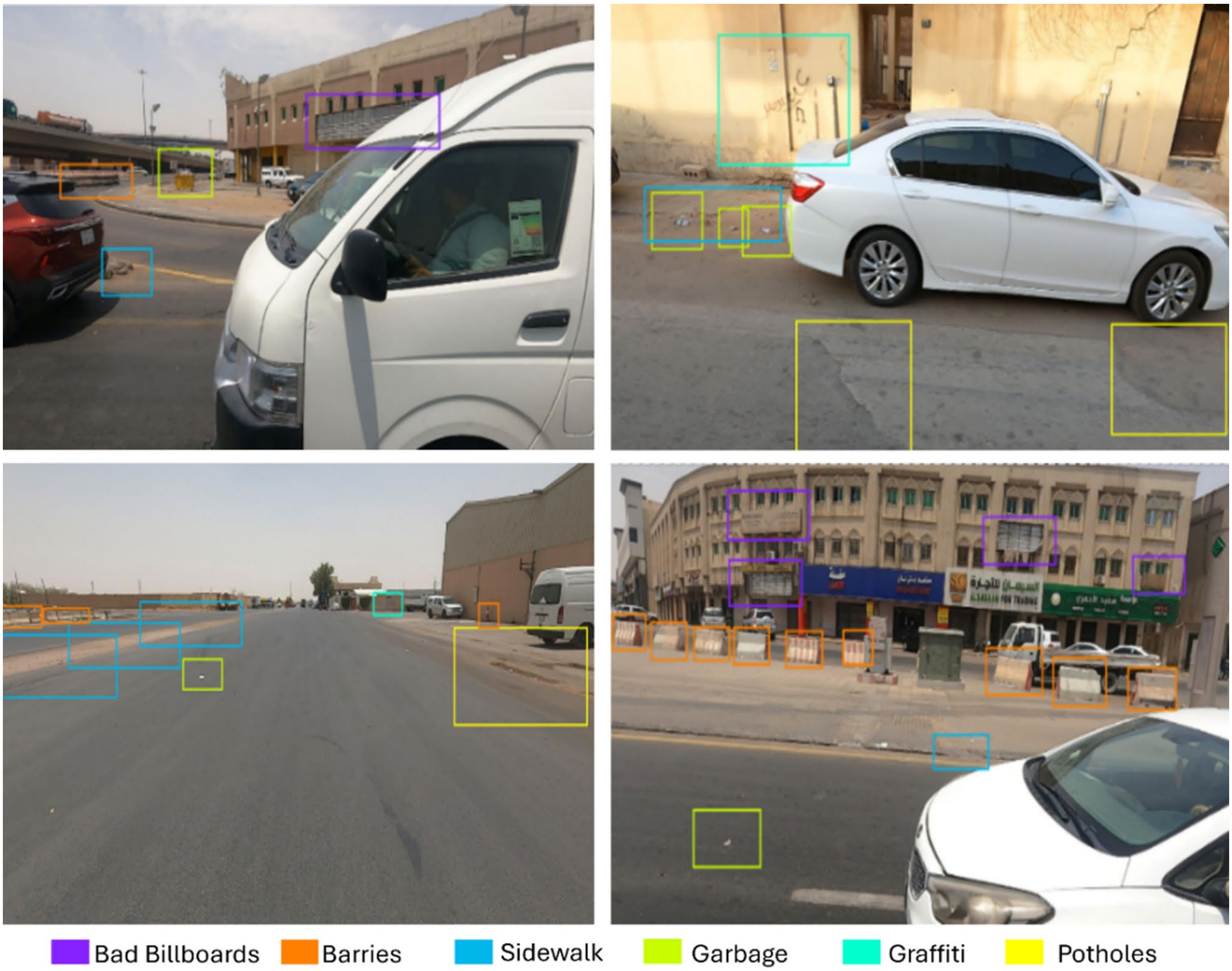


Fig. 8. Sample of the merged dataset.

Model performance

Yolo results

We used YOLOv5s because it is a lightweight model that offers faster performance with lower computational requirements, making it suitable for real-time VP detection. The training was conducted in three stages to progressively enhance performance.

In the first stage, YOLOv5s was trained for 50 epochs with a batch size of 16 using default pre-trained weights, serving as a baseline. In the second stage, training was extended to 250 epochs with a batch size of 32, using the Stochastic Gradient Descent (SGD) optimizer (learning rate = 0.01, momentum = 0.937, weight decay = 0.0005). Three warm-up epochs were introduced, and data augmentation techniques such as HSV shifts and rotations were applied, improving generalization. In the third stage, the model was trained for an additional 100 epochs using the best weights from stage two. The learning rate was reduced to 0.001, warm-up epochs increased to 5, and augmentation strength slightly lowered. Focal loss ($\gamma = 2.0$) was also introduced to emphasize hard-to-detect examples, stabilizing training, and further improving results.

The model was evaluated across all three datasets. The first dataset, with three classes (barriers, sidewalks, potholes), achieved the best results: mAP = 0.94, recall = 0.89, precision = 0.86, and F1-score = 0.87 (Table 5). Training and validation losses (Fig. 12) showed stable convergence, with object loss reaching 0.009 and box loss stabilizing around 0.022. The close alignment between training and validation curves indicates effective learning without overfitting, confirming strong generalization across VP classes.

The confusion matrix and the precision–recall curve, illustrated in Fig. 13, together with the prediction samples shown in Fig. 14, demonstrate that the model consistently achieved high performance across all classes. However, the pothole class showed the highest confusion with the background (0.38), indicating that potholes were the most challenging class for the model to detect reliably.

The second dataset demonstrated lower overall performance compared to the first dataset, with an average precision–recall of 0.56. Precision and recall were 0.60 and 0.53, respectively, resulting in an F1-score of 0.56. Some classes, such as clutter sidewalk and construction road, performed relatively well, whereas other classes,

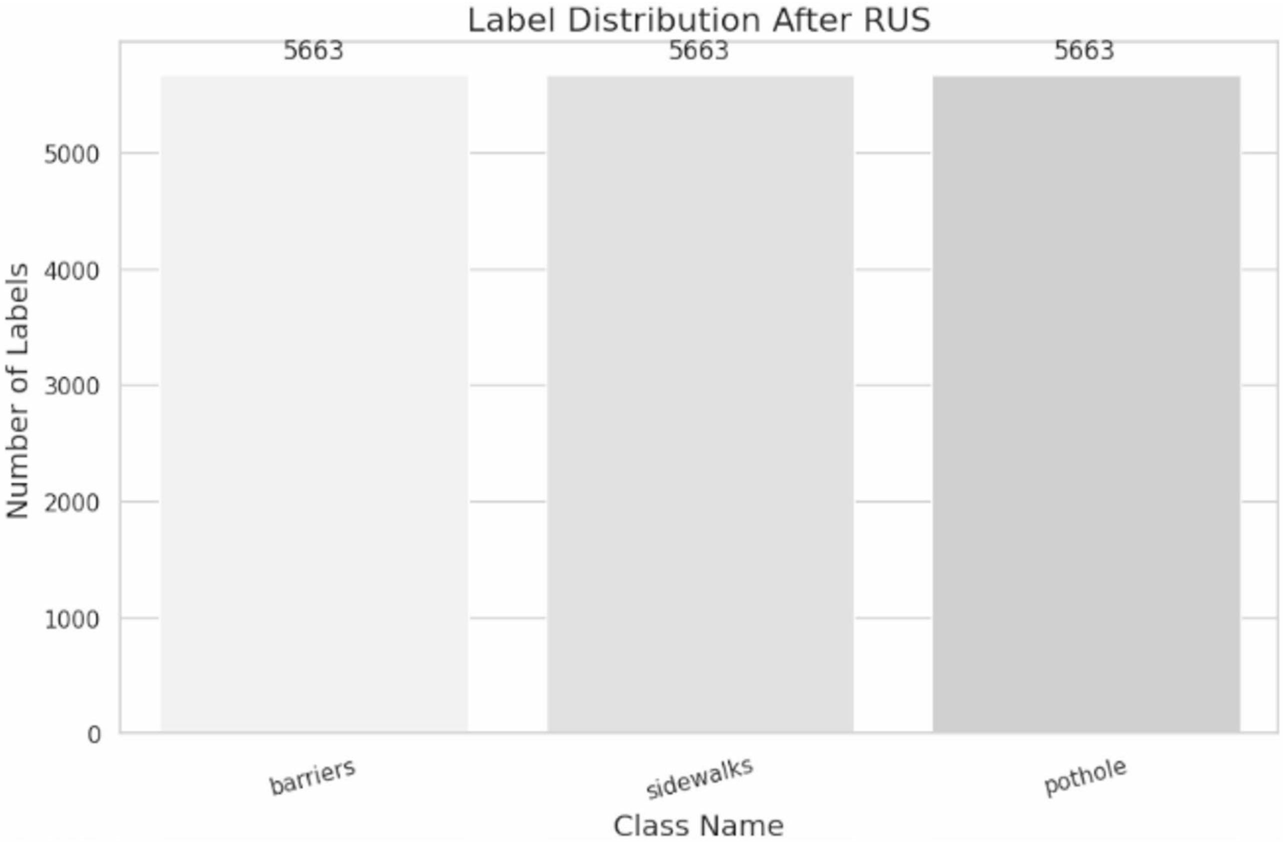


Fig. 9. Class distribution in the first dataset after RUS.

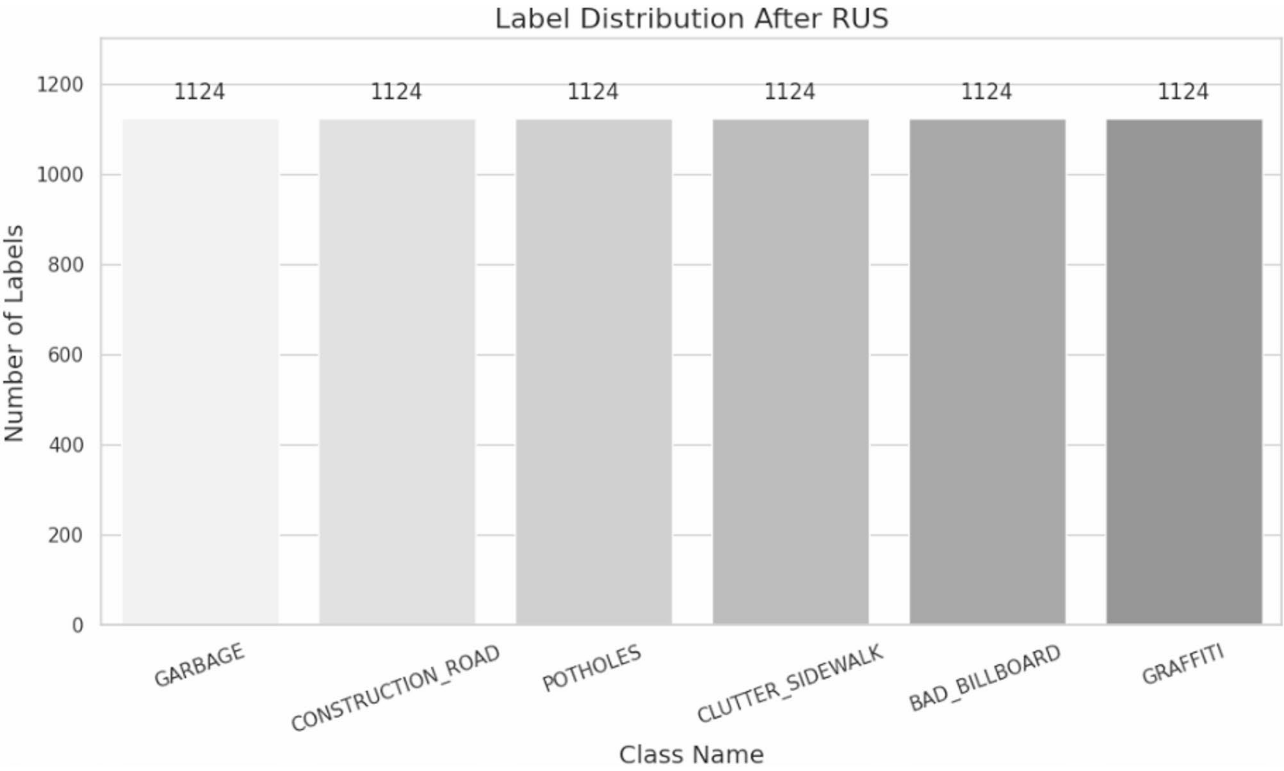


Fig. 10. Class distribution in the second dataset after RUS.

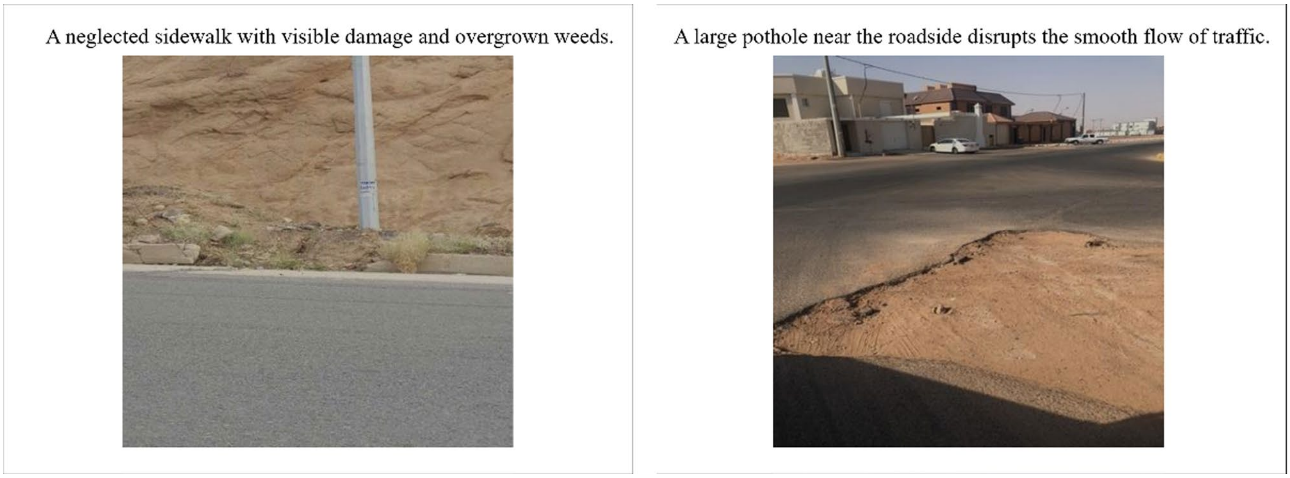

A neglected sidewalk with visible damage and overgrown weeds.
A large pothole near the roadside disrupts the smooth flow of traffic.

Fig. 11. Examples of enhanced image captions generated.

Dataset	mAP	Recall	precision	F1-socre
Dataset 1 (AlElaiwi, Ahmad et al. 2023)	0.94	0.89	0.86	0.87
Dataset 2 (Panigrahi[28])	0.56	0.52	0.60	0.56
Merged datasets	0.64	0.59	0.67	0.63

Table 5. Yolov5 results.

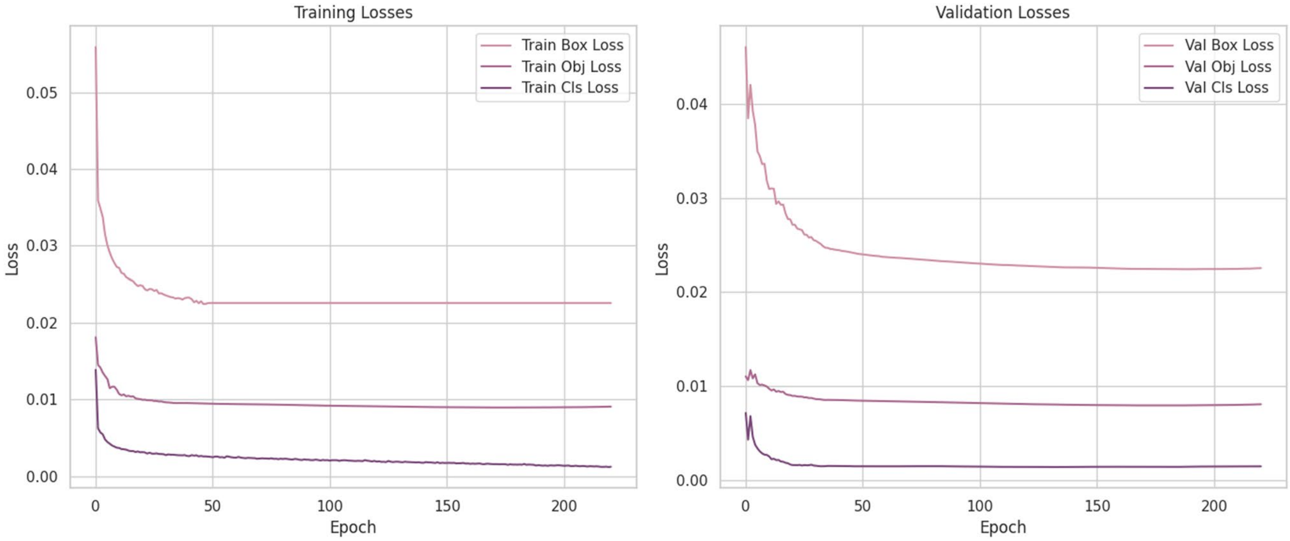


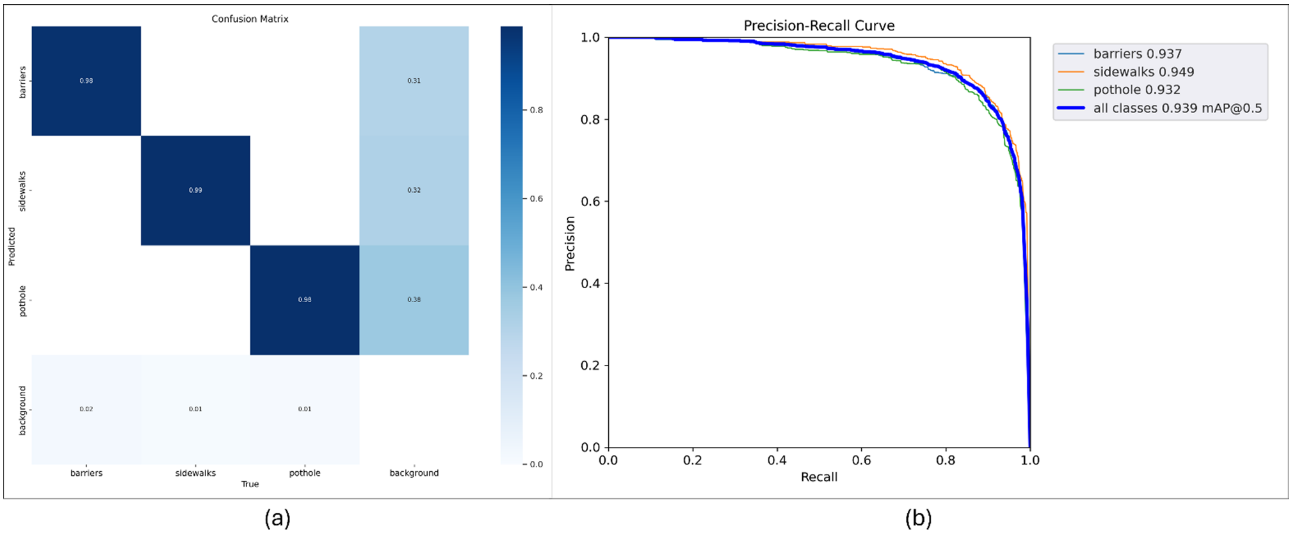
Fig. 12. YOLOv5 loss metrics on the first dataset.

including potholes and garbage, showed poor performance in both precision and recall, as illustrated in the confusion matrix and PR curve in Fig. 15.

The merged dataset achieved a moderate mAP of 0.64, which was higher than the second dataset but lower than the first. This outcome reflects the challenges of integrating diverse data sources, particularly inconsistencies in labeling quality and image conditions. As illustrated in Fig. 16, the model showed difficulty in distinguishing between visually similar classes such as fractures, barriers, and garbage, whereas graffiti achieved the highest mAP of 0.85.

EfficientDet results

The EfficientDet-D1 model was selected for its lightweight design and balanced trade-off between speed and accuracy. It was trained on the first dataset, which had previously yielded the best YOLOv5 results. Training was



(a)

(b)

Fig. 13. YOLOv5 evaluation on the first dataset, **a** confusion matrix, **b** Precision-Recall curve.

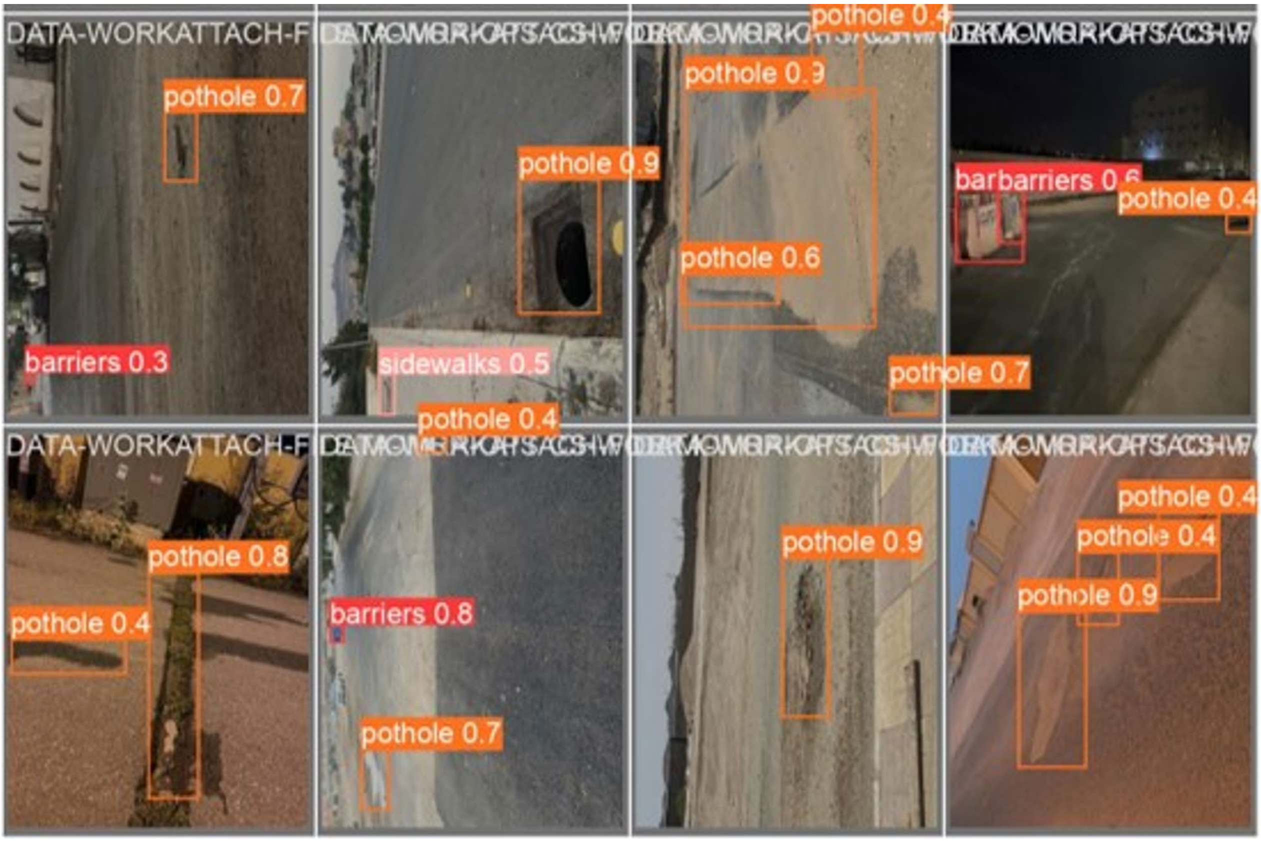


Fig. 14. YOLOv5 prediction result on the first dataset.

conducted in three phases. In the first phase, the model was trained on raw data using a learning rate of 1e−4, a batch size of 8, and 100 epochs. This initial attempt served as a baseline but yielded poor results due to class imbalance and the absence of prior knowledge. In the second phase, training was performed on the balanced dataset using transfer learning, where only the classification head was trained with pretrained EfficientDet-D1 weights. A higher learning rate of 5e−3, batch size of 32, and 10 epochs were used to quickly test the suitability of pretrained features for VP data. The results improved but remained limited. In the third phase, full fine-tuning was applied to the entire model with a lower learning rate of 5e−5, batch size of 16, and 50 epochs. These settings

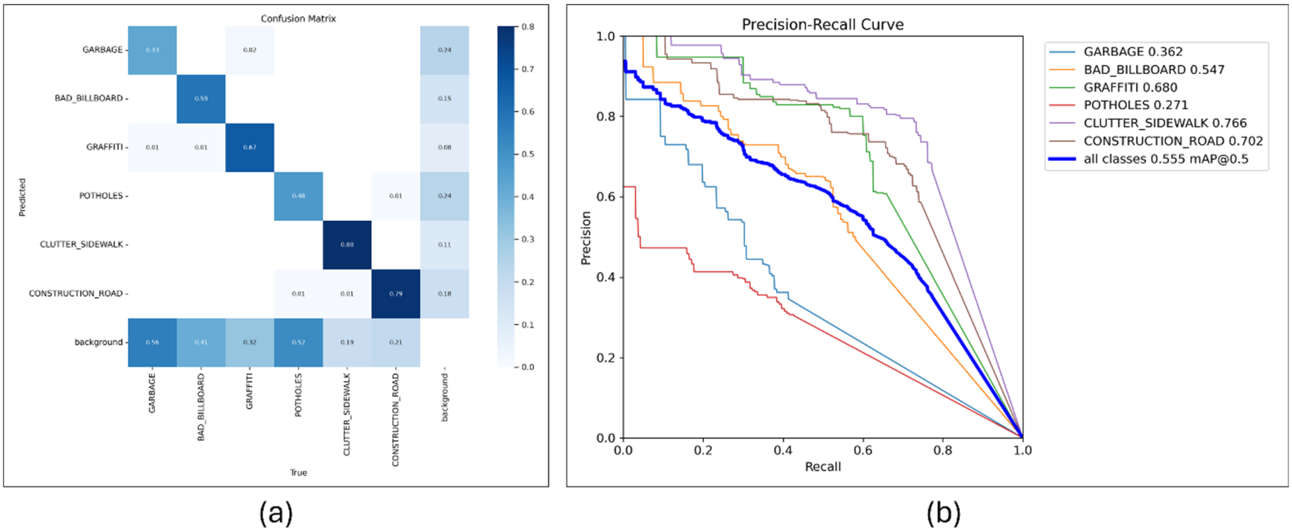


Fig. 15. YOLOv5 evaluation on the second dataset, **a** confusion matrix, **b** Precision-Recall curve.

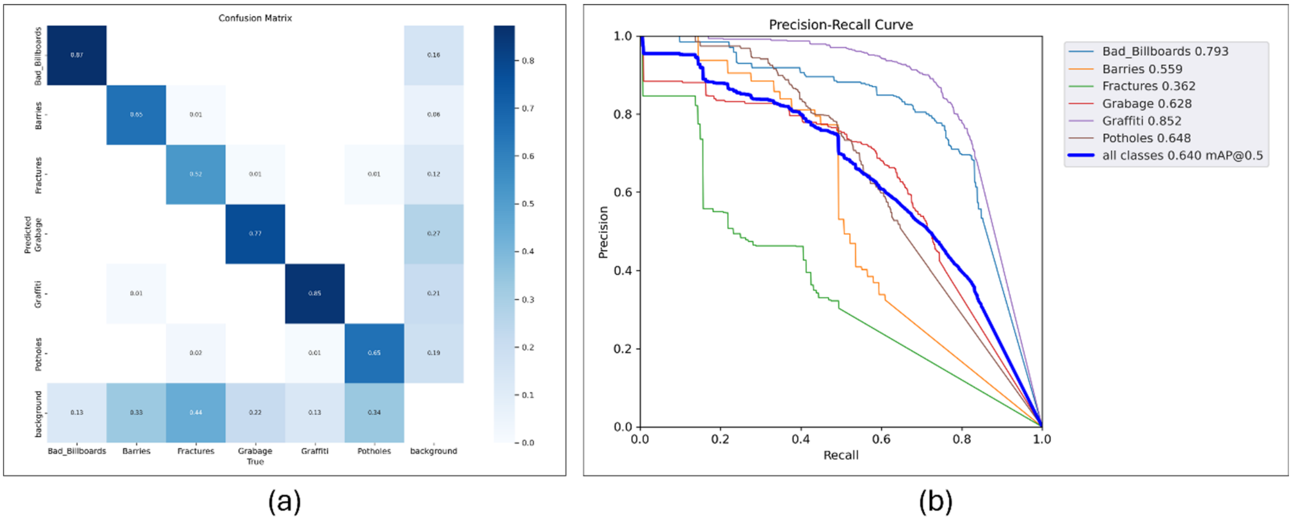


Fig. 16. YOLOv5 evaluation on the merged dataset, **a** confusion matrix, **b** Precision-Recall curve.

enabled gradual adaptation while preserving useful pretrained knowledge, resulting in the most stable training and the best overall performance.

Across all phases, classification loss consistently decreased for both training and validation sets, confirming effective learning without signs of overfitting or underfitting. The confusion matrix in Fig. 17 shows that the model effectively identified all classes, though it struggled with the background class—particularly misclassifying potholes at a rate of 0.41. Compared to YOLOv5 (Fig. 13), EfficientDet-D1 showed slightly better performance in detecting sidewalks, while both models exhibited similar challenges with potholes. This confusion may be attributed to multiple potholes appearing in single images, leading to overlapping detections and false positives.

The final performance metrics are presented in Table 6, while Fig. 18 illustrates sample predictions generated by the model. Overall, the results indicate strong performance, with confidence scores varying according to image quality and the visibility of the objects.

Ensemble results

To enhance detection performance and mitigate the limitations of individual models, an ensemble approach was employed by combining the outputs of YOLOv5s and EfficientDet-D1 using Non-Maximum Suppression (NMS). This integration leveraged YOLOv5s's speed in object localization and EfficientDet-D1's precision in feature extraction, resulting in a more robust detection system. As illustrated in Fig. 19, the ensemble model effectively detected various types of VP, such as potholes, damaged sidewalks, and barriers, achieving higher recall and precision compared to the individual models. The ensemble method yielded strong evaluation metrics, with a precision of 91%, recall of 95%, F1-score of 93%, and a mAP of 95%. Table 7 presents a comparison

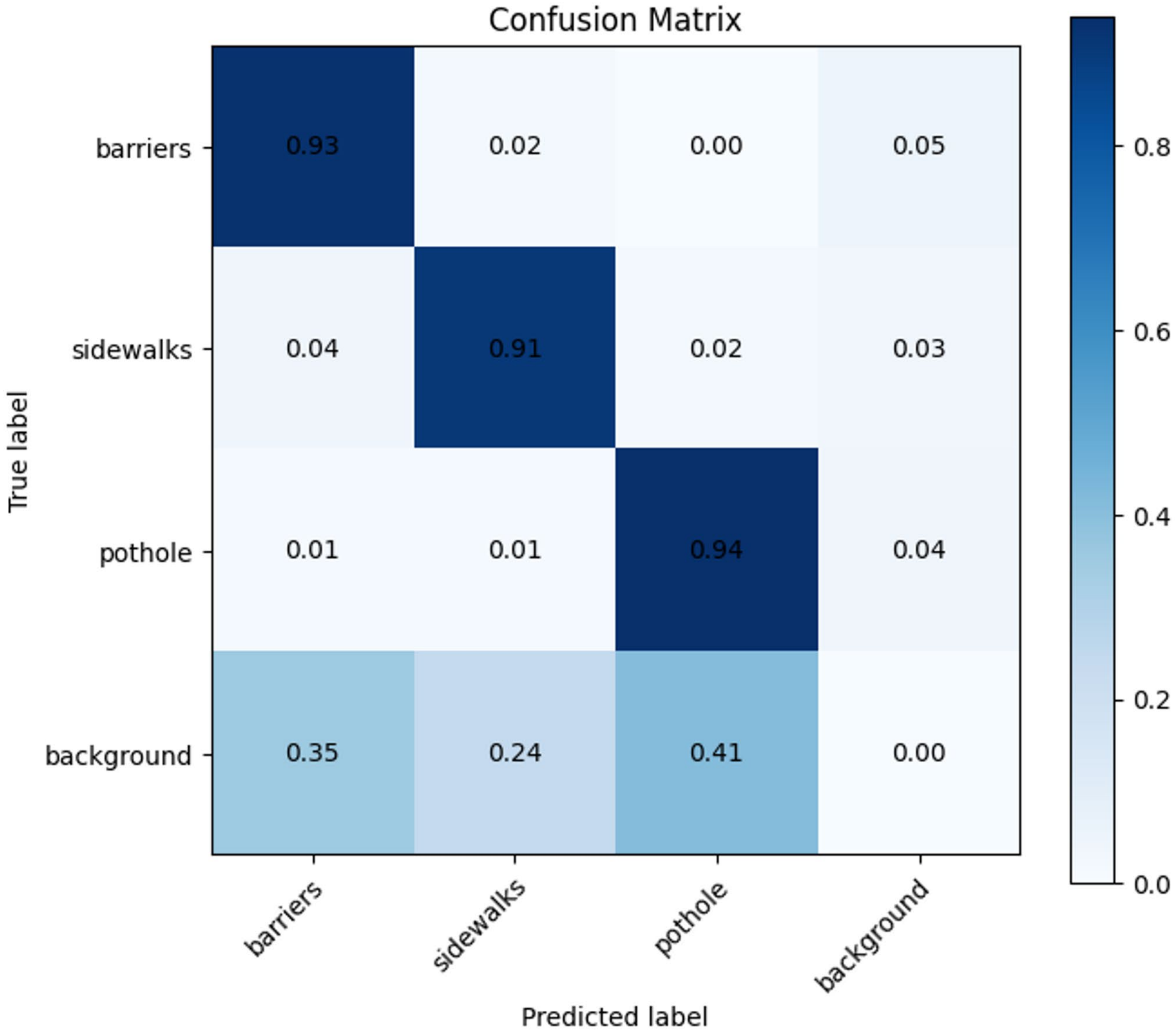


Fig. 17. Confusion matrix of EfficientDet.

Metric	Result
mAP	0.89
Recall	0.93
Precision	0.91
F1-Score	0.92

Table 6. EfficientDet-D1 results.

of the models' results. Ensemble post-processing settings: YOLOv5s confidence threshold = 0.25; EfficientDet confidence threshold = 0.25; boxes normalized to 0,1; class-wise NMS via ensemble_boxes.nms with IoU threshold = 0.50 (equal model weights; no score re-weighting).

Figure 20 presents examples of false positives, where the model misclassified objects such as waste bins or cars as barriers due to visual similarity, and in some cases mislabeled potholes. These errors highlight the need for further improvements in post-processing and data quality, though their impact on overall performance was minimal, confirming the effectiveness of the ensemble method in detecting VP.

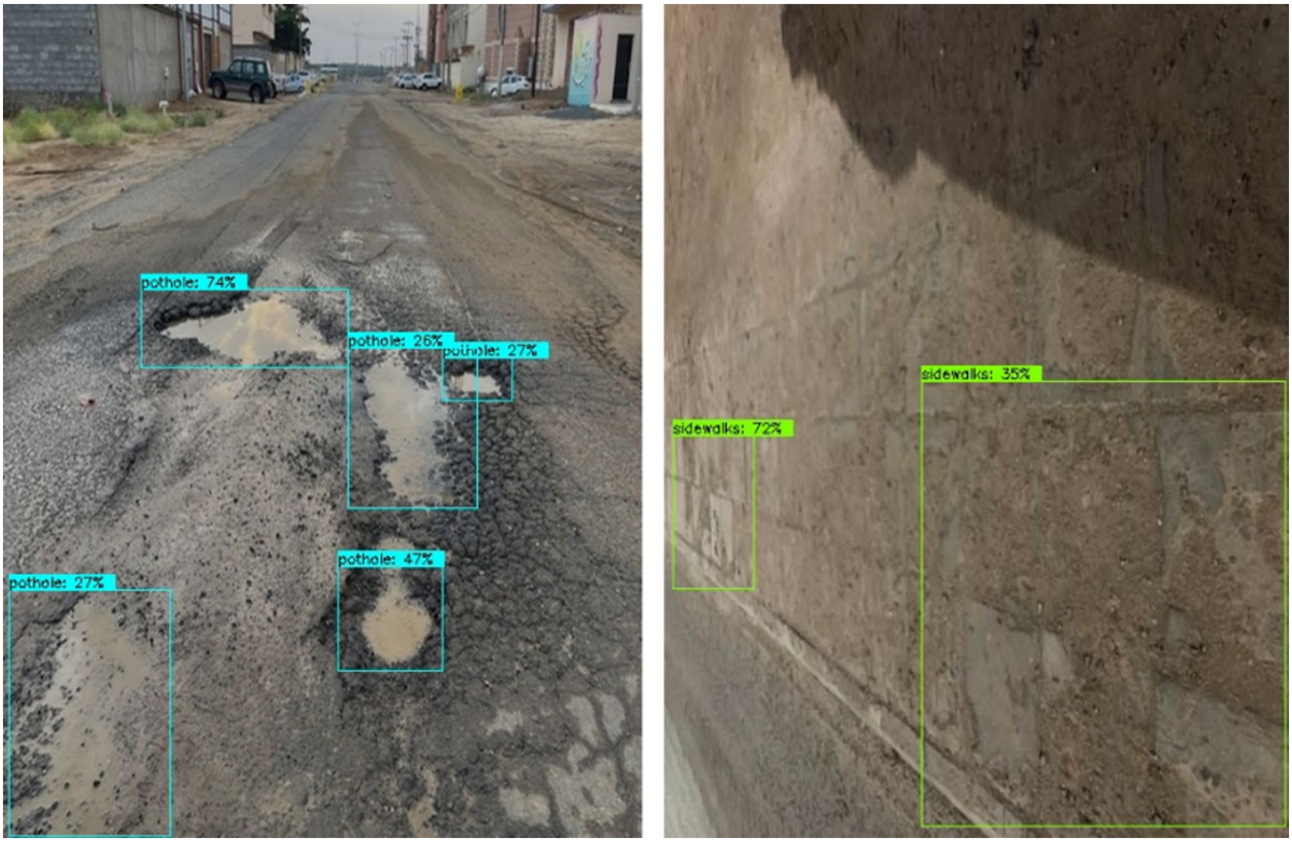


Fig. 18. EfficientDet-D1 prediction.

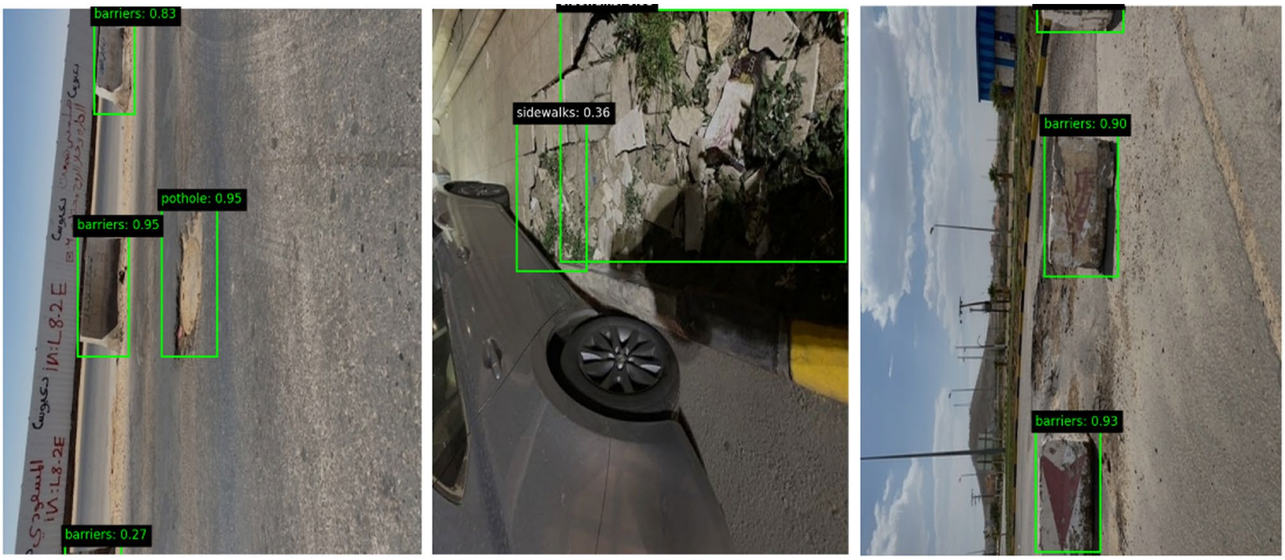


Fig. 19. Ensemble prediction.

BLIP-2 results

To generate image captions, two versions of the BLIP-2 model were tested. The first was the lightweight BLIP2-Flan-T5-XL, which produced Caption 1. The second, BLIP2-OPT-2.7B, was fine-tuned and generated Caption 2, as shown in Fig. 21.

For the first model, we used the following generation settings: max_new_tokens = 60, num_beams = 5, top_p = 0.9, temperature = 1.0, and early_stopping = True. A class-specific prompt was applied to reflect the detected object class and improve domain relevance.

Model	mAP	Recall	Precision	F1-score
Yolov5s	0.94	0.89	0.86	0.87
EfficientDet-D1	0.89	0.93	0.91	0.92
Ensemble (NMS)	**0.95**	**0.95**	**0.91**	**0.93**

Table 7. Final model evaluation results. Bold values indicate the best results among the models.

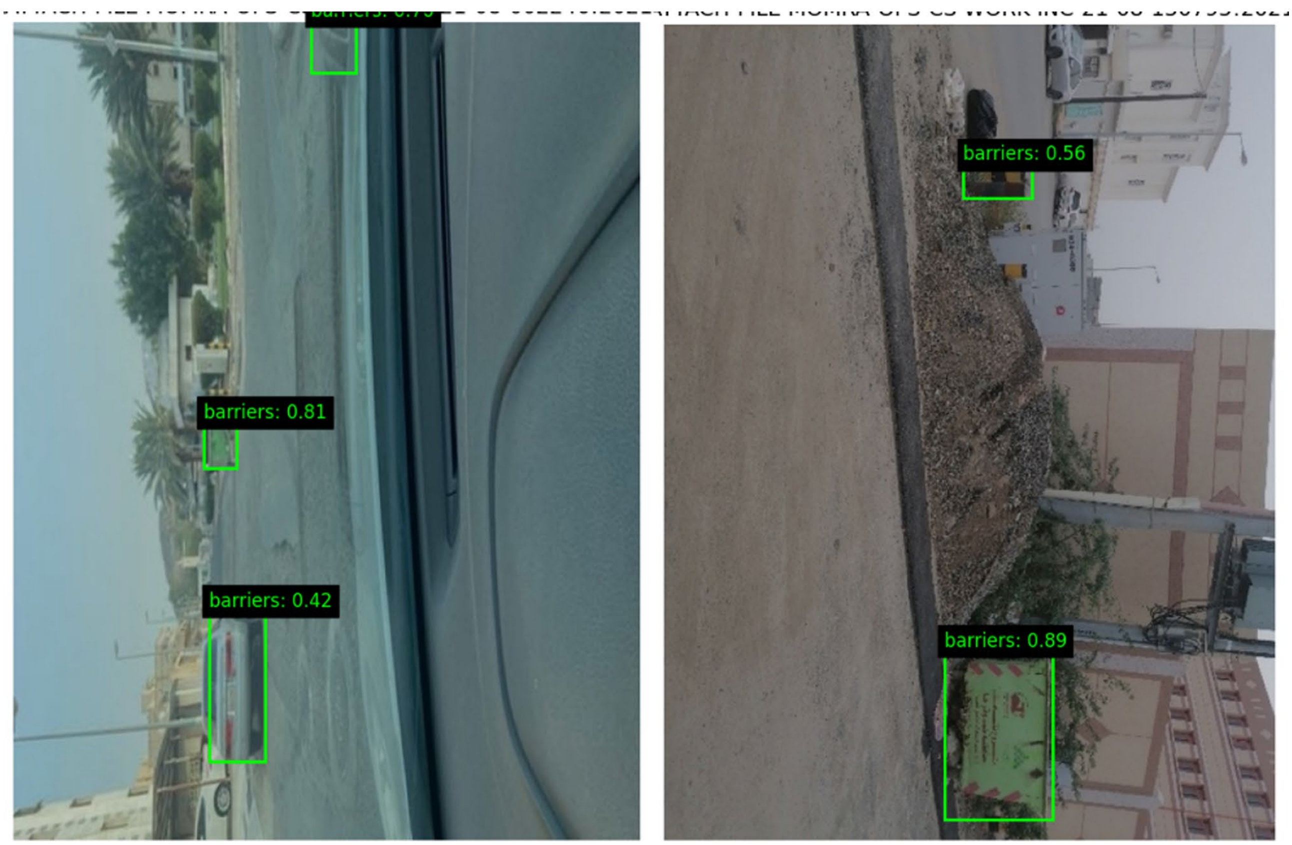


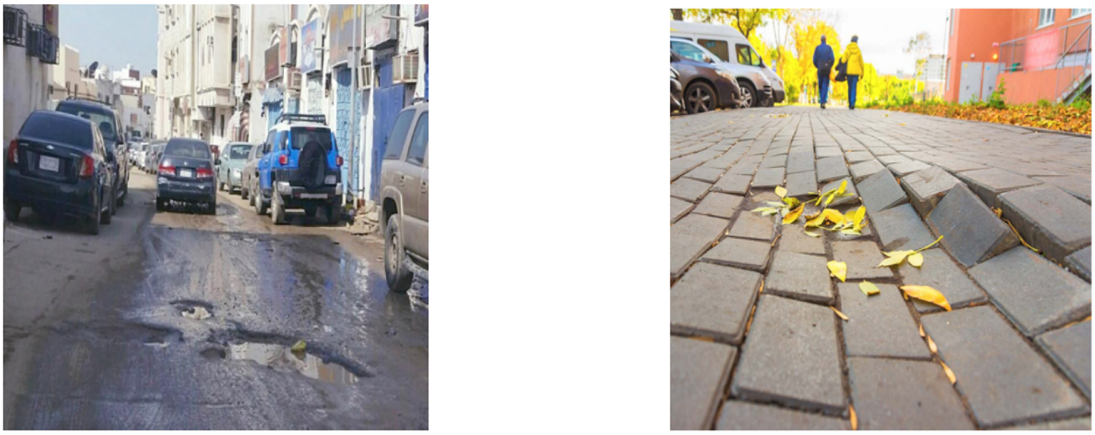
Fig. 20. Example of false positive detection by ensemble model.

The second model was fine-tuned using enhanced captions generated by ChatGPT. To reduce training costs while improving caption quality[37], Low-Rank Adaptation of Large Language Models (LoRA) was applied with $r = 16$, lora_alpha = 32, lora_dropout = 0.05, bias = "none", and target_modules = ["q_proj", "k_proj"]. In addition, bitsandbytes 8-bit quantization was employed to minimize GPU memory usage during training[38]. The fine-tuning was carried out with AdamW optimizer, a learning rate of 5e-5, 15 epochs, and a batch size of 8. Input tokenization used a maximum sequence length of 50 tokens, with padding and truncation applied. During inference, the generation setting was max_new_tokens = 50.

Human evaluation was conducted using a standardized scale[32] on 28 images, each rated by three evaluators experienced with the Madinaty application managed by the Riyadh Municipality. Each image was paired with two captions, and evaluators scored them on relevance and clarity from 1 (very poor) to 5 (excellent). We limited the evaluation to 28 images, as prior research indicates that increasing the number of questions can induce respondent fatigue and reduce accuracy in judgments[39]. Moreover, all three VP classes were represented, and it was observed that captions generated for the same class were highly consistent, further supporting the reliability of the evaluation despite the smaller sample size.

For analysis, the average score per caption was first calculated for each evaluator, then converted to percentages. Finally, results were averaged across the three evaluators, as summarized in Table 8. The results show that BLIP2-Flan-T5-XL achieved ~ 80% positive ratings, compared to ~ 65% for the fine-tuned BLIP2-OPT-2.7B. Consequently, the first model was selected for this study, while the second is considered for future improvement.

Additionally, the second model was automatically evaluated on 300 images from the fine-tuning dataset, which included captions refined through ChatGPT. The generated captions were compared with the refined references to assess wording and contextual alignment. The model achieved BLEU-1 to BLEU-4 scores of 0.215,



Caption 1: a road with a pothole in the middle of the road
Caption 2: an asphalt is covered in dust at the intersection of the road, showing signs of wear and damage.

Caption 1: a cracked sidewalk causes a lot of problems for pedestrians and motorists.
Caption 2: a man and a woman walk on the sidewalk in front of a damaged road.

Fig. 21. Sample captioning results for images used in the evaluation.

Evaluator	Caption 1 average %	Caption 2 average %
1	74.28	48.58
2	75	75.72
3	91.4	70.72
Average	**80.23**	**65.01**

Table 8. Human evaluation results. Bold values indicate the final average results from the human evaluation.

Metrics	Score
BLEU-1	0.215
BLEU-2	0.182
BLEU-3	0.116
BLEU-4	0.106
CIDEr	0.340

Table 9. Evaluation results for the fine-tuned BLIP2-OPT-2.7B model.

0.182, 0.116, and 0.106, respectively, and a CIDEr score of 0.340 (Table 9). These results partly align with the human evaluation, highlighting the need for further improvements in accuracy and contextual relevance.

Generating the report

To simulate VP reporting, a simple web interface was developed using Gradio. As shown in Fig. 22, users can upload an image, which is then processed by the ensemble model to detect VP objects. The BLIP2-Flan-T5-XL model generates a corresponding text description, and the interface displays the report details—including the image, VP type, date, time, and generated caption (Fig. 23). This prototype demonstrates the feasibility of integrating the proposed framework into automated reporting systems for government use.

Discussion and observations

The evaluation of the proposed models showed that the best performance came from the ensemble method, which combined YOLOv5s and EfficientDet-D1 using non-maximum suppression (NMS) as shown in Table 8. This approach achieved a mAP of 0.95, a recall of 0.95, a precision of 0.91, and an F1 score of 0.93, surpassing the individual models. YOLOv5s excelled on the first dataset with a mAP of 0.94, while EfficientDet-D1 had a high recall of 0.93. However, when the two datasets were combined, performance declined due to inconsistent annotations, varying image quality, and overlapping class meanings. The model faced class confusion, particularly between trash and barriers, as trash cans were often mistaken for barriers, especially when they looked like construction barrels. Misclassification also occurred with overflowing trash cans, depending on their context and shape. The 'garbage' had small sample sizes and limited visual variety, leading to misclassification and poor

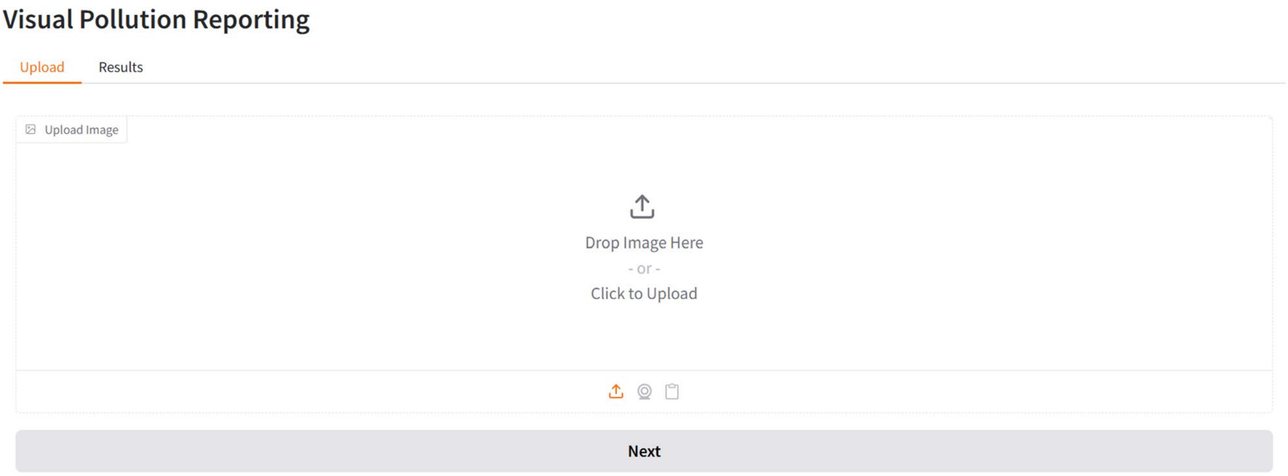
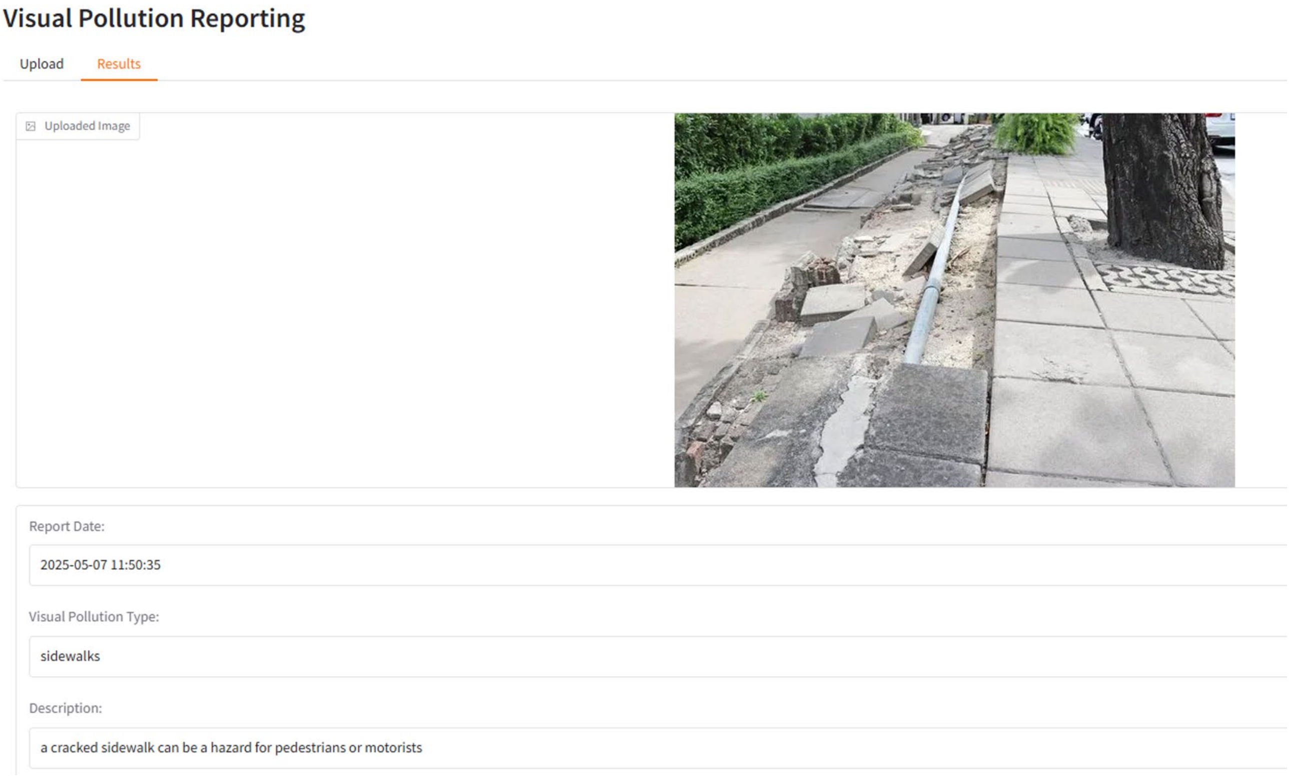


Fig. 22. Upload image interface.



Report Date:

2025-05-07 11:50:35

Visual Pollution Type:

sidewalks

Description:

a cracked sidewalk can be a hazard for pedestrians or motorists

Fig. 23. Auto-filled VP report interface.

generalization. These challenges emphasize the difficulties of merging datasets from different sources without consistent classification standards. Interestingly, these findings are different from those of (AlElaiwi, Al-antari et al. 2023)[8], who used YOLOv5 on a dataset similar to the first one in this study. While their results supported Yolo for urban scene understanding, their model achieved 93% mAP, 88% recall, 89% precision, and 89% F1-score. In contrast, this study achieved better accuracy and robustness through a multi-model ensemble approach. with 95% mAP, 95% recall, 91% precision, and 93% F1-score, which demonstrates a notable improvement. This improvement can be attributed to the multi-model integration and enhanced data preparation strategies, including random undersampling and targeted augmentation during the training stage, which helped balance class representation and reduce overfitting.

For image captions, according to the evaluators' results, the BLIP2-Flan-T5-XL model outperformed the fine-tuned BLIP2-OPT-2.7B model by achieving an average rating of 80% based on clarity and relevance. This performance is promising; additional gains are likely with domain adaptation and more targeted, category-specific prompting. By contrast, the fine-tuned BLIP2-OPT-2.7B did not meet the desired quality, mainly because the captioning dataset was small (3000 image–caption pairs) and included ChatGPT-generated captions used as labels, which can introduce noise and bias. Fine-tuning is expected to improve with a larger, domain-specific corpus of human-written captions and a modestly longer, better-tuned training schedule.

In real-world deployment, system outputs are reviewed by municipality staff before final reporting. If a report is identified as false, misleading, or irrelevant, it is immediately closed. This human-in-the-loop process helps mitigate the risks of false positives and model errors, ensuring the reliability and trustworthiness of the system in operational contexts.

So, the study achieved its goal by building an AI system that can detect, classify, and describe VP in city images without needing user intervention. Using both YOLOv5s and EfficientDet-D1 together helped improve detection accuracy, and BLIP-2 was used to write descriptions for the detected objects. In addition, a simple user interface was developed, as presented in Sect. "Generating the report", to simulate the reporting process and demonstrate the system's ability to generate complete VP reports. This helps reduce human mistakes and makes it easier for government organizations to understand and act on VP problems.

Conclusion

In conclusion, this study introduced a deep learning framework for detecting, classifying, and describing VP in urban environments using YOLOv5s, EfficientDet-D1, and BLIP-2. To the best of our knowledge, it is the first to combine YOLOv5 and EfficientDet for VP detection, while automatically generating descriptions with BLIP-2 to facilitate the reporting process. The methodology emphasized rigorous data preparation—including cleaning, balancing, and combining data from multiple sources—and demonstrated that an ensemble approach via Non-Maximum Suppression could achieve robust results (mAP = 0.95, F1-score = 0.93). In the captioning task, BLIP2-Flan-T5-XL achieved an average rating of 80% for clarity and relevance, outperforming the fine-tuned BLIP2-OPT-2.7B, which nonetheless showed promise for future domain-specific improvements.

Beyond reporting performance metrics, the implications of this work are significant for advancing smart city initiatives. By automating VP detection and reporting, the proposed system can reduce manual errors, streamline municipal workflows, and accelerate responses to urban issues. Despite these contributions, some limitations remain. The absence of an official dataset of human-written VP captions constrained evaluation, and the reliance on limited GPU resources restricted large-scale experimentation, such as k-fold validation. Moreover, occasional misclassifications between visually similar categories (e.g., trash and barriers) highlight the need for more balanced and representative data.

Looking ahead, several directions can enhance both impact and deployment. Expanding and diversifying datasets—particularly for underrepresented classes—will improve robustness and generalization. Introducing Arabic captioning is essential for practical adoption in Saudi municipalities, complementing the current English output. Additionally, incorporating Explainable AI methods (e.g., Grad-CAM) could increase transparency, while optimizing BLIP-2 prompts with domain-specific knowledge may further refine caption quality. Finally, integrating this framework into real-time mobile or edge-based applications would enable citizens to directly contribute to VP reporting, thereby scaling the system's reach and inclusivity.

Data availability

The datasets generated and/or analysed during the current study are available in the [Saudi Arabia Public Roads Visual Pollution Dataset] repository, [https://data.mendeley.com/datasets/bb7b8vtwry/5] and the [Urban Visual Pollution Dataset] repository, [https://www.kaggle.com/datasets/abhranta/urban-visual-pollution-dataset].

Received: 26 June 2025; Accepted: 21 August 2025
Published online: 26 August 2025

References

1. Quality of Life Program. Retrieved February 10 from https://www.vision2030.gov.sa/en/explore/programs/quality-of-life-program (2025).
2. Bibri, S. E. & Krogstie, J. Environmentally data-driven smart sustainable cities: Applied innovative solutions for energy efficiency, pollution reduction, and urban metabolism. *Energy Inf.* **3** (1), 29. https://doi.org/10.1186/s42162-020-00130-8 (2020).
3. Gao, H. et al. A systematic literature review and analysis of visual pollution. *Land*. **13** (7), 994. https://doi.org/10.3390/land13070994 (2024).
4. Mazroa, A. et al. Computer vision with explainable artificial intelligence for visual pollution detection in the Kingdom of Saudi Arabia. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3513696 (2024).
5. AlElaiwi, M. et al. Saudi Arabia public roads visual pollution dataset. https://doi.org/10.17632/BB7B8VTWRY.5 (2023).
6. Ministry of Municipalities and Housing. البلديات والإسكان تختتم العام 2024 بإنجازات وأرقام نوعية تعزز الارتقاء بجودة الخدمات في المدن السعودية. January 15 https://momah.gov.sa/ar/node/14892 (2025).
7. AlElaiwi, M. et al. Visual pollution real images benchmark dataset on the public roads. *Data Brief.* **50**, 109491. https://doi.org/10.1016/j.dib.2023.109491 (2023).
8. Kucharikova, Z. & Simko, J. *Visual Pollution Localization Through Crowdsourcing and Visual Similarity Clustering* (IEEE, New York, 2017).
9. Chmielewski, S. Towards managing visual pollution: A 3D isovist and voxel approach to advertisement billboard visual impact assessment. *ISPRS Int. J. Geo-Inf.* **10** (10), 656. https://doi.org/10.3390/ijgi10100656 (2021).
10. Ahmed, N., Islam, M. N., Tuba, A. S., Mahdy, M. R. C. & Sujauddin, M. Solving visual pollution with deep learning: A new nexus in environmental management. *J. Environ. Manag.* **248**, 109253. https://doi.org/10.1016/j.jenvman.2019.07.024 (2019).
11. Hossain, M. Y., Nijhum, I. R., Sadi, A. A., Shad, M. T. M. & Rahman, R. M. Visual pollution detection using Google street view and YOLO. In *2021 IEEE 12th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, UEMCON 2021* 433–440. https://doi.org/10.1109/UEMCON53757.2021.9666654 (2021).
12. Pham, V., Nguyen, D. & Donan, C. Road damages detection and classification with YOLOv7. http://arxiv.org/abs/2211.00091 (2022).

13. Arya, D. et al. Crowdsensing-based road damage detection challenge (CRDDC'2022). In *Proceedings –2022 IEEE International Conference on Big Data, Big Data 2022* 6378–6386. https://doi.org/10.1109/BigData55660.2022.10021040 (2022).
14. Pramanik, A., Bijoy, M. H. I. & Rahman, M. S. Detection of potholes using convolutional neural network models: A transfer learning approach. In *Proceedings of 2021 IEEE International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things, RAAICON 2021* 73–78. https://doi.org/10.1109/RAAICON54709.2021.9929623 (2021).
15. Mirajkar, R. et al. Enhanced pothole detection in road condition assessment using YOLOv8. In *Proceeding ICWITE 2024: IEEE International Conference Women Innovation Technology Entrepreneurship*. 429–433 https://doi.org/10.1109/ICWITE59797.2024.10502437 (2024).
16. Kumari, S., Gautam, A., Basak, S. & Saxena, N. YOLOv8 based deep learning method for potholes detection. In *Proceeding 2023 IEEE International Conference on Computer Vision and Machine Intelligence CVMI 2023*. https://doi.org/10.1109/CVMI59935.2023.10465038 (2023).
17. Tasnim, N. H., Afrin, S., Biswas, B., Anye, A. A. & Khan, R. Automatic classification of textile visual pollutants using deep learning networks. *Alexandria Eng. J.* **62**, 391–402. https://doi.org/10.1016/j.aej.2022.07.039 (2023).
18. Titu, M. F. S., Chowdhury, A. A., Haque, S. M. R. & Khan, R. Deep-learning-based real-time visual pollution detection in urban and textile environments. *Science*. **6** (1), 5. https://doi.org/10.3390/sci6010005 (2024).
19. AlElaiwi, M. et al. VPP: Visual pollution prediction framework based on a deep active learning approach using public road images. *Mathematics* **11** (1), 186. https://doi.org/10.3390/math11010186 (2023).
20. Alamri, E. R., Daym Alquradi, M. B., & Alsaeed, D. H. Building an automatic system for detecting manifestations of visual pollution using geospatial techniques and deep learning: An applied study on abandoned vehicles in Riyadh. *Arab. Int. J. Inform. Technol. Data* **4** (2), 7 (2024).
21. Banda, A. International image captioning using CNN and LSTM. *J. Res. Appl. Sci. Eng. Technol.* **9** (8), 2666–2669. https://doi.org/10.22214/ijraset.2021.37846 (2021).
22. Rashtchian, C., Young, P., Hodosh, M. & Hockenmaier, J. Collecting image annotations using amazon's mechanical Turk. 139–147. https://doi.org/10.5555/1866696.1866717 (2010).
23. Chen, H., Ding, G., Lin, Z., Zhao, S. & Han, J. *Show, Observe and Tell* (Attribute-driven Attention Model for Image Captioning, 2018).
24. Lin, T. Y. et al. Microsoft COCO: Common objects in context. http://arxiv.org/abs/1405.0312 (2014).
25. Ghosh, A., Acharya, A., Saha, S., Jain, V. & Chadha, A. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. http://arxiv.org/abs/2404.07214 (2024).
26. Wu, J., Gan, W., Chen, Z., Wan, S. & Yu, P. S. Multimodal large language models: A survey. http://arxiv.org/abs/2311.13165 (2023).
27. Anagnostopoulou, A., Gouvea, T. & Sonntag, D. Enhancing journalism with AI: A study of contextualized image captioning for news articles using LLMs and LMMs. http://arxiv.org/abs/2408.04331 (2024).
28. Rückert, J. et al. Overview of ImageCLEFmedical 2023--caption prediction and concept detection. In *Conference and Labs of the Evaluation Forum* (2023).
29. Zhou, W. et al. Transferring pre-trained large language-image model for medical image captioning. http://ceur-ws.org (2023).
30. Jeon, M., Ko, J. & Cheoi, K. Enhancing surveillance systems: Integration of object, behavior, and space information in captions for advanced risk assessment. *Sensors* **24** (1), 292. https://doi.org/10.3390/s24010292 (2024).
31. Ly, C., Mittal, U., Madaan, V. & Agrawal, P. Vehicle detection and classification using an ensemble of efficientdet and YOLOv8. *PeerJ Comput. Sci.* **10**, e2233. https://doi.org/10.7717/peerj-cs.2233 (2024).
32. Panigrahi, A. Urban visual pollution dataset. Retrieved February 18 from https://www.kaggle.com/datasets/abhranta/urban-visual-pollution-dataset (2025).
33. Li, J., Li, D., Savarese, S. & Hoi, S. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. https://github.com/salesforce/LAVIS/tree/main/projects/blip2 (2023).
34. Tan, M., Pang, R. & Le, Q. V. EfficientDet: Scalable and efficient object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 10778–10787. https://doi.org/10.1109/CVPR42600.2020.01079 (2020).
35. Nour, S. M., Shehab, R. S., Said, S. A. & Abdel Halim, I. T. Harnessing the power of an integrated artificial intelligence model for enhancing reliable and efficient dental healthcare systems. *Appl. Syst. Innov.* **8** (1), 7. https://doi.org/10.3390/asi8010007 (2025).
36. Seo, P. H., Sharma, P., Levinboim, T., Han, B. & Soricut, R. Reinforcing an image caption generator using off-line human feedback. http://www.conceptualcaptions.com/challenge (2020).
37. Hu, E. J. et al. LoRA: Low-rank adaptation of large language models. http://arxiv.org/abs/2106.09685 (2022).
38. Hoque, M., Hasan, R., Emon, S., Khalifa, F. & Rahman, M. M. Medical image interpretation with large multimodal models notebook for the CS_Morgan lab at CLEF 2024. https://github.com/Hasan-MdRakibul (2024).
39. Cernat, A., Sakshaug, J., Christmann, P. & Gummer, T. The impact of survey mode design and questionnaire length on measurement quality. *Sociol. Methods Res.* **53** (4), 1873–1904. https://doi.org/10.1177/004912412211401 (2024).

Acknowledgements

The authors would like to thank Saudi Electric University for its understanding.

Author contributions

Conceptualization, H.A. and N.A.; methodology, A.M. and N.A.; software, H.M.; validation, H.M. and N.A.; formal analysis, H.M. and N.A.; investigation, H.M. and N.A.; resources, H.Aand N.A.; data curation, H.M. and N.A; writing—original draft preparation, H.M and N.A.; writing—review and editing, N.A.; visualization, H.M and N.A.; supervision, N.A.; project administration, N.A; funding acquisition, N.A. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to N.A.

Reprints and permissions information is available at www.nature.com/reprints.