2026-08

Vispol: a real-world urban visual pollution dataset for environmental and infrastructure monitoring

Ahmad, Jubaer

Independent University, Bangladesh (IUB)

VisPol: A Real-World Urban Visual Pollution Dataset for Environmental and Infrastructure Monitoring

Prepared By:

Jubaer Ahmad (2221972)
Md. Faiji Akbar (2220717)
M.M. Simoon (2221681)

Dissertation submitted in partial fulfillment for the Degree of Bachelor of Science in Computer Science and Engineering

Department of Computer Science & Engineering
School Of Engineering, Technology & Sciences (SETS)
Independent University, Bangladesh

August 2026

Attestation

We understand the nature of plagiarism, and we are aware of the university's strict policy on the matter. We certify that this is the original work done by us. However, others' written work or software used in any part of this project is properly cited following internationally accepted academic guidelines .

Signature:

I. Jubaer Ahmad : _____

II. Md. Faiji Akbar : _____

III. M.M. Simoon : _____

Evaluation Committee

Supervisor

Name: …… Signature: ……

Internal Examiner 1

Name: …… Signature: ……

Internal Examiner 2

Name: …… Signature: ……

External Examiner

Name: …… Signature: ……

Office Use

…… ……

Program Coordinator Head of the Department

……

Director Graduate Studies, Research and Industry Relations

……

Library

Acknowledgement

The authors would like to express their sincere gratitude to Almighty Allah for granting them the strength, patience, and opportunity to complete this dissertation.

We are deeply grateful to our supervisor, Dr. Md. Rashedur Rahman, Assistant Professor, Department of Computer Science and Engineering, Independent University, Bangladesh, for his valuable guidance, support, and encouragement throughout this research. We also sincerely thank Mustaqueem Alam and J.M. Sadik-Ul Islam Smaron for their valuable guidance, technical support, and contributions to this work.

We gratefully acknowledge the Center for Computational & Data Sciences (CCDS), Department of Computer Science and Engineering, and the School of Engineering, Technology and Sciences (SETS), Independent University, Bangladesh, for providing the academic environment and resources necessary to conduct this research.

We also acknowledge the use of AI-based tools and Grammarly to improve the readability and clarity of the manuscript. AI-based image-generation tools were used to create selected illustrative visuals, where applicable and were reviewed and human-verified by the authors for relevance, accuracy, and consistency with the research content, and their use is explicitly indicated in the corresponding figure captions. No AI-based tools were used for data analysis, experimental evaluation, or interpretation of the research results.

Abstract

Rapid urbanization has made visual pollution an increasing challenge in cities such as Dhaka, Bangladesh, affecting urban appearance, environmental quality, and public safety. Common forms include unauthorized billboards, poster and banner clutter, roadside waste, exposed utility wires, damaged walls, and deteriorated roads. However, systematic monitoring is difficult because urban scenes are diverse and highly cluttered. Although deep learning–based object detection has advanced rapidly, realistic, deployment-oriented datasets for visual pollution detection remain limited. To address this gap, this study introduces VisPol, a real-world dataset containing 1,097 images and 15,933 annotated instances across six classes: Billboard, Poster/Banner, Waste, Wire, Damaged Wall, and Damaged Road. Images were collected from residential, commercial, and transportation areas across Dhaka and annotated using a standardized quality-assurance protocol.

To evaluate detection performance and computational efficiency, five YOLOv8 variants: YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x were trained under identical experimental settings. On the independent VisPol test set, YOLOv8x achieved the highest F1-score (0.3311) and mAP@50 (0.2486), while YOLOv8l achieved the highest precision (0.4180) and mAP@50–95 (0.1340). YOLOv8n provided the fastest inference at 4.7 ms per image, making it suitable for resource-constrained applications. Considering detection accuracy, computational cost, and model complexity together, YOLOv8s provides the strongest accuracy-efficiency trade-off, while YOLOv8m provides a comparable performance level at higher computational cost.

Class-level analysis showed that Waste and Poster/Banner were generally easier to detect, whereas Damaged Wall and Damaged Road were more challenging due to irregular boundaries, substantial visual variation, and ambiguous object regions. The models were also evaluated on the Visual Pollution Dhaka Streets and PlasticPol datasets for external validation. Performance decreased on both datasets, indicating limited cross-dataset generalization. However, qualitative analysis showed that this reduction was not attributable to domain shift alone. Differences in annotation protocols, object scale, scene composition, image acquisition conditions, class definitions, incomplete annotations, and object-grouping strategies also affected the evaluation results, causing some visually reasonable detections to be counted as incorrect.

Overall, this research contributes the VisPol dataset, a standardized annotation methodology, a comparative evaluation of five YOLOv8 architectures, and cross-dataset validation of model robustness under real-world conditions. The findings demonstrate the challenges of automated visual pollution detection in complex urban environments and highlight the importance of consistent annotation practices for reliable cross-dataset evaluation. VisPol and the proposed evaluation framework can support future research in visual pollution detection and practical applications in urban environmental monitoring and smart-city management.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Background

Visual pollution has become a growing concern in rapidly developing cities such as Dhaka, Bangladesh. This refers to the degradation of the visual environment caused by unwanted or poorly maintained elements such as unauthorized billboards, damaged infrastructure, tangled utility wires, and accumulated urban waste [1][2]. Besides, rapid urbanization and limitations in regulatory practices have contributed to the increasing presence of these problems [3][4]. Moreover, urban waste is also an important part of visual pollution because it can create risks for ecosystems and public health [5].

Computer vision and deep learning algorithms have opened new possibilities for monitoring urban environments automatically [6][7]. In conventional monitoring, visual pollution is often identified and documented through manual inspection, in which the process requires considerable time and human effort and may also lead to inconsistent observations [8]. Object detection is one of the commonly used computer vision techniques for locating and classifying objects in images. Its development has been strongly influenced by deep learning, particularly convolutional neural networks and more recent transformer-based approaches [9][10]. Such methods are proved to be used to detect and localize different forms of visual pollution from urban images.

AI-based environmental monitoring can therefore support more efficient urban management and planning [11]. However, the performance of an object detection system depends largely on the quality and characteristics of the dataset used for training and evaluation [12][13]. For visual pollution detection, a dataset needs to reflect the conditions found in actual urban environments rather than only controlled image acquisition settings. Many existing datasets do not fully capture factors such as changes in illumination, weather, camera distance, and complex urban backgrounds [14][15]. These differences can make it difficult to determine how well a trained detection model will perform when it is applied to real-world smart city monitoring scenarios.

1.2 Motivation

Visual pollution affects both the appearance and practical quality of urban environments. In developing regions such as Bangladesh, limited resources for infrastructure management can make these problems more difficult to address. Unauthorized advertising, deteriorating infrastructure and accumulated waste can create unsafe conditions for residents and may also hinder sustainable urban development [4].

Manual monitoring has several practical limitations. It depends on human inspectors, covers a limited geographic area, and may result in differences in how observations are recorded. AI based monitoring systems can provide wider coverage and more consistent detection, but their effectiveness in real-world settings still depends on the quality and diversity of the training data. Previous studies have shown that models with strong performance on benchmark test sets may perform less effectively when they are applied to unseen data with different environmental conditions and data distributions [8][9]. For this reason, datasets intended for deployment should represent the variation found in actual urban environments. This includes differences in lighting, weather, camera perspective, object occlusion, and background complexity [10][11]. Including these elements while creating the dataset can make sure that the models work better when they are used in life situations. It is important to think about these things when building the data. The models need to be ready for the world. That is why it is helpful to include these factors during the dataset development. It makes the models more dependable when they are put to use.

1.3 Problem Statement

Although deep learning and object detection have improved considerably, several problems still need to be addressed for visual pollution detection in real-world urban environments. The main problems considered in this research are:

- **Existing Research Problems:** Existing research does not provide a unified approach for evaluating visual pollution detection in resource limited urban environments, particularly in developing countries.
- **Dataset Limitations:** Publicly available datasets for visual pollution and urban waste detection often have limited diversity, imbalanced class distributions, controlled image acquisition conditions, or incomplete annotations. They may also contain fewer examples of real-world challenges such as occlusion, motion blur, and poor illumination.
- **Deployment-oriented Challenges:** Models trained on controlled datasets may lose performance when they are applied to actual urban scenes with different levels of noise, clutter, object scale and environmental variation. Therefore, performance on a conventional test set may not always reflect how a model will perform after deployment.
- **Technical Challenges:** Detecting small objects, dealing with class imbalance, and keeping computational requirements manageable remain important challenges. These issues become particularly relevant when object detection models are intended for deployment on resource constrained edge devices.

1.4 Research Objectives

- To develop a real-world visual pollution detection framework using a dataset that is suitable for deployment and multiple object detection models.
- Collect a dataset of visual pollution instances from different areas of Dhaka, Bangladesh, under varied real-world environmental conditions.
- Develop an annotation protocol for six pollution classes that aligns with the plan deployment-oriented models: billboard, poster/banner, waste, wire, damaged wall and damaged road.
- Analyze the dataset. Report its main statistics, including class distributions and potential class imbalance.
- Evaluate five YOLOv8 variants (YOLOv8n, YOLOv8s, YOLOv8m YOLOv8l and YOLOv8x) using the VisPol dataset under the same experimental settings.
- Performing validation using independent public datasets to examine cross-dataset generalization and model robustness.
- Compare the performance and computational requirements of the evaluated models under visual conditions, including object occlusion, and identify suitable models for deployment-oriented.

1.5 Research Questions

This research addresses the following questions:

- **RQ1:** How can a visual pollution dataset be constructed and annotated to represent the diversity and complexity of real-world urban environments?
- **RQ2:** How do the five YOLOv8 variants compare in detection performance, inference speed, and computational complexity when trained and evaluated on VisPol?
- **RQ3:** How well do models trained on VisPol generalize to independent visual pollution datasets with different environmental and data characteristics?
- **RQ4:** Which visual pollution classes are most challenging to detect, and what visual or data-related factors contribute to detection errors?
- **RQ5:** To what extent do differences in data distribution, annotation protocols, and object representation affect cross-dataset evaluation results?

- **RQ6:** Which YOLOv8 variant provides the most suitable trade-off between detection performance and computational efficiency for real-world visual pollution monitoring?

1.6 Research Contributions

This research makes the following contributions to computer vision and urban pollution monitoring:

- **VisPol Dataset:** A real-world visual pollution dataset comprising 1,097 images and 15,933 annotated object instances across six classes: Billboard, Poster/Banner, Waste, Wire, Damaged Wall, and Damaged Road, collected from diverse urban environments in Dhaka, Bangladesh.
- **Deployment-Oriented Data Collection:** A collection strategy designed to preserve real-world variation in object scale, viewing distance, occlusion, background complexity, and environmental conditions.
- **Standardized Annotation and Quality Assurance:** A structured bounding-box annotation procedure using Roboflow, followed by quality checking to improve annotation consistency across the six classes.
- **YOLOv8 Benchmark:** A systematic comparison of YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x using detection metrics, inference speed, and model complexity.
- **Cross-Dataset Validation:** Evaluation on independent public datasets to examine model generalization and investigate the influence of domain characteristics, annotation practices, and object representation on performance.
- **Deployment-Oriented Model Assessment:** Analysis of the relationship between detection accuracy and computational **cost** to identify the most practical YOLOv8 variant for real-world visual pollution monitoring.

1.7 Alignment with Sustainable Development Goals

This research contributes to the United Nations Sustainable Development Goals (SDGs), particularly those related to sustainable urban development, responsible environmental management, and climate action. Its primary relevance is to SDG 11, with additional connections to SDG 12 and SDG 13.

SDG 11 – Sustainable Cities and Communities: This research supports sustainable urban development by enabling automated detection of visual pollution in complex city environments. The VisPol dataset represents urban issues such as billboards, posters/banners, waste, exposed wires, damaged walls, and damaged roads, providing a foundation for data-driven urban monitoring and future smart-city applications [11].

SDG 12 – Responsible Consumption and Production: The Waste class provides a computer-vision-based means of identifying waste in urban environments. Such detection can support data-driven environmental monitoring and provide information for future urban waste-management systems. However, the present study is limited to waste detection and model evaluation and does not directly address waste collection, recycling, or resource recovery [2].

SDG 13 – Climate Action: The research contributes indirectly to climate action by developing computer-vision resources for monitoring environmental conditions in urban areas. VisPol can support future systems for continuous environmental assessment and data-driven urban management, contributing to broader efforts toward more sustainable and resilient cities.

2 Related Work

2.1 Visual Pollution Detection

Visual pollution has received less attention in computer vision research than several other types of environmental pollution [40]. Hossain et al. [40] developed an end-to-end visual pollution detection system for Dhaka, Bangladesh, using YOLOv5, Faster R-CNN, and EfficientDet. Their best result reached an mAP@50 of 0.85. The research used a dataset of 1,400 images collected from Google Street View. Although the system showed promising detection performance, the dataset had limited coverage of the different conditions found in real urban environments [40]. End-to-end artificial-intelligence systems have been proposed for analysing and detecting different forms of environmental pollution. Hossain et al. [44] developed an AI-based pollution analysis and object-detection system, demonstrating the potential of automated vision systems for environmental monitoring.

Al Mazroa et al. [42] later proposed MCVXAI-VPD, which combines YOLOv5 with explainable AI for visual pollution detection in Saudi Arabia. The model achieved an accuracy of 98.20% [42]. Their work also considered model interpretability, which can be useful when automated detection systems are applied in practical settings. This study was done in an area, so we need to look at data from other cities to see if the results are the same everywhere. We have to check if the findings from this study work in urban environments. The study was conducted in one place, so we need to do work to see how well the results apply to other cities.

2.2 Urban Waste Detection

Urban waste detection is another important application of deep learning-based object detection. Sayem et al. [2] developed the WaRP dataset, which contains 10,146 images covering 28 recyclable categories. They also proposed the 2S_DenseViT classification model, which achieved an accuracy of 83.11%. The model combined CNN and transformed-based approaches for waste classification [2].

Deep learning and street-view imagery can also be used to study how physical characteristics of urban streets influence visual perceptions. Huang et al. [41] examined physical urban-street features and visual walkability perception, providing additional evidence that image-based analysis can support understanding of urban environmental quality.

Real-time trash detection has also been investigated using CCTV-based monitoring and deep convolutional neural networks. Raza et al. [21] explored this approach for automated trash identification, demonstrating the relevance of computer vision to continuous urban cleanliness

monitoring. Federated learning has also been explored for trash classification in smart-city environments. Ghosh et al. [37] proposed an autonomous garbage-collection vehicle that integrates multiple sensors, deep-learning-based object recognition, GPS-based navigation, and an automated bin-lifting mechanism to enable efficient household waste collection with reduced human involvement. Khan et al. [22] investigated federated deep learning as an approach for improving automated waste-classification systems while considering distributed learning environments. Computer vision-based trash detection has also been extended beyond roads and streets to water environments. Tharani et al. [23] investigated trash detection in water channels, demonstrating the broader applicability of automated waste-monitoring techniques.

Deep learning has also been applied to integrated smart waste-management and classification systems. Hasan et al. [24] investigated deep-learning-based waste classification as part of smart-city waste management. Attention mechanisms have also been considered for improving automated trash detection in complex visual environments. Tharani et al. [25] proposed an attention neural network for detecting trash in water channels.

Dipo et al. [3] used YOLOv12 for waste classification and reported an mAP@50 of 0.78. Shrestha et al. [11] tested YOLOv8m on a dataset that has 2,350 images taken in Nepal and got an mAP@50 of 0.9825. These studies show that YOLO-based models can give results for waste detection and classification tasks [3][11]. In Bangladesh, Das et al [45] made a trash dataset, with 4,418 images. Said that using YOLOv5x gave an mAP@50 of 0.844. Their work also shows the importance of thinking about dataset features and the environment when creating waste detection models [45]. Machine learning is increasingly being investigated as a tool for supporting sustainable municipal waste management. Taweesan et al. [29] examined machine-learning-based approaches for identifying sustainable solutions in municipal waste-management systems.

The combination of YOLO-based detection and IoT technologies has become an important direction in smart waste-management research. Gelar et al. [38] reviewed YOLO and IoT applications in smart waste management, highlighting the growing integration of automated visual detection with smart-city systems. Recent studies have also explored real-time intelligent garbage monitoring using YOLO-based object detectors. Abo-Zahhad and Abo-Zahhad [48] investigated YOLOv8 and YOLOv5 for garbage monitoring and efficient collection, further supporting the use of real-time object detection in environmental sustainability applications.

2.3 Deep Learning Object Detection

The YOLO (You Only Look Once) family is widely used for real-time object detection. YOLOv8 is one of the commonly used versions of the YOLO family and includes improvements in feature

extraction, inference speed, and small-object detection [13][14]. Recent research has also applied YOLOv8-based instance segmentation to automated garbage recognition and sorting. Wang et al. [30] investigated YOLOv8-seg for waste recognition, illustrating the use of segmentation-based approaches when object-level localization alone may not be sufficient. Zhang and other researchers [13] created DDS-YOLO, which is based on YOLOv10n, and tested it on the RDD2022 dataset. The model reached 84.5% mAP@50. The model also cut down the number of parameters by 30%. Flops, by 37.8% compared with the baseline [13].

Ding et al. [15] developed SCD-YOLO for road crack detection by incorporating Shift-Wise Cross-Scale Dynamic Head modules. On the other hand, for the RDD2022 dataset, the model improved mAP by 4.1% compared with the baseline YOLOv8n [15]. Lin et al. [16] proposed YOLO-ROC, a lightweight detection model that reduced the number of parameters by 70.4% while maintaining competitive accuracy. Their results indicate that reducing model size can be useful when detection models are intended for edge deployment [16]. Edge intelligence provides another direction for deploying computer-vision systems in practical urban environments. Chowdhury et al. [32] investigated an edge-intelligence-based computer-vision solution, emphasizing the relevance of computationally efficient visual systems for real-world applications.

Deep learning has also been applied to detect road cracks and potholes as part of smart-city infrastructure monitoring. Chu et al. [17] demonstrated the use of deep learning for automated road-crack and pothole detection, highlighting the potential of computer vision for infrastructure condition monitoring. Road-defect detection can also contribute to assessing broader road safety risks and structural deterioration. Kek et al. [20] investigated crack detection together with structural deterioration assessment, showing how automated visual analysis can support road-infrastructure monitoring. YOLO-based models have also been applied specifically to real-time road-pavement damage detection. Sami et al. [43] investigated an improved YOLOv5 approach for road infrastructure management, demonstrating the suitability of real-time object detection for automated pavement monitoring.

Deep learning-based visual monitoring has also been applied to other smart-city problems in Bangladesh. Sukanto et al. [31] investigated unauthorized-vehicle detection, showing how object detection can support automated monitoring of urban environments. Computer vision has also been used for automatic monitoring of traffic-related events and vehicle activities. Karim et al. [34] investigated visual traffic-incident detection through automated monitoring, demonstrating another application of vision-based urban infrastructure and mobility analysis. Recent work has further explored AI-based detection of infrastructure faults using modern YOLO architectures. Rakin et al. [35] investigated YOLOv11 for infrastructure fault detection, supporting the growing use of object detection for automated urban infrastructure monitoring. Pavement-defect detection has also been

studied using efficient variants of newer YOLO architectures. Lin et al. [36] proposed YOLO11-WLBS for pavement-defect detection, further demonstrating the relevance of lightweight object-detection models to infrastructure monitoring.

Two-stage detection methods such as Faster R-CNN have also been evaluated for object detection tasks. Darisang et al. [49] compared YOLOv11, SSD-MobileNetV3, and Faster R-CNN with ResNet50 FPN for real-time waste detection and classification. YOLOv11 achieved the best overall detection performance, with 55.69% mAP@0.5:0.95, an average inference time of 14.1 ms, and 71.10 FPS, demonstrating its suitability for real-time waste monitoring. Some other studies have also used transformer-based and hybrid detection methods for this task [18][19]. When you want to use these models in the real-world, you have to think about how much computing power they need because that can be a big problem. YOLOv11, Faster R-CNN and these other methods all have their bad points.

2.4 Existing Datasets for Visual Pollution and Urban Monitoring

Several datasets have been developed for waste and infrastructure monitoring, but they differ in their coverage, size and intended applications [28]. The TACO (Trash Annotations in Context) dataset, introduced by Kumar et al. [39], contains 1,500 high-resolution images covering 28 waste categories and 60 sub-categories. The images were collected in diverse natural backgrounds. Although the dataset provides detailed waste annotations, its relatively small size and focus on litter in natural settings make it less suitable for broader urban visual pollution detection [39].

The Road Damage Dataset (RDD2022) [9][15] contains more than 47,000 images collected from six countries and was developed for road infrastructure monitoring. Its main focus is on road defects, so it does not cover other visual pollution categories such as waste, billboards, posters, wires, or damaged walls [9]. Dwivedi et al. [9] also experiment with cross-country domain shift through a federated learning experiment. Their results showed that the difference between countries can affect model performance, with India showing higher visual diversity and corresponding performance degradation [9]. Bangladesh-specific road-image datasets have also been developed for computer-vision research. The BRSSD10k dataset [46] provides a large-scale segmentation dataset focused on Bangladeshi road scenarios, offering locally relevant visual data for road and urban-environment analysis.

The GlobalWaste Data archive [26][27] combines 19 publicly available datasets into a collection of 89,807 images for waste classification. When we put together datasets that have types of things in them, it can cause a problem. Some types of things like paper and plastic are easy to find. Things like hazardous materials are not found as much.

The BRSSD10k dataset has a lot of pictures, 10,082 to be exact, from roads in Bangladesh. This dataset is useful for looking at things that happen on roads in that area. However, it is mostly used for finding objects in pictures and does not show all the different types of visual pollution that the BRSSD10k dataset could have shown. The BRSSD10k dataset is limited in what it can tell us about visual pollution categories. Waste-detection datasets have also been developed for aquatic environments rather than conventional urban streets. Cheng et al. [47] introduced the FLOW dataset and benchmark for floating-waste detection in inland waters, showing the expansion of computer-vision-based environmental monitoring to water pollution.

Table 2.1: Comparison of Existing Visual Pollution and Waste Detection Datasets.

Dataset	Images	Classes	Geographic Focus	Image Quality	Applicability
TACO	1,500	28 main/60 sub	Diverse (Global)	High	Litter in natural settings; limited urban focus
RDD2022	47,000+	4 (Road damage)	Multi-country (6)	Variable	Road infrastructure; not general visual pollution
GlobalWaste	89,807	14 main/68 sub	Global merge	Mixed	Waste classification; severe class imbalance
BRSSD10k	10,082	34	Bangladesh roads	High	Bangladeshi context; instance segmentation only
VisPol (Proposed)	1097	6 (Excluding signboard)	Dhaka, Bangladesh	Realistic/Diverse	Urban visual pollution; deployment-oriented

2.5 How VisPol Addresses Existing Limitations

Although existing datasets provide useful benchmarks for waste, road damage and other urban monitoring tasks, they do not fully cover the conditions required for real-world visual pollution detection [16-25, 28-37]. Vispol was developed to address these gaps:

- Controlled Environments: Unlike laboratory-collected datasets, VisPol images are captured in authentic urban settings under uncontrolled lighting, weather, and environmental conditions, ensuring realistic data representative of deployment scenarios.
- Clean Backgrounds: VisPol deliberately includes complex urban backgrounds with occlusions, multiple objects per image, and cluttered environments that reflect actual urban complexity.
- Multi-Range Imaging: The dataset includes images captured from diverse distances and angles, ranging from close-up detail views to wide-angle urban scenes.

- High Scene Complexity: VisPol encompasses diverse geographic locations across Dhaka, capturing various urban contexts, including commercial areas, residential neighborhoods, and industrial zones.
- High Deployment Realism: Every design decision, from acquisition methodology to annotation protocol to external validation prioritizes deployment readiness.

2.6 Research Gap Analysis

Despite recent advances in object detection and waste classification, existing research still has several gaps that need to be addressed [41,43,44,47-49]:

Table 2.2: Research Gap Analysis and VisPol Contributions.

Research Gap	Problem Statement	Existing Limitation	VisPol Solution
Methodological Gap 1	Lack of deployment-oriented evaluation frameworks	Datasets developed under controlled conditions fail in real-world deployment	Comprehensive environmental variability in data collection and evaluation
Methodological Gap 2	Absence of standardized annotation protocols for visual pollution	Inconsistent labeling across studies hinders reproducibility	Rigorous multi-stage QA annotation process using Roboflow
Dataset Gap 1	Limited region-specific datasets for developing countries	Most datasets focus on developed nations; limited applicability to South Asia	Dhaka-specific dataset with diverse urban contexts
Dataset Gap 2	Severe class imbalance in existing waste datasets	Rare pollution types underrepresented; poor model generalization	Natural class distribution across six visual pollution categories
Dataset Gap 3	Lack of multi-scale object instances in datasets	Poor performance on small objects; limited detection coverage	Diverse image scales and camera distances capturing multiple object sizes
Evaluation Gap 1	Limited cross-dataset generalization testing	Unclear how models transfer to different geographic contexts	External validation on independent public datasets with cross-domain analysis
Deployment Gap 1	Insufficient analysis of computational requirements	Unclear feasibility for edge deployment on resource-constrained devices	Detailed analysis of inference speed, parameters, and FLOPs for model variants
Deployment Gap 2	Lack of guidance on model selection for practical deployment	Practitioners unclear on which model to select given deployment constraints	Comprehensive comparison enabling informed model selection based on constraints

3 Materials and Methodology

3.1 Overall Research Framework

The overall research framework used in this study is shown in Figure 3.1. It summarizes the main steps of the research from the data collection and dataset preparation to model training, evaluation and external validations. The process started by selecting image collection locations from different areas of Dhaka, Bangladesh. The locations chosen include different types of urban environments. Images were then collected under real-world conditions with variations in viewpoint, object distance, illumination, occlusion and background complexity. These images formed the initial Vispol dataset.

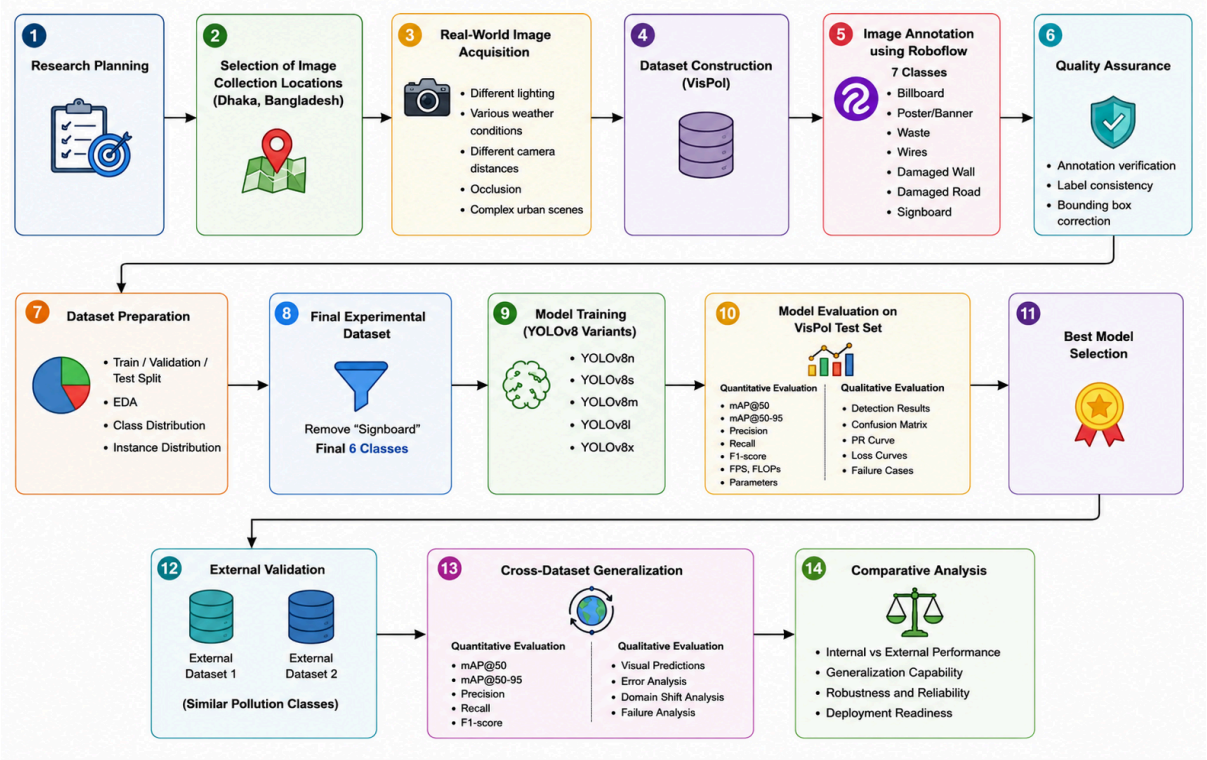


Figure 3.1: AI-generated illustration of the methodology and overall research framework.

After collection, the images were annotated using the Roboflow platform based on predefined annotation guidelines. At the initial stage, seven visual pollution classes were considered: billboard, poster/banner, waste, wire, damaged wall, damaged road, and signboard. The annotations were then reviewed for quality and consistency. Following this review, the Signboard class was removed due to several ambiguities of bill and poster overlap resulting in a final dataset containing six visual pollution classes. The prepared dataset was divided into training, validation, and test subsets. Exploratory data analysis was also performed to examine the dataset characteristics and class distributions.

The next step was to train and evaluate five YOLOv8 variants: YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x. Model performance was assessed using standard object detection metrics along with visual inspection of the predicted results. After the internal evaluation, the trained models were tested on two independent public datasets containing related visual pollution categories. We used this validation to see how well our model works with different sets of data and to find out if it behaves differently with each set of data.

Finally, we looked at the results from the external evaluations to see how well our model performs, if it can be used in different situations and if it is suitable to use in the real-world with the computer resources we have.

3.2 Study Area

3.2.1 Image Collection Locations

Images were collected from several urban areas of Dhaka, Bangladesh, covering residential, commercial and transportation environments. The main collection locations were Badda, Uttara, Bashundhara Residential Area, Khilkhet and Kuril. These areas were selected because they contain a range of urban scenes where visual pollution can be observed, including roadside advertising, construction activities, accumulated waste, and different types of urban infrastructure. The selected locations also provided variation in scene composition and environmental conditions, which helped in building a dataset that reflects real-world urban environments.

3.2.2 Image Collection Strategy

Images were collected in collaboration by a small research team from the selected areas of Dhaka. The locations were surveyed both on foot and by vehicle to cover different streets, intersections, commercial areas, residential neighborhoods and roadside environments. Images were taken from near, medium and far distances and from different viewing angles. This helped capture differences in object size and perspective within the collected images.

Unlike many existing benchmark datasets collected under controlled environments, the proposed VisPol dataset intentionally includes real-world complexities, such as cluttered backgrounds, heavy traffic, pedestrians, parked vehicles, partial occlusions, overlapping objects, construction activities, and varying urban layouts. Most images were captured during the daytime under natural lighting conditions to maximize image clarity while preserving realistic environmental variability. This collection strategy enhances the practical applicability of the dataset for deployment in smart city monitoring systems.

3.2.3 Collection Heatmap

Figure 3.2 shows the spatial distribution of collection locations for the images used in this study. We can observe from the heatmap that most of the collected images are concentrated on various selected areas of Dhaka, which implies that we have taken more than one urban location for our dataset rather than a single area. This geographic coverage has provided diversity in scene composition and urban environments and mitigated against reliance on observations from a particular location.

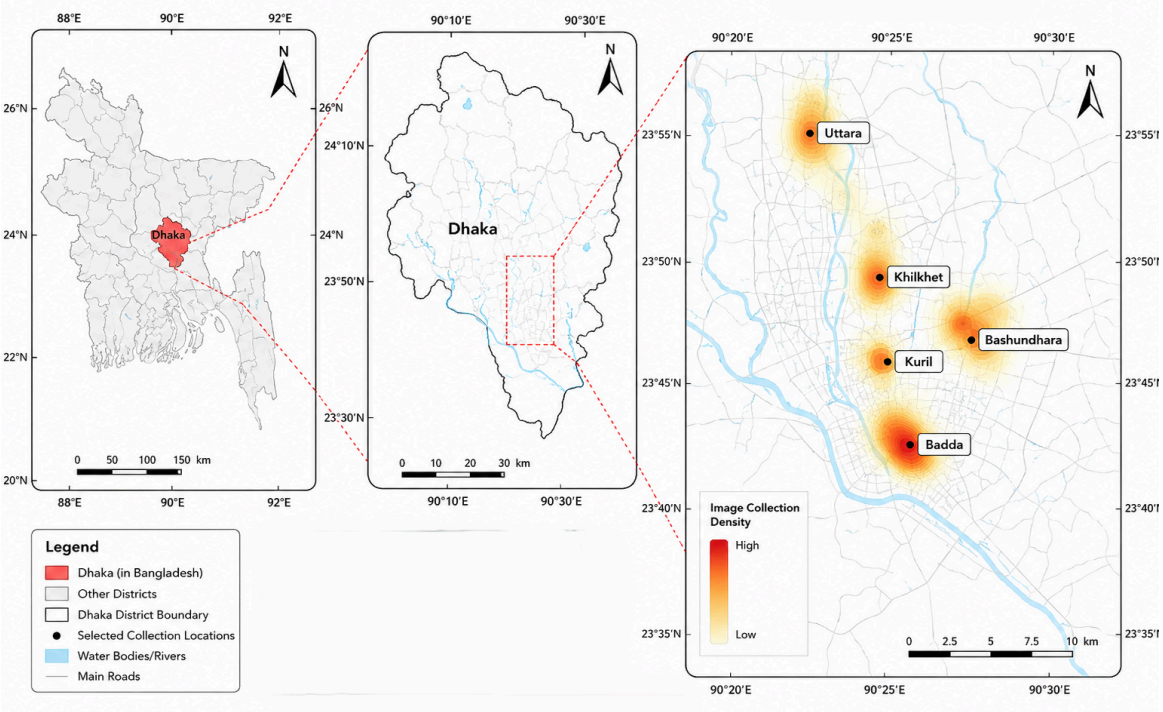


Figure 3.2: An AI Generated Heatmap that illustrates the Spatial Coverage of Image Collection Locations.

3.3 Dataset Collection

3.3.1 Image Acquisition Devices

Natural image quality and camera variation were introduced by imaging several of the most popular commercially available smartphones. And here are the most common devices used for data collection: Apple iPhone 16 Pro Max, Realme 8 and Xiaomi Redmi Note 14 Pro. The multi-device normalizes the imaging conditions in an actual deployment where images are taken from different camera sensors/imaging pipelines.

3.3.2 Image Capture Settings

Specifically, images were collected at near, medium, and far distances to diversify the dataset while keeping constant relative distance between camera angles and object orientation. This method actually collects large variations in object size, scale, viewing angle and perspective, which makes the task of object detection that is much more difficult while making the dataset resemble a real-life monitoring scenario even more closely.

3.3.3 Weather and Lighting Conditions

Most images were captured during daylight, primarily around midday and early afternoon, under generally favorable weather conditions with sufficient natural illumination. However, variations in sunlight intensity, shadows, reflections, and partial cloud cover were retained to reflect realistic outdoor environments. Such lighting variability is important for evaluating the robustness of vision-based infrastructure detection systems, as previous research has shown that reduced illumination can affect pothole and road-condition detection performance [33].

3.3.4 Real-World Challenges

The proposed VisPol dataset was intentionally collected under unconstrained real-world conditions to improve its deployment suitability. The dataset includes numerous challenging scenarios such as crowded streets, moving vehicles, pedestrians, background clutter, overlapping objects, partial occlusions, varying object scales, perspective distortion, damaged infrastructure, and diverse urban layouts. These complexities significantly increase the difficulty of visual pollution detection compared to existing datasets collected under relatively controlled environments, thereby providing a more realistic benchmark for evaluating object detection models.

3.4 Dataset Annotation

3.4.1 Roboflow Annotation Workflow

After image collection, the images were uploaded into the Roboflow platform for the dataset organization and manual annotation. Roboflow was used to create and manage the annotation classes, organize the collected images, prepare dataset versions, and export the annotations in a YOLO-compatible format. The platform also supported the annotation process by providing tools for reviewing annotations and maintaining different dataset versions.

3.4.2 Annotation Process

Images in the dataset were manually annotated by the use of axis-aligned boxes. Initially, seven categories were determined: Billboard, Poster/Banner, Waste, Wire, Damaged Wall, Damaged Road, and Signboard. Each instance of visible objects belonging to these categories was enclosed with a bounding box with minimal inclusion of the background and tight localization. All the pictures with several instances of object categories were annotated to ensure an accurate count of objects and their positions.

3.4.3 Annotation Protocol

The annotation protocol is established prior to the annotation process to facilitate the process of labelling. Among other things, the protocol includes standard definitions of classes, guidelines of placing bounding boxes, how to treat cases with partially visible objects, what to do with overlapping objects, how to assess the visibility of an object, and more.

3.4.4 Quality Assurance

To ensure the quality of annotation, a two-step process of quality assurance was implemented. The images received two rounds of independent checking after passing the initial stage of the annotation process. During this phase of quality verification, the captions of the bounding box, the labels of classes, the boundaries of objects, and the missing annotations were thoroughly checked. Any inconsistencies noticed during the quality checking process were fixed by performing manual verification before preparing the final version of the data set. The quality assurance process greatly improved the accuracy of the annotations and provided a template of a quality benchmark data set for further training and usage of the model.

3.4.5 Roboflow Annotation Interface

The Roboflow annotation interface used for data preparation is illustrated in Figure 3.3. The figure shows the workflow of the manual annotation of the bounding boxes used in this research, which includes choosing the class, localizing the object, and the features of the dataset management. First, it allowed efficient annotation methods; second, it provided the possibilities for quality checking and controlling the version of the dataset.

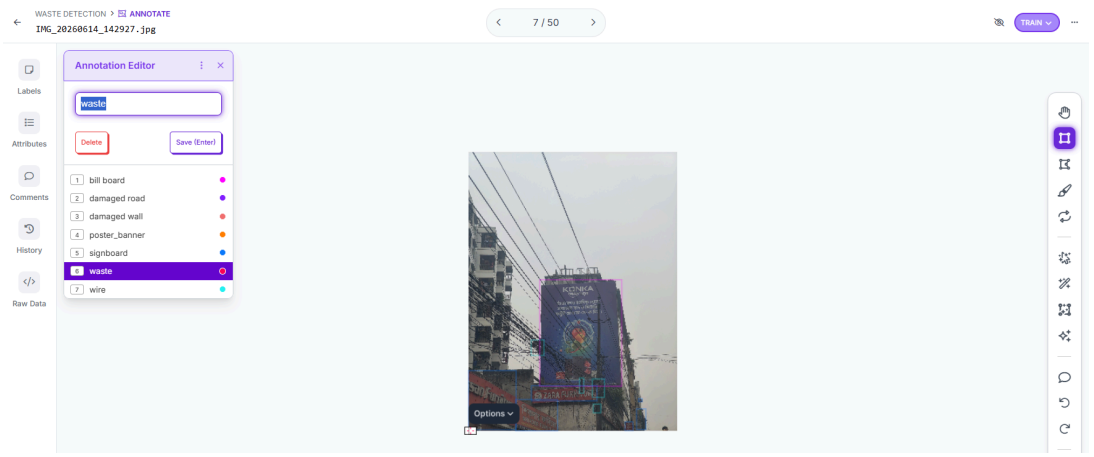


Figure 3.3: Roboflow Annotation Interface Used for Manual Annotation and Dataset Management.

3.5 Dataset Statistics

In this section, all essential statistical information regarding the VisPol dataset is provided. The section outlines the relevant statistics, such as the overall composition of the dataset, object instance statistics, class distribution, and class imbalance. All collected statistics are valuable information regarding the differences and complexity of the dataset, and they form the basis for further model development.

3.5.1 Image Statistics

Figure 3.4 depicts the composition of the dataset. The dataset consists of 1,097 annotated images and includes at least one object related to visual pollution. The dataset comprises 769 images for training purposes, 219 images for validation, and 109 images for the testing purposes of the model. This classification contributes to maintaining the independence between the testing and validation samples while ensuring the required number of training images is maintained. Overall, the dataset includes 15,933 object annotations that signify the large size of the database of manually annotated images collected in Bangladesh.

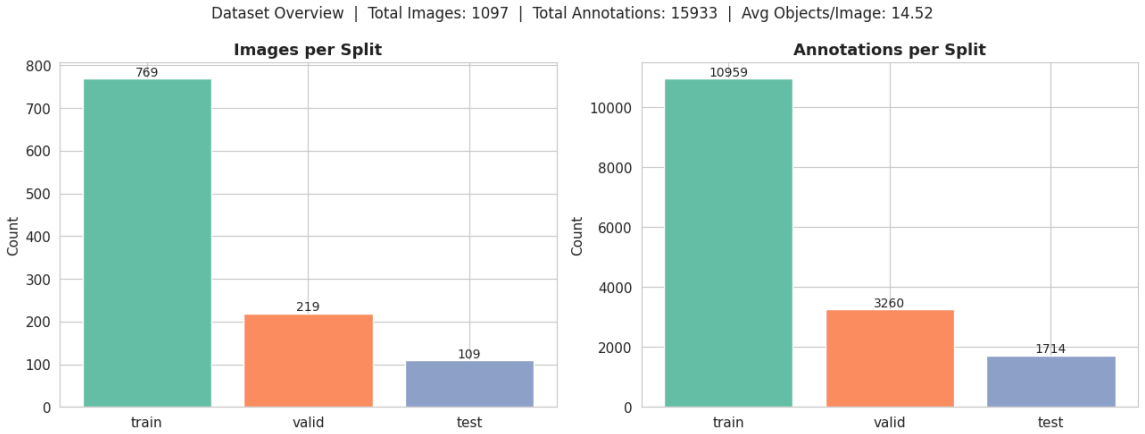


Figure 3.4: Overview of the VisPol Dataset Statistics.

3.5.2 Object Instance Statistics

Table 3.1: Class-Wise Object Instance Statistics of the VisPol Dataset.

Class	Labels	Images Containing Class	Average Objects per Image
Billboard	150	92	1.63
Damaged Road	375	205	1.83
Damaged Wall	283	149	1.90
Poster/Banner	5,081	907	5.60
Waste	8,428	865	9.74
Wire	1,616	464	3.48

The amount of images with respect to every class is also quite different, showing high variation in occurrence of these objects throughout the dataset. Moreover, the mean number of items in the images shows that Waste and Poster/Banner often appear in high densities, while Billboard, Damaged Wall, and Damaged Road appear less frequently in one image. These features make the VisPol dataset truly representative of the real urban environment with multiple pollution objects appearing simultaneously.

3.5.3 Class Distribution

The VisPol dataset was initially labeled with the help of manual annotations of seven object categories, including:

- Billboard
- Signboard
- Waste
- Damaged Wall
- Damaged Road
- Poster/Banner
- Wire

The categories which are shown in Figure 3.5 were taken by taking feedback from a preliminary field survey done in different urban areas of Dhaka that checked for the visual pollution forms and the poor urban infrastructure.

However, during the experimentation process, the Signboard category was dropped from the dataset. This is because unlike the other categories, signboards are mainly utilized for other purposes like traffic direction and commercial identification, thus not conforming to the definition of visual pollution adopted in this study.

Moreover, due to their resemblance to billboards and poster/banner objects, the signboards create significant confusion on the distinction between different classes.

As a result, the dataset called VisPol has ended up with six classes indicating visual pollution and therefore providing clearer evaluation datasets.

(a) The pink color box represents billboard, the blue color boxes represent signboards and the aqua color boxes represent tangled wire.

(b) The red color box represents waste and the orange color boxes represent posters/banners.

(c) The red color boxes represent waste, the orange color boxes represent posters/banners and the purple color boxes represent broken roads.

(d) The orange color boxes represent poster/banner, and the coral-colored boxes represent damaged walls.

Figure 3.5: Representative Object Annotations in the VisPol dataset.

The examples from Figure 3.5 are focused on the initial seven-class label assignment stage, including the Signboard category that was later disregarded. This depiction gives insight into the various visual pollution categories that the dataset covers. There are frequently multiple pollution items in one frame, resulting in complex environments with numerous objects that overlap, cover each other, or vary in scale.

3.5.4 Class Imbalance Analysis

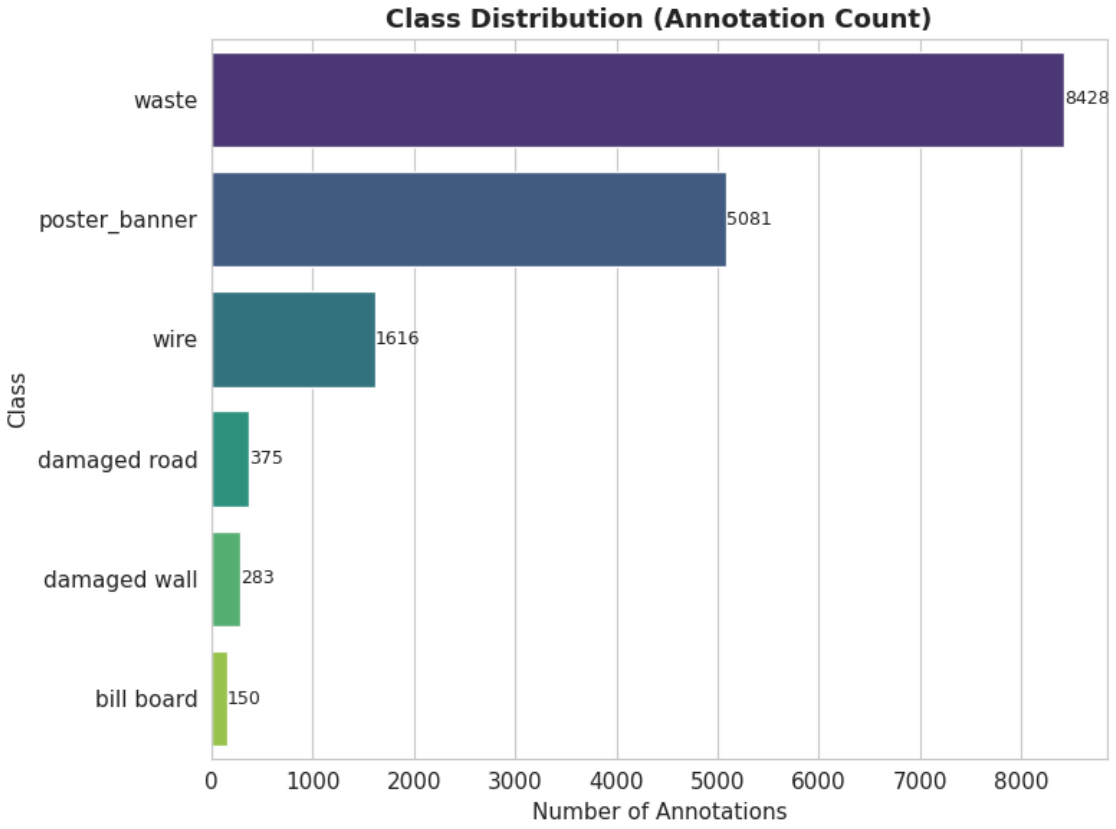


Figure 3.6: Class Frequency Distribution of the VisPol Dataset.

In the distribution of object annotations among the various experimental classes as shown in Figure 3.6, there is a clear indication of class imbalance throughout the survey. The Waste and Poster/Banner classes dominate the object count, while the number of instances of Baby Billboard, Damaged Wall and Damaged Road is much lower. Thus, the imbalance reflects real conditions rather than artificially being created. The preservation of this actual distribution allows for the dataset to better mirror deployed conditions in real life situations along with complicating the object detection task when it comes to minority classes.

3.6 Exploratory Data Analysis

In order to obtain a clearer picture of the statistical properties of the relevant dataset, exploratory data analysis (EDA) was performed. The analysis includes the following: statistics for the image level, an analysis of the object instance distributions, bounding box parameters, object sizes, aspect ratios, and frequency of classes.

3.6.1 Image Distribution

Table 3.2: Split Statistics of VisPol Dataset.

Metric	Value
Total Images	1,097
Total Label Files	1,097
Total Valid Annotations	15,933
Training Images	769
Validation Images	219
Test Images	109
Training Annotations	10,959
Validation Annotations	3,260
Test Annotations	1,714

The information in Table 3.2 illustrates the allocation of visual pollution photos in the dataset which was split into three sets that are the training, validation, and test partitions with numbers of visual pollution images in these sets being approximately 70%, 20%, and 10% respectively. The given results confirm the use of the 70/20/10 strategy for the data split. This way of partitioning leads to sufficient numbers of samples for training the model while at the same time providing independent subsets for tuning and evaluation of the performance of the model thus eliminating the chance for the evaluation bias.

3.6.2 Bounding Box Analysis

The data presented in Figure. 3.7 is concerned with distributions of bounding boxes dimensions. The analysis proves that there is greater variability of the object dimensions which gives rise to completely different physical sizes and different capturing distances of the objects of visual pollution that are part of the dataset. The data on bounding boxes show that images are equally distributed in the photos thus eliminating position bias during the training of the model. To summarize the given findings it can be said that the dataset has great variability of bounding boxes which proves its suitability for assessing object detection models.

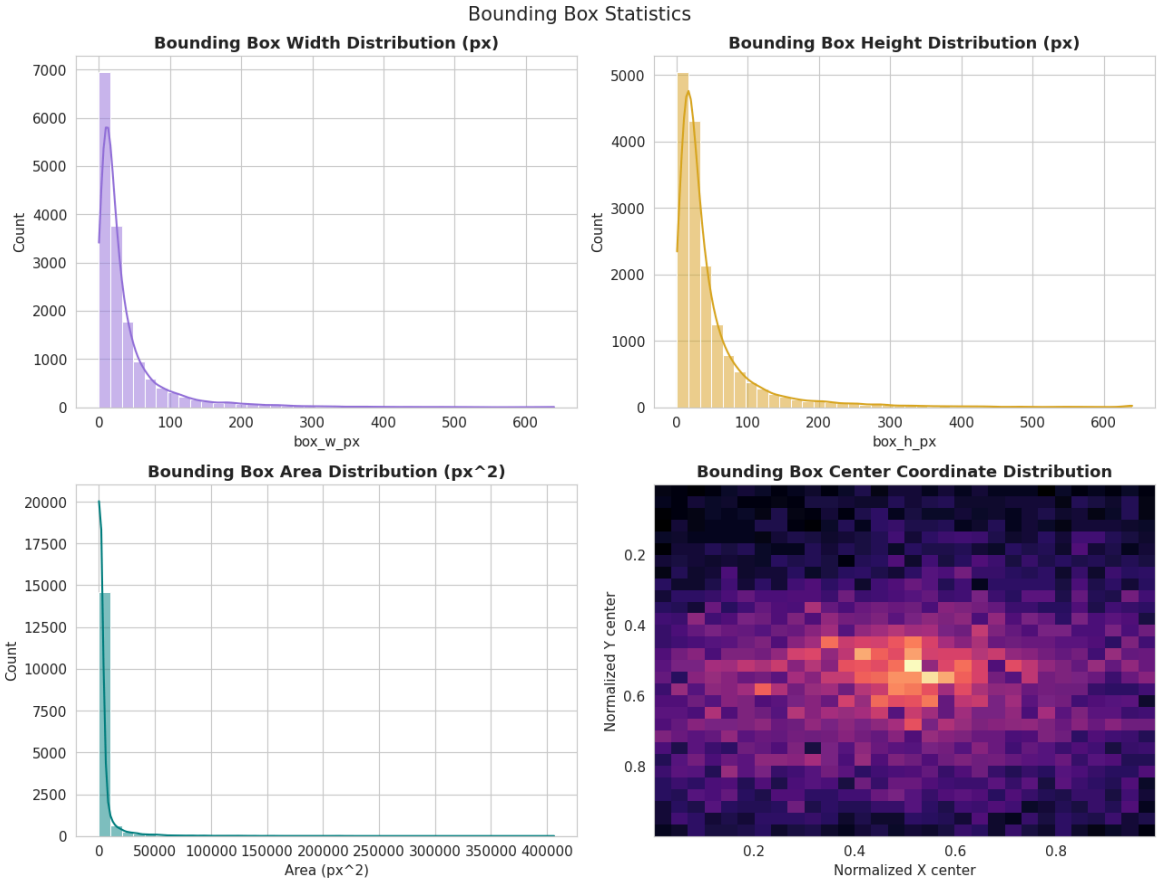


Figure 3.7: Statistical Analysis of Bounding Box Annotations.

3.6.3 Object Size Distribution

The VisPol dataset contains objects across a wide range of scales. Small objects, such as distant posters and utility wires, occupy only a few pixels, whereas large billboards and extensive road segments can cover substantial portions of an image. This scale variation increases the difficulty of object detection and provides a useful basis for evaluating detector performance under diverse object-size conditions.

3.6.4 Aspect Ratio Analysis

The annotated objects also exhibit substantial variation in shape and aspect ratio. Billboards and posters generally appear as wide rectangular regions, while damaged roads and walls may have irregular boundaries. Utility wires are typically long and extremely narrow. This diversity in object geometry requires detectors to effectively handle both elongated and compact structures with regular and irregular shapes.

3.7 External Validation Datasets

In order to validate the performance of the VisPol-trained models, two additional datasets were employed for verification. Both datasets are linked to the concept of visual pollution while being different from the data in terms of geography, annotations, and complexity.

3.7.1 External Dataset 1: Visual Pollution of Dhaka Streets

External Dataset 1, called "Visual Pollution Dhaka Streets" [50], is an open-object detection dataset created and made available via Kaggle and has not been a part of the VisPol project. The data has been collected only from the VisPol city, but it has not been carried out for the purpose of the VisPol project as it follows its own guidelines in data capture, use of camera equipment, and data annotation. Thus, the data collected has succeeded in creating a truly independent benchmark for the purpose of inter-dataset comparison. The dataset consists of three categories of visual pollution, which are billboards, street litter, and wire. The rest of the classes in VisPol (that is, posters/banners, damaged walls, or damaged roads) do not exist in the external data, and therefore, the data could not be used for inter-dataset comparison in this benchmark process. The statistics of external dataset 1 are shown in table 3.3.

Table 3.3: Dataset Statistics and Class Mapping of External Dataset 1.

Metric	Value
Total images	722
Total annotations	961
Number of classes	3 (Billboards, Street Litter, Wire)
Mapped VisPol Classes	Billboard, Waste, Wire
Annotation format	YOLO bounding box

Compared to VisPol (961 vs 15,933 annotations), the dataset has fewer annotations, with around 1.3 objects per image, much less than the 14.52 objects per image average in VisPol. From this, we can conclude that External Dataset 1 mainly contains sparse single- or sparse few-object scenes and is not composed of the densely populated, multi-object street scenes like the ones in VisPol.

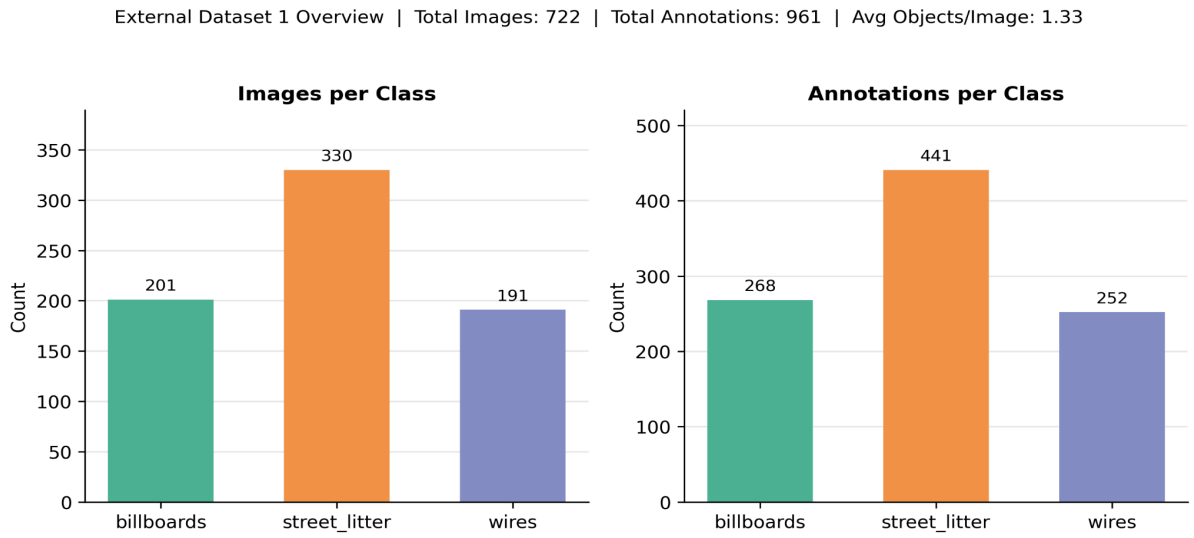


Figure 3.8: Object Annotation Distribution Across Training, Validation, and Test Sets of External Dataset 1.

Qualitative Comparison

There are various structural distinctions between External Dataset 1 and VisPol that are as follows:

Cleaner backgrounds: The image captures more the target object (billboard, wire pile, litter pile, etc.) in a more isolated way, with fewer surrounding structures in comparison to the cluttered images taken by VisPol.

Closer camera distances: The images depict the object at a relatively closer distance than in VisPol, where a combination of images from various ranges (near, medium, distant) is used.

Fewer occlusions: There are fewer instances of partial occlusion in External Dataset 1 images because they have fewer other objects in the picture, and hence fewer overlapping boxes than in VisPol.

Annotative inconsistencies: The differences in the annotation approach employed (for example, how tightly objects are bounded by boxes, how the litter is grouped versus placed in separate boxes) create a misalignment in the meaning of the labels that do not completely agree with VisPol.

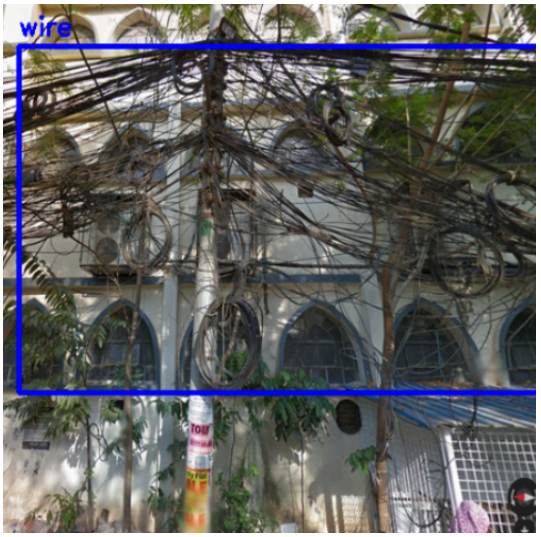
Different object scales: The distribution of images in External Dataset 1 differs from VisPol in size, with fewer representations of the very small, distanced objects presented in the latter. These variations together make a shift in the inbound training versus the evaluation or test distribution. The reason is that the model was trained using a collection of dense, complicated, multi-scale imagery that results in the learning of features pertaining specifically to that distribution. When being tested on images from Dataset 1's clean background images at the close distances with little annotations, the model is going

to perform poorly in the terms of overall precisions and recalls for the classes that were mapped due to the fact that it is going to differ from the representative in terms of stated above attributes.

(a) Billboard sample from External Dataset 1.

(b) Waste sample from External Dataset 1.

(c) Wire sample from External Dataset 1.



(d) Wire cluster sample from External Dataset 1.

Figure 3.9: Representative Annotated Samples From the External Dataset 1.

3.7.2 External Dataset 2: PlasticPol Dataset

External Dataset 2, known as the "PlasticPol Dataset" [51], is a freely available binomial object detection data set that specializes solely in the detection of waste/plastic pollution. Unlike External Dataset 1, this dataset does not contain any classes of visual pollution related to infrastructure (such as billboards, wires, and broken roads/walls), and the only annotated class is linked to VisPol Waste. Key statistics of External Dataset 2 are outlined in Table 3.4.

Table 3.4: Dataset Statistics and Class Mapping of External Dataset 2.

Metric	Value
Total Images	2418
Number of classes	1 (waste)
Mapped VisPol class	waste
Annotation format	YOLO bounding box

Since there is only one annotated class in External Dataset 2, the class imbalance analysis cannot be carried out in the same way as it was done for VisPol and External Dataset 1. Therefore, the most pertinent comparison should focus on the density and framing of waste instances with respect to the Waste class found in VisPol which has a total of 8,428 annotations for 865 images (9.74 objects/image).

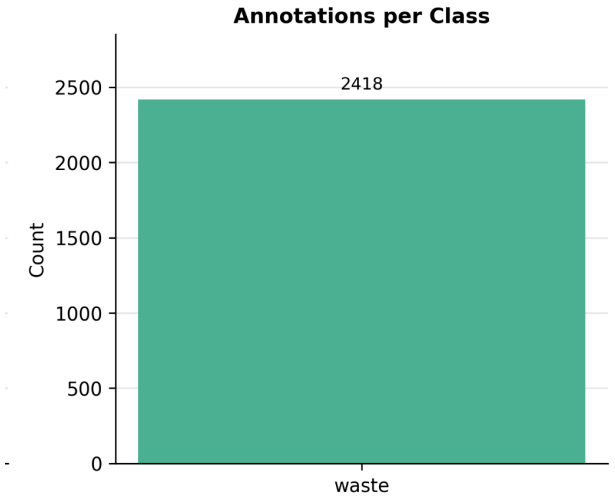


Figure 3.10: Waste Annotation Distribution in External Dataset 2.

In addition to this, as depicted in Figure 3.11, the PlasticPol dataset consists of some representative examples of images that are targeted towards waste and have little urban context around the subjects in close range.

(a) The green bounding box represents waste from External Dataset 2.

(b) The green bounding box represents waste from External Dataset 2.

Figure 3.11: Representative Annotated Samples from the PlasticPol dataset.

These samples reflect the object-based nature of the PlasticPol dataset as opposed to VisPol, which has a lot of clutter from other elements.

Qualitative comparison:

Waste-centric scenarios: Almost every piece of image emphasizes waste, in comparison to VisPol's incidental capturing of waste in street photos. Close-range photography: The objects in the images are mostly taken from a very close distance, which creates larger and more centralized boxes compared to VisPol's random collection of images from different distances.

Fewer infrastructure objects: The lack of billboards/wire/advertisements/posters as well as abandoned road/wall pictures means that the model cannot be evaluated using other non-waste detectors in this dataset. Limited variety of environment: The data looks like it is gathered from a narrow range of contexts (less different urban places, light conditions, places etc.) compared to VisPol's collections from various sites in Dhaka. Differences in the rules of annotation: The methods of defining what's considered a single "waste" object (e.g., not differentiating between the group of waste and items) may vary from VisPol's guidelines. All these factors make the assessment based on this dataset measure the waste detector's ability to operate in another visual environment rather than test other types of urban pollution detection. In cases where the photographs are less cluttered around the natural street scenes, performance differences observed here can be attributed largely to shifts in scene composition and camera framing rather than to occlusion or multi-object complexity. These findings complement the domain-shift insights drawn from External Dataset 1.

4 Experimental Setup

The chapter provides a thorough account of the experimental configuration used to test the proposed VisPol dataset. The details of the computing environment, model architectures, training configuration, optimization, loss functions, and evaluation metrics used in the present study are given. The same experimental procedure is adopted throughout the experiment to ensure that a fair comparison between all models was conducted in this work [52].

4.1 Experimental Environment

The whole experiment was conducted using the cloud platform Kaggle Notebook, which offers high-performance GPU resources for deep learning purposes.

The implementations were executed in Python 3.11 using the Ultralytics YOLOv8 framework based on PyTorch.

Table 4.1: Experimental Environment and Configuration.

Component	Specification
Platform	Kaggle Notebook
Operating System	Linux Ubuntu
GPU	Kaggle GPU
RAM	Approximately 30 GB
Python	Python 3.11
Deep Learning Framework	PyTorch
Detection Framework	Ultralytics YOLOv8
Annotation Tool	Roboflow
Libraries	OpenCV, NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn, Albumentations

4.2 Detection Models

Five variants of YOLOv8 were evaluated: **YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l,** and **YOLOv8x.**

Table 4.2: Model Training Configuration.

Parameter	Value
Image Size	640 × 640 pixels
Batch Size	16
Epochs	100
Optimizer	AdamW
Initial Learning Rate	0.001
Learning Rate Scheduler	Cosine Decay
Weight Decay	0.0005
Momentum	0.937
Confidence Threshold	0.25
IoU Threshold	0.70
Early Stopping	Disabled
Mixed Precision	Enabled
Random Seed	42

4.2.1 YOLOv8n (Nano)

- **Parameters:** 3.2 million
- **FLOPs:** 8.7 GFLOPs
- **Strengths:** Very fast inference, low memory, suitable for edge devices
- **Limitations:** Lower accuracy for small objects

4.2.2 YOLOv8s (Small)

- **Parameters:** 11.2 million
- **FLOPs:** 28.6 GFLOPs
- Strengths: Good balance, better small-object detection
- Limitations: Limited for extremely complex scenes

4.2.3 YOLOv8m (Medium)

- **Parameters:** 25.9 million

- **FLOPs:** 79.3 GFLOPs
- Strengths: Better localization, improved multi-scale detection
- Limitations: Increased GPU memory, slower inference

4.2.4 YOLOv8l (Large)

- **Parameters:** 43.7 million
- **FLOPs:** 165 GFLOPs
- Strengths: Excellent feature extraction, strong performance on complex scenes
- Limitations: High computational cost

4.2.5 YOLOv8x (Extra Large)

- Parameters: 68.2 million
- FLOPs: 258 GFLOPs
- Strengths: Highest detection capability, superior localization
- Limitations: Slowest inference, highest computational requirements

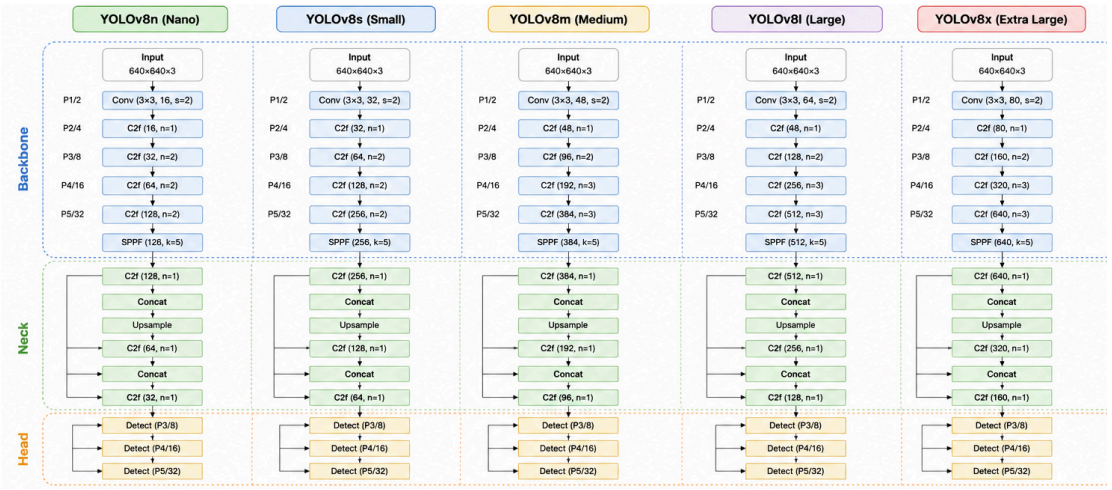


Figure 4.1: AI-Generated Illustration of the Architecture Comparison of YOLOv8 Model Variants (YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x), Showing the Backbone, Neck, and Detection Head with Progressive Scaling of Network Depth and Channel Width.

The five versions of the YOLOv8: YOLOv8n (Nano), YOLOv8s (Small), YOLOv8m (Medium), YOLOv8l (Large), and YOLOv8x (Extra Large) are compared under the architecture shown in Figure 4.1. For all models, the architecture is divided into three phases comprising the Backbone, Neck and Detection Head, although the models are different in depth, width and computational requirements. Backbone produces a number of features used by the model by way of convolutional layers and other modules. The Neck has a feature pyramid structure that allows fusing of low-level spatial and

high-level semantic features. Finally, the Detection Head shall accomplish object detection via three stages, thus allowing to locate all the objects of various sizes. As one moves from YOLOv8n to YOLOv8x, the number of channels and C2f blocks increases. So, YOLOv8n is the lightest and quickest model as per the fitting of the task with limitations of the resource and of the real time applications while YOLOv8x is the most efficient model to achieve the required level of detection to those tasks that rely heavily on the accuracy. Other variations (YOLOv8s, YOLOv8m, YOLOv8l) will have the intermediate levels in the context of detection efficiency and in terms of the work done by the model, thus enabling the specialists to select the model according to the quantity of available resources and technical requirements of the task.

4.3 Training Configuration

In order to ensure fairness in the comparisons made among the variants of YOLOv8, the process of training that was followed throughout the experiments was the same.

4.3.1 Data Augmentation

The augmentations applied in this study are mosaic, horizontal flipping, scaling, translation, HSV colour augmentation, and random perspective.

4.4 Loss Functions

The system used in YOLOv8 makes use of Binary Cross-Entropy (BCE) Loss, Distribution Focal Loss (DFL), and Complete Intersection over Union (CIoU) Loss.

4.4.1 Binary Cross-Entropy (BCE) Loss

The Binary Cross-Entropy Loss helps in the computation of the difference between the probability of the class that was predicted and the true one.

$$LBCE = -[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})] \quad (1)$$

Where y refers to the true class (0 or 1) and $\hat{y}$ signifies the predicted class probability.

4.4.2 Classification Loss

The total classification loss is derived from the total of all the losses in BCE through YOLOv8's anchor-free feature of detecting:

$$Lcls = (1 / N) \sum LBCE(y, \hat{y}) \quad (2)$$

where N is the number of the positive anchors in the batch.

4.4.3 Distribution Focal Loss (DFL)

This means that the Distribution Focal Loss is focusing on making the bounding box's localization all the more exact through string treatment of the prediction as a probability function:

$$LDFL = -\sum (1 - p)\gamma \log(p) \text{.....(3)}$$

where p is the probability predicted for certain classes, γ is a focusing feature (commonly taking a value of 2.0).

4.4.4 Complete IoU (CIoU) Loss

The Complete Intersection over Union Loss that's applied to the bounding box regression considers the factors of overlap, normalized distance, and aspect ratio:

$$LCIoU = 1 - IoU + (\rho 2(b, bgt) / c2) + (1 - IoU) \times \alpha \text{.....(4)}$$

where IoU stands for the kind of area defined by Intersection over Union, ρ represents the distance in Euclidean terms of the box centers, c stands for the diagonal of the enclosing rectangle; α is the difference between aspect ratios.

4.4.5 Total Loss Function

The final loss is a combination of all components:

$$Ltotal = \lambda 1\ Lcls + \lambda 2\ LDFL + \lambda 3\ LCIoU \text{ .....(5)}$$

4.5 Evaluation Metrics

The performance of detection was evaluated with the help of standard metrics used in computer vision.

4.5.1 Precision

Precision refers to the ratio of correct positive detection:

$$Precision = TP / (TP + FP) \text{ .....(6)}$$

Here, TP are True Positives indicating correct detections and FP are False Positives indicating wrong detections. The precision can take values between 0 and 1.

4.5.2 Recall

Recall refers to the ratio of actual positives detected successfully:

$$\text{Recall} = TP / (TP + FN) \quad \text{.....(7)}$$

FN refers to false negatives indicating failed detections.

4.5.3 F1-Score

The F1-Score is the harmonic mean of Precision and Recall:

$$\text{F1-Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad \text{.....(8)}$$

Useful when there is uneven class distribution. Range: 0 to 1.

4.5.4 Intersection over Union (IoU)

IoU measures the overlap between predicted and ground truth bounding boxes:

$$\text{IoU} = \text{Area}(B_{pred} \cap B_{gt}) / \text{Area}(B_{pred} \cup B_{gt}) \quad \text{.....(9)}$$

where Bpred is the predicted bounding box and Bgt is the ground truth box. Range: 0 to 1.

4.5.5 Average Precision (AP)

Average Precision is the area under the Precision-Recall curve:

$$AP = \int_0^1 P(r)\, dr \quad \text{.....(10)}$$

where P(r) is precision as a function of recall r. Computed using 11-point interpolation or exact curve.

4.5.6 Mean Average Precision at IoU 0.50 (mAP@50)

Mean Average Precision at IoU 0.50 is the average of AP values at IoU threshold 0.50 across all classes:

$$\text{mAP@50} = (1 / N_{classes}) \sum AP_i(@IoU=0.50) \quad \text{.....(11)}$$

where classes is the total number of classes. Less strict than mAP@50-95.

4.5.7 Mean Average Precision at IoU 0.50-0.95 (mAP@50-95)

Mean Average Precision at IoU 0.50-0.95 averages AP values at multiple IoU thresholds (0.50 to 0.95, increments of 0.05):

$$\text{mAP@50-95} = (1 / 10) \sum \text{mAP}(@IoU) | IoU=0.50:0.05:0.95 \quad \text{.....(12)}$$

More strict and comprehensive. Standard metric in COCO and modern benchmarks.

4.5.8 Confusion Matrix

The Confusion Matrix shows the distribution of TP, FP, TN, and FN:

$$CM = [TN \;\; FP] \quad [FN \;\; TP] \text{.....(13)}$$

Essential for understanding model behavior on individual classes and identifying specific error sources.

4.5.9 Floating Point Operations (FLOPs)

FLOPs refer to the number of arithmetic calculations performed during the forward pass of the model (expressed in GFLOPs, which is 10^9 calculations):

$$FLOPs = \sum 2 \times h \times w \times cin \times cout \times k^2 \text{ .....(14)}$$

In this case, the h, w denote the feature map sizes, cin, cout represents the channels, and k represents the kernel size. This measure shows the computational complexity of the deep learning model.

4.5.10 Number of Parameters

The total number of all learnable weights and biases in all layers of the model:

$$\text{Total Parameters} = \sum (wi + bi) \text{ .....(15)}$$

Having large parameters' numbers means the model has more possibilities but requires larger memory and more time for inference.

4.5.11 Frames Per Second (FPS)

FPS measures the speed of inference (the number of images processed in one second):

$$FPS = N / T \text{ .....(16)}$$

Here, N is the number of images processed, T is the total time in seconds. This metric is of importance for apps that require real-time processing.

4.5.12 Inference Time

Inference time is defined as the time needed for processing one image (measured in milliseconds):

$$\text{Inference time (ms)} = (T / N) \times 100 \text{ .....(17)}$$

This parameter is important for deployments in environments with limited resources and involves the time for preprocessing, conducting the model, and post-processing.

5 Results and Discussion

The section covers a thorough analysis of the proposed data set and benchmarks all the experiments done to analyze all the models of YOLOv8. The analysis includes the results obtained by cross-validation, by testing the independent test set, and the analysis through the process of error identification qualitatively.

5.1 Initial Validation and K-fold results

A three-fold cross-validation procedure was performed, with the initial validation results reported separately from the three cross-validation folds.

5.1.1 YOLOv8n Cross-Validation

The table 5.1 shows the results of the performance of every class according to YOLOv8n and it can be seen that the model produced stable results on such high-volume classes as Poster/Banner and Waste while the results of low-volume classes are not that well since there are not many instances present in those classes.

5.1.2 YOLOv8s Cross-Validation

Table 5.2 contains the results of the analysis of classes for YOLOv8s. Thus, the size of the network is increased, leading to the results being a bit more stable and results more certain.

5.1.3 YOLOv8m Cross-Validation

The results of YOLOv8m are depicted in Table 5.3. Results indicate that the medium version achieved a significant improvement in precision in localized categories like Poster/Banner, Waste, and Wire across fold.

5.1.4 YOLOv8l Cross-Validation

YOLOv8l's evaluation results are summarized in Table 5.4. The model performed well in precision but the recall metrics are limited in the case of certain sparse structural defects (Damaged Wall), mainly due to limited representation, irregular and diffuse object boundaries and substantial visual variability.

Table 5.1: Class-Wise 3-Fold Cross-Validation Performance Metrics for YOLOv8n.

Class	Fold	Images	Precision	Recall	mAP@50	mAP@50–95
Billboard	Initial Validation	12	0.487	0.182	0.192	0.119
	Fold 1	38	0.358	0.278	0.25	0.147
	Fold 2	23	0.172	0.114	0.0614	0.0363
	Fold 3	23	0.357	0.256	0.229	0.133
Damaged Road	Initial Validation	33	0.134	0.107	0.033	0.015
	Fold 1	62	0.153	0.168	0.0538	0.0188
	Fold 2	60	0.123	0.118	0.0624	0.0239
	Fold 3	61	0.123	0.113	0.0268	0.0106
Damaged Wall	Initial Validation	32	0.0609	0.0196	0.0123	0.0026
	Fold 1	30	0	0	0.0001	0
	Fold 2	51	0	0	0.0017	0.0004
	Fold 3	53	0.0714	0.018	0.0024	0.0006
Poster/Banner	Initial Validation	173	0.394	0.315	0.261	0.138
	Fold 1	275	0.396	0.331	0.266	0.141
	Fold 2	273	0.388	0.316	0.244	0.124
	Fold 3	274	0.405	0.336	0.258	0.143
Waste	Initial Validation	175	0.391	0.341	0.269	0.125
	Fold 1	258	0.371	0.331	0.244	0.114
	Fold 2	266	0.365	0.414	0.323	0.16
	Fold 3	249	0.321	0.385	0.237	0.113
Wire	Initial Validation	88	0.308	0.221	0.157	0.0728
	Fold 1	157	0.353	0.151	0.121	0.0539
	Fold 2	133	0.395	0.129	0.117	0.0558
	Fold 3	130	0.325	0.155	0.132	0.0627

Table 5.2: Class-Wise 3-Fold Cross-Validation Performance Metrics for YOLOv8s.

Class	Fold	Images	Precision	Recall	mAP@50	mAP@50–95
Billboard	Initial Validation	12	0.425	0.273	0.253	0.164
	Fold 1	38	0.377	0.303	0.248	0.134
	Fold 2	23	0.294	0.2	0.149	0.0973
	Fold 3	23	0.389	0.233	0.238	0.158
Damaged Road	Initial Validation	33	0.146	0.2	0.0998	0.0395
	Fold 1	62	0.157	0.112	0.094	0.0383
	Fold 2	60	0.171	0.155	0.0717	0.0295
	Fold 3	61	0.222	0.113	0.0467	0.0171
Damaged Wall	Initial Validation	32	0.0856	0.0784	0.0225	0.0076
	Fold 1	30	0.0984	0.0182	0.0078	0.0009
	Fold 2	51	0.108	0.0233	0.0193	0.0085
	Fold 3	53	0	0	0.0008	0.0002
Poster/Banner	Initial Validation	173	0.483	0.345	0.287	0.175
	Fold 1	275	0.428	0.374	0.306	0.168
	Fold 2	273	0.396	0.345	0.267	0.133
	Fold 3	274	0.485	0.332	0.291	0.16
Waste	Initial Validation	175	0.422	0.347	0.284	0.138
	Fold 1	258	0.369	0.398	0.288	0.134
	Fold 2	266	0.385	0.432	0.334	0.167
	Fold 3	249	0.365	0.364	0.25	0.114
Wire	Initial Validation	88	0.483	0.221	0.187	0.0899
	Fold 1	157	0.441	0.168	0.144	0.0694
	Fold 2	133	0.406	0.195	0.171	0.0857
	Fold 3	130	0.463	0.198	0.174	0.0844

Table 5.3: Class-Wise 3-Fold Cross-Validation Performance Metrics for YOLOv8m.

Class	Fold	Images	Precision	Recall	mAP@50	mAP@50–95
Billboard	Initial Validation	12	0.153	0.273	0.222	0.163
	Fold 1	38	0.481	0.333	0.312	0.188
	Fold 2	23	0.307	0.286	0.203	0.135
	Fold 3	23	0.527	0.256	0.292	0.153
Damaged Road	Initial Validation	33	0.149	0.12	0.0662	0.0235
	Fold 1	62	0.243	0.159	0.133	0.0622
	Fold 2	60	0.401	0.0909	0.0973	0.0492
	Fold 3	61	0.128	0.139	0.0657	0.0345
Damaged Wall	Initial Validation	32	0.46	0.018	0.0098	0.0051
	Fold 1	30	0	0	0.0002	0
	Fold 2	51	0.0738	0.0116	0.018	0.004
	Fold 3	53	0.0311	0.009	0.0057	0.0027
Poster/Banner	Initial Validation	173	0.42	0.403	0.349	0.203
	Fold 1	275	0.483	0.389	0.341	0.197
	Fold 2	273	0.52	0.348	0.315	0.174
	Fold 3	274	0.436	0.39	0.317	0.176
Waste	Initial Validation	175	0.459	0.351	0.317	0.153
	Fold 1	258	0.492	0.302	0.282	0.133
	Fold 2	266	0.538	0.328	0.324	0.16
	Fold 3	249	0.398	0.37	0.278	0.126
Wire	Initial Validation	88	0.468	0.238	0.224	0.114
	Fold 1	157	0.382	0.247	0.21	0.0942
	Fold 2	133	0.403	0.184	0.156	0.0816
	Fold 3	130	0.35	0.305	0.239	0.117

Table 5.4: Class-Wise 3-Fold Cross-Validation Performance Metrics for YOLOv8l.

Class	Fold	Images	Precision	Recall	mAP@50	mAP@50–95
Billboard	Initial Validation	12	0.531	0.318	0.227	0.15
	Fold 1	38	0.439	0.241	0.253	0.174
	Fold 2	23	0.0931	0.29	0.181	0.12
	Fold 3	23	0.251	0.163	0.125	0.0872
Damaged Road	Initial Validation	33	0.391	0.133	0.11	0.0449
	Fold 1	62	0.157	0.215	0.109	0.0548
	Fold 2	60	0.174	0.0909	0.0413	0.0252
	Fold 3	61	0.171	0.0783	0.038	0.0197
Damaged Wall	Initial Validation	32	0	0	0.0126	0.0045
	Fold 1	30	0.0282	0.0364	0.0016	0.0008
	Fold 2	51	0	0	0.0012	0.0006
	Fold 3	53	0.579	0.009	0.005	0.001
Poster/Banner	Initial Validation	173	0.45	0.394	0.338	0.201
	Fold 1	275	0.485	0.388	0.325	0.188
	Fold 2	273	0.469	0.324	0.263	0.143
	Fold 3	274	0.435	0.363	0.289	0.168
Waste	Initial Validation	175	0.478	0.32	0.31	0.155
	Fold 1	258	0.494	0.282	0.273	0.13
	Fold 2	266	0.445	0.325	0.268	0.133
	Fold 3	249	0.36	0.327	0.203	0.096
Wire	Initial Validation	88	0.341	0.232	0.204	0.0994
	Fold 1	157	0.454	0.247	0.206	0.0968
	Fold 2	133	0.437	0.146	0.126	0.0666
	Fold 3	130	0.468	0.174	0.166	0.0871

5.1.5 YOLOv8x Cross-Validation

In Table 5.5, the performance metrics for YOLOv8x, the largest member of its family, are reported. As it has the highest number of parameters, it achieved the best mAP@50-95 for the major classes, which encountered a problem with the few classes.

Table 5.5: Class-Wise 3-Fold Cross-Validation Performance Metrics for YOLOv8x.

Class	Fold	Images	Precision	Recall	mAP@50	mAP@50–95
Billboard	Initial Validation	23	0.249	0.257	0.263	0.187
	Fold 1	38	0.385	0.321	0.282	0.214
	Fold 2	23	0.312	0.274	0.235	0.178
	Fold 3	23	0.404	0.311	0.278	0.205
Damaged Road	Initial Validation	60	0.104	0.0762	0.0367	0.0181
	Fold 1	62	0.168	0.089	0.046	0.0275
	Fold 2	60	0.132	0.065	0.0345	0.0198
	Fold 3	61	0.171	0.081	0.045	0.0274
Damaged Wall	Initial Validation	51	0	0	0.0016	0.001
	Fold 1	30	0	0	0	0
	Fold 2	51	0	0	0	0
	Fold 3	53	0	0	0	0
Poster/Banner	Initial Validation	273	0.484	0.386	0.309	0.18
	Fold 1	275	0.518	0.412	0.345	0.212
	Fold 2	273	0.472	0.375	0.311	0.184
	Fold 3	274	0.513	0.407	0.34	0.204
Waste	Initial Validation	266	0.51	0.335	0.276	0.146
	Fold 1	258	0.455	0.362	0.258	0.134
	Fold 2	266	0.418	0.321	0.229	0.112
	Fold 3	249	0.453	0.355	0.254	0.132
Wire	Initial Validation	133	0.439	0.231	0.16	0.0862
	Fold 1	157	0.584	0.235	0.208	0.124
	Fold 2	133	0.531	0.201	0.175	0.098
	Fold 3	130	0.586	0.227	0.199	0.123

5.1.6 Overall Cross-Validation Comparison

Table 5.6 presents the overall cross-validation figures for all the models. YOLOv8m achieved the highest overall mAP@50 (0.1992), where the trade-off between capabilities and stability of the models can be seen. YOLOv8x achieved the highest mAP@50-95 (0.1110), which indicates the accuracy of the bounding boxes.

Table 5.6: Summary of 3-Fold Cross-Validation Performance Across YOLOv8 Variants.

Model	mAP@50	mAP@50-95	Precision	Recall
YOLOv8n	0.145955	0.074244	0.260107	0.199896
YOLOv8s	0.172245	0.088626	0.316827	0.217154
YOLOv8m	0.199210	0.104883	0.344350	0.230880
YOLOv8l	0.159983	0.088518	0.329000	0.205752
YOLOv8x	0.18	0.111000	0.339000	0.224000

5.2 Performance Evaluation on the Independent Test Set

The effectiveness of the models was assessed with the use of the withheld VisPol testing dataset including 109 images and 1714 instances. The performance across classes, overall detection metrics, and efficiency are detailed in tables 5.7 and 5.8.

5.2.1 Class-Wise Test Performance

The outcome of the evaluation of YOLOv8 models is depicted in table 5.7. In contrast to the outcome of cross-validation tests, which are indicative of model performance on the already known images, these results show the model's performance on images unseen before.

The results corroborate the previous finding which says that the most detectable classes are Waste and Poster/Banner classes, while the classes with fewer representation can be detected only with difficulty owing to the lack of sufficient training.

Table 5.7: Class-Wise Object Detection Results of YOLOv8 Model Variants on the VisPol Test Set.

Model	Class	Images	Instances	Precision	Recall	mAP@50	mAP@50–95
YOLOv8n	Billboard	8	18	0.606	0.428	0.405	0.203
	Damaged Road	22	43	0.195	0.14	0.0639	0.0252
	Damaged Wall	15	31	0.0886	0.0323	0.036	0.0247
	Poster/Banner	85	485	0.387	0.305	0.274	0.156
	Waste	92	971	0.407	0.365	0.281	0.133
	Wire	44	165	0.321	0.255	0.167	0.078
YOLOv8s	Billboard	8	18	0.557	0.556	0.476	0.203
	Damaged Road	22	43	0.169	0.186	0.102	0.046
	Damaged Wall	15	31	0	0	0.0049	0.0024
	Poster/Banner	85	485	0.506	0.377	0.35	0.2
	Waste	92	971	0.456	0.347	0.293	0.14
	Wire	44	165	0.511	0.228	0.244	0.132
YOLOv8m	Billboard	8	18	0.236	0.444	0.332	0.23
	Damaged Road	22	43	0.399	0.163	0.14	0.0614
	Damaged Wall	15	31	0.325	0.0323	0.034	0.0256
	Poster/Banner	85	485	0.496	0.412	0.374	0.212
	Waste	92	971	0.536	0.329	0.334	0.157
	Wire	44	165	0.494	0.17	0.176	0.0903
YOLOv8l	Billboard	8	18	0.549	0.389	0.388	0.242
	Damaged Road	22	43	0.409	0.129	0.107	0.0629
	Damaged Wall	15	31	0.15	0.0645	0.0408	0.0149
	Poster/Banner	85	485	0.484	0.408	0.364	0.207
	Waste	92	971	0.53	0.307	0.32	0.154
	Wire	44	165	0.386	0.256	0.235	0.125
YOLOv8x	Billboard	8	18	0.385	0.444	0.381	0.22
	Damaged Road	22	43	0.569	0.186	0.165	0.0738
	Damaged Wall	15	31	0.0525	0.0323	0.0144	0.0075
	Poster/Banner	85	485	0.4	0.396	0.352	0.209
	Waste	92	971	0.421	0.444	0.34	0.168
	Wire	44	165	0.231	0.418	0.239	0.115

5.2.2 Overall Efficiency and Accuracy Trade-Offs

In table 5.8, overall performance metrics are listed in terms of per image GPU inference time. YOLOv8n showed impressive performance in inference with the speed of 4.7 ms/image, which is why it can be used on the edge. YOLOv8x has the highest mAP@50 (0.2486) and F1 score (0.3311) but requires a processing time of 65.7 ms/image. In the case of YOLOv8s, it shows a reasonable compromise (0.2452 mAP@50 obtained at 11.3 ms).

Table 5.8: Overall Performance Metrics and Inference Speed Comparison of YOLOv8 Models on the VisPol Test Set.

Model Architecture	Precision	Recall	F1-Score	mAP@50	mAP@50–95	Inference Speed (per image)
YOLOv8n	0.3341	0.254	0.2886	0.2047	0.1033	4.7 ms
YOLOv8s	0.3662	0.2823	0.3188	0.2452	0.1206	11.3 ms
YOLOv8m	0.4142	0.2585	0.3183	0.2318	0.1294	23.2 ms
YOLOv8l	0.418	0.2589	0.3198	0.2426	0.134	39.6 ms
YOLOv8x	0.343	0.3201	0.3311	0.2486	0.1323	65.7 ms

5.3 Cross-Dataset Performance Evaluation and Generalization Analysis

To test the generalizability of the models, the authors trained them on VisPol and tested on two datasets that were not used in training: External Dataset 1 (Visual Pollution Dhaka Streets) and External Dataset 2 (PlasticPol).

5.3.1 Aggregate Cross-Dataset Metrics

Table 5.9 shows the performance of YOLOv8 models trained using VisPol when measuring their detection performance on the VisPol dataset as well as on the two external datasets. As seen, there is a noticeable performance drop when models are tested on external datasets, proving that there is a domain shift due to differences in image acquisition, annotation techniques, scene composition, and object distributions. Still, the suggested models show a decent detection capacity, thus proving the potential of the suggested VisPol dataset for cross-dataset generalization while demonstrating the difficulties of real-life application.

Table 5.9: Comparative Overall Cross-Dataset Performance Between the VisPol Test Set and External Evaluation Datasets.

Model	Evaluation Split	Images	Precision	Recall	F1-Score	mAP@50	mAP@50–95
YOLOv8n	Internal Test	109	0.3341	0.254	0.2886	0.2047	0.1033
	External Test 1	72	0.19	0.272	0.2237	0.101	0.0327
	External Test 2	242	0.157	0.16	0.1585	0.0603	0.0334
YOLOv8s	Internal Test	109	0.3662	0.2823	0.3188	0.2452	0.1206
	External Test 1	72	0.0687	0.236	0.1064	0.0515	0.0158
	External Test 2	242	0.114	0.126	0.1197	0.0409	0.0257
YOLOv8m	Internal Test	109	0.4142	0.2585	0.3183	0.2318	0.1294
	External Test 1	72	0.134	0.214	0.1648	0.0681	0.021
	External Test 2	242	0.184	0.165	0.174	0.095	0.0553
YOLOv8l	Internal Test	109	0.418	0.2589	0.3198	0.2426	0.134
	External Test 1	72	0.267	0.198	0.2276	0.114	0.0371
	External Test 2	242	0.125	0.155	0.1384	0.0452	0.0249
YOLOv8x	Internal Test	109	0.343	0.3201	0.3311	0.2486	0.1323
	External Test 1	72	0.0609	0.329	0.1028	0.0482	0.0188
	External Test 2	242	0.14	0.15	0.1448	0.0686	0.0416

5.3.2 Class-Wise Cross-Dataset Generalization

Table 5.10 shows a classification comparison between the internal VisPol test data and external validation datasets. From the results, it can be seen that classes that have highly differing annotation policies and object representations, especially Wire and Waste, show the highest performance degradation rates. The results of this analysis highlight the importance of distinguishing annotation granularity, object size, and scene complexity in cross-dataset performance. Further, the analysis demonstrates that external validation must be conducted to evaluate the ability of object detection models to operate beyond their training data.

Table 5.10: Class-Wise Cross-Dataset Generalization Comparison (mAP@50 / mAP@50-95).

Model	Class	Test Set (mAP@50/	External 1	External 2
YOLOv8n	billboard	0.405/0.203	0.242/0.075	—
	damaged road	0.0639/0.0252	—	—
	damaged wall	0.036/0.0247	—	—
	poster_banner	0.274/0.156	—	—
	waste	0.281/0.133	0.0533/0.0218	—
	wire	0.167/0.078	0.00823/0.00119	—
YOLOv8s	billboard	0.476/0.203	0.0905/0.0203	—
	damaged road	0.102/0.046	—	—
	damaged wall	0.00486/0.00243	—	—
	poster_banner	0.35/0.2	—	—
	waste	0.293/0.14	0.0567/0.0247	0.0409/0.0257
	wire	0.244/0.132	0.00764/0.00224	—
YOLOv8m	billboard	0.332/0.23	0.0912/0.0257	—
	damaged road	0.14/0.0614	—	
	damaged wall	0.034/0.0256	—	—
	poster_banner	0.374/0.212	—	—
	waste	0.334/0.157	0.0945/0.033	0.095/0.0553
	wire	0.176/0.0903	0.0187/0.0039	—
YOLOv8l	billboard	0.388/0.242	0.143/0.05	—
	damaged road	0.107/0.0629	—	—
	damaged wall	0.0408/0.0141	—	—
	poster_banner	0.364/0.207	—	—
	waste	0.32/0.154	0.195/0.0597	0.0452/0.0249
	wire	0.235/0.125	0.00567/0.00171	—
YOLOv8x	billboard	0.381/0.22	0.0614/0.0237	—
	damaged road	0.165/0.0733	—	—
	damaged wall	0.0144/0.00751	—	—
	poster_banner	0.352/0.209	—	—
	waste	0.34/0.168	0.078/0.0307	0.0686/0.0416
	wire	0.239/0.115	0.00518/0.00185	—

5.4 Training Performance and Qualitative Analysis of the YOLOv8l Model

Diagnostic evaluation of the YOLOv8l model provides insight into instance-level classification errors and feature convergence. Figure 5.1 shows the confusion matrix for the YOLOv8l model.

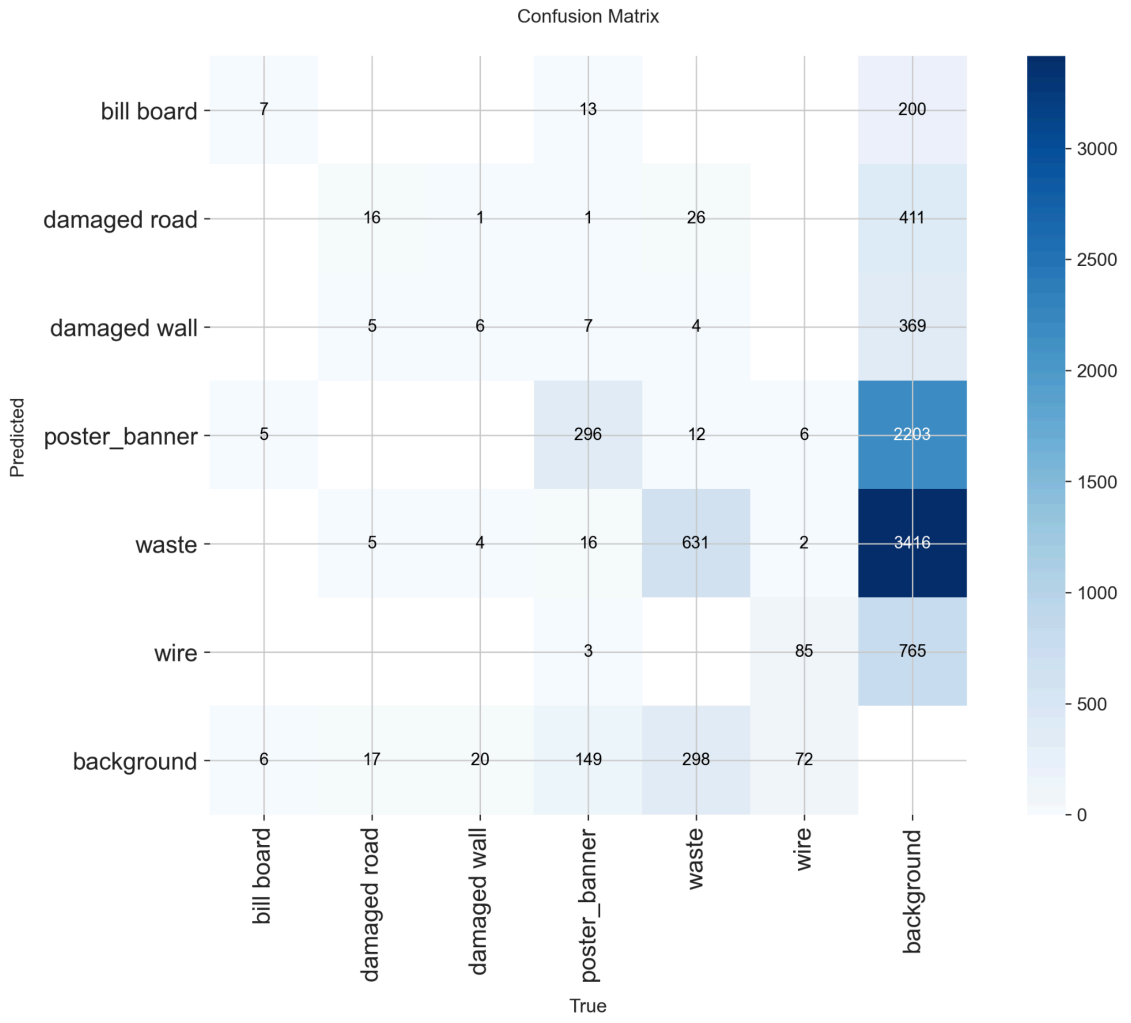
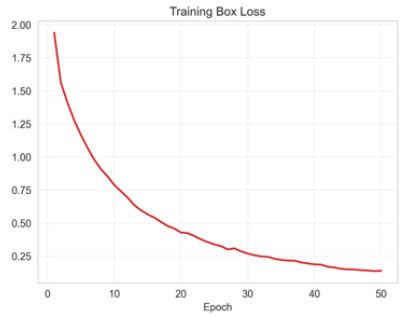


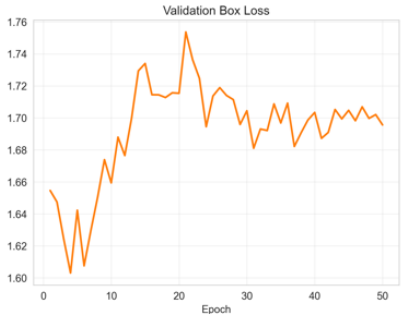
Figure 5.1: Confusion Matrix Demonstrating Instance-Level Classification and Confusion Rates for the YOLOv8l Model.

5.4.1 Training Dynamics

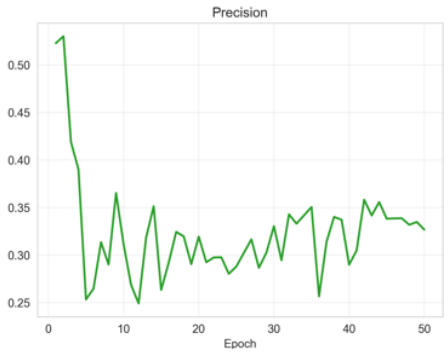
The training and validation loss curves (Box Loss, Distribution Focal Loss, and Classification Loss) have stabilized over a period of 100 epochs. Furthermore, Peak Precision, Peak Recall, and Peak mAP metrics have attained their stable values by not showing any signs of dramatic overfitting during training, which indicates that performance limits have been determined mainly by the complexity of the images and the distribution of classes rather than by optimization instability.



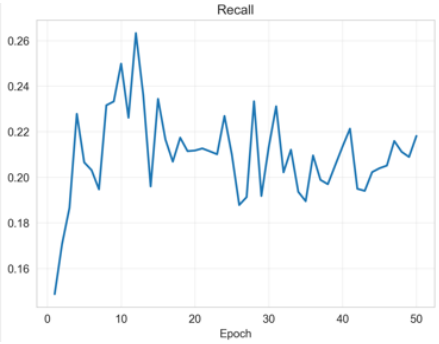
(a) Training Box Loss.



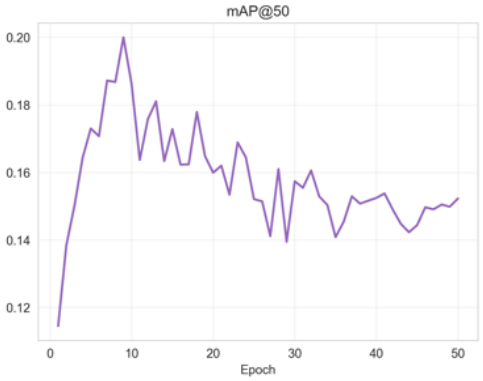
(b) Validation Box Loss.



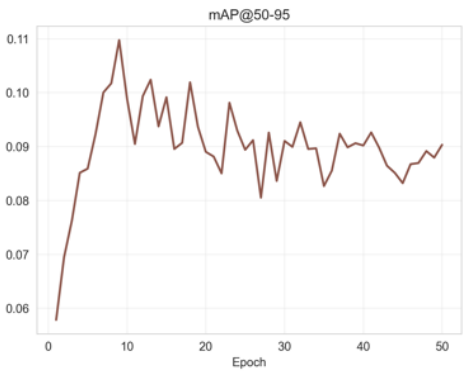
(c) Precision Curve.



(d) Recall Curve.

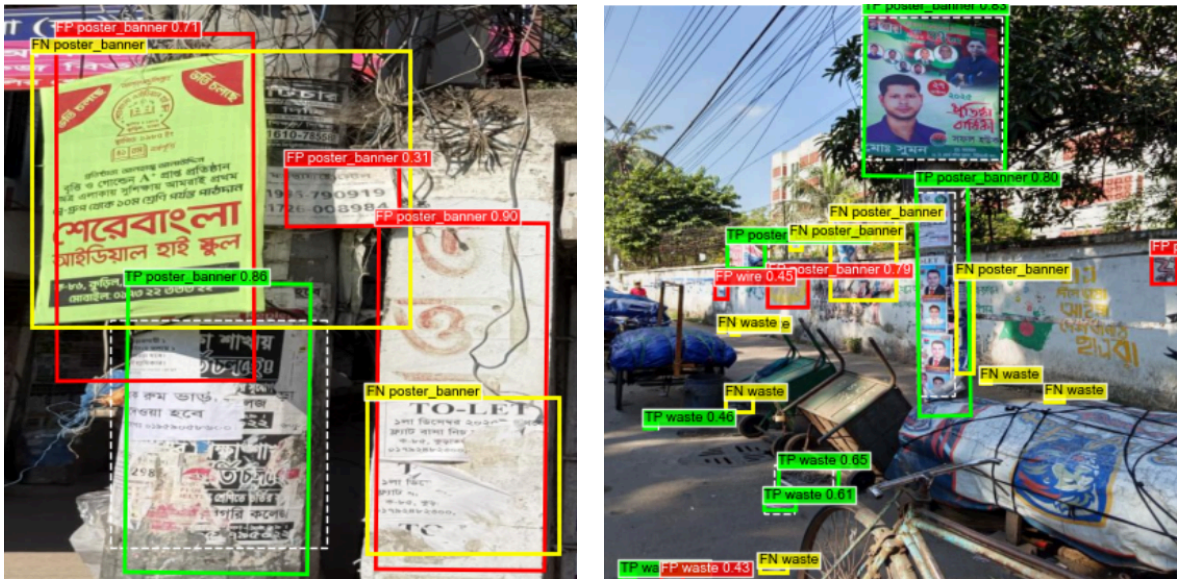


(e) mAP@50 curve.



(f) mAP@50-95 curve.

Figure 5.2 Training Loss and Validation Metric Convergence Curves for the YOLOv8l Detector.



(a) Red box represents False Positive, Yellow box represents False Negative, Green box represents True Positive.

(b) Red box represents False Positive, Yellow box represents False Negative, Green box represents True Positive.

Figure 5.3: Qualitative Detection Analysis Illustrating True Positive, False Positive, and False Negative Error Categories.

5.4.2 Failure Mode Categorization

Qualitative analysis shows that there are three main types of errors made in the urban areas:

- False Positive (FP): Some background features like quality signage/logo of the shop, walls that have clean surfaces like weathered ones, and other such structured facades were wrongly classified in some instances by the system.
- False negatives (FN): Small and occluded objects that are present at a distance like a small sign/poster and garbage that is far away are generally not recognized by the system, particularly in case of high visibility from the point of view of the poster/image.
- Confusion among various classes: This confusion is mainly present in Billboard and Poster and Banner classification where detection is based on the size of the posts and their distances.

5.5 Comparative Analysis of Internal and External Test Results

The important difference in scores of the data with respect to internal and external tests is qualitative in nature and mainly accounts for different ways adopted in the process of annotation in both cases.

(a) Blue box represents waste of the VisPol Dataset.

(b) Green box represents waste of the external 2 dataset.

Figure 5.4: Ground Truth Bounding Box Visualizations for Waste Class: VisPol Dataset vs. External Dataset 2.

5.5.1 Waste Density and Grouping Mismatches

Waste is also affected by a similar issue in the external datasets (e.g., YOLOv8l has 0.320 mAP@50 at the internal level versus 0.195 and 0.045 for External 1 and External 2, respectively):

Micro-Level Localization (VisPol): Detections of individual litter pieces, separate types of waste, and localized accumulation zones.

Macro-Level Region Grouping (External Datasets): Assembly of loosely spread waste piles into regions with broad bounding regions.

As a result, models that are trained to detect particular target instances generate localized bounding boxes that don't meet the conditions of spatial overlap with macro ground truth boundaries. This structural discrepancy reveals the influence of the differences in the annotation policy on cross-dataset evaluation.

(a) Pink box represents wire of the VisPol Dataset.

(b) Blue box represents the wire class of the external dataset 1.

Figure 5.5: Ground Truth Wire Annotation Scale Mismatch: Micro-Level Granularity (VisPol) vs. Macro-Level Single Bounding Box Framing (External Dataset 1).

5.5.2 Micro-Level vs. Macro-Level Annotation Scale Discrepancies

As illustrated in Figure 5.5, the drop in model performance for the "Wire" class on External Dataset 1 (mAP@50 = 0.00567) compared to the VisPol test set (mAP@50 = 0.235) is driven by spatial annotation scale mismatches:

1. **VisPol Annotation Strategy:** VisPol employs micro-level annotations, marking individual wire bundles, loops, and line segments with compact, tight bounding boxes.
2. **External Dataset Strategy:** External Dataset 1 uses macro-level annotations, enclosing entire overhead utility wire networks within single, expansive bounding boxes.

Because standard mAP calculation relies on strict Intersection over Union (IoU) thresholds (IoU $\geq$ 0.50), fine-grained predictions generated by VisPol-trained models fail to meet spatial overlap thresholds when evaluated against macro ground-truth regions. The evaluation pipeline records these localized predictions as false positives while flagging the macro ground truth as missed, resulting in a low mAP score despite accurate visual identification.

(a) Blue box represents waste of the VisPol Dataset.

(b) Green box represents waste of the external Dataset 2.

Figure 5.6: Ground Truth Waste Annotation Density Comparison: Fine-Grained Localized Bounding Boxes (VisPol) vs. Large Refuse Area Grouping (External Dataset 2).

According to Figure 5.6, the reason for the drop in performance of the YOLOv8l model on the external dataset (mAP@50 = 0.195) compared to the internal dataset (mAP@50 = 0.320) is due to the bounding box scaling problems and also due to the misunderstanding of the semantic class between the two datasets.

The main reason for this difference in performance is the difference in the granulation of the annotation. From the green bounding boxes shown in Figure 5.6, we can see that in the external dataset, the annotation follows the macro approach of bounding boxing, meaning that many street waste piles of large size are placed into one bounding box. At the same time, in the internal dataset that is represented by the blue bounding boxes, the annotation takes place at the micro level, meaning that the model identifies small litter ingredients that are present under stalls at the market. As a consequence of this large difference in scale between the two datasets, the model's IoU metrics are severely affected during cross-dataset evaluation.

5.5.3 Interpretation of External Validation Results

Results from external assessments should be regarded in terms of generalization difficulties rather than as evidence of VisPol-trained detectors' inability to identify the necessary concepts. External Dataset 1 is not similar in structure to VisPol, even though they are associated with Dhaka. The Dataset shows generally sparser footage, different framing of the objects, and different annotation standards. External Dataset 2 shows an even greater distributional shift since it is a single

classification waste-specialized dataset with relatively short-range footage. Therefore the drop in mAP from external datasets reflects both visual domain shift and evaluation policy mismatch.

Thus, the comparison shows why the use of a single internal test split may not provide a complete understanding of deployment suitability. A model may show average internal mAP while yielding considerably worse results due to the change in scale, grouping strategy, or scene organization. The difference is important as in the application of municipal monitoring the images that are to be deployed are likely to differ from the distribution of the training benchmark.

5.6 Overall Discussion

Several pieces of evidence show that VisPol is not a detection dataset that lends itself easily to use but rather a very difficult benchmark to work with. Overall performance in the experiment is moderate, and there is a noticeable difference in how the majority and minority categories are classified. This is because the aim is to create realistic urban scenes with a variety of clutter and many different-sized objects.

Having a model with high capacity did not necessarily guarantee its superior productivity. Model YOLOv8m had the greatest performance according to the mAP@50 parameter during cross-validation testing since it reached the value of 0.1992. On the other hand, the highest mAP@50–95 result during cross-validation testing belonged to model YOLOv8x, whose score was 0.1110. In terms of performance in the independent testing stage, model YOLOv8x demonstrated the best results according to the mAP@50 metric, as it reached the value of 0.2486 and the highest F1 score of 0.3311. As for the precision of the model, YOLOv8l turned out to be better with the value of 0.418. However, the modelless YOLOv8n did not take much time at all, with its running time of 4.7 ms per image. Interestingly, the results illustrate the trade-off between productivity and speed. Hence, it is important to understand what actual limitations developers may face while implementing this model instead of merely looking at the numbers of the model, i.e., the number of parameters. Also, taking into account the per-class analysis of the models, it is noticeable that Waste and Poster/Banner received good results with many training images. On the contrary, Damaged Wall and Damaged Road.

The findings from the comparison across different datasets demonstrate that external validation is necessary. The consistency with which precision, recall, and the mAP drop when applied to external datasets indicates that generalization depends not only on the semantic similarity of classes but also on acquisition conditions, framing of the object, granularity of annotations, and density of the scene. The main message from the external experiments is therefore methodological: the quality of the benchmark should be assessed in relation to both the performance in the within-dataset evaluations and the out-of-distribution behavior.

6 Threats to Validity

While the Vispol dataset and its experimental methodology were engineered to establish a realistic benchmark for urban visual pollution detection, certain factors could affect the validity and generalizability of the outcomes. Adhering to standard empirical research protocols in computer vision, this chapter examines potential threats to internal, external and construct validity alongside the mitigation strategies implemented.

6.1 Internal Validity

Internal validity refers to the extent to which the observed experimental results can be attributed to the proposed methodology rather than uncontrolled variables. In this research, several factors may have influenced the detection performance of the evaluated models.

6.1.1 Annotation Errors

Supervised object detection relies heavily on accurate annotation. Minor annotation errors may persist even when a standardised annotation procedure and several rounds of manual verification are implemented. Key issues involve minor bounding-box misalignments, missed micro-objects, ambiguous label assignments and inconsistent handling of occlusion. These challenges are especially pronounced in structural classes such as wire, damaged wall and billboard due to their irregular contours and frequent partial visibility.

To minimize annotation-related bias, every image in the VisPol dataset underwent a two-stage quality assurance process using Roboflow. To ensure uniform labeling across all contributors explicit annotation guidelines were established prior to data collection. Before finalizing the dataset, all bounding boxes and classes labels underwent manual inspection, refinement and verification. While minor annotation noise cannot be entirely eliminated, this multi-stage verification significantly improved overall dataset reliability and precision.

6.1.2 Class Imbalance

Because the Vispol mirrors the real-world occurrence of visual pollution in metropolitan areas, it inherently exhibits class imbalance. The majority of annotations are concentrated in the waste and poster. banner categories, whereas significantly fewer instances are present in billboards, damaged walls and damaged roads.

Such imbalance can skew training toward majority categories leading to reduced precision and recall for minority classes. This trend is corroborated by the experimental findings in chapter %, where models achieved higher detection metrics for waste and poster/banner but struggled to effectively generalize on less frequent structural defects.

Rather than artificially rebalancing the dataset through aggressive resampling or class trimming the natural class distribution was preserved to ensure realistic performance evaluations in deployment-oriented scenarios. While keeping object frequencies realistic, we used training-level strategies, like data augmentation. We also did an evaluation using many different metrics to help with the problems caused by class imbalance.

6.1.3 Training Bias

Training bias may arise from the characteristics of the collected data and the adopted experimental configuration. Since data collection was geographically confined to Dhaka, specific local architectural traits, infrastructure designs, signage formats and waste disposal patterns are heavily emphasized. Consequently, models risk overfitting to site-specific visual cues rather than acquiring broader domain-agnostic features representative of visual pollution.

To reduce this bias, images were collected from multiple residential, commercial, and transportation areas under varying viewpoints, scales, and scene complexities. In addition, five different YOLOv8 model variants were evaluated under identical experimental settings to ensure that the reported findings were not dependent on a single model architecture. Evaluating the models on external datasets further allowed us to quantify the impact of location-specific bias and assess domain transferability.

6.1.4 Hyperparameter Influence

Object detection performance is sensitive to training hyperparameters such as learning rate, batch size, optimizer, image resolution, confidence threshold, and data augmentation strategy. Different hyperparameter combinations may produce different performance outcomes. A standardised training strategy was enforced across all YLOv8 model variants to ensure comparative fairness. By keeping hyperparameter configurations such as the optimizer , learning rate schedule , batch size , input size , augmentation pipeline and early stopping rules strictly constant, training-induced variance was minimized. Consequently, performance variations can be attributed to the intrinsic structural differences of each architecture.

Nevertheless, further hyperparameter optimization could potentially improve individual model performance and therefore remains a potential source of internal validity threat.

6.2 External Validity

External validity concerns the extent to which the findings of this research can be generalized beyond the experimental setting.

6.2.1 Generalization to Other Cities

Data collection for Vispol was geographically restricted to Dhaka, Bangladesh. While the chosen locations encompass a broad range of urban scenarios ,they may not adequately represent the architectural idioms, transportation infrastructure and meteorological variations characteristic of other worldwide metropolitan areas.

Therefore, when deployed in locations with significantly differing visual features, models trained exclusively on VisPol may exhibit lower accuracy. To partially address this limitation, external evaluation was performed using two independent public datasets collected under different acquisition protocols and annotation standards. The observed performance degradation across these datasets highlights the importance of considering geographic diversity when developing deployment-oriented object detection systems.

6.2.2 Different Weather Conditions

The VisPol dataset primarily consists of images that were taken during the day under clear daytime conditions. These images show some changes like the varying sunlight intensities, shadows and clouds. The VisPol dataset does not have many images of adverse environmental conditions such as heavy precipitation, thick fog, or scenes taken at night. It also does not have images with a lot of haze in the air.

Environmental factors such as severe weather alter scene contrast and feature visibility in images and make it harder to see objects. So the VisPol dataset models may not work well in bad weather compared to the benchmark results reported in this dissertation. Consequently, models trained on VisPol may experience degraded detection performance when processing images captured under adverse weather conditions. Future dataset expansion incorporating seasonal and weather diversity would further strengthen model robustness.

6.2.3 Camera Variations

Data collection was executed using a diverse range of mobile devices from prominent manufacturers, including Apple, Xiaomi, and Realme. This setup introduced realistic variations in sensor characteristics, image resolution, exposure, and color processing.

However, many imaging devices used in practical smart-city applications including CCTV cameras, surveillance systems, vehicle-mounted cameras, and UAV platforms possess substantially different optical properties and image quality. Consequently, performance may vary when the trained models are deployed using alternative imaging systems. The use of multiple smartphones partially mitigates this limitation but does not completely eliminate hardware-related domain differences.

6.2.4 Domain Shift

Domain shift represents one of the most significant challenges for deployment-oriented object detection. Differences in annotation protocols, camera viewpoints, scene composition, object scale, environmental complexity, and class definitions may substantially influence detection accuracy.

The application of VisPol-trained models on external datasets showed a decline in performance, as explained in Chapter 5. Even though a qualitative inspection revealed that the issues of diversity in annotation criteria, groupings of classes, scale difference, and scene complexity made cross-domain transfer difficult. This suggests that the VisPol benchmark offers an all-encompassing evaluation while still leaving open the question of domain adaptation as one of the crucial areas of investigation for the future.

6.3 Construct Validity

Construct validity evaluates whether the adopted experimental methodology accurately measures the intended research objectives.

6.3.1 Choice of Evaluation Metrics

Model evaluation in this research relied on standard object detection metrics , including precision, recall , F1-score, intersection over union (IoU) and mean average precision across standard thresholds (mAP@50 and mAP@50–95). These metrics systematically quantify classification accuracy, bounding-box localization precision and general model robustness. However, quantitative metrics alone are insufficient to gauge deployment readiness, as operational reliability, annotation consistency and environmental resilience remain difficult to measure numerically. To address this, quantitative

evaluations were supplemented with qualitative error analysis, failure mode investigations and cross-dataset validations.

6.3.2 Dataset Representativeness

Although VisPol contains diverse urban scenes collected across multiple regions of Dhaka, the dataset cannot fully represent every possible form of visual pollution encountered worldwide. Certain rare infrastructure defects, seasonal environmental conditions, and region-specific urban characteristics remain absent.

Nevertheless, the dataset intentionally captures substantial diversity in viewing distance, object density, background complexity, illumination, and occlusion, making it considerably more representative of practical deployment scenarios than many existing benchmark datasets.

6.3.3 Benchmark Selection

Five YOLOv8 variants were selected as the primary benchmark models because they represent one of the most widely adopted real-time object detection frameworks in current computer vision research. The evaluation also included cross-dataset validation using two independent public datasets to assess model robustness beyond the training distribution.

Despite this comprehensive benchmarking, the study does not include recently proposed transformer-based detectors such as RT-DETR, DINO, or Grounding-DINO, nor does it evaluate vision-language detection frameworks. Consequently, future comparisons involving these emerging architectures may provide additional insights into the strengths and limitations of the proposed VisPol dataset.

6.4 Threats and Mitigation Strategies

Table 6.1 presents a summary of the identified validity threats and the mitigation strategies employed in this study. The implementation of these measures enhances the reliability and robustness of the findings of our experiments while bearing in mind the practical limitations of urban visual pollution detection in the real-world. Although some limitations are inherent to computer vision research in field applications, a number of potential threats were effectively addressed by strict annotation procedures, benchmarking, cross-validation, and evaluation on external datasets performed in this research. This study gives an overview of what is good and what is bad regarding currently available detectors, allowing us to bridge the gap between their theoretical functioning and effectiveness in real city conditions.

Table 6.1: Summary of Threats to Validity and Mitigation Strategies.

Threat Category	Potential Threat	Possible Impact	Mitigation Strategy
Internal Validity	Annotation inconsistencies	Incorrect labels and localization errors	Standardized annotation protocol, Roboflow-based annotation, two-stage manual quality assurance
Internal Validity	Class imbalance	Reduced performance on minority classes	Preserve natural distribution, data augmentation, cross-validation, class-wise evaluation
Internal Validity	Training bias	Location-specific feature learning	Multi-location image collection, standardized training, evaluation of five YOLOv8 variants
Internal Validity	Hyperparameter sensitivity	Performance variation across configurations	Unified experimental settings for all models
External Validity	Geographic limitation	Reduced transferability to other cities	External validation using independent public datasets
External Validity	Limited weather diversity	Lower robustness under adverse weather	Collection under naturally varying illumination and recommendation for future seasonal expansion
External Validity	Camera differences	Performance variation across imaging devices	Multi-device image acquisition using different smartphone cameras
External Validity	Domain shift	Cross-dataset performance degradation	Independent benchmark evaluation and qualitative domain-shift analysis
Construct Validity	Metric limitations	Deployment aspects not fully reflected	Combination of quantitative metrics and qualitative analysis
Construct Validity	Dataset representativeness	Limited coverage of all visual pollution scenarios	Diverse urban locations, complex scenes, multi-scale object instances
Construct Validity	Benchmark selection	Limited comparison with emerging detectors	Evaluation across five YOLOv8 variants and recommendation for future transformer-based comparisons

7 Conclusion and Future Work

7.1 Conclusion

This research introduced VisPol, a deployment-oriented visual pollution dataset and benchmarking framework for automated urban monitoring. The dataset contains 1,097 images and 15,933 annotated object instances across six classes and captures the variability of real urban environments in Dhaka, including clutter, occlusion, varying object scales, and complex backgrounds. Five YOLOv8 variants were systematically evaluated using standard detection metrics, cross-validation, an independent test set, and external datasets. The results showed that larger models generally provided stronger localization performance, while minority classes such as Damaged Wall and Damaged Road remained challenging due to limited representation, irregular and diffuse object boundaries and high visual variability. Cross-dataset evaluation further demonstrated that model performance can decline substantially when data distributions, scene structures, object representations, and annotation practices differ. These findings highlight that benchmark performance alone does not fully reflect real-world effectiveness and that realistic datasets and external validation are essential for reliable deployment-oriented evaluation. Overall, VisPol provides a foundation for studying visual pollution detection under realistic urban conditions and supports future development of automated urban monitoring systems.

7.2 Future Work

Future research should extend VisPol in terms of geographic coverage, class diversity, model capability, and deployment scenarios. Expanding the dataset across additional cities in Bangladesh and other countries would reduce geographic bias and improve assessment of cross-domain generalization, while adding categories such as graffiti, abandoned vehicles, and drainage problems would broaden its practical scope. Future studies can also evaluate transformer-based detectors and Vision-Language Models, alongside semi-supervised, self-supervised, and active learning approaches to improve detection and reduce annotation requirements. Moving from still-image detection to video-based monitoring could enable real-time applications using traffic cameras, vehicle-mounted systems, and UAVs, while integrating visual data with GIS, GPS, environmental sensors, and textual information could provide richer urban analysis. Finally, model optimization through quantization, pruning, and hardware acceleration could facilitate efficient Edge AI deployment on roadside cameras, IoT devices, and UAVs, moving VisPol toward scalable and real-time intelligent urban environmental monitoring.

References

[1] Sumon, S. I., Chowdhury, M. E., Chowdhury, J. U. K., Ashraf, A., Kashem, S. B. A., Majid, M. E., ... & Kunju, A. K. A. (2026). Floating waste detection using deep learning: a comparative study of YOLO, RT-DETR, and faster R-CNN. Neural Computing and Applications, 38(8), 293.

[2] Sayem, F. R., Islam, M. S. B., Naznine, M., Nashbat, M., Hasan-Zia, M., Kunju, A. K. A., ... & Chowdhury, M. E. H. (2025). Enhancing waste sorting and recycling efficiency: robust deep learning-based approach for classification and detection. Neural Computing and Applications, 37, 4567–4583.

[3] Dipo, M. H., Farid, F. A., Mahmud, M. S. A., Momtaz, M., Rahman, S., Uddin, J., & Karim, H. A. (2025). Real-Time Waste Detection and Classification Using YOLOv12-Based Deep Learning Model. Digital, 5(2), 19.

[4] Roy, P. K., Islam, M. R., & Rafi, M. J. A. (2024). Non-biodegradable plastic waste detection and classification using deep learning: Bangladeshi environmental scenario. In 2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT) (pp. 1-6). IEEE.

[5] Anika, S. C., Hossain, M. R., Tamjeed, A., & Jahan, M. (2024). Roadway Impairment Identification in Asphalt Road Using Yolov8. 3rd International Conference on Computing Advancements (ICCA 2024), 954–961.

[6] Hossen, R., Mistry, D., Rahman, M., Hridoy, W. A. S. A. R., Saha, S., & Ibrahim, M. (2025). Road Damage and Manhole Detection using Deep Learning for Smart Cities: A Polygonal Annotation Approach. arXiv preprint arXiv:2510.03797.

[7] Safyari, Y., Mahdianpari, M., & Shiri, H. (2024). A Review of Vision-Based Pothole Detection Methods Using Computer Vision and Machine Learning. Sensors, 24(17), 5652.

[8] Kumar, P. S. (2025). Road condition assessment: A framework for automatic detection of surface flaws. Department of CSE, VNR VJIET, Hyderabad, Telangana, India.

[9] Dwivedi, S. K., Arya, D., & Sekimoto, Y. (2024). Scaling Limits of Federated Learning for Road Damage Detection Across Countries. SSRN Preprint.

[10] Fotovvatikhah, F., Ahmedy, I., Noor, R. M., & Munir, M. U. (2025). A Systematic Review of AI-Based Techniques for Automated Waste Classification. Sensors, 25(10), 3181.

[11] Shrestha, S., Shrestha, S., Paudel, P., Gurung, S., Gairee, S., & Adhikari, S. (2024). Deep learning for waste management: Leveraging yolo for accurate waste classification. In 15th IOE Graduate Conference.

[12] Pati, B. M., Khadka, B., Thapa, U., Pal, S. K., & Pyakurel, D. Unveiling Waste and its Ripple Effect: Analyzing the Impact of Waste of Water Resources Quality.

[13] Zhang, K., Wu, Y., Li, N., Guo, X., Cheng, R., & Zhu, L. (2025). Lightweight Road Defect Detection Algorithm Based on Improved YOLOv10. Engineering Letters, 33(5), 1193-1202.

[14] Wang, C., & Sun, M. (2025). Improved road defect detection algorithm for YOLOv10n. In Eighth International Conference on Artificial Intelligence and Pattern Recognition (AIPR 2025) (Vol. 13993, pp. 1002-1010). SPIE.

[15] Ding, K., Ding, Z., Zhang, Z., Yuan, M., Ma, G., & Lv, G. (2024). SCD-YOLO: A novel object detection method for efficient road crack detection. Multimedia Systems, 30(6), 351.

[16] Lin, Z. & Pan, W. (2025). YOLO-ROC: A High-Precision and Ultra-Lightweight Model for Real-Time Road Damage Detection. Journal of Real-Time Image Processing.

[17] Chu, H. H., Saeed, M. R., Rashid, J., Mehmood, M. T., Ahmad, I., Iqbal, R. S., & Ali, G. (2023). Deep Learning Method to Detect the Road Cracks and Potholes for Smart Cities. Computers, Materials & Continua, 75(1), 1863–1881.

[18] Saeed, Z., & Raza, A. (2025). CrackNet: Pavement Crack Detection and Classification Based on Deep Learning Models. Intelligent Methods in Engineering Sciences, 4(3), 74–84.

[19] Aamir, M. S. B., Ilyas, M. H., Khalique, F., Bashir, B., Mahmood, S., & Khalil, R. A. (2025). Autonomous Road Defects Segmentation Using Transformer-Based Deep Learning Models With Custom Dataset. IEEE Access, 13, 174177–174199.

[20] Kek, S. L., Lim, F. P., & Yap, H. K. (2025). Prediction of road safety risks through crack detection and structural deterioration assessment. Mechatron. Intell Transp. Syst, 4(4), 198-209.

[21] Raza, S. M., Hassan, S. M. G., Hassan, S. A., & Shin, S. Y. (2021). Real-time trash detection for modern societies using CCTV to identifying trash by utilizing deep convolutional neural network. arXiv preprint arXiv:2109.09611.

[22] Khan, H. A., Naqvi, S. S., Alharbi, A. A. K., Alotaibi, S., & Alkhathami, M. (2024). Enhancing trash classification in smart cities using federated deep learning. Scientific Reports, 14(1), 11816.

[23] Tharani, M., Amin, A. W., Rasool, F., Maaz, M., Taj, M., & Muhammad, A. (2021). Trash detection on water channels. In International Conference on Neural Information Processing (pp. 379-389). Springer.

[24] Hasan, M. K., Khan, M. A., Issa, G. F., Atta, A., Akram, A. S., & Hassan, M. (2022). Smart waste management and classification system for smart cities using deep learning. In 2022 International Conference on Business Analytics for Technology and Security (ICBATS) (pp. 1-7). IEEE.

[25] Tharani, M., Amin, A. W., Maaz, M., & Taj, M. (2020). Attention neural network for trash detection on water channels. arXiv preprint arXiv:2007.04639.

[26] Ijaz, M., Khan, S. U. R., Rehman, A. U., Asif, T., Vollmer, S., Dengel, A., & Asim, M. N. (2026). GlobalWasteData: A Large-Scale, Integrated Dataset for Robust Waste Classification and Environmental Monitoring. arXiv preprint arXiv:2602.07463.

[27] Ahmad, G., Aleem, F. M., Alyas, T., Abbas, Q., Nawaz, W., Ghazal, T. M., ... & Ibrahim, A. M. (2025). Intelligent waste sorting for urban sustainability using deep learning. Scientific reports, 15(1), 27078.

[28] Nunkhaw, M., Chitwatkulsiri, D., & Miyamoto, H. (2025). Enhancing River Waste Detection with Deep Learning and Preprocessing: A Case Study in the Urban Canals of the Chao Phraya River. Water, 17(22), 3193.

[29] Taweesan, A., Koottatep, T., Kanabkaew, T., Eamrat, R., Pussayanavin, T., & Polprasert, C. (2025). Unveiling sustainable solutions for municipal waste management through machine learning. Environmental Challenges, 101333.

[30] Wang, Z., Zhang, Z., Wang, T., Zhang, X., Ren, Z., & Zhang, Y. (2026). Garbage Recognition based on YOLOv8-seg: An Instance Segmentation Approach for Automated Waste Sorting. International Core Journal of Engineering, 12(4), 267-278.

[31] Sukanto, S. D., Roy, D., Shakil, F., Singha, N., Asik, A., Joarder, A., ... & Ibrahim, M. (2025). Detecting unauthorized vehicles using deep learning for smart cities: A case study on bangladesh. arXiv preprint arXiv:2510.26154.

[32] Chowdhury, R. I., Anjom, J., & Hossain, M. I. A. (2024). A novel edge intelligence-based solution for safer footpath navigation of visually impaired using computer vision. Journal of King Saud University-Computer and Information Sciences, 36(8), 102191.

[33] Zanevych, Y., Yovbak, V., Basystiuk, O., Shakhovska, N., Fedushko, S., & Argyroudis, S. (2024). Evaluation of Pothole Detection Performance Using Deep Learning Models Under Low-Light Conditions. Sustainability, 16(24), 10964.

[34] Karim, A., Raza, M. A., Alharthi, Y. Z., Abbas, G., Othmen, S., Hossain, M. S., ... & Mercorelli, P. (2024). Visual detection of traffic incident through automatic monitoring of vehicle activities. World Electric Vehicle Journal, 15(9), 382.

[35] Rakin, R. Z., Rahman, M., Borsa, K. F., Farid, F. A., Rahman, S., Uddin, J., & Karim, H. A. (2025). Towards safer cities: AI-powered infrastructure fault detection based on YOLOv11. Future Internet, 17(5), 187.

[36] Lin, J., Wang, P., Ruan, Y., & Sun, Y. (2026). YOLO11-WLBS: An efficient model for pavement defect detection. Scientific Reports.

[37] Ghosh, S., Roy, A., & Majumdar, A. Smart Garbage Collector Vehicle: A Steps towards Swachh Bharat.

[38] Gelar, T., Fitriani, S., & Rachmat, S. (2025). A systematic literature review of YOLO and IOT applications in smart waste management. Green Intelligent Systems and Applications, 5(2), 123-139.

[39] Kumar, A., Sudarshan, L. B. N., Kumar, A., & Kumar, R. (2025). A review of computer vision applications in litter and cleanliness monitoring. In Proceedings of the Institution of Civil Engineers-Waste and Resource Management (Vol. 178, No. 1, pp. 30-50). Emerald Publishing Limited.

[40] Hossain, M. Y., Nijhum, I. R., Shad, M. T. M., Sadi, A. A., Peyal, M. M. K., & Rahman, R. M. (2023). An end-to-end pollution analysis and detection system using artificial intelligence and object detection algorithms. Decision Analytics Journal, 8, 100283.

[41] Huang, G., Yu, Y., Lyu, M., Sun, D., Dewancker, B., & Gao, W. (2024). Impact of physical features on visual walkability perception in urban commercial streets by using street-view images and deep learning. Buildings, 15(1), 113.

[42] Al Mazroa, A., Maray, M., Alashjaee, A. M., Alotaibi, F. A., Alzahrani, A. A., Alkharashi, A., ... & Alnfiai, M. M. (2024). Computer vision with explainable artificial intelligence for visual pollution detection in the Kingdom of Saudi Arabia. IEEE Access, 12, 193014-193027.

[43] Sami, A. A., Sakib, S., Deb, K., & Sarker, I. H. (2023). Improved YOLOv5-based real-time road pavement damage detection in road infrastructure management. Algorithms, 16(9), 452.

[44] Hossain, M. Y., Nijhum, I. R., Shad, M. T. M., Sadi, A. A., Peyal, M. M. K., & Rahman, R. M. (2023). An end-to-end pollution analysis and detection system using artificial intelligence and object detection algorithms. Decision Analytics Journal, 8, 100283.

[45] Das, D., Deb, K., Sayeed, T., Dhar, P. K., & Shimamura, T. (2023). Outdoor trash detection in natural environment using a deep learning model. IEEE Access, 11, 97549-97566.

[46] Baig, M. N., Patwary, M. M., Chowdhury, H. A., & Rahman, M. S. BRSSD10k: A Segmentation Dataset of Bangladeshi Road Scenario.

[47] Cheng, Y., Zhu, J., Jiang, M., Fu, J., Pang, C., Wang, P., ... & Bengio, Y. (2021). Flow: A dataset and benchmark for floating waste detection in inland waters. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10953-10962).

[48] Abo-Zahhad, M. M., & Abo-Zahhad, M. (2025). Real time intelligent garbage monitoring and efficient collection using Yolov8 and Yolov5 deep learning models for environmental sustainability. Scientific Reports, 15(1), 16024.

[49] Darisang, L. A., Dwisetio, S. K., Edbert, I. S., & Aulia, A. (2025). Real-time waste detection using YOLO, SSD, and Faster R-CNN integrated with CNN-based classification. In Proc. Int. Conf. on Research and Innovations in Information and Engineering Technology (RITECH) (pp. 76-83).

[50] Almalki, H., & Algethami, N. (2025). Ensemble deep learning with image captioning for visual pollution detection, classification, and reporting. Scientific Reports, 15(1), 31516.

[51] Titu, M. F. S., Chowdhury, A. A., Haque, S. R., & Khan, R. (2024). Deep-learning-based real-time visual pollution detection in urban and textile environments. *Sci*, *6*(1), 5.

[52] Smaron, J. M., Alam, M., & Islam, M. T. (2025). Deep Learning on Dental RGB and OPG Images for Public Health and Clinical Decision Support (Doctoral dissertation, Independent University, Bangladesh).