

Vision for Pollution: Deep Learning to detect visual pollution

Fergus Gault



4th Year Project Report
Computer Science
School of Informatics
University of Edinburgh
2026

Abstract

Visual pollution, caused by the presence of man-made elements such as advertisements, infrastructure, and urban clutter, negatively impacts perceived environmental quality and human wellbeing. Despite increasing interest in this area, existing approaches to visual pollution detection are often limited by small, localised datasets, manual assessment methods, and a narrow set of pollutant categories.

This work presents a scalable framework for detecting and quantifying visual pollution using deep learning and large-scale street-view imagery. A global dataset comprising over 2 million images from 179 countries was constructed using multiple data sources. Nine visual pollutant classes were defined, extending beyond commonly studied categories to include finer distinctions such as shop signs and mobile advertisements. A YOLO-based object detection model was trained on this dataset to identify pollutant instances across diverse geographic contexts.

To quantify visual pollution, the Visual Pollution Index (VPI) was used, enabling regional comparisons. The system was evaluated through multiple experiments, including analysis of regulatory impact in cities such as São Paulo and Polish municipalities, as well as land-use-based variation within Edinburgh.

Results demonstrate that the proposed approach can capture meaningful spatial and temporal patterns in visual pollution. However, limitations such as uneven data availability, class imbalance, and labelling ambiguity introduce challenges in global generalisation.

Overall, this work contributes a globally diverse dataset, an expanded pollutant taxonomy, and a scalable detection framework, providing a foundation for future research and practical applications in urban analysis and planning.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics Committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Fergus Gault)

Acknowledgements

I would like to express my gratitude to my supervisor, Karthik Mohan, for his support and guidance throughout this project. His insight and encouragement provided valuable perspectives that helped overcome challenges encountered during the work.

I would also like to thank Sohan Seth and the members of the SHAPE research group for their feedback and discussions, which were beneficial to the development of this project.

Finally, I would like to thank my family, friends, and my girlfriend for their continued support throughout this project and my degree. Their encouragement helped me navigate challenging periods, and this work would not have been possible without them.

I acknowledge the use of generative AI tools, primarily GPT-5 for assistance with structure and phrasing, and GitHub Copilot for code assistance, including generating mapping scripts and refactoring. These tools were used in accordance with University guidance, and all ideas, results, and analyses presented in this work are my own.

Table of Contents

1	Introduction	1
1.1	Motivation	1
1.2	Contributions	1
1.3	Structure	2
2	Background	3
2.1	Defining Visual Pollution	3
2.2	Historical and Societal Context	4
2.3	Assessing Visual Pollution	5
2.4	Computer Vision	6
2.5	Previous Work	7
2.5.1	Ahmed et al. (2019)	8
2.5.2	Gupta et al. (2024)	8
2.5.3	Rahman et al. (2026)	9
3	Visual Pollutants	11
3.1	Billboards	11
3.2	Bins	12
3.3	Utility Poles	12
3.4	Graffiti	12
3.5	Mobile Advertisement	12
3.6	Shop signs	13
3.7	Road signs	13
3.8	Barriers	13
3.9	Potholes	13
4	Dataset Construction	14
4.1	Data sources	14
4.1.1	Mapillary	14
4.1.2	KartaView	15
4.1.3	OpenStreetMap	15
4.1.4	Limitations	15
4.2	Data Collection Pipeline	16
4.2.1	Region creation	16
4.2.2	API Querying	17
4.2.3	Data cleaning	17

4.3	Geographic Sampling Strategy	17
4.4	Annotation Process	19
4.5	Dataset Statistics	20
4.5.1	Regions	20
4.5.2	Images	21
4.5.3	OSM Features	22
4.5.4	Detections	22
5	System Architecture	23
5.1	System Overview	23
5.2	Data storage	23
5.3	Detection Pipeline	24
5.4	Pollution Scoring	24
6	Model Training	26
6.1	Model Architecture	26
6.2	Training Setup	26
6.2.1	Hyperparameters	27
6.2.2	Data Augmentation	28
6.3	Evaluation Metrics	29
6.4	Initial Model	30
7	Experimental Results	31
7.1	Scores	31
7.1.1	Scores without OSM features	31
7.1.2	Scores with OSM features	32
7.2	São Paulo, Brazil	33
7.2.1	Results	33
7.3	Poland	34
7.3.1	Results	34
7.4	Pollution by land-use in Edinburgh, Scotland	35
7.4.1	Results	36
7.5	Additional Experiments	36
8	Discussion	37
8.1	Limitations	37
8.1.1	Data availability	37
8.1.2	Labelling	38
8.2	Future Work	39
9	Conclusion	40
	Bibliography	41
A	Additional experiments	46
A.1	Country-level Analysis	46
A.2	Top Regions per Pollutant Class	47

A.3	Heatmaps	49
B	Maps	50
C	Training Figures	52
D	Visual Pollutants	54
D.1	Billboards	54
D.2	Bins	55
D.3	Utility Poles	55
D.4	Graffiti	56
D.5	Mobile Advertisement	56
D.6	Shop Signs	57
D.7	Road Signs	57
D.8	Barriers	58
D.9	Potholes	58
E	Graphs	59
F	Codebase Features	60
F.1	collect_worldcities.py	60
F.2	create_dataset.py	61
F.3	label.py	61
F.4	run_inference.py	61
F.5	score_regions.py	62
F.6	train.py	62
F.7	download_data.py	62
F.8	Scripts	63

Chapter 1

Introduction

1.1 Motivation

Visual pollution refers to the degradation of a visual environment by man-made elements such as billboards, wires, poles, and litter [1]. Although it does not have the same direct environmental or physical health impacts as other forms of pollution, it has been shown to negatively affect wellbeing, perceived liveability, and mental health [2].

Some governments have recognised this issue and introduced regulatory measures. For example, São Paulo implemented the *Clean City Law*, which banned most forms of outdoor advertising [3], while Poland introduced the *Landscape Act*, enabling municipalities to regulate signage and advertisements [4]. Despite these efforts, visual pollution remains relatively under-addressed globally. In many developing regions, dense above-ground wiring and infrastructure contribute significantly to visual clutter, whereas wealthier regions are often able to mitigate such issues through underground infrastructure and stricter regulation.

Existing research has attempted to detect and quantify visual pollution; however, several limitations persist. These include reliance on localised datasets, manual or semi-automated quantification methods, and a narrow focus on a limited set of pollutant categories. Such constraints make large-scale, globally representative analysis challenging.

Recent work, such as the Visual Pollution Index proposed by Gupta et al. [5], has begun to address aspects of these challenges. Building on this, the present work aims to further advance the field by constructing a large-scale global dataset and expanding the range of pollutant classes considered.

1.2 Contributions

This project makes three primary contributions to the field of visual pollution detection using deep learning, with a particular focus on dataset construction and generalisation.

1. **Global dataset** – Prior work has largely focused on city- or country-level analysis, typically using imagery from a single geographic region. In contrast, this project constructs a large-scale global dataset using street-view imagery services, comprising data from 179 countries. This enables the study of visual pollution across diverse geographic and cultural contexts.
2. **Expanded class definitions** – Existing approaches often focus on a limited set of pollutant categories, such as billboards, wires, and litter. This project extends the scope to nine classes: billboards, utility poles, bins, mobile advertisements, road signs, shop signs, potholes, barriers, and graffiti. In particular, the introduction of shop signs and mobile advertisements refines the broader category of advertising into more granular and analytically meaningful subclasses.
3. **Globally trained detection model** – Many existing models are trained on localised datasets, and their ability to generalise across regions is therefore unclear. This project develops a detection model trained on globally sourced imagery, with the aim of improving robustness and generalisation performance across diverse environments.

1.3 Structure

This report is organised into nine main chapters:

1. **Introduction** – Provides an overview of the project, including its motivation and key contributions.
2. **Background** – Introduces relevant terminology and key concepts, and reviews prior work alongside its limitations.
3. **Visual Pollutants** – Describes the pollutant categories defined and used throughout the project.
4. **Dataset Construction** – Details the development of the global dataset, including data sources, the collection pipeline, and dataset statistics.
5. **System Architecture** – Explains the overall system design, including data storage, inference, and scoring methods.
6. **Model Training** – Describes the model architecture, training setup, and evaluation methodology.
7. **Experimental Results** – Presents the results of experiments, including regulatory impact analysis and geographic visualisations.
8. **Discussion** – Analyses the limitations of the project, interprets the results, and outlines directions for future work.
9. **Conclusion** – Summarises the key findings and overall contributions of the project.

Chapter 2

Background

This chapter examines the historical origins and evolving definitions of *visual pollution*, tracing the term’s evolution over the past century. It reviews previous work on defining and quantifying the phenomenon and discusses modern efforts to build datasets and apply computer vision techniques for its detection.

2.1 Defining Visual Pollution

The term *visual pollution* initially referred to the visible presence of polluting substances [6]. One of the earliest known appearances in print occurs in the *27th Annual Report of the New York State Department of Health* (1906, published 1907) [7], which discusses proposed sewage system improvements and notes that “by the introduction of fine screening practically all visual pollution could be eliminated from the sewage.” In more recent decades, however, the term has come to denote the degradation of a scene by unappealing or intrusive objects that diminish the visual quality or aesthetic value of an environment [2, 8, 1]. Among the most frequently studied sources of visual pollution are billboards [9, 1]. However, researchers have also examined the visual impacts of cell towers [8], overhead cables [10], and street litter [11].

The semantic evolution of the term reflects a broader shift in environmental discourse. Early uses described physically observable contaminants, whereas later applications draw on aesthetic and psychological perspectives. As urbanisation accelerated, planners and environmental psychologists began using the term to refer to visual intrusions, such as signage, cluttered skylines, and unregulated advertising, that disrupt visual harmony and reduce the perceived quality of urban spaces [8, 1]. This shift reflects growing recognition that environmental quality encompasses perceptual and psychological well-being, not only physical health.

To better understand how and why visual pollution feels “disruptive”, many researchers interpret the visual quality of a scene as tied to its degree of order or coherence. Weber’s work on architectural aesthetics emphasises that spatial regularity, symmetry, alignment, and visual continuity enhance aesthetic judgment [12]. Scenes exhibiting high compositional harmony are generally perceived as of higher visual quality. In

contrast, scenes lacking a governing structure or organising principle tend to appear chaotic or cluttered, yielding lower perceived quality [13]. From this perspective, visual pollution may be conceived as visual disorder, excessive contrast, incongruity, or unpredictability that undermines perceptual harmony.

Despite growing interest and numerous empirical studies, no single, universally accepted definition of visual pollution exists. Most researchers share a common underlying premise: that visual pollution arises from elements that disrupt aesthetic harmony or visual order, but each adjusts the definition to suit their context or analytical goals. The systematic review by Gao et al. [2] notes that many definitions centre on “unappealing or obtrusive elements that negatively impact visual quality”, often encompassing visibility reduction, emotional disturbance, and spatial incongruity. Meanwhile, other works emphasise the experiential dimension: “elements in a landscape that are visually uncomfortable or disharmonious, negatively affecting their perception and experience” [14].

This lack of standardisation complicates efforts to build consistent datasets and establish reliable evaluation criteria. To understand how visual pollution evolved from an abstract aesthetic concept into a tangible urban and environmental concern, the following section explores its historical and societal development. It traces key milestones such as the mid-twentieth-century regulation of billboards and São Paulo’s *Clean City Law*, illustrating how visual pollution became increasingly linked to sustainability, urban liveability, and overall quality of life.

2.2 Historical and Societal Context

By the mid-twentieth century, rapid urbanisation and economic growth had transformed the visual character of cities worldwide. In 1950, urban areas in industrialised nations already housed well over half of the population, for example, approximately 65% in the United States [15]. At the same time, the rise of mass consumer culture led to significant shifts in marketing and advertising. Companies increasingly adopted visually striking campaigns that used bright colours, large imagery, and persuasive slogans to attract attention and promote new products [16]. These developments not only altered the visual character of urban environments but also raised concerns about their impact on human wellbeing. Visual pollution has been linked to negative mental health outcomes, with cluttered and chaotic scenes contributing to stress and reduced comfort in urban spaces [2]. In addition, certain forms of visual clutter, such as billboards, have been associated with distracted driving and increased safety risks [17].

As a result, policymakers and planners began to recognise the broader societal impact of visual clutter on public spaces. Early regulations, such as the *Bonus Act of 1958*, offered financial incentives to U.S. states to control outdoor advertising within 660 feet (200 metres) of interstate highways [18]. Later legislation, most notably the *Highway Beautification Act of 1965*, required states to establish and enforce standards for billboard size, spacing, and lighting [19, 20]. Together, these acts reflect growing governmental awareness of the consequences of unchecked visual advertising.

Elsewhere, cities sought to control the urban spread of advertisements. In 2006, São

Paulo, Brazil, enacted the *Lei Cidade Limpa* (Clean City Law), which banned nearly all forms of outdoor advertising [3]. Mayor Gilberto Kassab justified the policy by stating that “the Clean City Law came from a need to combat pollution [...] pollution of water, sound, air, and the visual. We decided that we should start combating pollution with the most conspicuous sector – visual pollution” [21]. As a result, approximately 15,000 billboards were removed, revealing previously hidden architecture and encouraging residents to improve the image of São Paulo [21].

However, billboards and advertisements are not the only visual pollutants to consider. Overhanging telephone and utility wires may be less of an issue in developed countries, where many have been placed underground. Still, in developing countries such as India, Pakistan, and Bangladesh, they remain a significant source of visual clutter [11, 2]. In urban centres, utility poles and overhead wires are regularly listed among the top contributors to skyline disorder, obstructing views and creating chaotic visual environments [22]. In Indian case studies, utility wiring is often found alongside hoardings and unplanned signage, constituting a major visual pollutant [23]. Earlier work on overhead wires also regards them as a long-recognised source of aesthetic disturbance [24], and more recent environmental studies have quantified their negative visual impacts in both urban and rural landscapes [10].

Taken together, these developments illustrate how visual pollution has evolved from a primarily aesthetic concern into a recognised urban and environmental issue shaped by technological, economic, and policy forces. Understanding this historical trajectory highlights the need to evaluate visual pollution systematically. The following section, therefore, examines how researchers have attempted to *assess* visual pollution - initially through subjective surveys and expert evaluations, and more recently through spatial and computational approaches aimed at objective measurement.

2.3 Assessing Visual Pollution

Efforts to assess visual pollution face persistent challenges, mainly because visual perception is inherently subjective and shaped by individual experiences and cultural context [25, 1, 11]. Nevertheless, reliable assessment requires some degree of quantification to minimise subjectivity and enable consistent comparison across studies. To this end, researchers have employed a wide range of approaches: from human judgement and expert evaluation [25, 26, 9] to computer-based analysis and automated detection [27, 11], to evaluate and quantify visual pollution.

Traditional approaches often rely on surveys and perceptual studies. For example, Shang et al. used open-ended questionnaires to investigate visual thresholds, including detection, recognition, and perceived impact of objects, by systematically varying characteristics such as size, contrast, object type, and landscape context [28]. Bishop et al. conducted a closed-ended online questionnaire to assess the visual impact of wind turbines but noted that respondents’ worldviews, such as attitudes toward renewable energy, could influence their perceptions of visual effects [29]. Similarly, Ladenburg et al. examined how prior experience shaped perceptions of offshore wind farms, finding that coastal residents with direct views of turbines rated their visual impact

more negatively than participants living in urban areas [30]. These studies captured how individual experience and context influence visual evaluation. Yet, they did not provide a quantitative framework for assessing the visual impact of a specific object within a given scene.

More recent efforts have sought to reduce subjectivity by developing frameworks for systematically assessing visual pollution within a scene. Wakil et al. introduced a Visual Pollution Assessment (VPA) tool that utilises expert ranking and weighting of visual pollution objects to assist planners and developers in quantifying visual impact. Their study reported an average agreement of 83% among experts when evaluating scenes using this method [25]. Building on their work, Wakil et al. later proposed a Spatial Decision Support System (SDSS) designed to assist stakeholders, including development authorities, advertisers, billboard owners, and the public, in optimising billboard placement in accordance with regulatory constraints [26]. Gupta et al. [5] introduced the Visual Pollution Index (VPI) to computationally quantify visual pollution. It uses parameters such as the number of pollutants in a scene and predefined weights for each class to calculate a score.

With the advent of modern computer vision techniques, numerous studies have successfully trained models to detect forms of visual pollution [27, 31, 11, 32, 33, 34, 35, 36, 37]. Many of these efforts address narrow applications, such as pothole detection, by adapting data collection and augmentation methods to their specific context [31, 34, 35]. Others adopt a broader perspective, training models on a wider range of visual pollutants using either existing datasets, such as the Kingdom of Saudi Arabia's Ministry of Municipal and Rural Affairs and Housing (MOMRAH) dataset [38, 36], or custom datasets compiled from sources like Google Images and Google Maps [27, 11].

While these methods have achieved high success rates in detecting visual pollution, each presents distinct limitations. Manual approaches are constrained by the time and effort required to evaluate scenes and produce consistent scores. In contrast, computational methods often lack standardised quantification frameworks and rely on limited or region-specific datasets. As a result, models trained on such data may struggle to generalise across different environments. Developing large, diverse, and representative datasets is therefore essential to enable models that can accurately detect a broad range of visual pollutants across varied geographic and cultural contexts.

2.4 Computer Vision

While humans can easily recognise and classify instances of visual pollution, programming a system to do so explicitly is highly impractical due to the vast variability in appearance, scale, and context of visual pollutants. Computer vision algorithms address this challenge by learning from example data and generalising to new, unseen images, a process enabled by deep learning [39]. In this paradigm, models automatically learn hierarchical representations of visual features, making them well-suited for complex scene understanding tasks such as visual pollution detection.

Most approaches to visual pollution detection rely on supervised learning, where a model is trained on labelled images to predict the correct category for new inputs.

However, collecting and annotating large, diverse datasets is resource-intensive and often impractical. To mitigate this, researchers increasingly employ *Deep Active Learning*, which combines deep learning with active learning principles. Active learning allows a model to achieve high accuracy with fewer labelled samples by selectively querying the most informative data points for annotation [40]. In a Deep Active Learning framework, a small set of labelled images is first used to train a model, which then identifies additional unlabelled images with high-confidence predictions for further training [41]. This approach was demonstrated by AlElaiwi et al. (2022), who applied Deep Active Learning to the MOMRAH dataset of public road images in Saudi Arabia, achieving high precision in detecting visual pollutants [37].

Computer vision has therefore become a vital tool for identifying and classifying visual pollution, valued not only for its speed and scalability but also for its ability to generalise across diverse environments. These capabilities make it an essential component in developing large-scale, automated systems for visual pollution detection and analysis.

Common frameworks for object detection, which underpin many visual pollution detection systems, include *You Only Look Once* (YOLO), a real-time method that simultaneously evaluates bounding boxes and class probabilities in a single network pass [42]. Another widely used approach is *Faster R-CNN*, a two-stage detector that first proposes candidate regions likely to contain objects and then classifies the contents of each bounding box [43]. More recently, *EfficientDet* introduced a family of models designed to optimise both accuracy and computational efficiency, achieving 4–9× smaller model sizes and 13–42× fewer FLOPs than comparable detectors at the time of release [44]. These frameworks form the foundation for many contemporary systems capable of detecting a wide range of visual pollutants in complex, real-world environments.

In the context of visual pollution detection, AlElaiwi et al. reported that YOLO outperformed other architectures on the MOMRAH dataset [37], while Titu et al. similarly found that YOLOv5 achieved the highest detection accuracy in their study of urban visual clutter [27]. Conversely, Tasnim et al. observed that EfficientDet achieved superior performance in their experiments [32]. These varying results highlight that model performance is often application- and dataset-dependent, influenced by factors such as object scale, visual complexity, and environmental variability.

2.5 Previous Work

Research applying deep learning specifically to the detection of visual pollution remains limited. While several studies have explored this area, most focus on a narrow subset of visual pollutants or are restricted to region-specific contexts. This section reviews a selection of existing work, highlighting their methodologies, key findings, and limitations, and identifies the remaining research gaps that motivate the present study.

2.5.1 Ahmed et al. (2019)

Ahmed et al. (2019) [11] proposed a model for automated visual pollutant classification, identifying four categories of pollutants: billboards and signage, telephone and communication wires, network and communication towers, and street litter. The study was conducted in Bangladesh, which limits the model's generalisability beyond similar regional and socio-economic contexts.

Training data were collected from Google Images using search queries constructed from combinations such as "Bangladesh visual pollution + [category]". This approach provided the advantage of readily available, pre-labelled images but introduced significant drawbacks. There is no guarantee that the retrieved images accurately reflect the intended classes, as source-level mislabelling can introduce noise into the dataset. Consequently, manual verification of all samples remained necessary. In addition, the number and quality of images for each pollutant type were limited. Only 800 images were collected, 200 per category, which is a relatively small dataset. While sufficient for a proof of concept, this sample size would likely be inadequate for generalisation to other regions. Moreover, the dataset did not explicitly account for variations in lighting, weather, or viewing conditions, which could reduce model robustness in real-world scenarios.

The authors did not employ a standardised deep learning framework but instead implemented a custom convolutional neural network (CNN) architecture. Data augmentation was applied to expand the dataset. The model consisted of five convolutional and max-pooling layers, each with ReLU activation functions, except for the final layer, which utilised a softmax classifier.

The resulting model achieved a validation accuracy of 85%, which is notably high given the dataset's limited size. Overall, the model performed well in detecting the four selected categories. However, the authors acknowledged that the inherent subjectivity of the concept constrains the development of a universal visual pollutant classifier. This is because the boundaries between classes are not clearly defined, and categories can be subdivided almost indefinitely (e.g., distinguishing between commercial and political billboards).

2.5.2 Gupta et al. (2024)

Gupta et al. (2024) [5] proposed a Visual Pollution Index (VPI) to quantify visual pollution within a scene, alongside a YOLOv6-based object detection model. Their dataset was collected using drone imagery captured in Kolkata, India, under controlled conditions with favourable weather and lighting. While this approach ensures high-quality data, it may limit generalisation to real-world scenarios with more variable lighting and weather conditions.

The dataset comprises 11 visual pollution categories, with approximately 600–1250 images per class. This provides relatively balanced class representation and sufficient data for effective model training.

The YOLOv6 model achieved Intersection-over-Union (IoU) scores ranging from 0.71

to 0.92, with an average of 0.81. These results indicate strong localisation performance, suggesting that the dataset and training approach were effective for the defined task.

The proposed VPI consists of two primary components: the Counting Categories Ratio (CCR) and the Severity-Weighted Score (SWS). The CCR measures the proportion of pollutant categories present in a scene, with values ranging from 0 (no pollutants detected) to 1 (all categories present). This captures the intuition that multiple distinct pollutants may collectively degrade visual quality more than a single pollutant.

The SWS incorporates severity by assigning each category a score between 0 and 1. It is defined as:

$$SWS = \frac{\sum(\text{severity value} \times \text{number of instances per category})}{\text{total instances across all categories}}$$

This formulation enables the scoring system to account for both the quantity and perceived severity of pollutants, and can be adapted depending on the application.

The final VPI is computed as the product of CCR and SWS. This combined metric balances category diversity and severity, providing a more nuanced assessment than approaches based solely on object counts. However, the framework does not explicitly account for spatial extent or pixel-level coverage of pollutants. Subsequent work has addressed this limitation by introducing a pollution ratio (PR), which measures the proportion of image area occupied by pollutants in segmentation-based approaches [45].

While PR is not directly applicable to object detection, the VPI formulation remains valuable, particularly when severity weights are adapted to the specific task. It can provide meaningful insight into relative pollution levels within a region. However, when applied at larger geographic scales, such as comparing cities globally, additional considerations are required. In particular, variations in data availability may bias the index, as regions with fewer images could receive disproportionately high scores if those images contain a high density of pollutants. Therefore, further normalisation or sampling strategies would be necessary to ensure fair comparison across regions.

2.5.3 Rahman et al. (2026)

Rahman et al. (2026) [46] present a review of deep learning approaches for urban visual pollution and propose a real-time monitoring framework based on a visual pollution index (VPI). The study aims to identify commonly used models and datasets, assess their potential to inform the design of a real-time system, and highlight key implementation challenges.

The authors note that deep learning significantly improves the automation and scalability of environmental monitoring systems, with YOLO-based models being among the most frequently adopted approaches. However, a major limitation identified is the lack of a generalised benchmark dataset, as most existing datasets are small, region-specific, and focused on a limited set of pollutant categories.

To address these challenges, a multi-layered conceptual framework is proposed. The first layer focuses on data collection and recommends using diverse sources, including

CCTV, drone imagery, video streams, street-view imagery, and smartphone captures. Each source provides complementary benefits: street-view imagery enables large-scale coverage, CCTV offers temporal monitoring, drone imagery provides alternative viewpoints, and mobile captures allow for crowdsourced data collection.

For detection, the framework employs a YOLOv8 object detector with COCO-style annotations, chosen for its real-time inference capability and suitability for deployment on edge devices such as Raspberry Pi. This choice aligns with prior work in the domain, in which YOLO-based models are commonly used for efficient visual pollution detection [27].

The framework further incorporates segmentation to obtain pixel-level representations of pollutants, enabling fine-grained analysis such as estimating the proportion of an image occupied by different objects. A classification stage is also included to provide higher-level semantic interpretation, enabling distinctions between visually similar yet contextually distinct objects (e.g., legal versus illegal posters) and supporting actionable insights.

A scoring mechanism is proposed based on tunable parameters, including pollutant density, area ratio, and recurrence frequency. While this provides a flexible approach, it does not explicitly account for differences in the severity of various pollutant types. In contrast, Gupta et al. [5] propose a weighted scoring system that assigns different importance to each category. Incorporating such weighting could improve the accuracy and interpretability of the final visual pollution score.

Overall, the work presents a comprehensive and practical framework for real-time visual pollution monitoring. It addresses several limitations of previous studies, particularly in terms of scalability and data diversity, although aspects such as dataset standardisation and pollutant weighting remain open challenges.

Chapter 3

Visual Pollutants

A total of nine visual pollutant categories were defined for this project. These categories include both commonly studied forms of visual pollution and additional sub-categories designed to better capture variations observed in street-level imagery. Examples of the different categories are provided in Appendix D.

3.1 Billboards

Billboards are a well-established form of visual pollution and are among the most widely studied categories in the literature [26, 11].

In this project, a distinction was made between static billboards, such as Figure D.1, and advertisements displayed on vehicles, such as Figure D.5. This distinction is important because vehicle-based advertisements introduce a temporal component when using street-view imagery, and therefore do not reliably represent the persistent visual environment of a location.

Road signs were treated as a separate category to maintain a clear functional distinction, as they serve essential roles within urban infrastructure. Similarly, shop signs were separated from billboards. In many countries, including the United Kingdom, businesses are required to display signage [47]. Although shop signs may contribute to visual clutter, their functional and legal necessity justifies their treatment as a distinct category.

However, this distinction becomes less clear in regions where such regulations are weaker or inconsistently enforced, and shop signage may overlap with billboard-style advertising. For this reason, shop signs were retained as a separate class rather than excluded entirely.

Finally, the definition of a billboard was intentionally kept broad to capture a wider range of advertising formats, including posters and banners. While this improves coverage, it also introduces some ambiguity, which is discussed further in Section 8.1.

3.2 Bins

Although bins are necessary for waste management, they can contribute to visual pollution when poorly maintained, for example, when overfilled or placed carelessly. Previous work has focused specifically on the impact of overflowing bins [23].

However, because street-view imagery is temporally constrained, it is redundant to assess whether a bin is overflowing at the time of capture. As a result, this project considers the presence of bins in general, rather than attempting to detect their condition.

Detecting bins in this way enables analysis of the spatial distribution and density of waste infrastructure, which may still provide useful insights into differences between urban environments, including comparisons between developed and developing regions.

3.3 Utility Poles

Utility poles are treated as an umbrella category, encompassing structures such as telephone poles, electricity poles, and pylons.

These structures have been considered in previous studies of visual pollution [24, 11, 27], while other work has examined the visual impact of overhead wiring [25]. In this project, the focus is placed on utility poles rather than wires, as poles are more consistently visible and easier to detect in street-view imagery.

To ensure consistency, only poles with visible wires were classified as utility poles. As a result, the presence of poles can be interpreted as a proxy for overhead wiring, although wires themselves were not explicitly analysed. Street lamps and lighting infrastructure were excluded from this category, as they are not typically considered visual pollutants in the same way as utility infrastructure supporting overhead wires.

3.4 Graffiti

Graffiti has been widely identified as a form of visual pollution in existing studies [48, 49]. It is included in this project due to its association with perceived disorder and reduced quality of life in urban environments.

By detecting graffiti in street-view imagery, this project enables analysis of its spatial distribution, including potential relationships with socioeconomic factors such as levels of urban deprivation.

3.5 Mobile Advertisement

Mobile advertisement, defined as spatially non-permanent advertising (e.g. on buses or other vehicles), has not been extensively studied as a form of visual pollution. However, when using street-view imagery, it is important to distinguish mobile advertisements from static billboards. This distinction is necessary because mobile advertisements are transient and do not represent the persistent visual environment of a location. As a

result, they were treated as a separate category in this project, enabling their exclusion from visual pollution scoring and analysis.

3.6 Shop signs

Shop sign detection has been explored in prior work outside the context of visual pollution [50, 51]. In this project, shop signs were included as a separate category primarily to maintain a clear distinction from billboards.

As previously discussed, shop signage is legally required in many countries. Although some shop signs may contribute to visual clutter, their functional and regulatory role justifies treating them separately from other forms of advertising.

3.7 Road signs

Road sign detection is a well-established area of research, largely due to its importance in automated driving systems. The Vienna Convention on Road Signs and Signals [52] established international standards for the design and appearance of road signs. This standardisation enables models trained on relatively limited datasets to generalise effectively across multiple countries that are signatories to the convention.

Road signs are not considered visual pollutants in this project, as they serve essential functional roles in navigation and safety within urban environments. However, because of their visual similarity to billboards in size, shape, and placement, they are included in a separate category. This reduces misclassification and helps prevent false positives in billboard detection.

3.8 Barriers

Barriers, such as concrete roadside barriers, have been considered in previous studies of visual pollution [37] and are prominently featured in datasets such as MOMRAH [38].

In addition to permanent barriers, temporary structures, such as construction barriers and traffic cones, were included. Although these objects are not permanent, they can remain in place for extended periods and therefore contribute to the visual environment of a location. However, because of their temporary nature, these objects were assigned a lower weight in the overall visual pollution scoring.

3.9 Potholes

Potholes are a widely studied form of visual pollution [36, 35, 31] and are often associated with deteriorating road conditions and poor infrastructure.

Detecting potholes in this project enables analysis of their spatial distribution, including potential relationships with other visual pollutants, such as graffiti, and correlations with socioeconomic indicators, such as urban deprivation.

Chapter 4

Dataset Construction

For this project, a large dataset was required in order to train and evaluate models effectively. Previous work in this area has often suffered from limited geographic generalisation because datasets are concentrated in a small number of regions. To address this limitation, it was necessary to collect imagery from a wide range of locations worldwide.

Street-view imagery provides an effective source of such data. These images are typically geo-tagged, providing precise geographic coordinates, and they capture street-level scenes where many forms of visual pollution occur. Additionally, several platforms provide programmatic access to large volumes of street-view imagery through public APIs, enabling large-scale automated data collection.

4.1 Data sources

Three primary data sources were used: Mapillary, KartaView, and OpenStreetMap (OSM). Mapillary and KartaView provide street-level imagery captured by contributors worldwide, whereas OpenStreetMap provides geographic information on infrastructure and objects in the environment.

Using multiple imagery sources helped improve geographic coverage. In areas where one platform had limited imagery, the other could often provide additional images. OpenStreetMap data was used to obtain the locations of known objects, which could provide useful contextual information for regions where few images were available.

4.1.1 Mapillary

Mapillary is a crowdsourced street-view imagery platform operated by Meta. Users upload geo-tagged images captured using smartphones, dash cameras, or other recording devices. These images are then made publicly available through the Mapillary API.

Data was collected using the <https://graph.mapillary.com> API endpoint, and for each request, several metadata items were retrieved, including the image identifier, geographic coordinates, image URL, and capture timestamp.

To reduce dataset redundancy, at most two images were retrieved for each queried location. In many cases, requesting a larger number of images returned long sequences of near-identical frames captured by the same camera while travelling along a road. Limiting the number of near-duplicate images per location increased the dataset's scene diversity.

4.1.2 KartaView

KartaView, formerly known as OpenStreetCam, is another crowdsourced street-view imagery platform operated by Grab Holdings. The platform provides street-level imagery uploaded by contributors worldwide. KartaView has particularly strong coverage in parts of Southeast Asia due to Grab Holdings' wider operations in the region. As with Mapillary, contributors upload geo-tagged images which can be accessed programmatically through the KartaView API.

Data was collected using the <https://api.openstreetcam.org/2.0> endpoint. Similar metadata to that retrieved from Mapillary was collected from each response, including image identifiers, geographic coordinates, capture timestamps, and image URLs.

Unlike Mapillary, only a single image was requested for each query location. Preliminary experiments showed that requests for multiple images often returned long sequences of near-identical frames captured along the same road segment. Restricting requests to a single image per location reduced redundancy and increased the dataset's scene diversity.

4.1.3 OpenStreetMap

OpenStreetMap (OSM) is an open, collaboratively maintained geospatial database operated by the OpenStreetMap Foundation. It provides structured geographic data describing roads, infrastructure, and points of interest contributed by volunteers worldwide.

Data was retrieved using the `overpass` Python library, which provides bindings for the Overpass API and enables programmatic querying of OSM features. Because OSM contains a very large number of feature types, queries were restricted to those relevant to visual pollution analysis. The selected features included bins, billboards, power infrastructure, barriers, and traffic signs.

This data complemented the street-view imagery sources used in the dataset. In areas with limited or no street-view imagery coverage, OSM features provided additional geographic information, thereby improving spatial coverage and ensuring that relevant infrastructure elements were still represented in the dataset.

4.1.4 Limitations

Although combining three data sources generally allowed imagery to be obtained for most regions, geographic coverage was uneven. Coverage was substantially greater in high-income regions, where camera ownership may be higher and more affordable,

while areas such as parts of Africa, India, and China had significantly fewer available images. This imbalance reflects differences in street-view coverage and volunteer data contributions, thereby limiting the dataset's ability to represent all geographic regions equally.

Issues were also encountered with the reliability of some image links returned by the KartaView API. Although the API provided URLs for image access, a proportion of these links did not resolve to valid images. Consequently, an additional validation step was required to verify that each returned link corresponded to an accessible image before it was included in the dataset.

Another limitation concerns image quality. There was no scalable mechanism to automatically verify that retrieved images were suitable for analysis. In some cases, images were blurred, poorly framed, or primarily comprised close-up views of surfaces such as walls, which are not useful for object detection tasks. However, these cases represented only a small proportion of the dataset and were therefore retained.

Finally, street-view imagery may be captured at different times and dates. As a result, certain features such as billboards or bins may no longer be present or may have changed location since the image was recorded. Although capture timestamps were stored where available, temporal changes in infrastructure were not explicitly analysed in this work.

4.2 Data Collection Pipeline

To support large-scale and reliable data collection, an automated pipeline was developed to minimise manual intervention and ensure consistent processing across many geographic regions.

4.2.1 Region creation

The pipeline begins by receiving either a single city name or a file containing a list of city names. Once the system has initialised its components, it uses the `osmnx` library to retrieve a shapefile describing the city's administrative boundary from OpenStreetMap.

In some cases, a boundary could not be retrieved, or the resulting geometry was excessively large for practical processing. In such situations, the system instead generated a fallback region defined by a bounding box of predefined size centred on the city's geographic coordinates.

Each region is assigned a unique identifier and a name derived from its bounding box coordinates. Metadata stored for each region includes the region identifier, name, bounding box, creation timestamp, associated city and country, the city's population (when available), and a boolean flag indicating whether OpenStreetMap feature data was collected for that region.

Dividing cities into manageable regions was important to ensure consistent geographic coverage during data collection and analysis. Associating regions with city names also simplified later analysis and interpretation of results. Additionally, the use of fallback

bounding boxes ensured that smaller cities, which may not always return a shapefile from OpenStreetMap, could still be included in the dataset.

4.2.2 API Querying

To improve data-collection speed, all three APIs were queried in parallel using the `ThreadPoolExecutor` class from Python's `concurrent.futures` module. For each API, between 20 and 40 worker threads were used to issue multiple requests concurrently.

Each region was uniformly divided into a number of smaller subregions before querying the APIs. Mapillary regions were divided into 10,000 subregions, KartaView regions into 1,000 subregions, and OpenStreetMap regions into 60 subregions. Each subregion was queried individually, enabling the system to achieve dense, evenly distributed coverage across the region. Subregions were implemented as smaller geographic bounding boxes derived from the original region bounding box. Dense scanning was also used to analyse specific cities, thereby multiplying the number of subregions by 10.

Experiments were conducted to evaluate concurrent querying of multiple regions using threading; however, this approach reduced the number of successful responses returned by the APIs. As a result, regions were processed sequentially while requests within each region were issued in parallel.

4.2.3 Data cleaning

Two forms of data cleaning were performed during data collection.

The first ensured that images remained within the boundaries of the region shapefile, if available. During collection, images were retrieved using the bounding box that encloses the shapefile, so some returned images may lie outside the true administrative boundary. A spatial filtering step was therefore applied to remove any images whose coordinates fell outside the shapefile.

The second validation step ensured that the returned image URL actually resolved to a valid image. Some URLs returned by the API did not resolve to valid images, likely due to removed or inaccessible content. This was achieved by issuing a `GET` request to each image URL and verifying the returned content. During this request, the image width and height were also recorded, enabling their use later in the pipeline's annotation stage.

4.3 Geographic Sampling Strategy

To achieve broad geographic coverage while keeping the dataset manageable, a global city dataset provided by Simplemaps [53] was used. This dataset contains approximately 48,000 entries, including city–country pairs, geographic coordinates (latitude and longitude), and population estimates.

To ensure that the data collection process remained computationally feasible, the dataset was filtered to include only cities with populations exceeding 100,000. This reduced the dataset to 6,135 cities, which were used as the starting points for the data collection pipeline. A population threshold of 100,000 was chosen to balance large urban centres with smaller yet significant cities, while maintaining a high likelihood of sufficient imagery availability. After processing, a total of 5,414 regions were successfully collected.

A limitation of this approach is that the resulting dataset remained unevenly distributed across countries. After filtering, China, India, and the United States together accounted for 1,619 of the selected cities, representing approximately 30% of the total. However, many of these cities, particularly in China and India, yielded very few images from the available street-view sources and therefore contributed little to the final dataset after subsequent filtering and validation.

Figure 4.1 illustrates the initial distribution of selected cities across countries. However, when applying a minimum threshold of 300 images per city (Figure 4.2), the number of usable cities decreases substantially. This reduction is particularly pronounced in countries such as China. Although 681 cities were initially selected, only 15 met the minimum image threshold.

These results highlight a significant imbalance in global street-level imagery availability, which introduces geographic bias into the dataset. Regions with limited image coverage are more likely to be excluded under strict filtering criteria, reducing their representation in subsequent analysis.

To mitigate this issue, OpenStreetMap (OSM) features were incorporated as an additional data source. When considering cities with either at least 300 images or available OSM features (Figure B.1), the number of usable cities in China increases to 357. This substantially improves geographic coverage and enables analysis in regions where image-based observations alone are insufficient.

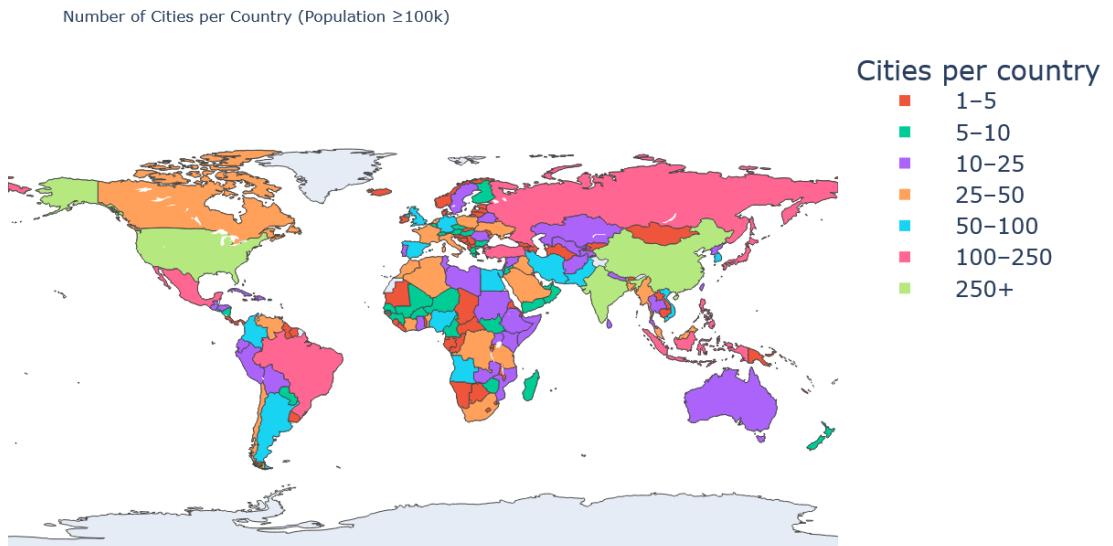


Figure 4.1: Number of selected cities per country with population above 100,000.

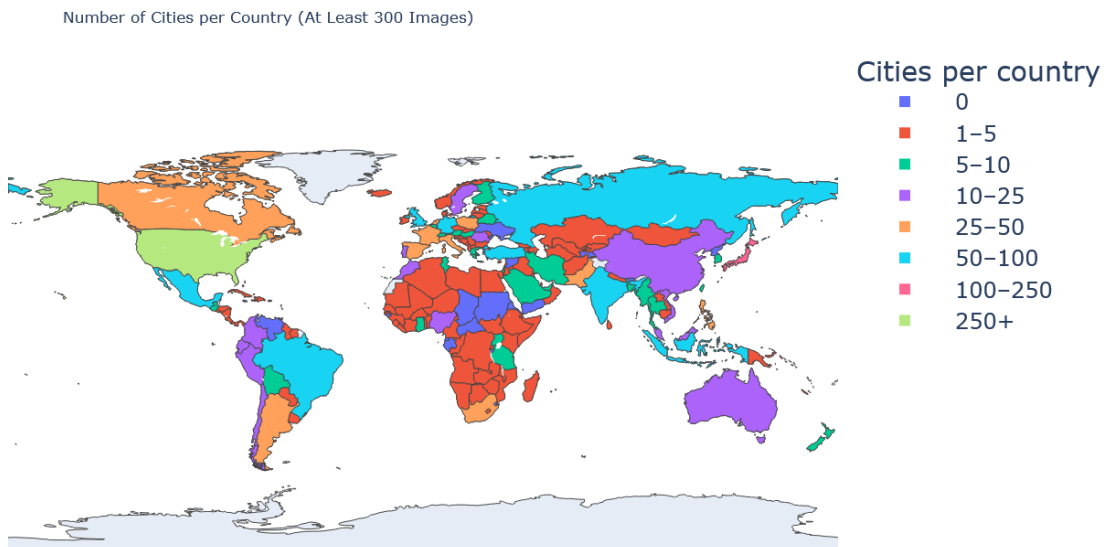


Figure 4.2: Number of cities per country with at least 300 available images.

4.4 Annotation Process

Images were annotated using Label Studio [54], which was used to manage and organise annotation tasks. For images with existing detections, predicted bounding boxes from the model described in Section 6.4 were converted into Label Studio format and attached

to the corresponding images. This facilitated efficient annotation by allowing predictions to be reviewed and corrected, rather than created from scratch.

To ensure geographic diversity, images were sampled at the country level. For each country in the dataset, 10 images were randomly selected for manual annotation. This approach provided broad global coverage while maintaining a manageable annotation workload.

In total, approximately 1,500 images were annotated. A subset of images was excluded due to poor quality or the absence of relevant objects.

For the final dataset, model detections were exported to a `.ndjson` file. This format enables efficient reuse of detections for large-scale annotation and supports the development of more generalised models in future work.

4.5 Dataset Statistics

The following dataset statistics do not include regions scanned with dense scanning, as these were handpicked cities for analysis rather than the general worldwide scan. The cities analysed are listed in Chapter 7.

4.5.1 Regions

In total, 5,414 regions across 179 countries were scanned. Table 4.1 shows the countries with the highest number of scanned regions. As expected, these correspond closely to countries with large populations, since the sampling strategy selected cities with populations exceeding 100,000.

This indicates that the dataset is biased towards highly urbanised and densely populated countries. While this improves data availability and coverage in major urban areas, it may under-represent less populated regions and rural environments.

A map view of the regions can be found later in Section 8.1.

Country	Regions
China	681
India	527
United States	411
Brazil	238
Japan	201
Philippines	165
Turkey	164
Russia	161
Mexico	141
Indonesia	132

Table 4.1: Top 10 countries by regions scanned.

4.5.2 Images

In total, 2,199,851 images were collected. Although most countries had at least some regions that returned images, the distribution was highly uneven. The majority of images originated from Western countries, with the United States alone contributing over 350,000 images.

Table 4.2 illustrates the top 10 countries by number of images. This imbalance reflects differences in street-view coverage and data availability across regions, rather than the underlying distribution of visual pollution.

	Images
United States	350,774
Japan	167,460
Germany	117,998
Brazil	98,858
Mexico	84,949
Russia	79,824
United Kingdom	68,134
Indonesia	66,152
Canada	60,123
India	58,572

Table 4.2: Top 10 countries by images collected.

Figure 4.3 shows the distribution of image capture dates across the dataset, grouped by source. Both Mapillary and KartaView exhibit a notable increase in activity from around 2015 onwards, reflecting the growing adoption of crowd-sourced street-level imagery platforms. Mapillary consistently contributes a higher volume of images throughout this period.

A small number of images are recorded with capture dates prior to 2010. These are likely attributable to incorrect device timestamps rather than representing true capture times, and therefore should be interpreted with caution.

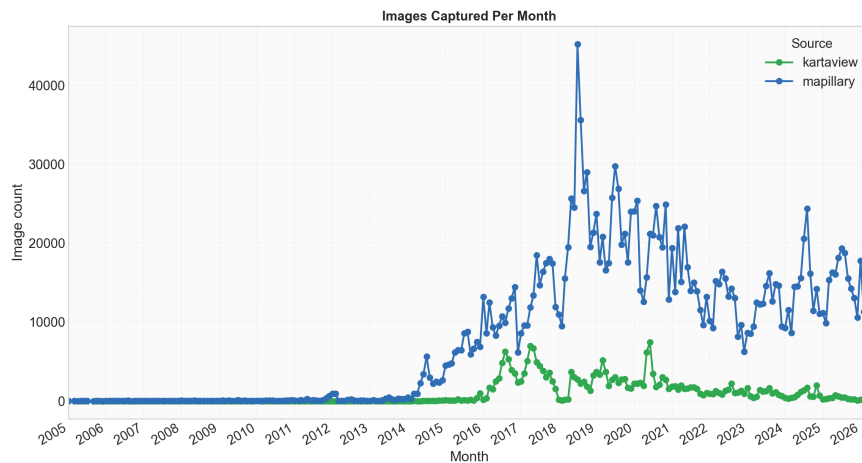


Figure 4.3: Distribution of image capture dates by source.

4.5.3 OSM Features

In total, 11,637,746 OpenStreetMap (OSM) features were collected. The majority of these were categorised under the `power` tag, which includes electrical infrastructure such as utility poles and related components.

A large number of `traffic_light` features were also retrieved. This is partly due to the query design, as the `traffic_signals` tag encompasses not only traffic lights but also other roadside infrastructure, including signage and related elements.

Over one million bin-related features were identified, providing useful complementary data to the image-based detections. This supports cross-validation between visual observations and map-based infrastructure data.

Overall, OSM features provide additional context for analysing visual pollution, particularly in regions with limited or no image coverage.

Feature	Count
power	7,576,862
traffic_light	2,470,268
bin	1,093,306
other	397,588
billboard	99,204
street_light	458
traffic_sign	30
barrier	30

Table 4.3: Top OSM feature types by count.

4.5.4 Detections

In total, 2,347,126 detections were produced. Of these, approximately 48% (1,136,271) correspond to the `utility_pole` class, followed by `road_sign`. This distribution is expected, as both categories represent essential infrastructure that is prevalent in most urban environments. The next most frequent class is `billboard`, with 446,580 detections. The remaining classes appear significantly less frequently, indicating a clear class imbalance within the dataset.

Feature	Count
power	7,576,862
traffic_light	2,470,268
bin	1,093,306
other	397,588
billboard	99,204
street_light	458
traffic_sign	30
barrier	30

Table 4.4: Top OSM feature types by count.

Chapter 5

System Architecture

5.1 System Overview

The basic flow of the system is shown in Figure 5.1. It shows the flow of data from the APIs through the collection pipeline and the process the data goes through. A more detailed description of the codebase’s features can be found in Appendix F.

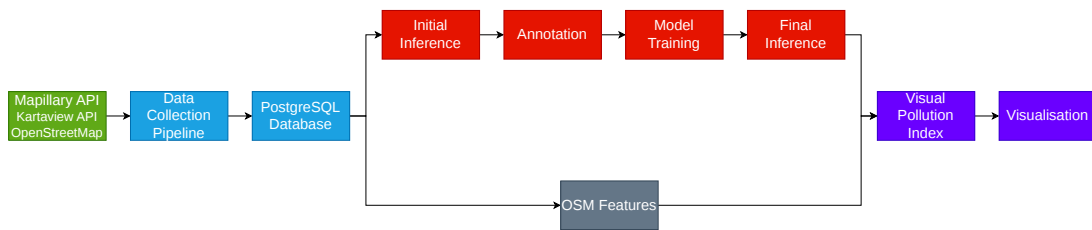


Figure 5.1: System diagram

5.2 Data storage

All collected data was stored in a PostgreSQL database. PostgreSQL was chosen because it supports concurrent access reliably, allowing multiple components of the pipeline to read from and write to the database during data collection, cleaning, and annotation. This was particularly important because the system used parallel API querying and multiple processing stages operating over the same dataset.

The database schema consisted of four main tables: `regions`, `images`, `detections`, and `osm_features`. Together, these tables represented the geographic areas being scanned, the street-view imagery collected from those areas, the object detections associated with each image, and the OpenStreetMap features retrieved for each region.

The relationships between these tables were straightforward. Each region may contain multiple images and OpenStreetMap features, while each image may contain multiple detections. This structure allowed the pipeline to efficiently link imagery, geographic context, and model predictions during later analysis. Table 5.1 summarises the main database tables and their roles within the pipeline.

Table	Primary key	Purpose / key fields
regions	id	Stores each geographic region, including bounding box coordinates, city, country, population, creation timestamp, and whether OSM data has been fetched.
images	id	Stores metadata for collected street-view images, including source image ID, coordinates, capture timestamp, source platform, URL, review status, width, and height. Each image belongs to one region.
detections	id	Stores object detections for each image, including predicted label, confidence, bounding box, detection timestamp, and manual review flag. Each detection belongs to one image.
osm_features	id	Stores OpenStreetMap features associated with a region, including OSM ID, feature type, coordinates, name, and creation timestamp. Each feature belongs to one region.

Table 5.1: Summary of the PostgreSQL database schema used in the data collection pipeline.

5.3 Detection Pipeline

To detect visual pollution objects (VPOs) within the collected images, inference was performed immediately after image collection. Parallel processing was implemented using Python’s `ThreadPoolExecutor`, with 20 worker threads processing images in batches of 32.

Images were temporarily downloaded to a local directory for inference and deleted afterwards. This approach enabled processing large numbers of images without requiring permanent local storage.

Once inference was complete, the YOLO model’s detection results were extracted and converted to a format compatible with the database schema. For each detection, the stored fields included `image_id`, `label`, `confidence`, and `bbox`. The `image_id` field links each detection to its corresponding image, allowing multiple detections to be associated with a single image.

5.4 Pollution Scoring

Scoring of pollution was performed in two distinct ways: (1) scoring individual images, and (2) aggregating scores at the region level. Image-level scoring was used to analyse local variation within a region, whereas region-level scoring was used to compare regions.

The scoring method is based on the Visual Pollution Index (VPI) proposed by Gupta et al. [5], with severity weights derived from Wakil et al. [25].

Pollutant	Severity
Billboard	0.9
Graffiti	0.4
Utility Pole	0.3
Bin	0.2
Pothole	0.1
Barrier	0.1
Mobile Advertisement	0.1
Shop Sign	0.1
Road Sign	0.0

Table 5.2: Severity weights used for scoring. Bold values derived from Wakil et al.

Additional pollutant weights were assigned based on their perceived visual impact relative to other categories. Bins were assigned a slightly lower score than utility poles, as they are typically smaller and less visually dominant within a scene. Road signs were assigned a score of zero due to their functional necessity in urban environments. Shop signs and mobile advertisements were assigned low, non-zero values to account for both their occasional visual impact and minor model confusion with billboards (see Figure C.2). Barriers were also assigned a low score, as they are generally temporary and therefore less representative of persistent visual pollution within a region.

The VPI is computed as:

$$\begin{aligned}
 \text{CCR} &= \frac{\text{Number of Distinct Categories Detected}}{\text{Total Number of Categories}} \\
 \text{SWS} &= \frac{\sum (\text{Severity Value} \times \text{Number of Instances for Each Category})}{\text{Total Number of Instances}} \\
 \text{VPI} &= \text{CCR} \times \text{SWS}
 \end{aligned} \tag{5.1}$$

Directly applying this formulation at the region level introduces challenges. Regions in the dataset contain varying numbers of images, leading to inconsistent sample sizes. As a result, regions with fewer images may yield unstable or unrepresentative scores, whereas regions with more images provide more reliable estimates.

To mitigate this, only regions containing more than 300 images were included in region-level comparisons. This threshold was empirically determined because regions with at least 300 images exhibited more stable, consistent scoring behaviour, thereby reducing the influence of sampling noise.

In addition to image-based detections, OpenStreetMap (OSM) features were incorporated into the scoring process. These features were assigned severity weights according to Table 5.2, allowing structured spatial data to complement image-based observations and improve overall coverage, particularly in regions with limited imagery.

Chapter 6

Model Training

6.1 Model Architecture

A YOLO26 [55] model from the Ultralytics YOLO family was selected for object detection. YOLO-based architectures are widely used for visual pollution detection due to their strong performance in real-time object detection and a favourable balance between accuracy and computational efficiency [56]. Previous work has shown that YOLO frequently outperforming other common architectures, often with measurably greater accuracy [45, 37, 27].

YOLO26, released in 2026, introduces architectural and optimisation improvements over earlier YOLO variants, with a particular emphasis on efficiency and small-object performance. This is especially relevant for this project, as many visual pollution objects, such as utility poles and distant signage, occupy only a small portion of the image and are therefore challenging to detect.

The medium-scale variant, `yolo26m`, was selected as it provides a practical trade-off between detection accuracy (mAP50–95) and inference speed. Larger variants offer improved accuracy but incur significantly higher computational cost, while smaller variants reduce latency at the expense of detection performance. Given the dataset’s scale and the requirement to process millions of images, `yolo26m` was deemed an appropriate compromise.

6.2 Training Setup

Training was performed on a Google Compute Engine virtual machine of type `g2-standard-4`, which provides 4 vCPUs and 16 GB of system memory. The system was equipped with an NVIDIA L4 GPU with 24 GB of GDDR6 memory, enabling significantly faster training compared to CPU-based training.

The dataset was divided into training, validation, and test sets using splits of 70%, 20%, and 10%, respectively. Care was taken to ensure that the distribution of classes was maintained across each split, so that all classes were adequately represented in

the training, validation, and test sets. This approach ensured that sufficient data was available for model training while retaining samples for validation and final evaluation.

6.2.1 Hyperparameters

Training was performed using the hyperparameters shown in Table 6.1. These parameters control optimisation behaviour, data loading, and augmentation strategies during training.

Parameter	Value	Description
Epochs	220	Total training epochs
Image size	1024	Training input resolution
Batch size	16	Images per batch
Initial LR (lr_0)	0.003	Initial learning rate
Final LR factor (lrf)	0.01	Final LR scaling
Warmup epochs	5	LR warmup duration
Mosaic	0.5	Mosaic augmentation probability
Close mosaic	15	Disabled in final epochs
Patience	50	Early stopping threshold
Workers	8	Data loader workers

Table 6.1: Training hyperparameters used for the YOLO model.

Training was conducted for 220 epochs to ensure sufficient model convergence. Performance metrics, including precision, recall, and mAP50, stabilised between epochs 180 and 190, indicating that the model had largely converged while avoiding premature termination.

An input image resolution of 1024×1024 was selected to preserve fine-grained details present in street-view imagery. This is particularly important for small or thin objects, such as utility poles and distant road signs, which may be difficult to detect at lower resolutions.

A batch size of 16 was chosen to balance computational efficiency and GPU memory constraints. Larger batch sizes would improve training stability but were not feasible given the available hardware, whereas smaller batch sizes would yield noisier gradient updates.

The learning rate schedule was configured with an initial learning rate of 0.003 and a final scaling factor of 0.01, allowing for rapid initial learning followed by gradual convergence. A warmup period of 5 epochs was used to stabilise early training and prevent large gradient updates.

Mosaic augmentation was applied with a probability of 0.5 (see Section 6.2.2), and was disabled during the final 15 epochs to allow the model to fine-tune on more natural image distributions. Early stopping with a patience of 50 epochs was used to prevent overfitting while ensuring that training was not terminated prematurely.

Finally, 8 data loader workers were used to maximise data throughput and minimise input bottlenecks during training.

6.2.2 Data Augmentation

Data augmentation was applied during training using the `Albumentations` library. Data augmentation improves model generalisation by introducing controlled variability in lighting, image quality, and object geometry, thereby reducing overfitting to specific capture conditions. The augmentations used in the training pipeline and their associated parameters are shown in Table 6.2.

Augmentation	Parameters	Probability
Random brightness / contrast	brightness ± 0.2 , contrast ± 0.2	0.30
Hue / saturation / value shift	hue ± 6 , saturation ± 15 , value ± 12	0.20
JPEG compression	quality 55–95	0.15
Motion blur	kernel size 3–5	0.10
Gaussian noise	std 0.02–0.08	0.10
Random safe crop	size 1024×1024	0.30
Affine transform	scale 0.9–1.15, rotation $\pm 8^\circ$, shear $\pm 4^\circ$	0.35
Mosaic	4-image composition	0.50
Horizontal flip	—	0.50

Table 6.2: Image augmentations applied during training.

The selected augmentations were designed to reflect the variability present in street-view imagery. Photometric transformations, including brightness/contrast and hue/saturation/value adjustments, simulate changes in lighting conditions caused by time of day, weather, and camera exposure settings. This improves robustness to illumination differences across geographic regions.

JPEG compression and Gaussian noise model variations in image quality arising from different sensors, compression pipelines, and API sources. These augmentations reduce sensitivity to artefacts such as compression loss and sensor noise, which are common in large-scale, crowdsourced imagery.

Motion blur reflects the capture conditions of street-view platforms, where images are often acquired from moving vehicles. Incorporating blur helps the model maintain performance when object edges are less well defined.

Geometric transformations, including random cropping and affine transformations, improve spatial robustness. Cropping encourages the model to recognise partially visible objects and reduces reliance on full object context, while affine transformations (scaling, translation, rotation, and shear) promote invariance to object position, orientation, and scale within the image.

Mosaic augmentation was applied during training; mosaic combines four images into a single composite image, increasing scene diversity and introducing significant variation in object scale and visibility. This is particularly beneficial for detecting small or partially occluded objects, as it exposes the model to a wider range of contexts within a single training example.

However, mosaic images are less representative of natural scenes, as they introduce artificial boundaries between image regions. To address this, mosaic augmentation was

disabled during the final 15 epochs of training, allowing the model to fine-tune on more realistic image distributions.

Finally, horizontal flipping introduces additional viewpoint variation, particularly for objects with no inherent directional bias, further improving generalisation.

6.3 Evaluation Metrics

Model performance was evaluated using standard object detection metrics, including precision, recall, F1 score, and mean Average Precision (mAP). In particular, mAP@0.5 (mAP50) and mAP@0.5:0.95 (mAP50–95) were used to assess detection accuracy across different Intersection-over-Union (IoU) thresholds.

The YOLO26m model achieved an overall mAP50 of 0.81 and mAP50–95 of 0.71 across all nine pollutant classes. Performance was highest for well-defined, structured objects such as barriers and road signs (mAP > 0.87), whereas more ambiguous or visually variable categories, such as potholes and graffiti, exhibited lower performance (mAP $\approx$ 0.67–0.71).

Table 6.3 shows the overall and per-class performance metrics for the model. Precision remained consistently high across all classes (greater than 0.9 at the optimal confidence threshold), while recall plateaued at approximately 0.72. This indicates that the model prioritises reducing false positives at the expense of missing some true instances. The F1 score peaked at approximately 0.81 at a confidence threshold of 0.39, which was therefore selected as the operating threshold for inference.

Class	Precision	Recall	F1 Score	mAP@50
Overall (Macro Avg)	0.902	0.733	0.805	0.812
Barrier	0.975	0.780	0.866	0.903
Billboard	0.839	0.776	0.806	0.844
Bin	0.944	0.739	0.828	0.864
Graffiti	0.946	0.625	0.752	0.709
Mobile Advertisement	0.947	0.783	0.857	0.847
Pothole	0.929	0.591	0.721	0.669
Road Sign	0.853	0.841	0.847	0.873
Shop Sign	0.860	0.688	0.765	0.777
Utility Pole	0.828	0.780	0.803	0.822

Table 6.3: Overall and per-class performance metrics.

Figure C.1 shows the distribution of annotated bounding boxes. Most bounding boxes are relatively small, which can be attributed to the high frequency of objects such as utility poles and road signs, which are typically thin, distant, or occupy a small proportion of the image. The dataset also exhibits class imbalance, with categories such as potholes, mobile advertisements, and bins having significantly fewer instances compared to more common classes such as utility poles, road signs, and billboards. Despite this imbalance, the model demonstrates strong overall performance.

Figure C.2 presents the normalised confusion matrix. Strong diagonal values indicate that most classes are correctly identified, particularly for structured categories such as road signs and barriers. However, confusion is observed between visually similar classes, including road signs, shop signs, and billboards, reflecting overlapping visual characteristics such as rectangular shapes, text, and placement within the scene.

Additionally, a notable proportion of true objects are misclassified as background across multiple classes. This is particularly evident for smaller or less salient objects, such as potholes and distant utility poles. This behaviour is consistent with the observed recall plateau and suggests that the model struggles with detecting small, thin, or partially occluded objects.

6.4 Initial Model

An initial YOLOv12m model was trained on a small dataset to accelerate the annotation process for the final dataset. This dataset consisted of approximately 300 images, with a class distribution roughly reflecting that of the final YOLO26 model. The model was trained for 100 epochs using similar hyperparameters.

The primary purpose of this model was to perform inference on a subset of the collected images prior to annotation. These predictions were then imported into Label Studio, so existing bounding boxes were used rather than being annotated from scratch. This significantly reduced annotation time, particularly for frequently occurring classes such as billboards, where only minor adjustments were often required.

However, the model achieved poor performance. This was largely due to the dataset splitting strategy, which, unlike the final model, did not enforce class-balanced splits. As a result, dominant classes such as billboards and utility poles were well represented, while rarer classes such as potholes and barriers were underrepresented or absent in the validation set. Consequently, the model was unable to learn robust representations for these minority classes.

Although YOLOv12 was the most recent model available at the time, the release of YOLO26 during the project prompted a transition to the newer architecture. This decision was made to take advantage of improved performance and to align with current state-of-the-art approaches.

Chapter 7

Experimental Results

7.1 Scores

7.1.1 Scores without OSM features

When considering only image-based detections within each region, the VPI defined in Section 5.4 produced the results shown in Table 7.1. These results highlight the cities with the highest and lowest estimated levels of visual pollution based solely on image data.

To improve the reliability of the results, only regions containing more than 300 images were included in the analysis. Regions with smaller sample sizes were excluded, as limited observations can lead to high variance and unrepresentative scores. This threshold was chosen empirically, as regions with approximately this number of images exhibited stable and consistent scoring behaviour.

Highest Pollution Scores			Lowest Pollution Scores		
City	Country	Score	City	Country	Score
Kolkata	India	0.653	Shenzhen	China	0.071
Pune	India	0.568	Lahti	Finland	0.078
Barmhan Kalan	India	0.567	Mission Viejo	United States	0.082
Dhaka	Bangladesh	0.563	Siracusa	Italy	0.086
Kumasi	Ghana	0.562	Columbia	United States	0.089
Caloocan City	Philippines	0.560	Roseville	United States	0.099
Malang	Indonesia	0.538	Messina	Italy	0.110
Chennai	India	0.532	Pomona	United States	0.112
Purwokerto	Indonesia	0.531	Stockholm	Sweden	0.112
Vishakhapatnam	India	0.523	Serpukhov	Russia	0.113

Table 7.1: Cities with the highest and lowest visual pollution scores without OSM features.

The highest-scoring cities are concentrated in South Asia, Southeast Asia, and parts of Africa. These regions are often associated with high urban density, rapid development, and less restrictive control of outdoor advertising and street infrastructure. As a result,

scenes frequently contain a greater diversity and frequency of pollutant categories, which increases both the CCR and SWS components of the VPI.

In contrast, the lowest-scoring cities are primarily located in Europe, North America, and East Asia. These regions typically benefit from stricter urban planning policies, more consistent infrastructure management, and regulation of signage and advertising. Consequently, both the diversity and frequency of detected pollutants are reduced, leading to lower VPI values.

Figure 7.1 illustrates the global distribution of normalised region scores. A clear spatial pattern emerges, with higher scores clustering in parts of South and Southeast Asia, while lower scores dominate in Europe and North America.

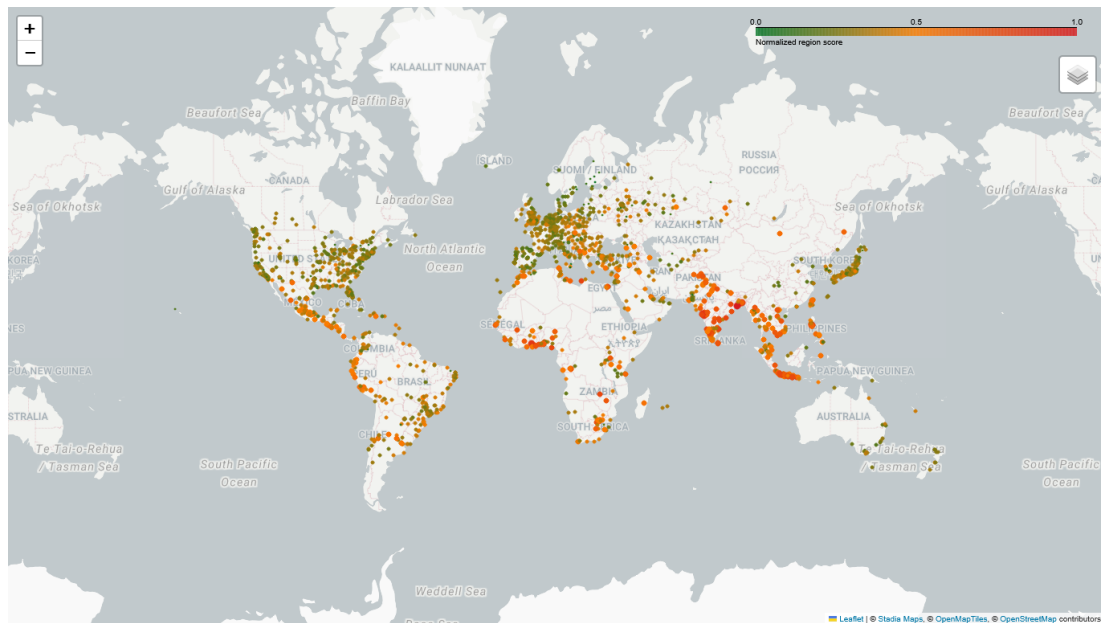


Figure 7.1: Normalised visual pollution scores across all regions.

However, this distribution must be interpreted with caution. Differences in data availability, street-view coverage, and image quality across regions may introduce bias into the results. In particular, regions with denser image coverage may produce more reliable estimates, whereas sparsely sampled regions may exhibit higher variance in their scores.

7.1.2 Scores with OSM features

To improve coverage and incorporate structured spatial data, OSM features were included in the scoring process. The resulting scores are shown in Table 7.2.

Highest Pollution Scores			Lowest Pollution Scores		
City	Country	Score	City	Country	Score
Kolkata	India	0.391	Shenzhen	China	0.050
Caloocan City	Philippines	0.368	Columbia	United States	0.066
Chennai	India	0.347	Mission Viejo	United States	0.071
Barmhan Kalan	India	0.343	Lipetsk	Russia	0.084
Cairo	Egypt	0.336	Pomona	United States	0.084
Dhaka	Bangladesh	0.335	Yonkers	United States	0.084
Dakar	Senegal	0.334	Burlington	United States	0.087
Malang	Indonesia	0.329	Orebro	Sweden	0.088
San Fernando	Philippines	0.329	Roseville	United States	0.088
Bekasi	Indonesia	0.328	Carlsbad	United States	0.091

Table 7.2: Cities with the highest and lowest visual pollution scores (with OSM features).

Including OSM features results in a noticeable compression of score values across all regions. This is expected, as structured features such as utility infrastructure and bins introduce additional detections across many regions, reducing relative variation between cities.

Despite this, the relative ordering of cities remains broadly consistent, with many high-scoring regions appearing in both tables. This suggests that image-based detections capture a significant portion of the underlying visual pollution signal, while OSM features provide complementary information rather than fundamentally altering rankings.

7.2 São Paulo, Brazil

The *Lei Cidade Limpa* (Clean City Law) implemented in São Paulo provides a natural case study for evaluating the impact of regulation on visual pollution. As the law led to the removal of a large number of billboards, it is expected that São Paulo would exhibit a lower proportion of billboard detections than other major Brazilian cities.

To evaluate this, São Paulo was compared with Rio de Janeiro, Belo Horizonte, and Fortaleza. The primary metric used was the proportion of images classified as billboards within each region:

$$\text{Billboard Ratio} = \frac{\text{Number of Billboard Detections}}{\text{Total Number of Images}} \quad (7.1)$$

This metric provides a simple proxy for the prevalence of large-scale advertising structures within each city. Differences in this ratio can indicate the effectiveness of regulatory interventions in reducing specific types of visual pollution.

7.2.1 Results

The results do not show a clear or consistent reduction in billboard density associated with regulation. While São Paulo exhibits a lower density than Rio de Janeiro and

Fortaleza, Belo Horizonte displays an even lower value, contradicting the expected pattern.

City	Images	Billboards	Density	Δ vs São Paulo
São Paulo	10,879	1,194	0.110	—
Rio de Janeiro	7,653	1,271	0.166	+51.3%
Belo Horizonte	10,052	911	0.091	-17.5%
Fortaleza	9,070	1,517	0.167	+52.3%

Table 7.3: Comparison of billboard density across selected Brazilian cities.

One plausible explanation for this discrepancy is the broad definition of “billboard” used in this project. The detection class includes posters, banners, and other forms of signage that are not strictly equivalent to the large outdoor advertising structures targeted by the Clean City Law. As a result, the measured densities may not accurately reflect the specific types of advertising regulated in São Paulo.

Consequently, without manual verification, it is difficult to determine the extent to which detected instances correspond to regulated billboards. This limits the ability to draw direct conclusions about the policy’s effectiveness from these results alone.

7.3 Poland

In 2015, Poland introduced the *Landscape Act*, which enabled municipalities to regulate the aesthetics of public spaces, including advertising and signage [4]. This provides an opportunity to evaluate changes in visual pollution over time under varying regulatory conditions.

As noted by Brecz-Kupiec (2024) [4], implementation of the act was decentralised, with municipalities introducing restrictions at different times, while some did not introduce regulations at all. This variation enables a natural comparison between pre- and post-regulation environments.

Three cities were selected for analysis: Kraków, Gdańsk, and Wrocław. Gdańsk introduced restrictions in 2018, while Kraków and Wrocław implemented regulations in 2020. For each city, two dense scans were collected, one prior to the introduction of restrictions and one following their implementation.

The analysis focuses on changes in billboard prevalence, measured using the Billboard Ratio defined previously. By comparing pre- and post-regulation data, it is possible to assess whether legislative changes correspond to measurable reductions in visual pollution.

7.3.1 Results

The results show a clear and consistent reduction in billboard density following the introduction of advertising regulations across all three cities.

City	Period	Images	Billboards	Density	$\Delta\%$
Kraków	2005–2019	3,600	876	0.243	-49.9
	2020–2026	7,941	967	0.122	
Gdańsk	2005–2017	1,901	550	0.289	-40.8
	2018–2026	2,059	353	0.171	
Wrocław	2005–2019	4,005	1,035	0.258	-38.7
	2020–2026	4,295	680	0.158	

Table 7.4: Billboard density before and after regulation in selected Polish cities.

Across all three cities, billboard density decreases substantially after the introduction of regulation, with reductions ranging from 38.7% to 49.9%. This consistent trend across independently sampled locations provides evidence that the observed effect is not isolated to a single city.

Differences in sample size are unlikely to explain the observed reductions. Gdańsk and Wrocław have comparable numbers of images across time periods, while Kraków exhibits a similar relative decrease despite a substantially larger post-regulation sample. This suggests that the reduction is not an artefact of sampling variation.

However, the results should be interpreted with caution, as factors such as image quality, spatial coverage, and camera perspective can influence detection. Consequently, the analysis reflects changes in observed billboard density within the dataset rather than the true underlying distribution.

Nevertheless, the consistent reductions across all three cities indicate that the approach can capture meaningful changes aligned with real-world regulatory interventions.

7.4 Pollution by land-use in Edinburgh, Scotland

To analyse how visual pollution varies across land-use types, Edinburgh was selected as a case study. OpenStreetMap data was retrieved using the Python package `osmnx`, which provided polygon boundaries and associated land-use classifications for different areas of the city.

Five land-use categories were considered: commercial, residential, industrial, retail, and park/leisure/green. Two metrics were defined to enable both overall and class-specific analysis:

$$\text{Pollution ratio}_{c,l} = \frac{\text{Number of detections of class } c \text{ in land-use } l}{\text{Number of images in land-use } l} \quad (7.2)$$

$$\text{Pollution ratio}_l = \frac{\text{Total detections in land-use } l}{\text{Number of images in land-use } l} \quad (7.3)$$

Equation 7.3 provides an aggregate measure of visual pollution across different land-use types, while Equation 7.2 enables analysis of how individual pollutant categories are

distributed within each land-use class. Detections that did not fall within a polygon were assigned to the nearest available polygon to ensure full spatial coverage.

7.4.1 Results

The results indicate that pollution ratios are highest in commercial and retail areas, while residential regions exhibit the lowest overall values. This trend is illustrated in Figure 7.2.

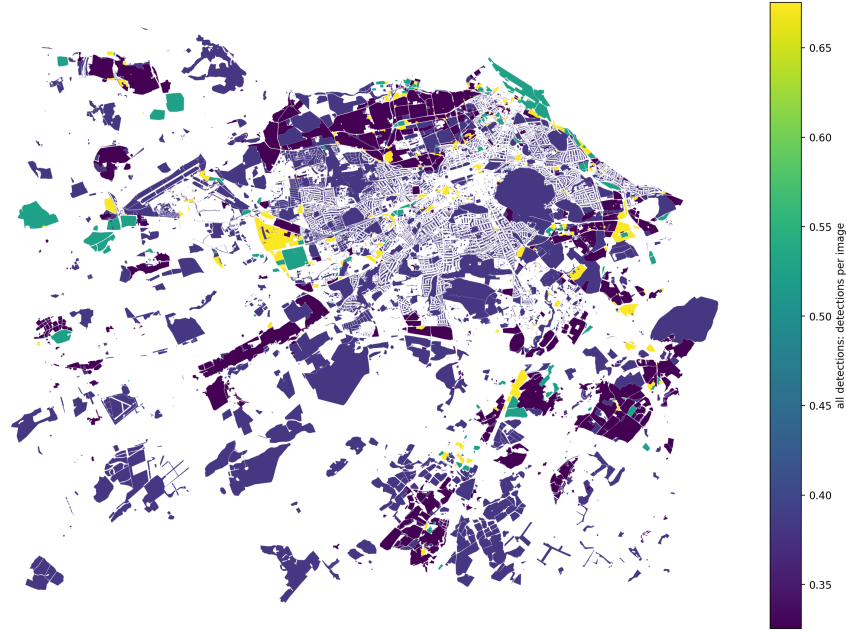


Figure 7.2: Distribution of detections across land-use categories

Further analysis reveals that certain pollutant types exhibit distinct spatial patterns. For example, bins are more frequently detected in residential areas and green spaces, as shown in Figure E.1. This aligns with expected urban infrastructure distribution, providing qualitative validation of the model's behaviour.

These findings demonstrate that the model is capable of capturing meaningful variation in visual pollution across different land-use types. This approach can be extended to other cities to compare how land-use influences visual pollution in different urban contexts.

7.5 Additional Experiments

Additional experiments, including country-level analyses, top regions per pollutant, and heatmaps, are presented in Appendix A.

Chapter 8

Discussion

This project demonstrates the effectiveness of deep learning for detecting and quantifying visual pollution at a global scale. The proposed approach enables analysis of how visual pollution varies across time and land-use types, with results that generally align with expected urban patterns.

However, interpretation of these results requires careful consideration of context. In particular, visual pollution scores do not directly correspond to aesthetic quality. For example, regions such as deserts may receive low scores due to the absence of detectable pollutants, rather than reflecting a visually desirable environment.

Despite this, the framework enables meaningful comparisons between regions with similar characteristics, particularly where data coverage is comparable. In such cases, the system can provide insight into which objects contribute most significantly to visual pollution within different urban environments.

8.1 Limitations

8.1.1 Data availability

A key limitation of this work is the uneven distribution of data across regions. This imbalance is illustrated in Figure 8.1, where regions in India and China are frequently shown in red, indicating low image availability, while Europe and North America exhibit substantially greater coverage.

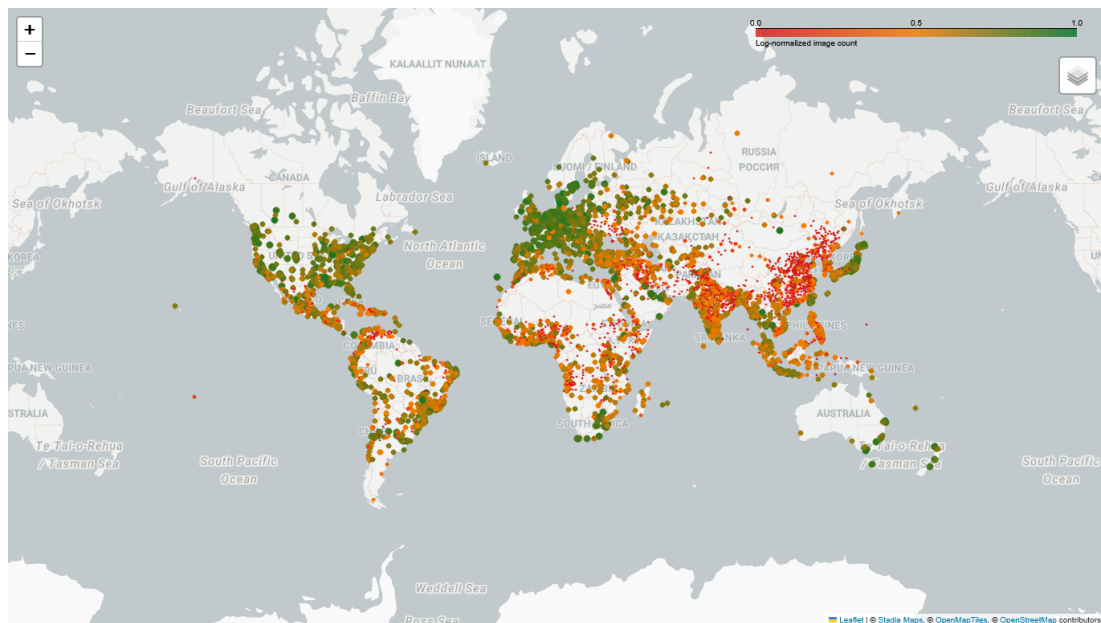


Figure 8.1: All regions collected, with a log-normalised colour scale representing the number of images per region.

Although OpenStreetMap (OSM) features were incorporated to mitigate this issue (see Figure B.1), a similar geographic imbalance persists. As shown in Figure B.2, Europe and North America still contain a significantly higher density of OSM features compared to many parts of Asia and Africa.

This uneven distribution introduces a structural bias into the analysis. Regions with higher data availability yield more stable and representative scores, whereas sparsely sampled regions are more likely to produce noisy estimates. Consequently, cross-regional comparisons may disproportionately favour well-covered areas, limiting the generalisability of the findings.

8.1.2 Labelling

Three primary limitations were identified in the labelling process: (1) ambiguity in category definitions, (2) cross-cultural and linguistic challenges, and (3) class imbalance.

The definition of a billboard was intentionally broad to capture a wider range of advertising formats, including posters and banners. While this improves coverage, it introduces inconsistency when comparing results against real-world regulations, where definitions are typically more precise. As a result, the detected “billboard” category does not always align directly with policy definitions.

Labelling across a global dataset also introduces challenges related to language and cultural context. In non-English-speaking regions, it is not always possible to accurately interpret signage, and visual styles vary significantly between countries. Consequently, some shop signs may have been misclassified as billboards.

This introduces label noise, which may bias the model towards over-detecting certain

categories in specific regions, affecting both detection reliability and cross-regional comparisons.

Finally, the dataset exhibits class imbalance, with some categories appearing far more frequently than others. This can lead to stronger performance on dominant classes and weaker performance on underrepresented ones, reducing the reliability of the overall scoring.

8.2 Future Work

Future work should focus on improving dataset quality and balance. Increasing the representation of underrepresented classes through targeted data collection and class-balanced training strategies would enable the model to learn more robust representations across all pollutant types. The dataset produced in this project, which converts model-generated detections into annotated images, provides a foundation for selectively expanding specific classes in future training data.

In addition, incorporating annotators from diverse geographic and cultural backgrounds would improve labelling consistency, particularly for categories influenced by local context. This would enhance the model's ability to generalise across regions.

Further improvements could be achieved by introducing automated filtering of low-quality images. Removing blurred or poorly exposed images would reduce noise in both training and evaluation, thereby improving model performance.

Expanding the geographic coverage of the dataset is another key direction. Performing dense scans across entire countries would enable more detailed spatial comparisons, similar to Figure A.1. Additional filtering strategies could be applied to analyse specific pollutant types, such as excluding road signs and utility poles or focusing solely on billboard detections. This approach could also support the development of country-specific models to improve detection accuracy.

Finally, future work could explore the integration of vision-language models to provide more interpretable outputs. Rather than relying solely on numerical scores, such models could generate descriptive explanations of visual pollution, offering more actionable insights for urban planners and policymakers.

Chapter 9

Conclusion

This project addressed key limitations in existing visual pollution detection work by leveraging large-scale data collection and expanding class definitions. A global dataset comprising over 2 million images from 179 countries was constructed, and a detection model was trained using this data, with annotations stored separately for reproducibility and further use.

In addition, the introduction of new classes, including shop signs and mobile advertisements, enabled a more granular analysis of advertising-related pollution. By separating these from the broader billboard category and assigning distinct weights within the Visual Pollution Index, the system allows for more nuanced regional comparisons. These categories remain underexplored in existing literature and therefore provide a useful foundation for future research.

The trained YOLO26m model achieved strong performance despite the challenges posed by noisy labels, class imbalance, and global variability in the dataset. This demonstrates the feasibility of applying deep learning to large-scale, geographically diverse visual pollution analysis.

However, several limitations remain, including uneven data availability and class imbalance. These factors affect both model performance and the reliability of regional comparisons, highlighting important directions for future work.

Overall, this project demonstrates a scalable approach to analysing visual pollution at a global level, providing both methodological contributions and practical insights into how visual pollution varies across regions.

Bibliography

- [1] Adriana Portella. *Visual Pollution: Advertising, Signage and Environmental Quality*. Routledge, 1st edition, 2014.
- [2] Hangyu Gao, Shamsul Abu Bakar, Suhardi Maulan, Mohd Johari Mohd Yusof, Riyadh Mundher, Yu Guo, and Benxue Chen. A systematic literature review and analysis of visual pollution. *Land*, 13(7), 2024.
- [3] Kurt Kohlstedt. Clean city law: Secrets of são paulo uncovered by outdoor advertising ban, February 2016. Accessed October 2025.
- [4] Sara Brecz-Kupiec. Evaluating a decade of landscape resolutions in poland: Progress, challenges, and effectiveness in public space management. *European Research Studies*, XXVII(3):888–907, December 2024.
- [5] Advait Gupta, Manan Padsala, Devesh Jani, Tanmay Bisen, Aastha Shayla, and Gargi Srivastava. *From Above and Beyond: Decoding Urban Aesthetics with the Visual Pollution Index*, page 101–119. Springer Nature Switzerland, Cham, 2024.
- [6] Oxford English Dictionary. “visual pollution”, n.d. Accessed October 2025.
- [7] New York State Department of Health. 27th annual report of the state department of health of new york, 1906. Published 1907.
- [8] John Copeland Nagle. Cell phone towers as visual pollution. *Notre Dame JL Ethics & Pub. Pol’y*, 23:537, 2009.
- [9] Szymon Chmielewski, Danbi J. Lee, Piotr Tompalski, Tadeusz J. Chmielewski, and Piotr Weżyk. Measuring visual pollution by outdoor advertisements in an urban street using intervisibility analysis and public surveys. *International Journal of Geographical Information Science*, 30(4):801–818, 2016.
- [10] Andreas Sumper, Oriol Boix-Aragones, Joan Rull-Duran, Joan Amat-Algaba, and Joachim Wagner. Assessment of the visual impact of existing high-voltage lines in urban areas. *Electricity*, 2(3):285–299, 2021.
- [11] Nahian Ahmed, M. Nazmul Islam, Ahmad Saraf Tuba, M.R.C. Mahdy, and Mohammad Sujauddin. Solving visual pollution with deep learning: A new nexus in environmental management. *Journal of Environmental Management*, 248:109253, 2019.
- [12] Rolf Weber. *On the Aesthetics of Architecture: A Psychological Approach to the*

- Structure and the Order of Perceived Architectural Space*. Routledge, 1st edition, 1995.
- [13] J.T. Lang. *Creating Architectural Theory: The Role of the Behavioral Sciences in Environmental Design*. Role of the Behavioural Sciences in Environmental Design. Van Nostrand Reinhold Company, 1987.
 - [14] Hangyu Gao, Shamsul Abu Bakar, Maulan Suhardi, Yu Guo, Mohd Johari Mohd Yusof, Riyadh Mundher, Yue Zhuo, and Jun Qi. Constructing a conceptual framework: Interpreting visual preference and visual pollution factors among viewers in highway landscapes. *Transportation Research Interdisciplinary Perspectives*, 31:101399, 2025.
 - [15] Hannah Ritchie, Veronika Samborska, and Max Roser. Urbanization. *Our World in Data*, 2024. <https://ourworldindata.org/urbanization>.
 - [16] A Lot Media. The evolution of advertising companies over the decades, October 2024. Accessed October 2025.
 - [17] Szymon Chmielewski, Marta Samulowska, Michał Lupa, Danbi Lee, and Bogdan Zagajewski. Citizen science and webgis for outdoor advertisement visual pollution assessment. *Computers, Environment and Urban Systems*, 67:97–109, 2018.
 - [18] Outdoor Advertising Association of America. Laws & regulations, n.d. Accessed October 2025.
 - [19] Federal Highway Administration. How the highway beautification act became a law, June 2017. Accessed October 2025.
 - [20] Scenic America. About the highway beautification act, n.d. Accessed October 2025.
 - [21] NYC Global Partners. Best practice: Clean city act, April 2011. Accessed October 2025; PDF.
 - [22] Eshetu Gelan. Assessing visual pollution: The impact of urban outdoor advertisements in addis ababa, ethiopia. *Architecture*, 5(1), 2025.
 - [23] Vasudha Gokhale. Examining impact of visual pollution on city environment: Case study of pune, india. *Spandrel Journal of School of Planning and architecture Bhopal*, 1:10, 01 2011.
 - [24] Robert J. Schodorf. A study of visual pollution from overhead wires and associated structures. Master's thesis, Western Michigan University, August 1973. Accessed October 2025.
 - [25] Khydija Wakil, Malik Asghar Naeem, Ghulam Abbas Anjum, Abdul Waheed, Muhammad Jamaluddin Thaheem, Muhammad Qadeer ul Hussnain, and Raheel Nawaz. A hybrid tool for visual pollution assessment in urban environments. *Sustainability*, 11(8), 2019.
 - [26] Khydija Wakil, Ali Tahir, Muhammad Qadeer ul Hussnain, Abdul Waheed, and Raheel Nawaz. Mitigating urban visual pollution through a multistakeholder

- spatial decision support system to optimize locational potential of billboards. *ISPRS International Journal of Geo-Information*, 10(2), 2021.
- [27] Md Fahim Shahoriar Titu, Abdul Aziz Chowdhury, S. M. Rezwanaul Haque, and Riasat Khan. Deep-learning-based real-time visual pollution detection in urban and textile environments. *Sci*, 6(1), 2024.
 - [28] HAIDONG SHANG and I.D. BISHOP. Visual thresholds for detection, recognition and visual impact in landscape settings. *Journal of Environmental Psychology*, 20(2):125–140, 2000.
 - [29] Ian D Bishop. Determination of thresholds of visual impact: The case of wind turbines. *Environment and Planning B: Planning and Design*, 29(5):707–718, 2002.
 - [30] Jacob Ladenburg. Visual impact assessment of offshore wind farms and prior experience. *Applied Energy*, 86(3):380–387, 2009.
 - [31] Christian Koch and Ioannis Brilakis. Pothole detection in asphalt pavement images. *Advanced Engineering Informatics*, 25(3):507–515, 2011. Special Section: Engineering informatics in port operations and logistics.
 - [32] Najia Hasan Tasnim, Sadia Afrin, Barna Biswas, Arifa Akter Anye, and Riasat Khan. Automatic classification of textile visual pollutants using deep learning networks. *Alexandria Engineering Journal*, 62:391–402, 2023.
 - [33] Ivana Marin, Saša Mladenović, Sven Gotovac, and Goran Zaharija. Deep-feature-based approach to marine debris classification. *Applied Sciences*, 11(12), 2021.
 - [34] Vung Pham, Chau Pham, and Tommy Dang. Road damage detection and classification with detectron2 and faster r-cnn. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 5592–5601, 2020.
 - [35] Ji-Won Baek and Kyungyong Chung. Pothole classification model using edge detection in road image. *Applied Sciences*, 10(19), 2020.
 - [36] Mohammad AlElaiwi, Mugahed A. Al-antari, Hafiz Farooq Ahmad, Areeba Azhar, Badar Almarri, and Jamil Hussain. Visual pollution real images benchmark dataset on the public roads. *Data in Brief*, 50:109491, 2023.
 - [37] Mohammad AlElaiwi, Mugahed A. Al-antari, Hafiz Farooq Ahmad, Areeba Azhar, Badar Almarri, and Jamil Hussain. Vpp: Visual pollution prediction framework based on a deep active learning approach using public road images. *Mathematics*, 11(1), 2023.
 - [38] Mohammad AlElaiwi, Hafiz Ahmad, Jamil Hussain, Mugahed Al-antari, Badar Almarri, and Areeba Azhar. Saudi arabia public roads visual pollution dataset, 2023.
 - [39] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
 - [40] Burr Settles. Active learning literature survey. 2009.

- [41] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021.
- [42] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015.
- [43] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.
- [44] Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection, 2020.
- [45] Yoseph Widistika Ranga Prakoso. Vision for pollution: Deep learning to detect visual pollution. Master’s thesis, University of Edinburgh, 2025.
- [46] Mohammad Masudur Rahman, Md Rashedur Rahman, Ashraful Islam, Saadia B. Alam, and M. Ashraful Amin. A scoping review of deep learning for urban visual pollution and proposal of a real-time monitoring framework with a visual pollution index. (arXiv:2602.09446), February 2026. arXiv:2602.09446 [cs].
- [47] GOV.UK. Signs, stationery and promotional material for limited companies, n.d. Accessed: 2026-03-18.
- [48] Chuanbo Hu, Shan Jia, Fan Zhang, Changjiang Xiao, Mindi Ruan, Jacob Thrasher, and Xin Li. Updexplainer: An interpretable transformer-based framework for urban physical disorder detection using street view imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 204:209–222, October 2023.
- [49] Javier Rozo Alzate, Marta S. Tabares, and Paola Vallejo. Graffiti and government in smart cities: a deep learning approach applied to medellín city, colombia. In *International Conference on Data Science, E-learning and Information Systems 2021, DATA’21*, page 160–165, New York, NY, USA, 2021. Association for Computing Machinery.
- [50] Mrouj Almuhaajri and Ching Y. Suen. Shop signboards detection using the shos dataset. In *Pattern Recognition and Artificial Intelligence: Third International Conference, ICPRAI 2022, Paris, France, June 1–3, 2022, Proceedings, Part II*, page 235–245, Berlin, Heidelberg, 2022. Springer-Verlag.
- [51] Bin Wang and Li Jing. *First-place Solution for Streetscape Shop Sign Recognition Competition*. January 2025.
- [52] United Nations. United nations treaty collection: Road traffic convention (1968), n.d. Accessed: 2026-03-18.
- [53] Simplemaps. World cities database, 2026. Accessed: 9 March 2026.
- [54] Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. Label Studio: Data labeling software, 2020-2025. Open source software available from <https://github.com/HumanSignal/label-studio>.
- [55] Glenn Jocher and Jing Qiu. Ultralytics yolo26, 2026.

- [56] Md. Yearat Hossain, Ifran Rahman Nijhum, Md. Tazin Morshed Shad, Abu Adnan Sadi, Md. Mahmudul Kabir Peyal, and Rashedur M. Rahman. An end-to-end pollution analysis and detection system using artificial intelligence and object detection algorithms. *Decision Analytics Journal*, 8:100283, 2023.

Appendix A

Additional experiments

A.1 Country-level Analysis

An experiment was conducted to analyse visual pollution at the country level. This involved generating a hexagonal map in which each cell was coloured according to the ratio of detected pollutants to the number of images.

This representation enables comparison of relative pollution levels across different regions within a country. Figure A.1 illustrates this for the United Kingdom. Higher values are observed in major urban areas, particularly around London, Birmingham, and Manchester.

However, these elevated values are largely driven by the high frequency of road sign detections, which are more prevalent in dense urban environments. As a result, the map reflects not only visual pollution but also underlying differences in infrastructure density, which should be considered when interpreting the results.

A more detailed view could be obtained by scanning the entire country, enabling relative comparisons across areas. However, that was not in the scope of this project.

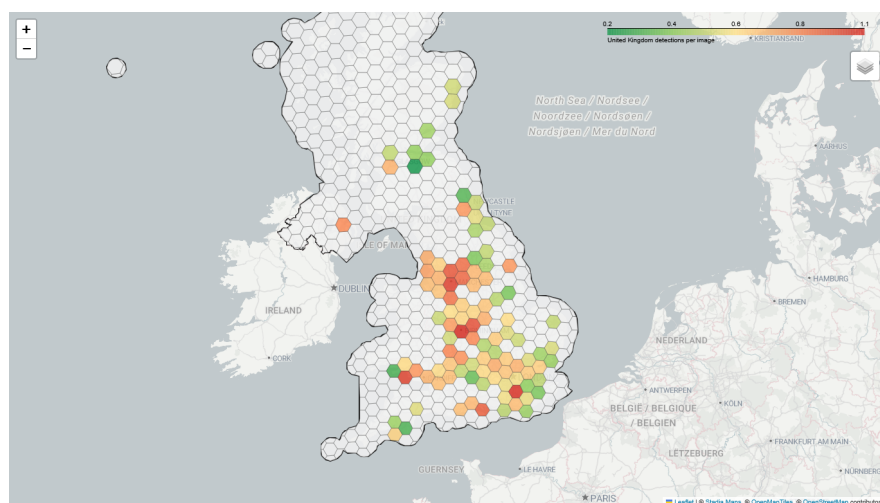


Figure A.1: Hex-map of visual pollution ratios in the United Kingdom.

A.2 Top Regions per Pollutant Class

To analyse regions with the highest concentration of each pollutant, a detection-to-image ratio was computed for each class within each region. Only regions containing more than 300 images were included to ensure statistical reliability. The resulting rankings are shown in Table A.1.

Several geographic patterns can be observed. Spanish cities dominate the top rankings for bins, suggesting consistent urban infrastructure and waste management visibility. In contrast, the highest billboard detection rates are concentrated in cities within India and Sri Lanka. This aligns with the VPI results presented in Table 7.1, where billboards were the primary contributors to elevated scores.

Graffiti detections are most prominent in Latin American cities, which may reflect the cultural prevalence of street art in these regions. Additionally, utility pole detections are particularly high in Japanese cities, likely due to the widespread use of above-ground electrical infrastructure.

These results highlight how different forms of visual pollution vary geographically and demonstrate the importance of class-specific analysis when interpreting regional pollution patterns.

Class	City (Country)	Images	Detections	Rate
Barrier	Ash Shihaniyah (Qatar)	333	89	0.267
	Nakhon Ratchasima (Thailand)	429	100	0.233
	Sejong (South Korea)	466	97	0.208
	Hamilton (New Zealand)	940	182	0.194
	Kissimmee (USA)	407	77	0.189
Billboard	Sri Jayewardenepura Kotte (Sri Lanka)	638	672	1.053
	Maharagama (Sri Lanka)	651	678	1.041
	Ranchi (India)	408	413	1.012
	Haldwani (India)	443	434	0.980
	Kolkata (India)	328	313	0.954
Bin	Santander (Spain)	1374	110	0.080
	Barcelona (Spain)	1869	146	0.078
	Sabadell (Spain)	975	74	0.076
	Badalona (Spain)	971	68	0.070
	Pamplona (Spain)	1110	77	0.069
Graffiti	Santiago (Chile)	2691	562	0.209
	Quito (Ecuador)	2176	320	0.147
	Temperley (Argentina)	352	44	0.125
	San Luis Potosi (Mexico)	3713	413	0.111
	Puebla (Mexico)	3164	345	0.109
Mobile Ad	Newcastle (South Africa)	340	41	0.121
	Faisalabad (Pakistan)	473	39	0.082
	Ahmedabad (India)	612	45	0.074
	Stara Zagora (Bulgaria)	429	30	0.070
	Shubra al Khaymah (Egypt)	347	21	0.061
Pothole	Kenema (Sierra Leone)	938	37	0.039
	Bandundu (DR Congo)	462	9	0.019
	Dera Ghazi Khan (Pakistan)	304	5	0.016
	Gulu (Uganda)	677	10	0.015
	Koidu (Sierra Leone)	961	14	0.015
Road Sign	Shenzhen (China)	477	624	1.308
	Lipetsk (Russia)	366	390	1.066
	Fuenlabrada (Spain)	484	443	0.915
	Ivanovo (Russia)	1551	1418	0.914
	Taiyuan (China)	356	323	0.907
Shop Sign	Sharjah (UAE)	1986	828	0.417
	Kaohsiung (Taiwan)	1702	674	0.396
	Ar Rusayfah (Jordan)	496	177	0.357
	Bursa (Turkey)	358	124	0.346
	Puyang (China)	1123	361	0.321
Utility Pole	Anjomachi (Japan)	977	1983	2.030
	Aguadilla (Puerto Rico)	350	710	2.029
	Okinawa (Japan)	1099	2219	2.019
	Ishizaki (Japan)	627	1228	1.959
	Koshigaya (Japan)	727	1334	1.835

Table A.1: Top five cities per class ranked by detection rate (detections per image).

A.3 Heatmaps

To examine the spatial distribution of pollution within regions, heatmaps were generated. These were constructed using the severity weights defined in Table 5.2, such that areas with higher weighted pollution scores are more prominently highlighted.

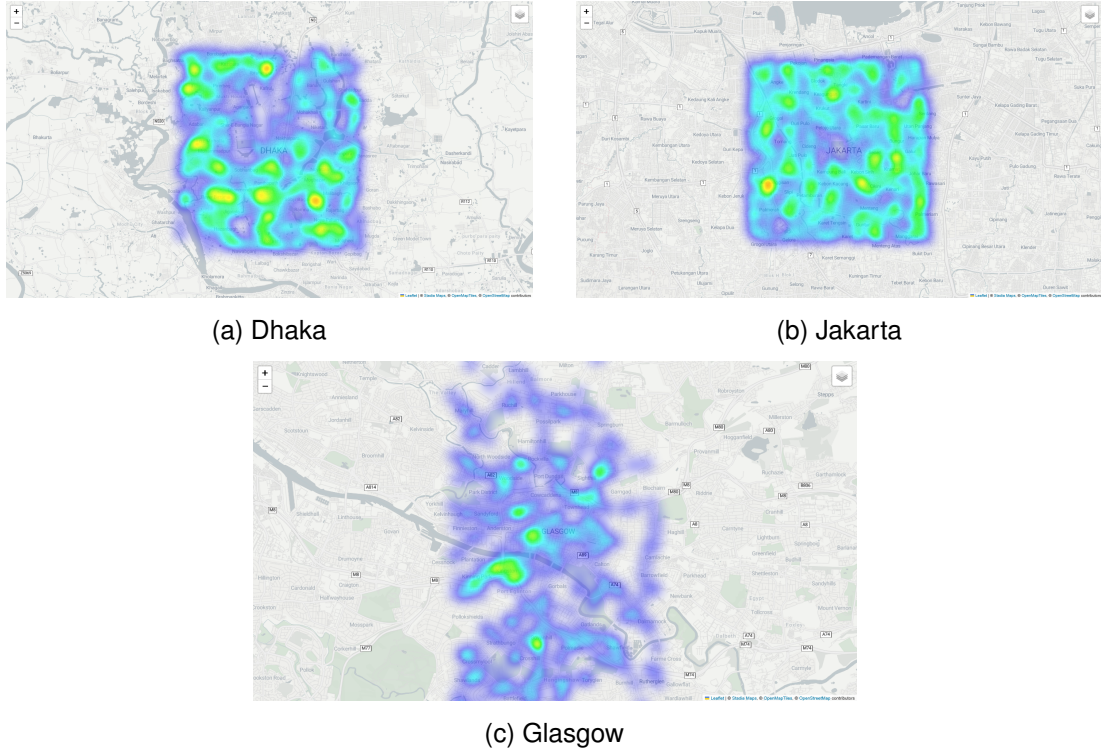


Figure A.2: Heatmaps illustrating the spatial distribution of visual pollution in selected cities.

Clear spatial patterns can be observed. In both Dhaka and Jakarta, lower pollution levels are visible in areas such as parks, large open spaces, and airport regions, where fewer detectable objects are present. In contrast, higher concentrations tend to occur in dense urban areas.

Glasgow shows elevated pollution scores concentrated around the city centre. This pattern may reflect genuinely higher levels of visual pollution, but it may also be influenced by uneven data coverage, with a greater number of images available in central areas.

Overall, these heatmaps demonstrate how spatial visualisation can help identify localised hotspots of visual pollution, providing a useful tool for targeted analysis and potential urban planning interventions.

Appendix B

Maps

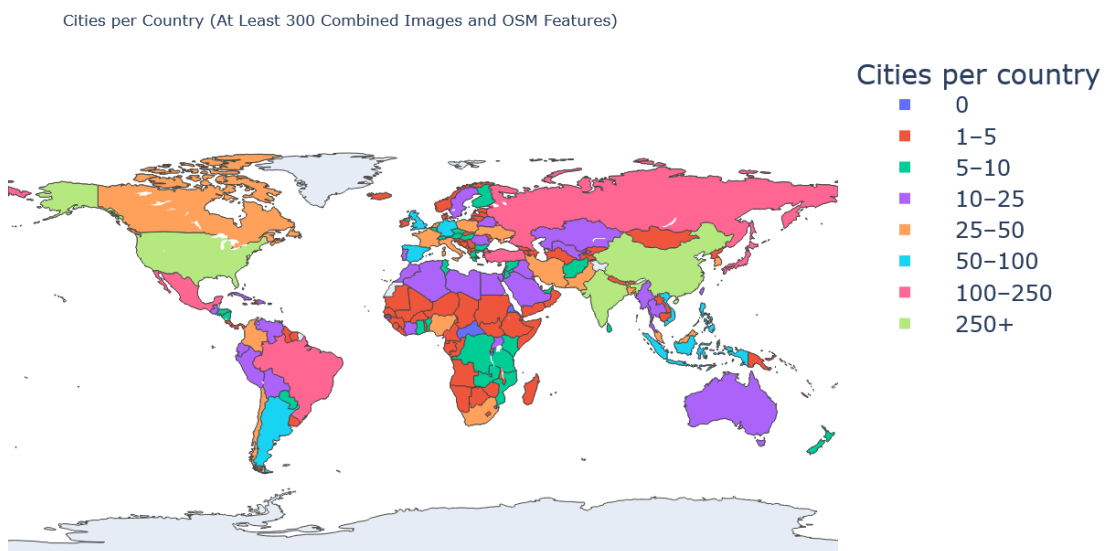


Figure B.1: Number of cities per country with at least 300 images and OSM features

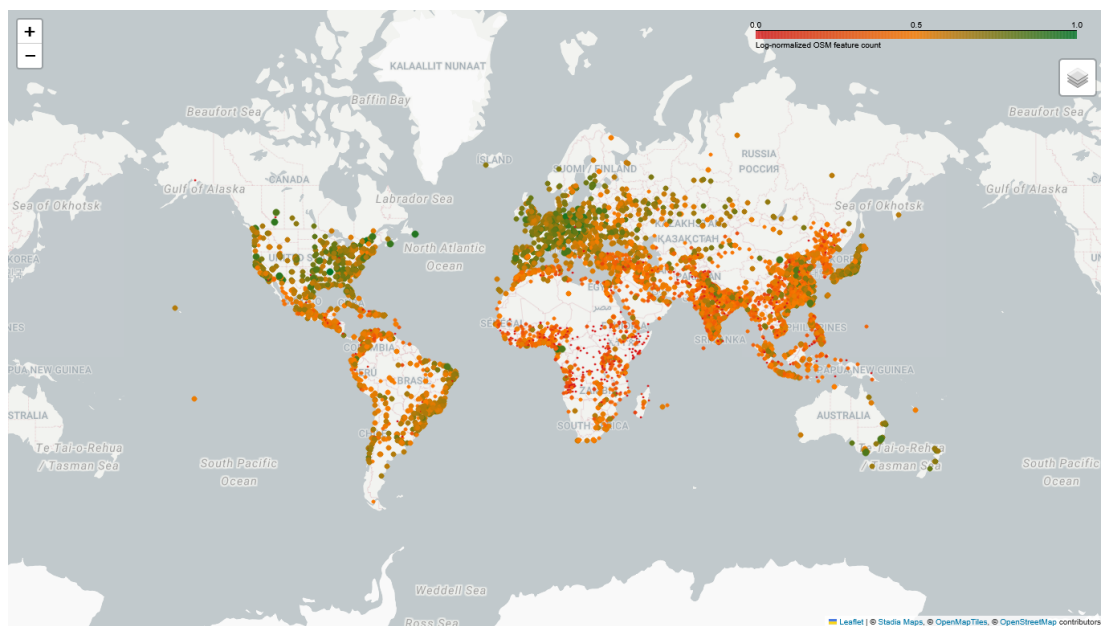


Figure B.2: OSM points fetched on a log-normalised colour scale representing the number of points per region.

Appendix C

Training Figures

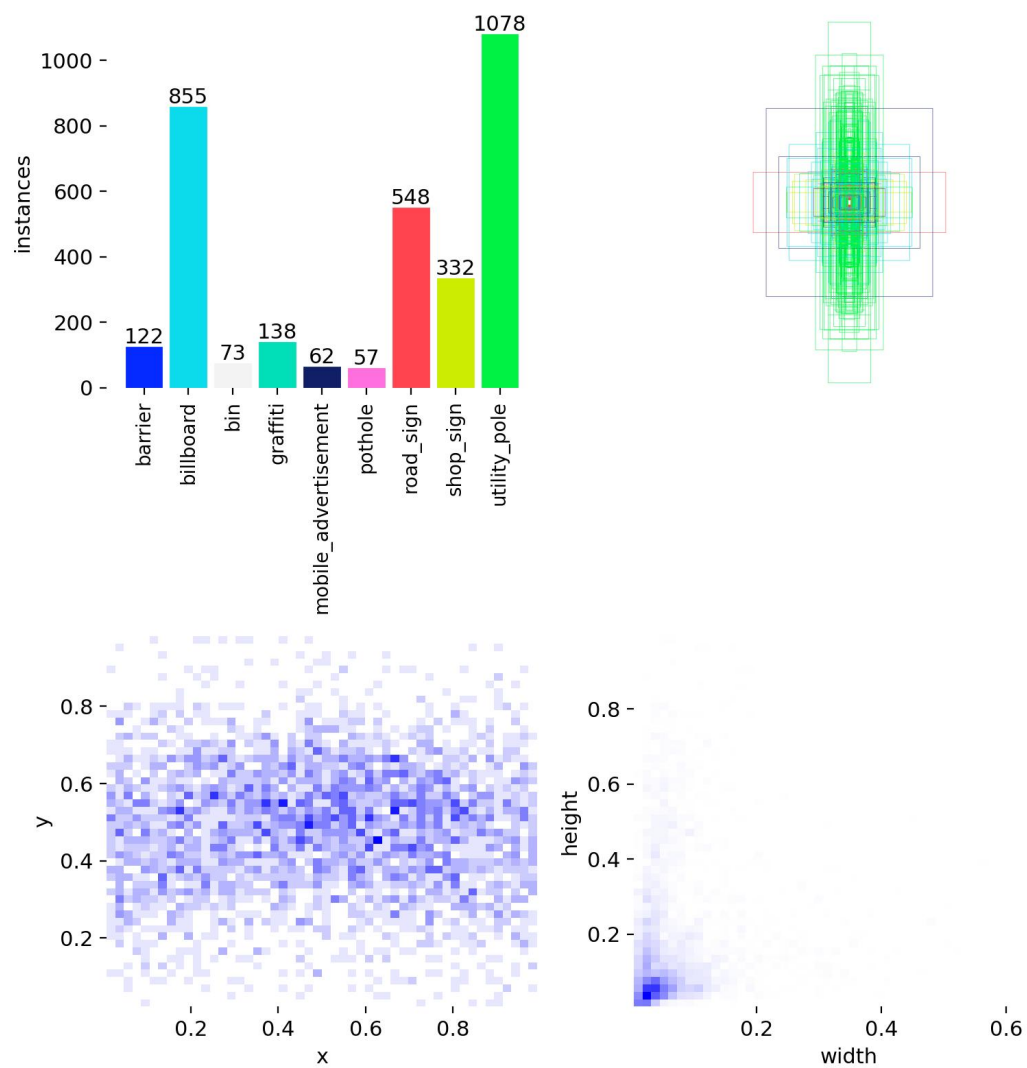


Figure C.1: Annotation statistics

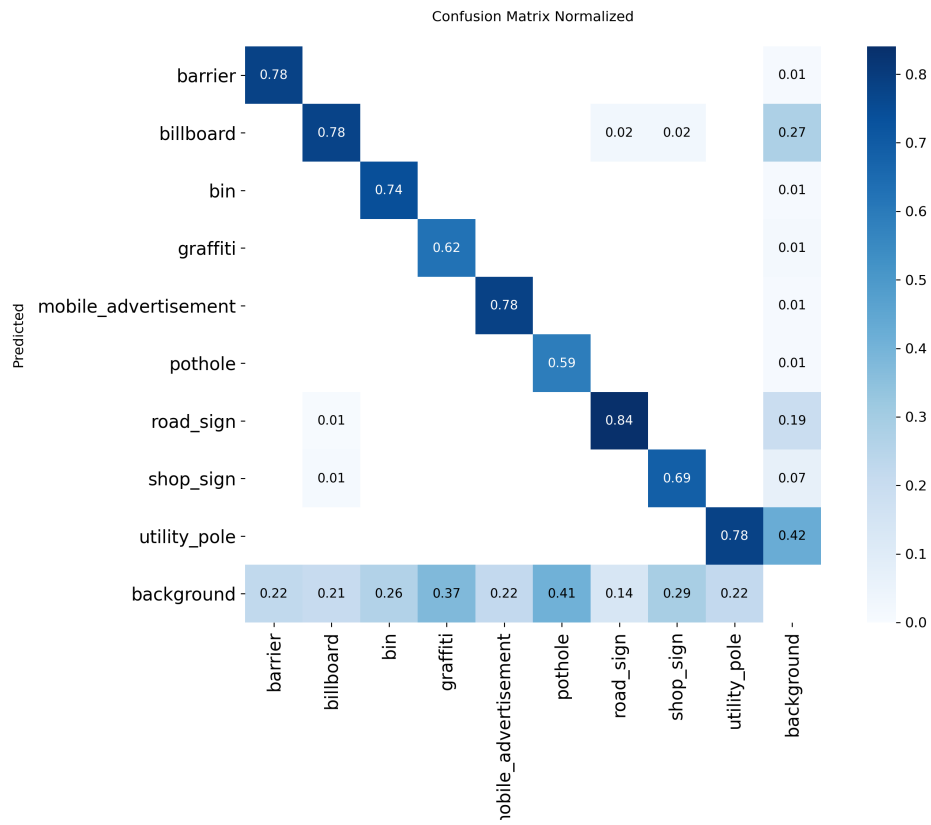


Figure C.2: Normalised confusion matrix

Appendix D

Visual Pollutants

D.1 Billboards



Figure D.1: Example images of objects included in the billboard class.

D.2 Bins



Figure D.2: Example images of objects included in the bin class.

D.3 Utility Poles



Figure D.3: Example images of objects included in the utility poles class.

D.4 Graffiti



Figure D.4: Example images of objects included in the graffiti class.

D.5 Mobile Advertisement



Figure D.5: Example images of objects included in the mobile advertisement class.

D.6 Shop Signs



Figure D.6: Example images of objects included in the shop sign class.

D.7 Road Signs

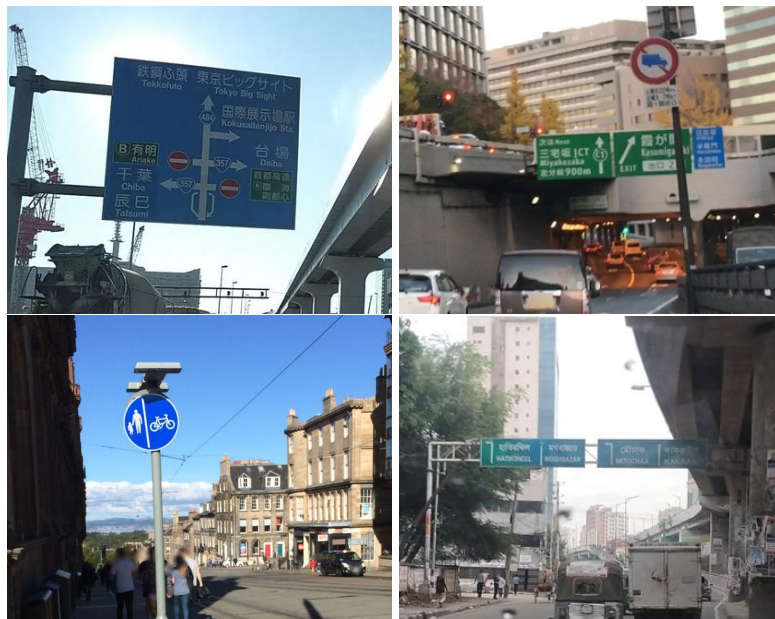


Figure D.7: Example images of objects included in the road sign class.

D.8 Barriers



Figure D.8: Example images of objects included in the barrier class.

D.9 Potholes



Figure D.9: Example images of objects included in the pothole class.

Appendix E

Graphs

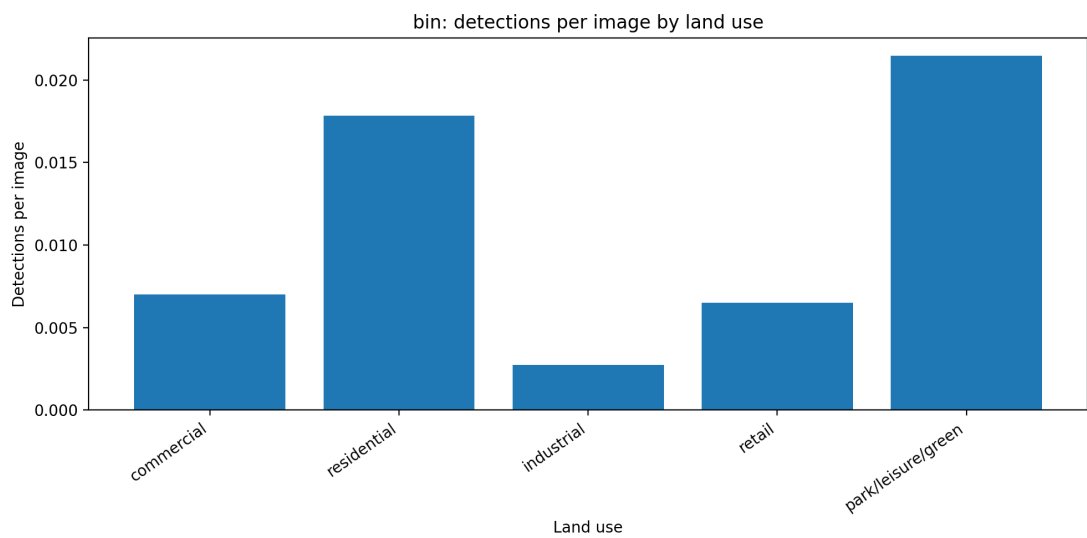


Figure E.1: Bin detections by land use

Appendix F

Codebase Features

This appendix provides an overview of the main top-level scripts and their roles within the system. The most recent version of the codebase can be found at <https://github.com/Fergus-Gault/visual-pollution-tool>.

F.1 `collect_worldcities.py`

Purpose. Scan, score, and run inference on all cities with a population greater than 100,000 listed in `data/worldcities.csv` from SimpleMaps [53].

Overview. The script filters `worldcities.csv` to include only cities exceeding the population threshold and stores the result in a temporary file. This filtered dataset is then processed using a multiprocessing version of the `Pipeline` class.

The `Pipeline` supports the use of multiple API keys to parallelise region scanning. Additionally, the number of concurrent region requests per API key can be configured via `REGION_WORKERS` in `config.py`. In practice, higher levels of parallelism produced inconsistent results, and this parameter is therefore set to 1 by default.

Pipeline Workflow. At a high level, the `Pipeline` performs the following steps:

- Scan the region
- Map images to the region
- Run inference (if a model is loaded)
- Score the region
- Map detections to the region

Region Scanning. Region scanning is handled by the `Scanner` class. For each region, the system:

- Creates a new region entry if it does not already exist
- Retrieves the region shapefile where available

- Falls back to a bounding box if the shapefile is unavailable or excessively large
- Divides the region into subregions and processes them in parallel using worker threads

All APIs (e.g. Mapillary and KartaView) are queried in parallel. Retrieved images are filtered to ensure they lie within the defined region (shapefile or bounding box), and each image URL is validated to confirm accessibility. All valid data are then stored in the database.

F.2 `create_dataset.py`

Purpose. Create a training dataset from annotated images exported via Label Studio.

Overview. This script requires Label Studio to be running. A `LabelStudioClient` is initialised to retrieve image metadata, including image IDs, URLs, task IDs, and associated annotations.

The `DatasetManager` converts bounding box annotations from Label Studio format into a YOLO-compatible format. For each image, a dataset record is created containing the image URL, dimensions, and annotation data.

Images are then downloaded and stored in a new directory under `datasets/`. The dataset is subsequently split into training, validation, and test sets according to predefined ratios, with an attempt to preserve class balance across splits.

Finally, all records are written to a `.ndjson` file, which can be used directly as input to `train.py`.

F.3 `label.py`

Purpose. Upload a sample of images from each country to Label Studio for annotation.

Overview. This script selects a configurable number of images per country from the database. For each image, a payload is constructed containing the image URL and identifier.

If prior inference results exist, the corresponding predictions are included in the payload. These predictions contain the model version, labels, and bounding boxes, allowing pre-annotations to be displayed in Label Studio.

Once uploaded, the images can be reviewed and annotated within Label Studio.

F.4 `run_inference.py`

Purpose. Run model inference on all images that have not yet been processed.

Overview. This script performs inference on images within one or more regions. It accepts optional city and country arguments; if neither is provided, inference is run

across all regions.

The script first retrieves all images that have not yet been processed (i.e. no existing inference results). The `InferenceManager` then spawns multiple worker threads, each responsible for downloading images to a temporary directory and running inference using the specified model.

Predictions are processed to extract bounding boxes, class labels, and confidence scores. These results are stored in the database and linked to the corresponding images. Finally, a detection map is generated and saved for each processed region.

F.5 `score_regions.py`

Purpose. Compute Visual Pollution Index (VPI) scores for one or more regions.

Overview. This script accepts optional city and country arguments; if none are provided, all regions are scored. Scores can be computed either with or without the inclusion of OSM features.

The `Scorer` class retrieves region identifiers and aggregates relevant data, including the number of images and associated detections per region. Regions with insufficient image counts are filtered out based on a predefined threshold to ensure reliability.

VPI scores are then computed using the Counting Categories Ratio and Severity Weighted Score. The resulting scores are written to a CSV file and stored in the database alongside the corresponding region.

F.6 `train.py`

Purpose. Train a YOLO model using a prepared dataset.

Overview. This script trains a YOLO model using a dataset generated by `create_dataset.py`. The model variant can be specified via command-line arguments.

The script requires the path to a `.ndjson` dataset file, which is automatically converted into YOLO's expected format at runtime. Training is then performed using the configured hyperparameters.

Weights & Biases (WandB) is used to track training progress, including metrics such as loss and performance over time. Training hyperparameters can be adjusted in `config.py`.

F.7 `download_data.py`

Purpose. Export all images and associated predictions from the database into a structured dataset.

Overview. This script operates similarly to `create_dataset.py`, but instead of retrieving annotated images from Label Studio, it exports all images and corresponding detections directly from the database.

Due to the large dataset size, the data is partitioned into shards, each stored as a `.tar` archive containing approximately 10,000 images. A corresponding `.ndjson` file is generated to map each image to its shard and to store associated prediction data.

For the dataset used in this project, the total size of the exported data was approximately 280 GB.

F.8 Scripts

`scripts/calculate_billboard_ratio.py` Computes the ratio of billboard detections to images for specified region IDs.

`scripts/delete_detections.py` Deletes all detections and resets image status to *unreviewed*. This is required when re-running inference with a new model.

`scripts/delete_region.py` Deletes a region and all associated data from the database.

`scripts/heatmap_detections_severity.py` Generates the heatmap shown in Figure A.2.

`scripts/map_cities_300_combined_images_osm.py` Generates the map shown in Figure B.1.

`scripts/map_cities_300_images.py` Generates the map shown in Figure 4.2.

`scripts/map_dataset.py` Generates the map shown in Figure 4.1.

`scripts/map_landuse_density.py` Generates the map shown in Figure 7.2.

`scripts/map_region.py` Generates maps showing image locations and corresponding detections within a region.

`scripts/rebalance_dataset.py` Rebalances class distributions across dataset splits, particularly after merging multiple `.ndjson` files.

`scripts/top_pollutant_regions.py` Produces the data presented in Table A.1.

`scripts/world_image.py` Generates the map shown in Figure 8.1.

`scripts/world_osm.py` Generates the map shown in Figure B.2.

`scripts/world_score.py` Generates the map shown in Figure 7.1.