# Green Sense

**Dor Halevy , Adi Inbar and Noam Arian**

**Abstract**

Monitoring the quality of urban green spaces is essential for public health, climate resilience, and cost efficient municipal maintenance, yet cities rarely exploit existing street camera networks for this purpose. Green Sense is a computer vision pipeline that automatically classifies urban vegetation into Healthy, Dried, and Contaminated conditions from street level imagery, targeting viewpoints and scene composition typical of municipal cameras. To overcome the lack of labelled street camera data, the study constructs a synthetic dataset using the FIBO generative model configured to mimic fixed camera geometry, validates perceptual realism with human screening and NIQE scores, and applies a shared evaluation protocol across all models. Three model families are examined: a ResNet50 transfer learning baseline, a ViT B16 Vision Transformer that is further analyzed using Qwen2.5 VL for semantic auditing of difficult cases, and a CLIP-DINOv2 hybrid that fuses vision language and self supervised embeddings with vegetation specific color and texture statistics in a Random Forest classifier. On the synthetic street camera test set, the ResNet50 baseline shows substantial performance degradation, while the ViT B16 markedly improves robustness, and the CLIP-DINOv2 hybrid achieves the strongest results, reaching about 0.97 accuracy and 0.998 ROC AUC with reduced confusion between Dried and Contaminated scenes. These findings indicate that carefully validated synthetic data combined with transformer and hybrid foundation model architectures can provide a practical basis for large scale, camera based monitoring of urban green space quality, which will be further assessed on real municipal feeds in future work.

**Keywords:** Urban green spaces; street level imagery; synthetic dataset; generative AI; vegetation quality classification; Healthy, Dried, Contaminated vegetation; Vision Transformers; CLIP; DINOv2; domain shift; environmental monitoring; municipal maintenance workflows.

## 1. Introduction

Urban green spaces are increasingly recognized as essential urban infrastructure that supports physical and mental health, social cohesion, and environmental resilience in dense cities. Parks and vegetated areas help mitigate urban heat islands, improve local air quality, and provide accessible places for routine physical activity and psychological restoration, making their condition a matter of public health as well as urban design. At the same time, municipalities face growing pressure to monitor and maintain these spaces systematically as they age, intensify in use, and experience stresses such as drought, pollution, and heavy foot traffic. Traditional monitoring practices rely on periodic manual inspections by municipal staff or contractors, which are time consuming, expensive, and difficult to scale consistently across an entire city. In parallel, remote sensing indicators such as the Normalized Difference Vegetation Index offer broad coverage but primarily reflect vegetation quantity and greenness, rather than critical quality attributes such as cleanliness, visible litter, or early signs of vegetation stress at the level experienced by residents. Recent research using eye level smartphone imagery and drone images has demonstrated that computer vision and deep learning can classify vegetation health and surface contamination with promising accuracy, but these approaches either depend on labor intensive manual data collection or specialized aerial surveys that do not readily translate into continuous, citywide operations.

Within this context, there is a clear methodological and practical gap between high level satellite indicators and small scale, manually collected image datasets. Cities already deploy extensive networks of fixed street level cameras, yet these feeds are rarely used to monitor the quality of urban green spaces in an automated way. Existing work on eye level images suggests that residents primarily value cleanliness, good maintenance, and the absence of litter or visible decay, and that these conditions can be framed as discrete visual classes such as Healthy, Dried, and Contaminated. However, models trained on close range smartphone photos may not generalize to the oblique and elevated viewpoints, variable lighting conditions, and mixed built and natural context that characterize typical municipal street cameras. Convolutional neural networks with strong local texture biases, such as ResNet50, can achieve near perfect performance on curated close up datasets, but are vulnerable to domain shift when confronted with synthetic or real street camera imagery where textures, scales, and visual noise differ substantially from the original training domain. There is therefore a need for a scalable and domain robust pipeline that can assess green space quality from street camera viewpoints and that can be integrated into municipal maintenance workflows.

The Green Sense project addresses this need by developing an end-to-end computer vision pipeline for automated assessment of urban green space quality from street level imagery, distinguishing vegetation into three operational categories: Healthy, Dried, and Contaminated. To overcome the lack of labelled street camera data, the project constructs a synthetic dataset using the FIBO visual generative AI model, configured through structured prompts to mimic typical municipal camera perspectives with variation in distance, angle, lighting, and surrounding built environment, and to depict each target condition in a controlled manner. All generated images are manually screened and annotated, and their perceptual realism is further assessed using the Natural Image Quality Evaluator, retaining only images that exceed a predefined quality threshold for model training. Methodologically, the work follows a staged design. First, a baseline classifier replicating the ResNet50 pipeline from the study Green Space Quality Analysis Using Machine Learning Approaches is trained on the synthetic dataset to provide a comparable reference point under the street camera domain. Second, a structured improvement pipeline is applied, combining hyperparameter tuning, evaluation of alternative architectures such as Vision Transformers that model global relationships between image patches, and integration of multimodal representations from models such as CLIP and self supervised visual backbones such as DINO to enhance robustness under domain shift. Third, the refined Green Sense model is evaluated on authentic street camera images provided by a municipal authority using the same labelling protocol, in order to quantify transfer from synthetic to real conditions and assess readiness for deployment in automated green space monitoring and maintenance support systems. Through this approach, the project aims to bridge the gap between experimental deep learning studies on curated imagery and scalable tools capable of continuously tracking the health and cleanliness of urban greenery at the level residents experience in daily life.

## 2. Literature Review

Urban green spaces are a critical component of modern cities, shaping residents' quality of life, social cohesion and public health. As these areas age and intensify in use, municipalities face growing pressure to systematically monitor their cleanliness, vegetation quality and overall maintenance level at scale. In parallel, advances in computer vision and spatial analysis have enabled new approaches to automatically classify and evaluate green spaces in civil environments, moving beyond traditional manual surveys. This literature review will first show the societal importance of clean and well maintained urban greenery and existing monitoring practices, and then examine the main methodological frameworks and models used to classify green spaces from visual and spatial data.

According to Bedimo Rung, Mowen and Cohen, urban green spaces and parks provide accessible settings for regular physical activity, which is associated with reduced risks of obesity, cardiovascular disease, diabetes and other chronic conditions. They also report psychological, social and environmental benefits of green spaces, including improved mood, lower stress and fewer symptoms of depression and anxiety, stronger neighborhood ties and social capital, and contributions to cleaner air and mitigation of urban heat [1].

Kaplan's evaluation of a "vest pocket" park in downtown Ann Arbor shows that even a very small green space in a dense business district functions as a valued refuge that offers quiet, rest and visual relief. Users reported that simply having this planted, shaded spot nearby improved their daily experience of the city, highlighting the strong restorative benefits of accessible greenery in urban cores [2].

According to Madureira et al., the most valued green space characteristics are cleanliness and good maintenance, richness in plant species, the presence of water bodies, sufficient benches and a generally tranquil atmosphere. Cleanliness and maintenance stand out as the single most important attribute across the three cities, more important than park size, parking or high visitor numbers [3].

Greenspaces are increasingly recognized as critical urban infrastructure rather than a "nice to have" amenity, because they support physical and mental health, reduce health inequalities and help cities adapt to challenges such as air pollution, heat and flooding. In response, municipalities and their partners now treat parks and wider green infrastructure as natural capital: they use spatial planning tools, inclusive design, community programs and formal valuation methods to demonstrate benefits and to prioritize investment. Funding typically comes from a mix of local authority budgets, national health and environment strategies, and dedicated mechanisms such as the Community Infrastructure Levy (a planning charge on new developments used to fund local infrastructure, including parks and green spaces) and Section 106 developer contributions (site specific legal agreements where developers fund or provide infrastructure to mitigate the impacts of their projects), as well as external grants from bodies like the National Lottery Heritage Fund and National Trust. While the review stresses that budgets are under pressure, it shows that targeted investment in high quality, accessible greenspace is seen as a relatively low cost way to deliver health, social and environmental priorities [4].

This paper [5] contrast conventional satellite based indicators of urban green cover, particularly the Normalized Difference Vegetation Index (NDVI), with a human scale measure of street greenery that better reflects residents' daily visual exposure. Using Google Street View imagery and a SegNet convolutional network, they segment vegetation at pixel level and compute a Green View Index (GVI) for each street segment, which is then classified into low, medium and high greenery based on expert judgements. This GVI based metric is integrated with space syntax measures of street accessibility to derive indicators of "daily accessed greenery," capturing where people are most likely to experience greenery during routine pedestrian and commuting trips. When compared with NDVI at planning area scale, the daily accessed greenery indicators show only partial correspondence, leading the authors to conclude that top down NDVI alone cannot represent the benefits actually enjoyed by city residents and should be complemented by street level, machine learning based assessments of visible greenery in planning practice.

In the study "Deep Green Diagnostics Urban Green Space Analysis Using Deep Learning and Drone Images," [6] the authors present an operational framework for fine grained assessment of urban vegetation health and surface contamination using high resolution aerial imagery. The work addresses the limitation of satellite based indices such as NDVI that only capture overall vegetation quantity and cannot distinguish between healthy, dry or polluted green areas. To overcome this, the authors construct a dataset of geo referenced RGB images acquired by a DJI Phantom 4 drone flown over four distinct urban environments in Mexico. The raw images are tiled into patches of 200 by 200 pixels, from which 9 901 samples are manually labelled into eight classes that jointly encode vegetation condition healthy, dry, unhealthy or no vegetation and the presence or absence of visible solid waste contamination. On this dataset they train a convolutional neural network followed by a multilayer perceptron classifier that learns discriminative visual features directly from the image patches and assigns each tile to one of the eight semantic categories. The resulting system achieves around 72 percent accuracy on a held out test set, with particularly strong performance for healthy vegetation and no vegetation classes, and demonstrates reasonable transferability when retrained on small samples from a new area. Overall, this article shows that deep learning applied to drone imagery can provide spatially detailed, operational information on both vegetation condition and urban cleanliness, and offers an open source prototype that municipal or health authorities can adapt for monitoring and decision support.

The article "Green Space Quality Analysis Using Machine Learning Approaches" [7] develops a full pipeline for assessing the visual quality of urban green spaces using machine learning on eye level images. It defines quality through conditions such as cleanliness, maintenance and dryness, builds a labelled dataset with three classes healthy, dried and contaminated, and compares several transfer learning image classification models, selecting ResNet50 as the best performing model and deploying it as a practical web tool for green space monitoring. Methodologically, the study follows the CRISP DM lifecycle, collecting 944 smartphone photos of Kuala Lumpur green spaces in 2022, manually annotating them into the three quality classes, and then addressing class imbalance through data augmentation using flips, rotations, brightness and contrast changes and CLAHE to obtain a balanced dataset of 986 images. The augmented images are resized to 224 by 224 pixels and split into training, validation and test sets, and nine ImageNet pretrained convolutional networks ResNet50, ResNet101, DenseNet201, VGG16, VGG19, XCeption, MobileNet, InceptionResNetV2 and EfficientNetB7 are fine tuned using Adam optimisation, categorical cross entropy loss and learning rate reduction, with performance evaluated via accuracy, precision, recall, F1 score, Cohen Kappa and ROC AUC. ResNet50 is chosen as the overall best model, reaching about 99 percent accuracy together with almost perfect precision, recall and ROC AUC, and therefore selected for deployment in a Streamlit web application that allows users to upload or capture a photo and receive an automatic quality label, while the authors conclude that such transfer learning pipelines on relatively small eye level datasets can already deliver operational tools for green space maintenance support and point to mobile deployment and segmentation based models as important future directions.

## 3. Methodology

This chapter outlines the methodological process of the Green Sense project as an end to end supervised image classification study designed to assess urban green space quality from street level municipal camera viewpoints. The chapter provides a procedural map of how the study is executed, from problem operationalization through dataset construction, model development, evaluation, and planned external validation. Detailed technical specifications and implementation level decisions for each component are presented in subsequent chapters.

The target task is three class classification of urban vegetation condition from RGB images. Each image depicts public green spaces within a realistic urban context that may include sidewalks, roads, buildings, and street furniture, reflecting the scene composition typically captured by municipal cameras. The model predicts one of three operational categories. Healthy denotes vegetation that appears vital and well maintained, with predominantly green coverage and no visible waste. Dried denotes vegetation showing clear water stress or decay, expressed through brown or yellow coloration, patchiness, exposed soil, or a generally withered appearance, without obvious solid waste. Contaminated denotes scenes where vegetation is present but visibly affected by litter or pollution on or directly adjacent to the green surface, including items such as bottles, bags, cans, or mixed debris. These operational definitions are used consistently for synthetic data labelling and for the planned real world validation set to ensure comparability across evaluation stages.

The methodology uses three dataset sources with distinct roles. A reference smartphone dataset from prior work is used as conceptual grounding and as a benchmark reference for baseline model design, representing close-range human perspective imagery and illustrating performance under a viewpoint domain different from municipal cameras. The primary dataset for Green Sense, which is a synthetic street camera dataset generated to emulate municipal camera geometry and scene composition. This dataset is constructed to provide balanced representation of the three classes under varied lighting, distance, and urban context, and serves as the main resource for training, internal validation, and controlled comparison between model families. A municipal street camera dataset is planned as an external test set and will consist of sampled frames from city cameras in which green spaces are visible and can be labelled under the same definitions. This dataset is reserved exclusively for final evaluation and is not used for model training or tuning, enabling an external validity assessment of transfer from synthetic to real world imagery.

Because labelled municipal street camera imagery is not available at scale, Green Sense constructs its training domain using synthetic generation. Image generation was performed using FIBO [8], a visual generative AI model developed by Bria.ai.

All candidate synthetic images undergo a two layer curation procedure combining human annotation and quantitative quality screening. First, images are manually screened and annotated, which provides decision rules for borderline cases and supports consistent application of the class definitions. Images that cannot be confidently labelled are removed rather than forced into a category, and when multiple annotators are involved, disagreements are resolved by consensus to reduce label noise. Second, perceptual image quality is assessed using the Natural Image Quality Evaluator metric. NIQE [9] is computed for each synthetic image to filter out low realism generations and severe artifacts, and only images meeting the predefined quality threshold are retained for modelling. This step reduces the likelihood that models learn synthetic failure modes rather than semantically meaningful cues associated with the target classes.

After curation, the synthetic dataset is partitioned using a stratified split to maintain balanced class proportions across training, validation, and internal test sets. The validation set is used for hyperparameter selection and model selection, while the internal test set is held out for final reporting of synthetic domain performance. In addition to the standard internal test set, a challenging synthetic subset is constructed to simulate difficult street camera conditions such as extreme illumination, stronger clutter, occlusions, and reduced visibility of contamination cues. This subset is used to evaluate robustness and to compare model behaviour under domain stress.

The first modelling stage replicates a ResNet50 based baseline consistent with prior literature. The ResNet50 model uses ImageNet pretraining and a three class classification head, and it is trained on the curated synthetic dataset using a standardized preprocessing and augmentation pipeline. Its performance is measured under the same evaluation metrics used for all subsequent models, establishing a controlled reference point and quantifying the extent of domain shift challenges introduced by street camera viewpoints.

The second modelling stage introduces a Vision Transformer architecture as an alternative that may improve robustness to viewpoint and texture shift by modelling global spatial relationships between patches.

The initial ViT model is fine tuned using the same train validation split and comparable optimization settings to ensure that differences in performance reflect architecture and representation rather than inconsistencies in evaluation. To guide targeted refinement, the study incorporates Qwen2.5 [10] VL as a semantic auditing tool applied to misclassified and borderline samples produced by the ViT model. These samples are analyzed using structured prompts that request scene interpretation and class cue attribution, with the objective of identifying systematic confusion patterns, generation artifacts, or annotation ambiguities, particularly between Dried and Contaminated. Insights from this audit inform revisions to data curation decisions and training procedures, after which the ViT model is fine tuned again under the same protocol. This produces an improved ViT configuration intended to better align learning with the operational class definitions and reduce domain specific failure cases.

In parallel to end to end deep learning classifiers, Green Sense evaluates a hybrid approach based on feature embeddings extracted from foundation models. CLIP embeddings provide semantic alignment between images and class descriptive text prompts, while DINOv2 embeddings provide robust self supervised visual representations that often transfer well across domains. These embeddings are supplemented with interpretable vegetation related statistics derived from color based masking. All features are fused into a single representation per image, and a downstream classical classifier is trained on these vectors to predict the three classes. This pathway provides a complementary modelling strategy that may generalize differently than fully fine tuned deep networks and enables additional interpretability through feature importance and embedding similarity patterns.

All model families are evaluated under a shared protocol to ensure comparability. Training uses early stopping based on validation performance, and model selection is performed using a balanced metric such as macro averaged F1 score or Cohen kappa to reflect operational reliability across all conditions and reduce sensitivity to class imbalance. Final internal evaluation is conducted on the held out synthetic test set and on the challenging synthetic subset using the same metrics across models, enabling a fair comparison between the baseline ResNet50, the refined ViT approach, and the hybrid CLIP DINO classifier. The output of this stage is a selected Green Sense model configuration supported by comparative evidence across these alternatives.

The final stage is designed to assess transfer from synthetic training to real municipal camera conditions. Frames containing visible public green spaces will be sampled from municipal camera feeds and processed under privacy preserving constraints, including cropping and anonymization where required. Images will be labelled using the same class definitions and annotation guide used for the synthetic dataset. The selected Green Sense model will then be applied to this real world dataset without additional training, and performance will be evaluated using the same metric suite. This step provides an external validity assessment and supports an evidence based estimate of readiness for deployment in municipal monitoring workflows.

## 4. Generative

This chapter describes the construction of a synthetic generative dataset that transforms the original baseline images into street camera view representations of urban green spaces. It then outlines the validation procedure used to ensure that the generated images remain realistic and that their inherited labels for Healthy, Dried, and Contaminated conditions remain consistent and reliable.

### *4.1 Synthetic generative Dataset*

FIBO was adopted as the underlying visual generation model due to a combination of controllability, determinism, scalability, and system-level integrability, which are critical for structured and repeatable visual generation workflows. FIBO provides high controllability over visual attributes, enabling precise manipulation of factors such as camera angle, composition, lighting, and color through text or JSON-formatted inputs. FIBO is the first JSON-native text-to-image model, where a short prompt is converted into a structured JSON schema that explicitly specifies each visual factor. In addition, FIBO supports an image-based inspiration mode, in which related visual outputs are generated from an input image. This is achieved by extracting structured guidance from the reference image while maintaining a controlled degree of visual variation. This mechanism enables balancing fidelity to the source image with creative diversity, without sacrificing controllability. Finally, FIBO is natively compatible with Vision-Language Models (VLMs). This capability allows visual generation to function as a component within broader multimodal decision-making pipelines, supporting effective reasoning for error handling and precise task execution.

The generation process is designed to approximate fixed municipal camera conditions by controlling viewpoint properties and scene content, including elevated camera height, downward oblique angle, wide field of view, and realistic urban surroundings so that vegetation appears at plausible scale relative to roads and sidewalks. Generation is class conditioned through prompt templates. Healthy prompts enforce well maintained vegetation with no visible debris, Dried prompts enforce vegetation stress characteristics such as discoloration and sparsity, and Contaminated prompts explicitly include visible litter objects near or on vegetation. For each of the different classes, we employ the image-based inspiration mode using reference images drawn from the dataset associated with the primary study. This approach enables the construction of a class-specific generated dataset in which the distributions of visual properties such as color composition, object types, and overall visual characteristics closely match those of the corresponding reference images. As a result, the generated datasets exhibit comparable visual distributions across classes, allowing for a more reliable and fair comparison of downstream results in subsequent evaluations.

Prompts are authored in both natural language form and structured JSON form to improve controllability over lighting, composition, palette, and object inclusion. Image generation proceeds in batches, candidate images are reviewed, and refinement prompts are applied when class attributes are insufficiently expressed or when unwanted artifacts dominate the scene. The output of this stage is a pool of candidate synthetic images for subsequent curation. Figure 1 illustrates examples of this class conditioned generation process, showing original close up scenes and their corresponding synthetic street camera views.

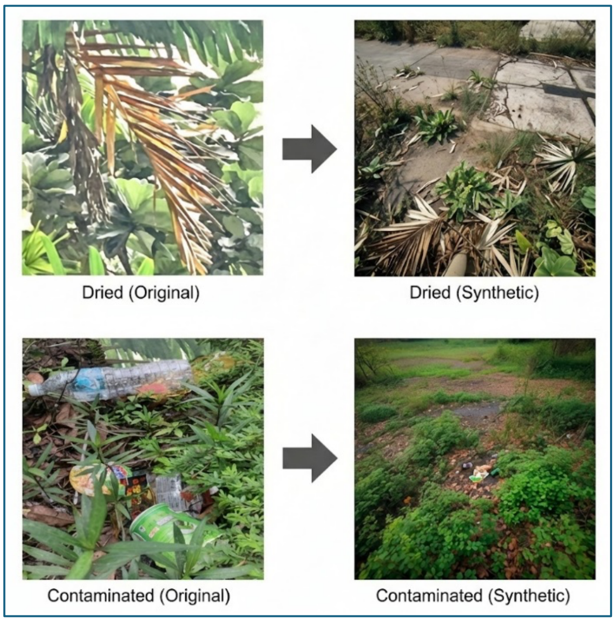

**Figure 1.** Examples of class conditioned image generation, showing original close up Dried and Contaminated scenes (left) and their corresponding synthetic street camera views (right).

### *4.2 Validation of Dataset*

Validating the perceptual quality of generated images is essential to ensure that the synthetic street view dataset remains visually realistic and that its assigned labels remain trustworthy, especially under the substantial domain shift from the original eye level close up imagery to a wider street camera viewpoint. In this work, validation is performed in two stages. A human screening step first verifies scene plausibility and label correctness, and NIQE, the Natural Image Quality Evaluator, is then applied as an objective no reference measure of image naturalness, enabling consistent threshold based filtering.

NIQE is a no reference image quality metric that estimates perceptual naturalness from the image itself, without relying on human opinion scores. The method assumes that real world photographs follow stable natural scene statistics and that artifacts or unrealistic textures shift these statistics in measurable ways. Practically, NIQE computes natural scene statistics features from local 96×96 luminance patches, fits these features with a multivariate Gaussian model, and measures the distance to a reference model learned from high quality natural images; lower scores indicate more natural looking images.

Table 4.1 reports NIQE statistics for each class in the original baseline images and in the corresponding synthetic street view images. Across all three classes, the synthetic images achieve lower mean and median NIQE scores than the original images, with substantially reduced variance, indicating higher statistical naturalness and more consistent perceptual quality. For example, the Dried class mean NIQE decreases from 10.70 in the original set to 8.26 in the synthetic set, while the Contaminated class mean decreases from 11.52

to 9.45. Combined with manual screening that confirms scene realism and label consistency, these results support that the synthetic dataset meets the defined quality standards and is suitable for reliable model training and evaluation under street camera conditions.

**Table 4.1: NIQE statistics by class and dataset type.**

| | Dataset type | Mean NIQE | Median NIQE | Std. dev. NIQE |
|---|---|---|---|---|
| *Healthy* | Original | 10.17 | 9.55 | 2.59 |
| | Synthetic | 8.94 | 8.96 | 1.05 |
| *Dried* | Original | 10.70 | 10.37 | 2.74 |
| | Synthetic | 8.26 | 8.21 | 0.65 |
| *Contaminated* | Original | 11.52 | 11.05 | 2.715 |
| | Synthetic | 9.45 | 9.49 | 1.01 |

## 5. Models

This chapter presents the complete model development pipeline for Green Sense, documenting the evolution from a baseline convolutional architecture through Vision Transformers to a hybrid foundation model approach. The progression is structured around a central challenge: achieving robust classification performance under the domain shift from close-up smartphone imagery to synthetic street camera viewpoints that differ substantially in scale, angle, and background complexity. Three model families are evaluated under identical data splits and metrics to enable direct comparison. Section 5.1 establishes the ResNet50 baseline and quantifies its sensitivity to viewpoint changes, Section 5.2 introduces Vision Transformers as a global context modeling alternative that improves robustness, and Section 5.3 presents a hybrid architecture that fuses CLIP and DINOv2 embeddings with interpretable vegetation statistics to achieve the strongest performance while retaining practical interpretability.

### *5.1 Baseline model*

ResNet50 is a 50-layer deep convolutional neural network built on residual learning, where skip connections enable effective gradient flow through very deep architectures. It is selected as the Green Sense baseline because it serves as the best-performing model in the reference study and is widely adopted as a transfer learning backbone, combining strong representational capacity with practical computational cost. The architecture consists of an initial stem layer followed by four stages of residual blocks (conv2_x through conv5_x) that progressively increase feature depth from 64 to 2048 channels while reducing spatial resolution, culminating in global average pooling that produces a 2048-dimensional feature vector for classification.

#### 5.1.1 Implementation and Training Protocol

The baseline uses ImageNet pretrained ResNet50 with only the final fully connected layer replaced to produce three class outputs (Healthy, Dried, Contaminated) with softmax activation. All convolutional layers retain their pretrained weights and are fine tuned end to end rather than frozen. Images are resized to 224×224 pixels and normalized using ImageNet channel statistics.

Data augmentation follows the reference study protocol and includes random horizontal flips, 90 degree rotations, brightness and contrast adjustments, and CLAHE histogram equalization. The dataset is split using stratified sampling into 70% training, 20% test, and 10% validation sets. Training uses the Adam optimizer with categorical cross entropy loss and a low learning rate (1e-4), with early stopping triggered by validation accuracy to prevent overfitting.

Performance is evaluated using accuracy, macro averaged precision, recall, F1 score, Cohen's kappa, and macro one vs rest ROC AUC. This metric suite matches the reference study and is applied consistently across all model families to enable direct comparison. The implementation is built in PyTorch using the torchvision ResNet50 module.

#### 5.1.2 Baseline behavior and limitations

On the original close up smartphone dataset, the reproduced ResNet50 baseline achieves very high accuracy with strong precision and recall across all three classes, confirming that the implementation matches the reference study. When the same model is applied to the synthetic street camera dataset, performance drops markedly, indicating sensitivity to changes in viewpoint, scale, and background context, a known challenge for convolutional networks under domain shift.

## *5.2 ViT model*

Vision Transformers [11] extend the Green Sense modelling pipeline beyond convolutional networks by replacing local convolutions with global self attention over image patches. In this chapter, Green Sense adopts the ViT B16 configuration as an alternative backbone to address the performance drop observed when the ResNet50 baseline is applied to synthetic street camera imagery.

### 5.2.1 Motivation and attention mechanism

The ResNet50 baseline performs strongly on the original close up image dataset, but its reliance on local texture patterns makes it sensitive to changes in viewpoint, scale, and background context introduced by synthetic street camera imagery. Vision Transformers offer a different inductive bias by representing each image as a sequence of patch tokens and updating these tokens through global self attention, so that every region can directly use information from any other region in the frame. In Green Sense, this property is used to model how vegetation, litter, and built elements co occur across the entire image, allowing the model to interpret cues such as dried grass next to sidewalks or contaminated vegetation near road edges as context dependent patterns rather than isolated textures.

Self attention computes how strongly each patch should attend to every other patch when forming its next representation, producing context aware features that integrate information from relevant regions anywhere in the image. This global interaction is particularly important under street camera conditions, where local texture statistics can shift due to scale changes or generative artifacts while spatial layout and object co occurrence remain informative.

### 5.2.2 ViT B16 architecture and training

Green Sense uses the ViT B16 configuration from torchvision, initialized with ImageNet pretrained weights and fine tuned for three class green space quality classification. For an RGB image resized to 224 by 224, the model partitions the input into non overlapping 16 by 16 patches, projects each patch to a 768 dimensional embedding, adds a learned class token and positional embeddings, and processes the resulting token sequence through 12 Transformer encoder layers with 12 attention heads. After the final encoder layer, the transformed class token serves as a global image representation and is passed to a project specific linear layer that maps 768 features to three outputs, followed by softmax to produce class probabilities for Healthy, Dried, and Contaminated conditions.

ViT B16 is trained as an end to end classifier on the curated synthetic street camera dataset using the same stratified split as the baseline, with 70 percent training, 10 percent validation, and 20 percent test. All images are resized to 224 by 224, normalized with ImageNet statistics, and augmented with flips, 90 degree rotations, brightness and contrast perturbations, and CLAHE to improve robustness to illumination and viewpoint variability. Optimization uses the Adam optimizer with cross entropy loss, a low learning rate, ReduceLROnPlateau scheduling, and early stopping based on validation accuracy to limit overfitting on synthetic data.

### 5.2.3 Semantic auditing with Qwen2.5 VL

Although ViT B16 improves robustness on synthetic imagery, borderline scenes still cause confusion, especially between Dried and Contaminated. Qwen2.5 VL, a pretrained vision language model, is therefore used as a semantic auditing tool to analyze misclassified or low confidence samples rather than as a replacement classifier. For selected images where ViT predictions disagree with ground truth or show diffuse probabilities, structured prompts request a scene description, visual evidence for each class, and plausible reasons for ambiguity such as shadows, clutter, or small high contrast objects.

The resulting explanations are reviewed for recurring patterns, for example dried grass that resembles scattered litter at street camera resolution or synthetic scenes with unrealistically uniform textures. These insights inform targeted refinements in data curation and training, including stricter labelling guidelines for Dried versus Contaminated and adjustments to synthetic generation prompts that produce visually ambiguous outputs.

### 5.2.4 Performance and role

Fine tuned ViT B16 achieves higher results on the synthetic street camera dataset than the ResNet50 baseline, with improved robustness under viewpoint and scale variation. Gains are particularly evident in cases where class evidence depends on global context rather than local texture alone. Together with the semantic auditing findings, these results support ViT B16 as the primary deep classification backbone in Green Sense, providing a stronger foundation before introducing hybrid representation frameworks in subsequent chapters.

## *5.3 CLIP and DINOv2 hybrid model*

T The CLIP [12] and DINOv2 [13] hybrid model forms the final Green Sense configuration, replacing a purely end to end deep classifier with a feature based pipeline built on pretrained foundation models. It combines CLIP's vision language embeddings, which align images with textual descriptions, with DINOv2's self supervised visual embeddings that capture robust scene structure without labels.

### 5.3.1 Motivation and background

T The ResNet50 baseline shows a sharp performance drop when moved from close range smartphone imagery to synthetic street camera views, while the fine tuned ViT B16 model remains strong on the same synthetic domain, indicating that global attention based backbones handle viewpoint and texture changes more reliably than conventional convolutional networks. Building on this, the CLIP and DINOv2 hybrid model is introduced not to replace ViT B16, but to explore whether large scale vision language and self supervised pretraining can further improve robustness and interpretability for greenspace monitoring.

CLIP provides joint image-text embeddings trained with a contrastive objective, which allows images of parks to be aligned directly with natural language descriptions such as healthy vegetation or park with visible trash and supports prompt based semantic scoring of each scene. DINOv2 contributes complementary self supervised visual embeddings learned from augmented image views without labels, capturing object structure and textures that transfer well across domains, including variations in lighting and camera geometry typical of street level imagery. In Green Sense, the hybrid model combines these semantic and visual embeddings with simple vegetation statistics so that classification can exploit both high level textual alignment and robust visual cues while remaining more interpretable than a single deep end to end network.

### 5.3.2 Training, prediction

During training, each labelled greenspace image is passed through a fixed feature extraction pipeline that is identical to the one used at prediction time. At scene level, CLIP produces an image embedding and similarity scores to prompts describing healthy, dried, and contaminated parks, while DINOv2 produces a second embedding that captures detailed visual structure without text supervision. Vegetation regions are then detected using color thresholds, cropped, and encoded by both CLIP and DINOv2, and these vegetation embeddings are pooled per image and combined with simple statistics such as mean green intensity, intensity variance, edge density, number of vegetation regions, and overall vegetation coverage. All scene level embeddings, vegetation embeddings, prompt similarity scores, and vegetation statistics are concatenated into a single feature vector, scaled with a normalization transform learned on the training set, and used to train a Random Forest classifier to predict Healthy, Dried, or Contaminated.

At prediction time, new images are processed through the same feature pipeline and normalized with the stored transform before being passed to the trained Random Forest, which outputs a class probability distribution. In parallel, CLIP prompt similarities are grouped into three class level semantic scores and normalized, and the final decision combines these semantic scores with the Random Forest probabilities: clear prompt consensus can override a low confidence classifier output, while in more ambiguous cases the two sources are blended into an ensemble distribution.

### 5.3.3 Performance

On the synthetic street camera dataset, this CLIP and DINOv2 hybrid model achieves the highest accuracy, macro F1 score, Cohen kappa, and ROC AUC among all tested configurations, slightly improving on ViT B16 and clearly outperforming the ResNet50 baseline. The fusion of semantic CLIP features, robust DINOv2 embeddings, and explicit vegetation statistics particularly reduces confusion between Dried and Contaminated scenes, and provides interpretable outputs such as vegetation coverage and prompt based scores that are useful for municipal greenspace monitoring.

## 6. Results

This chapter reports the performance of the best configuration from each model family on the synthetic street camera test set. All three models, ResNet50, ViT B16, and the CLIP DINOv2 hybrid, were trained and fine tuned on the synthetic street camera dataset before evaluation, ensuring that reported results reflect each architecture at its best under the same training conditions. The focus is on operational metrics that matter for municipal greenspace monitoring, including accuracy, macro precision, macro recall, macro F1, Cohen kappa, and macro one versus rest ROC AUC.

### *6.1 Models on synthetic street camera test set*

This subsection presents the results of all three Green Sense models on the held out synthetic street camera test set, showing how accurately each model classifies Healthy, Dried, and Contaminated vegetation under synthetic street camera conditions.

#### 6.1.1 ResNet50 baseline

On the synthetic street camera test set, the ResNet50 baseline reaches precision of 0.57 for Healthy, 0.80 for Dried, and 0.23 for Contaminated scenes, with corresponding recall values of 0.90, 0.56, and 0.16. This pattern shows that while the model retrieves most Healthy examples, it misses many Dried and especially Contaminated parks, and when it predicts Contaminated it is often incorrect, reflecting substantial confusion between the degraded classes.

#### 6.1.2 ViT B16

On the same synthetic test set, ViT B16 attains precision of 1.00 for Dried, 0.87 for Contaminated, and 0.86 for Healthy scenes, with recall values of 0.82, 0.89, and 1.00 respectively. In practice, this means the transformer both finds nearly all examples in each class and keeps false positives low, especially for the challenging Dried and Contaminated conditions where the ResNet50 baseline struggles.

#### 6.1.3 CLIP-DINOv2 hybrid model

On the synthetic street camera test set, the CLIP–DINOv2 hybrid model reaches precision of 0.99 for Healthy, 0.98 for Dried, and 0.95 for Contaminated scenes, with recall values of 1.00, 0.94, and 0.95 respectively.

#### 6.1.4 Summary

Across the synthetic street camera test set, performance improves stepwise from ResNet50, through ViT B16, to the CLIP and DINOv2 hybrid, which achieves the highest precision and recall for all three classes, as summarized in Table 6.1. Notably, ViT B16 also delivers strong results across all metrics, and both transformer based configurations clearly outperform the ResNet50 baseline.

**Table 6.1. Performance of Green Sense models on the synthetic street camera test set.**

| Model | Accuracy | Precision | Recall | F1 | Cohen kappa | ROC AUC |
|---|---|---|---|---|---|---|
| ResNet50 baseline | **0.581** | **0.604** | **0.573** | **0.557** | **0.375** | **0.796** |
| ViT B16 | **0.905** | **0.910** | **0.905** | **0.902** | **0.856** | **0.972** |
| CLIP-DINOv2 hybrid model | **0.972** | **0.972** | **0.972** | **0.972** | **0.958** | **0.998** |

### 6.2 *Evaluation on real world images*

This subsection introduces a small real image evaluation set, consisting of 33 manually labelled photos of urban vegetation collected from multiple sources and annotated as Healthy, Dried, or Contaminated using the same definitions as the synthetic dataset. It uses this set to test the ResNet50, ViT B16, and CLIP DINOv2 models without additional training, providing a first indication of how well a pipeline tuned on synthetic street camera images transfers to genuinely observed municipal like scenes.

#### 6.2.1 ResNet50 baseline

The ResNet50 baseline reaches precision of 0.63 for Healthy, 0.75 for Dried, and 0.60 for Contaminated scenes, with recall values of 0.86, 0.38, and 0.55 respectively. While the model recovers most Healthy examples, it misses a large portion of Dried and Contaminated parks, with several Contaminated scenes misclassified as Healthy, indicating that the baseline's texture dependent features do not transfer reliably to uncontrolled real world conditions.

#### 6.2.2 ViT B16

ViT B16 reaches precision of 0.65 for Healthy, 0.40 for Dried, and 0.73 for Contaminated scenes, with recall values of 0.79, 0.25, and 0.73 respectively. Contaminated scenes are classified with reasonable reliability, while Dried vegetation is the weakest class, with many instances misclassified as Healthy

#### 6.2.3 CLIP–DINOv2 hybrid model

On the 33 real world images, the CLIP DINOv2 hybrid model reaches precision of 0.92 for Healthy, 0.78 for Dried, and 0.91 for Contaminated scenes, with recall values of 0.93, 0.88, and 0.91 respectively. Despite no retraining on real images, these results reflect a meaningful transfer from the synthetic domain

#### 6.2.4 Summary

Across the 33 real world images, the CLIP DINOv2 hybrid model achieves the highest precision and recall for all three classes, as summarized in Table 6.2, clearly outperforming both ResNet50 and ViT B16, which reach similar and considerably lower accuracy on this set. Unlike the synthetic domain where ViT B16 showed a strong gain over ResNet50, both models struggle comparably under real uncontrolled conditions, while the hybrid model demonstrates a more robust transfer from synthetic training to genuine municipal like scenes.

**Table 6.2. Performance of Green Sense models on the synthetic street camera test set.**

| Model | Accuracy | Precision | Recall | F1 | Cohen kappa | ROC AUC |
|---|---|---|---|---|---|---|
| ResNet50 baseline | **0.636** | **0.661** | **0.592** | **0.600** | **0.419** | **0.777** |
| ViT B16 | **0.636** | **0.591** | **0.588** | **0.582** | **0.426** | **0.805** |
| CLIP-DINOv2 hybrid model | **0.849** | **0.855** | **0.848** | **0.848** | **0.770** | **0.899** |

## 7. Conclusions

Green Sense establishes a clear path for leveraging existing municipal camera networks to monitor urban green space quality at scale, transforming passive surveillance feeds into actionable insights about cleanliness, vegetation health, and maintenance needs. The project yields five key lessons with broader implications for computer vision in civic applications.

Synthetic data can serve as a reliable foundation for real world deployment when properly validated. By using FIBO to generate street camera like images under controlled conditions for viewpoint, lighting, and class attributes, and then applying human screening alongside NIQE filtering, Green Sense demonstrates that synthetic datasets can capture realistic domain characteristics without the cost and scarcity of manual labelling. This approach is particularly valuable for municipal settings where labelled camera imagery is limited.

Vision transformers excel in domain shift scenarios by prioritising global composition over local textures. Unlike convolutional networks that rely on fine grained textures and short range patterns, transformers model long range relationships between image patches, making them inherently more robust to changes in scale, angle, and background context typical of street cameras. In the Green Sense setting, this inductive bias aligns well with wide angle scenes where evidence for Healthy, Dried, or Contaminated conditions is distributed across vegetation, hardscape, and surrounding urban elements rather than confined to small local regions.

Vision language models illuminate the behind the scenes reasoning of classifiers through semantic auditing. When applied to misclassified synthetic images, models like Qwen2.5 VL articulate why Dried grass might be confused with Contaminated debris, for example shadows mimicking litter at oblique angles, or generation artifacts blurring class boundaries. This explanatory power turns statistical errors into interpretable insights, enabling targeted fixes to data, labels, or training protocols rather than indiscriminately changing architectures.

Vision language models enable flexible scene classification and detection through natural language prompts. CLIP's shared image text space allows querying feeds with evolving descriptions like "emerging litter near vegetation" or "drought stressed browning under current lighting," supporting dynamic adaptation to seasonal threats, public complaints, or policy priorities without model retraining. This capability extends Green Sense beyond a fixed three class classifier toward a more general scene understanding tool that can answer task specific questions about urban greenery as operational needs change over time.

Unsupervised feature extraction methods like DINOv2 improve hybrid models by providing domain robust visual representations. DINOv2 captures shape, texture, and structural invariances across viewpoints and lighting, complementing CLIP's semantics and color statistics in a fused pipeline classified by a lightweight Random Forest. This hybrid design yields peak performance with modest computational cost.

Taken together, these five principles, validated synthetic data, transformer architectures tuned for global structure, vision language models for diagnosis and querying, and unsupervised visual feature fusion, offer a practical blueprint for building operational vision systems in cities that face both data scarcity and complex, shifting visual domains.

## 8. Discussion and Future Work

Green Sense demonstrates that validated synthetic data, transformer architectures, and hybrid foundation model representations can jointly form a scalable and operational framework for municipal green space monitoring. While the current real-world evaluation is intentionally conservative in scope, the results already indicate meaningful transfer from synthetic street-camera training to genuine urban imagery, particularly in the CLIP–DINOv2 hybrid configuration. Rather than representing constraints, the identified limitations define a clear and promising trajectory for expansion. Scaling validation across diverse cities, seasons, and lighting conditions will enable statistically grounded deployment readiness. Incorporating domain adaptation and semi-supervised refinement will further reduce the synthetic–real gap while preserving data efficiency. Extending the three-class formulation toward multi-label and severity-aware modeling can enhance ecological sensitivity and maintenance prioritization. Integrating temporal modeling over continuous video streams will allow detection of gradual vegetation stress and emerging contamination events, while multimodal fusion with satellite indices and environmental signals may support composite urban green quality indicators. Together, these directions position Green Sense not merely as a classifier, but as a foundational platform for dynamic, multimodal urban environmental intelligence capable of evolving alongside real municipal infrastructures.

## 9. References

[1] P. A. J. M. P. D. A. C. M. Ariane L. Bedimo-Rung, "The Significance of Parks to Physical Activity and Public Health".

[2] R. Kaplan, "EVALUATION OF AN URBAN VEST-POCKET PARK °," USDA FOREST SERVICE.

[3] F. N. ,. J. V. O. a. T. M. Helena Madureira, "Preferences for Urban Green Space Characteristics: A Comparative Study in Three Portuguese Cities," enviornments.

[4] "Improving access to greenspace A new review for 2020," Public Health England.

[5] S. L. Donghwan Ki, "Analyzing the effects of Green View Index of neighborhood streets on walking time using Google Street View and deep learning," Landscape and Urban Planning.

[6] H. C. C. A. D. P. L.-J. I. A. V.-M. a. M. S. S.-C. Marco A. Moreno-Armendáriz, "Deep Green Diagnostics: Urban Green Space Analysis Using Deep Learning and Drone Images," sensors, 2019.

[7] Z. R. a. N. Z. Jaloliddin Rustamov, "Green Space Quality Analysis Using Machine Learning Approaches," sustainability.

[8] E. Gutflaish , E. Kachlon , H. Zisman , T. Hacham , N. Sarid , A. Visheratin, S. Huberman , G. Davidi , G. Bukchin, K. Goldberg and R. Mokady, "Generating an Image From 1,000 Words: Enhancing Text-to-Image With Structured Captions," 2025.

[9] A. Mittal, R. Soundararajan and A. C. Bovik, "Making a “Completely Blind” Image Quality Analyzer," 2013.

[10] T. Qwen, "Qwen2.5 Technical Report," 2025.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, "AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE," 2021.

[12] A. Radford, W. J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," 2021.

[13] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski and A. Joulin, "Emerging Properties in Self-Supervised Vision Transformers," 2021.

GitHub Project Link: https://github.com/halevydor/Green-Sense-CV-Project