UrbanWasteNet: A Hierarchical Multimodal Framework for Automated Street-Level Waste Detection Using Vision and Language Models

Siwei Zhang, Jun Ma*
Department of Urban Planning and Design, The University of Hong Kong, Hong Kong SAR
junma@hku.hk

Abstract

Improperly managed waste dumpsites pose severe environmental threats, yet their random distribution makes systematic detection challenging. We present UrbanWasteNet, a hierarchical multimodal framework integrating computer vision and large language models for automated street-level waste detection. Our three-phase architecture combines EfficientNet-based visual processing with spatial metadata analysis, followed by GPT-4o semantic verification. Experimental results on our UrbanDumpSight dataset demonstrate 96.6% detection accuracy and 86.2% classification accuracy across waste types. The dual-phase design significantly reduces false positives while enabling scalable processing of urban imagery, providing municipal authorities with an effective automated solution for waste management and environmental monitoring.

Introduction

Global solid waste generation continues to increase annually, presenting critical challenges for urban waste management (Lu et al., 2024). With waste production predicted to rise from 2.2 billion tons in 2019 to 3.4 billion tons by 2050 (Alves, 2023), effective disposal strategies become increasingly urgent. Compounding this challenge, over one-third of global waste ends up in unregulated dumpsites (World Bank, 2022), causing environmental degradation (Zhang et al., 2023), public health risks (Tomita et al., 2020) and resource waste (Bui et al., 2022). Consequently, developing effective strategies for managing improper urban waste disposal is essential.

A key challenge in controlling improper waste disposal is locating randomly dumped waste outside monitored areas. While regular monitoring and hotspot identification are crucial (Zhang and Ma, 2024), most cities lack systematic monitoring capabilities, particularly in under-developed regions. While some developed areas attempt regular monitoring initiatives, they still rely primarily on manual inspections, which prove labor-intensive and operationally inefficient given the extensive spatial coverage required in urban environments (Fraternali et al., 2024).

Recent automated detection approaches have explored urban sensing and computer vision methods for waste identification. However, deep learning models alone lack sufficient contextual understanding to determine whether dumping locations are improper or legitimate. Existing systems struggle to distinguish between unauthorized waste accumulation and controlled disposal sites, such as designated collection points or construction staging areas (Raphela et al., 2024). Furthermore, complementary multimodal data sources, including points of interest (POI), demographic information, and road network characteristics, can provide crucial contextual information that enhances classification accuracy. Therefore, a more comprehensive framework that integrates multiple data modalities and contextual reasoning capabilities is essential for effective real-world waste monitoring applications.

This paper proposes a novel multimodal framework that integrates computer vision with Large Language Models (LLMs) for enhanced improper dumpsite detection and analysis. Our approach combines street-view imagery, geospatial features, demographic information, and environmental context to create comprehensive understanding of waste disposal scenarios from multiple perspectives. The framework employs a three-phase architecture: first, a deep learning-based visual encoder extracts spatial and visual features from street-view imagery and auxiliary urban data; second, an LLM processes these features alongside contextual information to make nuanced decisions about waste disposal appropriateness, enabling human-like reasoning that distinguishes between various waste accumulation types based on environmental context and regulatory definitions; third, the LLM component generates automated analytical reports, significantly reducing time requirements for environmental assessment documentation.

The key contributions of this paper are as follows:

- We propose the first multimodal framework that systematically combines street-view imagery with comprehensive urban sensing data (demographic, spatial, and infrastructure information) for enhanced waste detection accuracy.

- We demonstrate the effectiveness of integrating LLMs with computer vision for improper dumpsite detection, achieving superior accuracy through advanced contextual understanding and semantic reasoning capabilities.
- We provide an end-to-end automated report generation system that streamlines environmental assessment workflows, enabling efficient decision-making for municipal waste management authorities.

Related Work

Large Vision Language Models. The rapid advancement of Large Language Models has catalyzed the development of Large Vision Language Models (LVLMs), which merge sophisticated linguistic reasoning with visual perception capabilities. These hybrid architectures extend traditional language models beyond text processing to handle complex multimodal scenarios, enabling more nuanced understanding across diverse data types.

The evolution of vision-language integration began with foundational works like CLIP (Radford et al., 2021), which pioneered cross-modal alignment through contrastive learning on massive image-text datasets. Subsequent developments have refined this approach through various architectural innovations. BLIP (Li et al., 2022) advanced the field by implementing dedicated unimodal encoders that process visual and textual inputs separately before fusion, incorporating specialized tokens for holistic sentence representation similar to transformer-based language models. More recent architectures have focused on efficient cross-modal bridging mechanisms. BLIP-2 (J. Li et al., 2023, p. 2) introduced intermediate query-based transformers to connect pre-trained visual encoders with frozen language models, while MiniGPT4 (Zhu et al., 2023) employed lightweight projection layers to align visual features with instruction-tuned language models like Vicuna. Contemporary frameworks such as ShareGPT4V (Chen et al., 2023) have emphasized high-quality training data generation, creating comprehensive datasets spanning spatial reasoning, object recognition, and aesthetic evaluation to enhance multimodal benchmark performance.

Advanced systems like Qwen-VL (Bai et al., 2023) demonstrate the integration of diverse visual tasks including document analysis, optical character recognition, and visual question answering within unified architectures, showcasing remarkable versatility across multiple application domains.

AI for Outdoor Waste Detection. Remote sensing platforms, including satellite imagery and unmanned aerial vehicles (UAVs), have emerged as predominant tools for large-scale waste site identification. Recent deep learning implementations have shown promising yet varied performance levels. Notable examples include the LKN-ME convolutional architecture achieving 0.44 average precision for waste categorization across 32 Chinese urban areas (H. Li et al., 2023), ResNet-FPN frameworks reaching 0.66 precision for binary classification using Italian satellite data (Torres and Fraternali, 2023), and specialized architectures like CascadeDumpNet) and BCA-net (Sun et al., 2023) designed for multi-class detection tasks. However, the inherent spatial resolution constraints of satellite imagery significantly limit these approaches' effectiveness in detecting small-scale dumping incidents critical for operational waste management activities.

Complementing overhead monitoring, street-level detection systems leverage fixed urban camera networks and mobile sensing platforms to capture illegal dumping activities at ground level. Smart city initiatives have deployed comprehensive camera infrastructures integrated with urban elements such as streetlight poles for continuous surveillance coverage. Event-driven detection frameworks utilize sophisticated object tracking algorithms to identify dumping behaviors in real-time, with some implementations demonstrating 97% accuracy across diverse environmental conditions including varying illumination scenarios (Lu, 2019) These systems provide automated geocoding capabilities and vehicle identification functionality upon detection of disposal events.

Current street-level computer vision applications primarily focus on distinguishing roadside waste accumulations from typical urban environments, though they necessitate extensive manual annotation efforts for training data preparation (Wu et al., 2023). Existing methodologies predominantly target individual litter detection rather than comprehensive dumpsite characterization, potentially generating misleading assessments influenced by sporadic debris rather than identifying substantial waste accumulation zones requiring municipal intervention.

Methods

In this study, we propose UrbanWasteNet for urban street level dumpsite identification, the definition of targeted dumpsite, the structure and function of each phase is shown below.

Dumpsite Definition and Categorization

Given the absence of standardized definitions for street-level improper dumpsites in existing literature, this study establishes comprehensive operational definitions with quantifiable thresholds for automated detection systems. We define improper dumpsites as unauthorized accumulations of discarded materials deposited in non-designated locations that exceed specific volume and spatial criteria.

Figure 1 Improper dumpsite types and exclusion scenarios for urban waste detection

Three essential criteria must be satisfied for improper dumpsite classification. The volume criterion requires substantial waste accumulation that exceeds incidental littering thresholds, indicating systematic improper disposal patterns (Zhang et al., 2025). Quantifiable thresholds include: (1) countable items comprising at least ten identifiable pieces, or (2) uncountable waste covering areas exceeding 0.5 square meters. The location criterion specifies waste deposited in unauthorized areas including streets, alleyways, vacant lots, or public spaces not designated for waste collection. The disorderliness criterion requires visual evidence of inappropriate disposal practices, distinguishing unauthorized dumpsites from legitimate, organized waste storage or collection points.

Certain scenarios may visually resemble improper dumpsites but do not satisfy classification criteria (Figure 1). Excluded categories include temporary yard waste from scheduled maintenance activities; active construction sites with proper machinery and safety barriers; and overflow situations at designated collection facilities that remain within authorized boundaries.

The framework categorizes improper dumpsites into three distinct types based on material composition and disposal characteristics (Figure 1). **Construction and demolition waste** (Type 1) encompasses building lifecycle materials containing significant recyclable components such as concrete, masonry, and timber. **Municipal solid waste** (Type 2) includes residential refuse and organic materials that pose health risks through bacterial proliferation if improperly managed. **Bulky waste** (Type 3) comprises oversized items unsuitable for standard collection services, including furniture and bicycles, requiring specialized handling due to size constraints and recycling complexity.

UrbanWasteNet Framework Architecture

Overall Framework Structure. UrbanWasteNet implements a three-phase hierarchical processing architecture designed to balance computational efficiency with detection accuracy for large-scale municipal deployment (Figure 2).

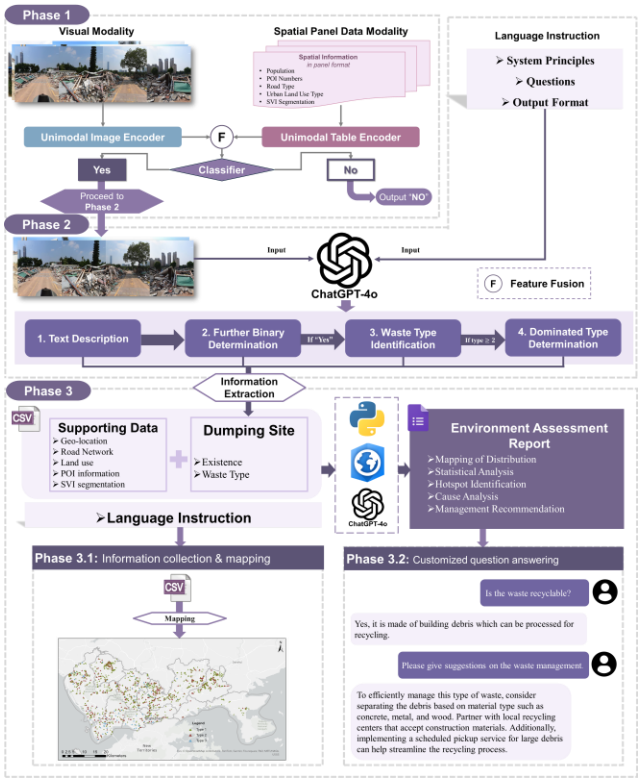

Figure 2 Framework of UrbanWasteNet

The framework employs a filtering approach where each successive stage applies more sophisticated analysis to progressively smaller datasets, optimizing resource utilization while maintaining comprehensive coverage.

The sequential design enables efficient processing of extensive urban imagery through initial high-throughput screening, followed by detailed semantic analysis only for identified candidates. This approach addresses the fundamental challenge of applying computationally intensive language models to citywide surveillance while preserving the contextual reasoning capabilities necessary for reliable waste detection in complex urban environments.

Phase 1: Multimodal Initial Detection. The primary screening stage employs a dual-branch deep learning architecture that processes street-view imagery and spatial metadata simultaneously to identify potential waste accumulation sites (Table 1). The visual processing branch utilizes EfficientNet-B0 as the feature extraction backbone, leveraging its mobile inverted bottleneck convolution blocks for efficient feature learning with reduced computational overhead. Squeeze-and-excitation attention mechanisms enhance feature representation by emphasizing relevant visual patterns while suppressing background noise.

The spatial processing branch implements a four-stage fully connected network that extracts contextual information

Vision branch baseline Network				
Stage i	**Operator** $\widehat{\mathcal{F}}_i$	**Output Shape** $\widehat{H}_i \times \widehat{W}_i \times \widehat{C}_i$	**#Layers** $\hat{L}_i$	**# Param**
1	Conv 3 × 3	1024 × 256 × 32	1	928
2	MBConv1, 3 × 3	1024 × 256 × 16	1	1,448
3	MBConv6, 3 × 3	512 × 128 × 24	2	16,714
4	MBConv6, 5× 5	256 × 64 × 40	2	46,640
5	MBConv6, 3× 3	128 × 32 × 80	3	242,930
6	MBConv6, 5× 5	128 × 32 × 112	3	543,148
7	MBConv6, 5× 5	64 × 16 × 192	4	2,026,348
8	MBConv6, 3× 3	64 × 16 × 320	1	717,232
9	Conv2dNormActivation	64 × 16 × 1280	1	412,160
10	AdaptiveAvgPool2D	1 × 1 × 1280	1	0
11	Dropout +Flatten	Flatten to 1D: 1280	1	0
Spatial branch baseline network				
Stage i	**Operator** $\widehat{\mathcal{F}}_i$	**Output Shape**	**#Layers** $\hat{L}_i$	**# Param**
1	Fully Connected Block (Feature Number, 64)	64	1	3,584
2	Fully Connected Block (64, 128)	128	1	8,576
3	Fully Connected Block (128, 256)	256	1	33,536
4	Fully Connected Block (256, 512)	512	1	133,632

Table 1 Network Parameters Summary for Phase 1Vision and Spatial Branches

from geographic and demographic variables including population density, point-of-interest distributions, road network characteristics, and land use classifications. Each processing stage incorporates batch normalization, dropout regularization, and ReLU activation functions to ensure training stability and prevent overfitting. Early fusion methodology concatenates visual and spatial features before final binary classification, enabling the model to learn integrated representations that leverage complementary information from both modalities.

Phase 2: LLM based Semantic Verification and Classification. Candidates identified in Phase 1 undergo verification through GPT-4o using structured prompting protocols that guide systematic environmental assessment. The verification process implements a sequential questioning framework that establishes contextual understanding, confirms waste presence, identifies material types, and determines dominant categories for mixed accumulations. Standardized output formatting ensures consistent data extraction while preserving semantic reasoning capabilities necessary for distinguishing between legitimate activities and improper disposal practices.

This stage serves dual functions by eliminating false positives through contextual analysis and performing accurate categorical classification based on semantic understanding rather than purely visual features. The approach addresses fundamental limitations of traditional computer vision systems in scenarios involving visual similarity between problematic accumulations and legitimate activities such as construction staging or authorized collection points.

Phase 3: Automated Analysis and Reporting. The final stage aggregates detection results to generate comprehensive assessment reports supporting municipal decision-making processes. Automated scripts extract statistical summaries, identify spatial patterns, and quantify accumulation distributions across geographic regions. The aggregated data undergoes analysis through language model processing that produces readable reports containing statistical summaries, critical area identification, causal analysis, and specific management recommendations tailored to local conditions and available resources.

This phase transforms individual detection results into actionable intelligence that supports strategic planning and resource allocation decisions while providing consistent reporting formats compatible with existing municipal management systems.

Experiments

Training Dataset

The *UrbanDumpSight* dataset was constructed to address the specific requirements of multimodal waste detection in urban environments. The dataset design prioritizes three key objectives: establishing consistent criteria for improper disposal identification, enabling categorical classification across waste types, and incorporating realistic urban complexity to minimize false positive rates in deployment scenarios.

The dataset employs a hierarchical labeling scheme applied to individual sampling locations. Each location receives binary classification labels indicating disposal presence, multi-label annotations identifying all observable waste categories within predefined taxonomies, and single-class assignments determining the predominant waste type when multiple categories coexist. This multi-level annotation approach supports both detection and classification tasks while providing flexibility for various analytical requirements.

Spatial sampling incorporates comprehensive visual coverage through four-directional imagery captured at 90-degree intervals, ensuring complete environmental documentation for each location. Beyond visual data, the dataset integrates extensive contextual metadata including demographic characteristics, infrastructure density measures, land use classifications, and built environment attributes. This multimodal approach enables models to incorporate spatial context and environmental factors that influence waste accumulation patterns, supporting more robust prediction capabilities across diverse urban settings.

Phase 1 Training Configuration

The initial detection model training utilized a stratified data partition allocating 70% of samples for training and 30% for validation, ensuring balanced representation across waste categories and environmental contexts. The computational environment consisted of an Intel Core i9-12900K processor operating at 3.20 GHz with NVIDIA GeForce RTX 3090 Ti graphics acceleration, implemented through PyTorch 2.0.1 framework integration with supporting libraries including NumPy, Pandas, and Scikit-learn for data processing operations.

The hybrid architecture combines EfficientNet-B0 feature extraction for visual processing with multilayer perceptron networks for spatial data integration. The visual component utilizes ImageNet pre-trained weights with architectural modifications replacing the final classification layer to extract feature representations suitable for fusion operations. Training parameters included 100-epoch execution with 0.0001 initial learning rate, Adam optimization, and StepLR scheduling implementing 0.1 decay factors at 20-epoch intervals.

Data preprocessing incorporated standard augmentation techniques including resizing, normalization, and stochastic transformations to enhance model generalization across varied lighting conditions and camera perspectives. Batch processing utilized size-8 configurations with persistent worker allocation to optimize data loading efficiency. Cross-entropy loss functions supported both binary detection and multi-class classification objectives depending on specific task requirements.

Phase 2 and 3 Implementation Framework

The semantic analysis and reporting phases leverage GPT-4o capabilities through API integration within Python scripting environments. This implementation approach ensures scalable processing for large-scale urban datasets while maintaining the contextual reasoning capabilities necessary for accurate waste classification and false positive reduction.

The verification protocols implement structured prompting sequences that guide systematic environmental assessment through sequential questioning frameworks. Output formatting specifications ensure consistent data extraction and integration with downstream analysis pipelines. The automated reporting system processes aggregated detection results through statistical analysis scripts that generate visualization outputs and summary metrics before language model synthesis into comprehensive assessment documents.

This multi-stage implementation design balances computational efficiency with analytical depth, enabling cost-effective processing of citywide imagery while preserving the semantic understanding necessary for reliable waste management decision support.

Results

Comparative Performance Evaluation

We conducted comprehensive experiments to evaluate the effectiveness of our proposed UrbanWasteNet framework across multiple dimensions. The evaluation encompasses three critical comparisons: backbone architecture selection for visual feature extraction, unimodal versus multimodal approaches, and single-phase versus dual-phase frameworks.

Visual Backbone Comparison. We evaluated three representative architectures for street-view image processing: Swin Transformer representing vision transformers, ResNet50 as a traditional CNN baseline, and EfficientNet-B0 as our proposed backbone. Table 2 presents the binary classification results across these architectures. EfficientNet-B0 achieves the highest performance with 0.945 accuracy and

Model Configuration	Model	Precision	Recall	F1-score	Accuracy
(a) Single Phase: Unimodal Visual Model	Swin Transformer	0.94	0.94	0.94	0.939
	ResNet50	0.93	0.91	0.92	0.922
	EfficientNe**t**	0.93	0.96	0.95	**0.945**
(b) Single Phase: Bimodal Visual-Spatial Model	Swin Transformer-based Bimodal model	0.95	0.96	0.95	0.953
	EfficientNet-based Bimodal model **(Our Phase 1)**	0.96	0.94	0.95	**0.955**
(c) Dual-Phase: Bimodal Visual-Spatial Model + LLM	EfficientNet-based Bimodal Model with LLM GPT-4o **(Our Phase 1+Phase 2)**	0.98	0.94	0.96	**0.966**

Table 2 Comparative Performance Metrics for Various Model Configurations in Binary Classification. *Note: Precision, Recall, and F1-score represent the performance metrics for the "label = yes" class.*

Model Configuration	Model	Precision				Recall				F1-score				Accuracy
		No	Type1	Type2	Type3	No	Type1	Type2	Type3	No	Type1	Type2	Type3	
(a) Single Phase: Unimodal Visual Model	Swin Transformer	0.91	0.74	0.51	0.36	0.94	0.76	0.49	0.13	0.93	0.75	0.50	0.19	0.790
	Efficient-Net	0.92	0.79	0.63	0.31	0.95	0.78	0.58	0.32	0.93	0.79	0.60	0.32	0.820
(b) Single Phase: Bimodal Visual-Spatial Model	Swin Transformer-based Bimodal model	0.93	0.76	0.58	0.52	0.96	0.79	0.52	0.33	0.95	0.78	0.54	0.41	0.822
	Efficient-Net-based Bimodal model	0.91	0.76	0.66	0.40	0.98	0.84	0.42	0.13	0.94	0.80	0.51	0.20	**0.826**
(c) Dual-Phase	**Efficient-Net-based Bimodal Model with LLM GPT-4o**	0.95	0.78	0.75	0.64	0.99	0.88	0.55	0.23	0.97	0.83	0.63	0.34	**0.862**

Table 3 comparative Performance Metrics for Various Model Configurations in Four Class Classification

0.95 F1-score, demonstrating superior feature extraction capabilities for urban waste detection scenarios. The balanced precision-recall trade-off (0.93 precision, 0.96 recall) indicates robust detection performance across diverse environmental conditions.

Multimodal Integration Impact. To assess the contribution of spatial information integration, we compared unimodal visual models with their bimodal counterparts incorporating tabular spatial features. Both Swin Transformer and EfficientNet-based bimodal models showed consistent improvements over their unimodal versions. Our EfficientNet-based bimodal model achieves 0.955 accuracy, representing a 1.0% improvement over the unimodal baseline. This demonstrates that spatial context significantly enhances discrimination between legitimate activities and improper waste disposal.

Dual-Phase Framework Effectiveness. The integration of GPT-4o as a semantic verification component yields substantial performance gains. Our complete dual-phase framework achieves 0.966 accuracy with 0.98 precision, representing critical improvements for real-world deployment where false positive reduction is paramount. Given that actual dumpsites constitute less than 1% of urban sampling

points, the precision enhancement from 0.96 to 0.98 translates to a 50% reduction in false positives requiring manual verification.

Multi-Class Classification Results. Beyond binary detection, we evaluated the framework's capability for waste type classification across four categories: no waste, construction debris, domestic waste, and bulky items. Traditional deep learning approaches struggle with this task due to visual similarity between waste categories and contextual dependencies for accurate classification.
Table 3 demonstrates that single-phase models achieve moderate performance with accuracy around 0.820, while our dual-phase framework significantly improves classification accuracy to 0.862. The 5.1% improvement highlights the critical role of semantic reasoning in distinguishing waste categories that appear visually similar but differ in context, scale, or environmental setting.

Ablation Analysis. We conducted systematic ablation studies to isolate the contribution of individual components. Spatial feature integration provides consistent 1-2% accuracy improvements across all backbone architectures, with EfficientNet-based models achieving 1.0% gains (0.945 to 0.955) and Swin Transformer models showing 1.4% improvements (0.939 to 0.953). This validates the importance of geographic and demographic context in distinguishing legitimate activities from improper waste disposal.The GPT-4o verification stage contributes the most significant performance enhancement, delivering 1.1% accuracy improvement in binary classification (0.955 to 0.966) and 4.3% gains in multi-class scenarios (0.826 to 0.862). More critically, semantic verification provides precision enhancement from 0.96 to 0.98, representing a 50% reduction in false positives—crucial for operational efficiency where actual dumpsites constitute less than 1% of sampling points.

The results demonstrate complementary contributions from each component, with spatial integration addressing contextual limitations of visual approaches and LLM verification enabling semantic reasoning for distinguishing visually similar waste categories. The combined framework achieves precision requirements necessary for practical municipal deployment.

Case Study

To demonstrate the practical applicability of *UrbanWasteNet*, we deployed our framework across Shenzhen, a major metropolitan area in southern China, to map the spatial distribution of improper dumpsites and analyze waste type patterns across urban landscapes. The case study encompasses comprehensive street-level monitoring across Shenzhen's urban districts, utilizing our dual-phase framework to automatically detect and classify improper waste accumulations from thousands of street-view locations.

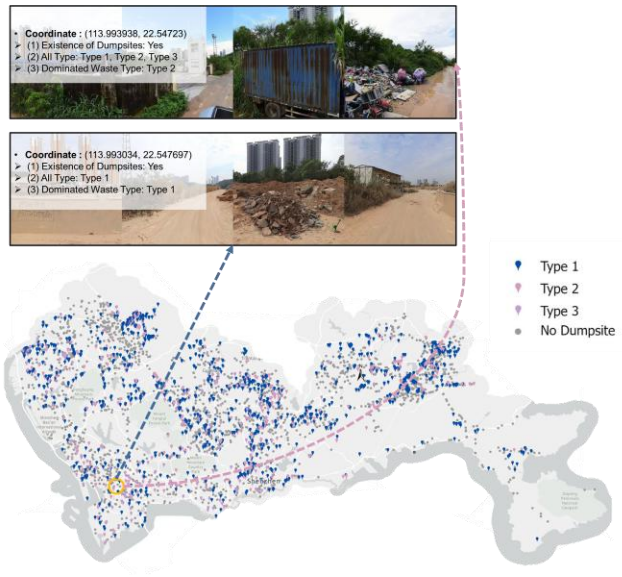

Figure 3 Spatial distribution of detected improper dumpsites across Shenzhen urban area with waste type classification

Figure 3 illustrates the detected dumpsite distribution across Shenzhen, revealing distinct spatial clustering patterns of different waste types. Type 1 (construction and demolition waste) shows the highest prevalence, represented by blue markers concentrated in developing urban peripheries and construction zones. Type 2 (municipal solid waste) appears in pink markers, primarily distributed in high-density residential areas and commercial districts, while Type 3 (bulky waste) displays scattered distribution patterns across various urban zones indicated by purple markers.

Conclusion

This paper presents UrbanWasteNet, a novel dual-phase framework combining computer vision and large language models for automated street-level waste detection and classification. Our approach achieves 0.966 binary classification accuracy and 0.862 multi-class accuracy, representing significant improvements over traditional single-phase methods. The integration of spatial contextual features and semantic reasoning addresses key limitations of purely visual approaches, particularly in distinguishing visually similar waste categories and reducing false positives. The Shenzhen case study demonstrates practical deployment capabilities for city-wide waste monitoring. Our framework provides municipal authorities with an efficient tool for automated waste management, enabling targeted interventions and optimized resource allocation for urban cleanliness maintenance.

References

Alves, B., 2023. Global waste generation - statistics & facts [WWW Document]. Statista. URL https://www.statista.com/topics/4983/waste-generation-worldwide/ (accessed 7.30.24).

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J., 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. https://doi.org/10.48550/arXiv.2308.12966

Bui, T.-D., Tseng, J.-W., Tseng, M.-L., Lim, M.K., 2022. Opportunities and challenges for solid waste reuse and recycling in emerging economies: A hybrid analysis. Resources, Conservation and Recycling 177, 105968. https://doi.org/10.1016/j.resconrec.2021.105968

Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D., 2023. ShareGPT4V: Improving Large Multi-Modal Models with Better Captions. https://doi.org/10.48550/arXiv.2311.12793

Fraternali, P., Morandini, L., Herrera González, S.L., 2024. Solid waste detection, monitoring and mapping in remote sensing images: A survey. Waste Management 189, 88–102. https://doi.org/10.1016/j.wasman.2024.08.003

Li, H., Hu, C., Zhong, X., Zeng, C., Shen, H., 2023. Solid Waste Detection in Cities Using Remote Sensing Imagery Based on a Location-Guided Key Point Network With Multiple Enhancements. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 16, 191–201. https://doi.org/10.1109/JSTARS.2022.3224555

Li, J., Li, D., Savarese, S., Hoi, S., 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. https://doi.org/10.48550/arXiv.2301.12597

Li, J., Li, D., Xiong, C., Hoi, S., 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. https://doi.org/10.48550/arXiv.2201.12086

Lu, M., Zhou, C., Wang, C., Jackson, R.B., Kempes, C.P., 2024. Worldwide scaling of waste generation in urban systems. Nat Cities 1, 126–135. https://doi.org/10.1038/s44284-023-00021-5

Lu, W., 2019. Big data analytics to identify illegal construction waste dumping: A Hong Kong study. Resources, Conservation and Recycling 141, 264–272. https://doi.org/10.1016/j.resconrec.2018.10.039

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning Transferable Visual Models From Natural Language Supervision. https://doi.org/10.48550/arXiv.2103.00020

Raphela, T., Manqele, N., Erasmus, M., 2024. The impact of improper waste disposal on human health and the environment: a case of Umgungundlovu District in KwaZulu Natal Province, South Africa. Front. Sustain. 5. https://doi.org/10.3389/frsus.2024.1386047

Sun, X., Yin, D., Qin, F., Yu, H., Lu, W., Yao, F., He, Q., Huang, X., Yan, Z., Wang, P., Deng, C., Liu, N., Yang, Y., Liang, W., Wang, R., Wang, C., Yokoya, N., Hänsch, R., Fu, K., 2023. Revealing influencing factors on global waste distribution via deep-learning based dumpsite detection from satellite imagery. Nat Commun 14, 1444. https://doi.org/10.1038/s41467-023-37136-1

Tomita, A., Cuadros, D.F., Burns, J.K., Tanser, F., Slotow, R., 2020. Exposure to waste sites and their impact on health: a panel and geospatial analysis of nationally representative data from South Africa, 2008–2015. The Lancet Planetary Health 4, e223–e234. https://doi.org/10.1016/S2542-5196(20)30101-7

Torres, R.N., Fraternali, P., 2023. AerialWaste dataset for landfill discovery in aerial and satellite images. Sci Data 10, 63. https://doi.org/10.1038/s41597-023-01976-9

World Bank, 2022. Solid Waste Management [WWW Document]. World Bank. URL https://www.worldbank.org/en/topic/urbandevelopment/brief/solid-waste-management (accessed 7.4.23).

Wu, T.-W., Zhang, H., Peng, W., Lü, F., He, P.-J., 2023. Applications of convolutional neural networks for intelligent waste identification and recycling: A review. Resources, Conservation and Recycling 190. https://doi.org/10.1016/j.resconrec.2022.106813

Zhang, S., Ma, J., 2024. CascadeDumpNet: Enhancing open dumpsite detection through deep learning and AutoML integrated dual-stage approach using high-resolution satellite imagery. Remote Sensing of Environment 313, 114349. https://doi.org/10.1016/j.rse.2024.114349

Zhang, S., Ma, J., Jiang, F., 2025. Monitoring street-level improper dumpsites via a multi-modal and LLM-based framework. Resources, Conservation and Recycling 218, 108227. https://doi.org/10.1016/j.resconrec.2025.108227

Zhang, S., Ma, J., Zhang, X., Guo, C., 2023. Atmospheric remote sensing for anthropogenic methane emissions: Applications and research opportunities. Science of the Total Environment 893. https://doi.org/10.1016/j.scitotenv.2023.164701

Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M., 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. https://doi.org/10.48550/arXiv.2304.10592

Supplementary Materials

1. Dataset Description and Access

Study Region and Data Acquisition. We selected Shenzhen, Guangdong Province, as our study area due to its rapid urban development and complex waste management challenges. As a major Chinese megacity with extensive urban villages, Shenzhen provides an ideal testbed for street-level waste detection research.

Our data collection process involved systematic sampling across Shenzhen's road infrastructure. We extracted road network information from OpenStreetMap (February 2023) and established sampling points every 50 meters along five road categories: main arterials, secondary roads, tertiary roads, internal roads, and elevated highways. Street view imagery was obtained via Baidu Street View API (October 2023), primarily capturing scenes from July 2021, with some remote area images dating to 2016. At each point, we collected four directional images (90° intervals, 512×512 resolution) and concatenated them into panoramic views (2048×512 resolution), as shown in **Fig. S1 (d)**

We enriched each sampling point with contextual metadata (**Fig. S1(c)**), incorporating demographic and geographic attributes from multiple sources. Manual inspection of all collected imagery revealed waste dumpsites at only 0.85% of locations (2,082 positive samples). Given the multi-category nature of most dumpsites, we implemented comprehensive labeling for all waste types present while identifying the predominant category.

Fig. S1(a) displays the spatial distribution of annotated samples, while **Fig. S1(b)** illustrates waste type frequencies, with construction debris representing the largest category in Shenzhen. To address class imbalance, we randomly sampled 2,082 negative examples, creating a balanced dataset of 4,164 total samples.

The complete UrbanDumpSight dataset, including panoramic images and metadata spreadsheets, is publicly available through Science Data Bank.

2. Phase 1: Bimodal Detection Framework

Phase 1 implements a dual-input supervised learning approach combining street-view imagery with geospatial tabular data to identify illegal dumping locations. The framework architecture (**Fig. S2a**) features two parallel processing streams that extract complementary feature representations.

Vision Processing Stream: We employ EfficientNet-B0 as the image feature extractor, leveraging its optimized depth-width-resolution scaling for computational efficiency. The network processes panoramic street images through eight progressive stages utilizing Mobile Inverted Bottleneck Convolution (MBConv) modules (**Fig. S2b**). Each MBConv block employs dimension expansion via 1×1 convolutions, lightweight depthwise processing, and dimension reduction,

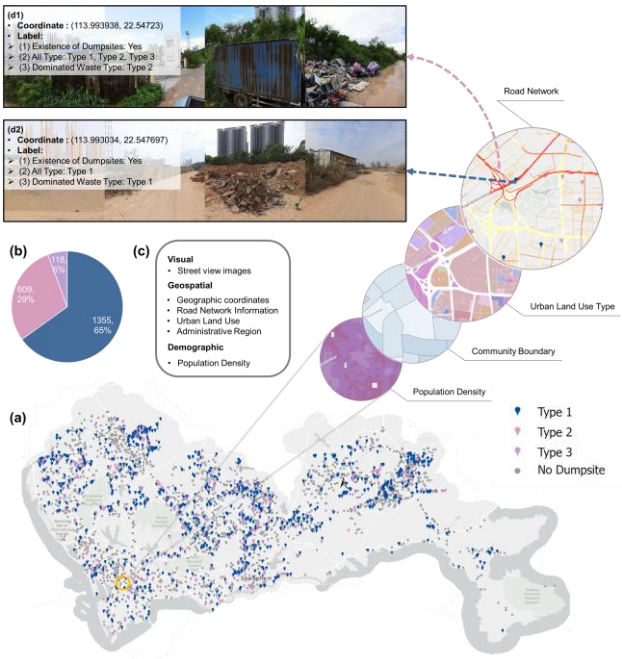

Fig. S1 Overview of the UrbanDumpSight Dataset. (a) Distribution of sample points with equal amounts of positive and negative data points in the study area. **(b)** Breakdown of waste categories among positive sample points. **(c)** Metadata associated with each sample point, including geospatial and demographic information. (**d)** Examples of street view images containing waste, along with their labels.

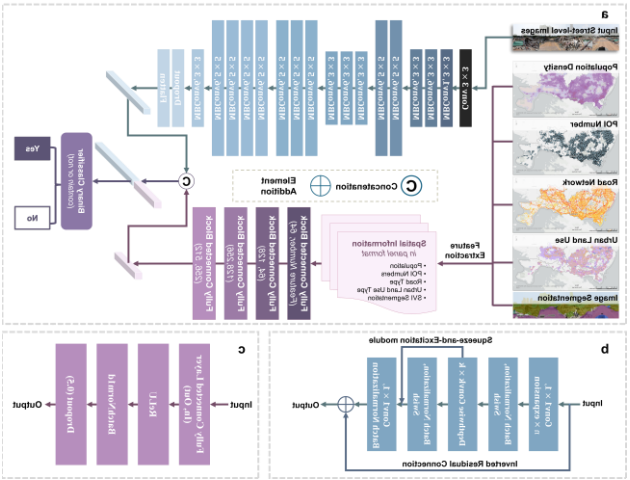

Fig. S2 Architecture of the Phase 1 Bimodal Model: (a) Overall architecture of Phase 1. (b) Structure of the vision branch's basic module: MBConv k×k. (c) Structure of the spatial branch's basic module: Fully Connected Block.

incorporating squeeze-and-excitation attention mechanisms for feature refinement.

Spatial Processing Stream: Geographic and demographic attributes are processed through a four-stage fully connected architecture with progressively expanding dimensions (64→128→256→512 neurons). Each stage combines linear transformation, ReLU activation, batch normalization, and dropout regularization (**Fig. S2c**).

Feature Integration: The streams are merged via early fusion concatenation, combining visual representations (1280-d) with spatial embeddings (512-d) into unified feature vectors. This integrated representation feeds into a binary classifier with cross-entropy optimization, enabling the model to leverage both visual patterns and contextual geographic information for enhanced dumpsite detection accuracy.

3. Prompt for LLM-based Classification

Samples flagged as positive in Phase 1 undergo secondary analysis using GPT-4o with structured prompting. Following established practices for sequential reasoning guidance, we employ a multi-step query framework to enhance classification precision.

Prompt Architecture: The instruction set comprises system definitions, sequential queries, and standardized output formatting (**Table S1**).

Reasoning Workflow: The four-question sequence establishes environmental context before proceeding to binary classification, then detailed categorization if positive. Each query includes explicit formatting constraints to ensure systematic output parsing for downstream analysis.

Expert Role Assignment: The model operates under environmental monitoring specialist persona to leverage domain-specific reasoning patterns and maintain consistency across evaluations.

Category	Details
System Principles	The overall task is to analyze urban improper dumpsites based on the concatenated street view images from four angles of a sample point. The definition of improper dumpsites is as follows: (1) ***Volume***: Significant waste accumulation, distinguishable from incidental littering. Must show a pattern of improper disposal. - Countable items (e.g., bulky waste, domestic garbage): At least 10 items. - Uncountable items (e.g., construction debris, sand/gravel piles): Area > 0.5 square meters. (2) ***Location***: Waste in non-designated areas (e.g., streets, alleys, vacant lots, public spaces). (3) ***Disorderliness***: Waste must be disorderly and inappropriately dumped. Exclude temporary yard waste or ongoing construction activities. Pretend you are an expert in environmental monitoring of improper dumpsites and use your knowledge to answer the following questions.
Question 1	Describe the overall environment depicted in the street view image and determine the likely functional zone.
Question 2 Format 2	Does an improper dumpsite exist at the specified location? Answer with “yes” or “no” only.
Question 3 Format 3	If an improper dumpsite is identified, list all visible waste types. List all applicable categories from the following and separate them with commas: - Type 1: Construction and renovation waste, building debris. - Type 2: Small household trash. - Type 3: Bulky waste.
Question 4 Format 4	If more than one type of waste is identified, indicate the most dominant category. Indicate the most dominant category as "Type 1", "Type 2", or "Type 3".
Overall Output Format	Please format your response as follows: 1. **Functional zones**: [describe without hyphens] 2. **Existence of Improper Dumping**: [Yes/No] 3. **Types of Waste Visible (All)**: [Type 1, Type 2, Type 3] 4. **Types of Waste (Dominant)**: [Type 1/Type 2/Type 3]

Table S1 Language Instructions Applied in Phase 2

4. Limitation and Future Work

Cross-City Generalization: Our evaluation is constrained to a single custom dataset from one metropolitan area, limiting assessment of cross-urban transferability. Different cities exhibit distinct cultural practices and waste management systems that significantly influence dumping patterns and collection behaviors. Future research must validate model robustness across diverse urban contexts to establish broader applicability.

Temporal and Environmental Variability: Street view imagery presents inherent temporal limitations, as data collection timing affects waste visibility. Seasonal variations (e.g., snow coverage masking winter dumping) and collection schedule dependencies create detection inconsistencies. Enhanced data collection protocols incorporating multi-temporal sampling and complementary sensing modalities could address these constraints.

Class Distribution Imbalance: The dataset exhibits substantial inter-class disparities, particularly underrepresentation of bulky waste categories. This imbalance manifests in suboptimal multi-class performance metrics, with notably low recall and F1-scores for minority classes. Standard accuracy measurements may obscure imbalance-sensitive behaviors without cost-aware evaluation frameworks.

Future Directions: Priority areas include: (1) developing domain adaptation techniques for cross-city deployment, (2) implementing temporal data fusion strategies, (3) applying advanced sampling methods and cost-sensitive learning approaches to address class imbalance, and (4) establishing standardized evaluation protocols that account for real-world deployment constraints and varying municipal waste management practices.