Detection and Mitigation of Urban Heat Island Effect Using Vision-Language Models

Ayeshmantha S K S*, Kumara B D A N*, Silva G M S S*, Madhuwantha G K O*,
Vishan Jayasinghearachchi*, Kaushalya Rajapakse*, Rajitha de Silva†
*Dept. of Software Engineering, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka
†Lincoln Center for Autonomous Systems, University of Lincoln, Lincoln, United Kingdom
{it21219320, it21256264, it21802126, it21802058}@my.sliit.lk, {vishan.j, kaushalya.r}@sliit.lk, odesilva@lincoln.ac.uk

***Abstract*—Urban Heat Islands (UHIs) intensify urban temperatures, contributing to environmental degradation and public health concerns. Mitigating these effects is crucial for creating more sustainable and livable cities. In this paper, we address the problem of detecting UHI-prone areas and generating context-specific mitigation strategies using street-level imagery. We propose HeatScape, an AI-powered framework that combines semantic segmentation and Vision-Language Models (VLMs) to identify heat-retaining surfaces from user-captured images. Unlike traditional approaches reliant solely on satellite data, HeatScape uses ground-level visual inputs to offer fine-grained urban analysis. VLMs then interpret these segmented scenes to recommend urban planning actions for new and existing building projects. Our experiments show a UHI detection accuracy of 94.25% using logistic regression, with VLM-generated suggestions aligning with established urban sustainability practices. HeatScape demonstrates a scalable, low-cost tool for data-driven environmental assessment and urban design.**

***Index Terms*—Urban Heat Island, Semantic Segmentation, Visual Language Models, AI in Urban Planning, Mitigation Strategies**

I. Introduction

Urban areas are becoming increasingly vulnerable to the Urban Heat Island (UHI) effect, where certain man-made surfaces trap more heat than surrounding areas, resulting in elevated temperatures. This phenomenon amplifies energy consumption, deteriorates public health, and accelerates climate-related degradation. Traditional large-scale mitigation strategies often lack precision, ignoring local environmental and structural contexts. An intelligent system that integrates visual perception and reasoning, scalable to different urban settings through modular design, and accessible via common mobile devices without specialized sensors, is essential for enabling localized, sustainable heat mitigation strategies.

In this paper, we investigate the problem of accurately detecting UHI-prone regions from street-level imagery and generating meaningful, context-sensitive mitigation strategies through Vision-Language Model (VLM)-based scene understanding.

In general, addressing UHI detection requires identifying materials that contribute to increased heat retention, such as exposed concrete, dark rooftops, bare metal surfaces, and glass facades, and assessing their spatial distribution in urban environments. Although previous work relies heavily on thermal satellite imagery and GIS-based approaches [1]–[3], these methods often lack ground-level granularity and are not easily accessible to the public, as high-resolution thermal satellite imagery is often commercially restricted, updated infrequently, and requires specialized tools for interpretation. This limits public awareness and the ability of communities or local planners to assess UHI conditions in a timely and localized manner. Our approach leverages semantic segmentation for fine-grained identification of urban surface types and applies Vision-Language Models (VLMs) to interpret these visual scenes and suggest urban planning interventions.

The main contribution of this paper is a modular AI framework, HeatScape, that combines deep semantic segmentation and VLM-based reasoning to detect UHI-prone areas from street-level imagery and generate context-aware mitigation strategies. By leveraging visual information from widely available mobile and street-level imagery, our system offers a more accessible and adaptable alternative to conventional satellite or thermal imaging methods. We developed a semantic segmentation pipeline capable of identifying a comprehensive set of urban surface types—including heat-contributing materials such as asphalt, concrete, and metal, as well as heat-mitigating features like vegetation. Based on this detailed surface mapping, a Vision-Language Model (VLM) generates context-sensitive mitigation strategies such as installing green roofs or vertical gardens, replacing impervious surfaces with permeable light-colored pavers, applying reflective roof coatings, and introducing tree-lined shading along exposed walkways. Our experimental results demonstrate a UHI detection accuracy of 94.25%. The mitigation strategies generated by the Vision-Language Model—such as introducing green roofs, tree-lined shading, reflective roof coatings, and permeable pavement—were qualitatively consistent with evidence-based best practices in sustainable urban design. These strategies align with validated research showing measurable improvements in thermal performance, human comfort, and heat-related risk reduction. For example, green roofs have been shown to lower rooftop temperatures by up to 30°C and reduce ambient heat through evapotranspiration [1], [2]; reflective pavements can reduce surface temperatures by 2–5°C compared to conventional asphalt [4]; and strategic urban tree planting is associated with improved pedestrian comfort and reduced emergency healthcare demand due to extreme heat [5], [6]. This consistency between VLM outputs

and established mitigation practices supports the relevance of our model's recommendations in real-world UHI planning contexts. Our findings indicate that HeatScape serves as a low-cost and extensible decision support tool for both urban planners and citizens. Unlike traditional systems that rely on high-resolution satellite or thermal imagery, which may be expensive, infrequently updated, or restricted by licensing, our framework operates on readily available RGB images captured by mobile phones or street-level cameras. This makes the system affordable to deploy at scale and adaptable across diverse urban contexts, including low-resource or data-limited environments.

II. Literature Review

Urban Heat Island (UHI) effects have emerged as a growing concern in climate-resilient urban planning due to their disproportionate impact on health, energy use, and microclimatic degradation [1], [2]. The increasing urbanization and reduction of natural land cover have been shown to intensify localized surface temperatures. While traditional studies relied on remote sensing [1], [7] and thermal infrared imagery [8] for detection, recent advances in computer vision and machine learning have enabled finer-grained, cost-effective alternatives [9], [10].

Several deep learning models have proven effective in UHI-relevant segmentation tasks [11]. Fully Convolutional Networks (FCNs) introduced pixel-wise classification for urban scene understanding [12], while DeepLabV3+ improved boundary precision through atrous spatial pyramid pooling [13]. Dense semantic segmentation techniques, such as those applied by Zhou et al. [14], enabled accurate mapping of impervious surfaces. Similarly, Zhang et al. [5] demonstrated how CNN-based segmentation could classify reflective rooftops and green spaces—both crucial indicators of thermal behavior. These works show strong performance for specific urban features, but they often rely on large annotated datasets and may lack generalizability across cities. Our method addresses this by combining YOLO's real-time object detection with SAM's zero-shot segmentation, reducing dependence on retraining.

Recent research has shown promise in employing more adaptive, data-efficient segmentation models like YOLO [15] and Meta AI's SAM [16]. While YOLO enables real-time detection of urban elements, SAM allows general-purpose segmentation with minimal retraining. Such models offer flexibility across cities with varying visual characteristics. Our work combines both to achieve high-resolution, adaptable segmentation pipelines without extensive annotation requirements.

Beyond image interpretation, structured environmental metadata has been leveraged to predict UHI intensity. Tan et al. [6] used decision tree classifiers and support vector machines (SVM) to correlate features like surface reflectivity, temperature, and humidity with UHI risk. Their results confirmed that a hybrid of vegetation indices and built-up surface ratios could approximate thermal anomalies with high accuracy. More recent studies have incorporated ensemble methods such as XGBoost to improve generalizability across seasons and locations [17]. However, these approaches still depend heavily on structured environmental measurements and often lack integration with spatial layout analysis. Our contribution combines spatial interpretation with VLM-guided suggestion, enabling an end-to-end mitigation system.

Generating targeted mitigation strategies has remained a challenge. ENVI-met (Environmental Meteorology) simulations [4] and GIS-based zoning tools provide detailed impact modeling, but they often require expert knowledge and computational resources. An alternative is the use of generative AI to bridge perception and policy. Vision-Language Models (VLMs), such as LLaVA (Large Language and Vision Assistant) [18] and GPT-4V [19], allow multimodal reasoning and suggestion generation from both images and structured inputs. Although their adoption in climate and urban studies is still nascent, Zhang et al. [20] show how prompt engineering can align VLM outputs with sustainability goals.

Complementary systems have explored integrating AI into participatory planning tools or environmental dashboards [21], [22], demonstrating the potential of AI to assist both experts and citizens. Our work extends this trajectory by combining fine-grained segmentation, UHI prediction, and VLM-guided recommendation into an accessible, low-cost urban resilience tool.

In summary, while prior efforts have addressed individual challenges in detection or mitigation, our proposed framework uniquely unifies segmentation, metadata-based UHI detection, and VLM-driven urban adaptation guidance in a modular pipeline.

III. Methodology

The methodology of this research is designed to bridge the gap between environmental sensing and actionable urban intervention through a modular AI-powered framework. Our approach is structured around two primary components: (1) an image segmentation pipeline for detailed urban material analysis, and (2) UHI detection and a vision language model (VLM) driven engine for generating tailored mitigation strategies. This section elaborates on each component, the supporting system architecture, and the validation measures employed.

A. Component 1: Image Segmentation Pipeline

Image Acquisition and Prepossessing: To capture the diverse and complex nature of urban environments, we developed a custom mobile application that enables users to photograph cityscapes from various points of view. This ensures that our data set includes a wide range of urban forms, from dense commercial districts to residential neighborhoods and green spaces. Each image undergoes a standardized preprocessing workflow, including resizing, normalization, and color correction, to optimize data quality for downstream analysis.

Object Detection and Semantic Segmentation: The initial phase focuses on the robust identification and precise delineation of objects of interest within a single 2D image. This

process leverages a multi-stage computational pipeline that integrates advanced deep learning models to achieve coarse-to-fine object localization. The pipeline begins with the application of a You Only Look Once (YOLOv8) model, pre-trained on the Cityscapes dataset, to perform initial object detection. This model processes the input image in a single forward pass, generating class-labeled bounding boxes for relevant urban entities such as 'building', 'road','sidewalk','vegetation' and 'wall'. While effective for localization, these bounding boxes lack the geometric precision required for detailed material analysis. To address this, the bounding box coordinates serve as prompts for the subsequent segmentation phase, where the Mobile Segment Anything Model (MobileSAM) is employed to generate high-fidelity, pixel-level masks for each detected instance. By leveraging these coordinates, MobileSAM accurately isolates each object from its background and adjacent instances. A post-processing step involving morphological operations is applied to remove noise and fill small holes, ensuring clean, contiguous object representations that are critical for subsequent analysis.

Material Classification: Following the successful segmentation of objects, the methodology shifts to classifying their constituent materials, addressing the challenge of material heterogeneity in large structures like buildings. To achieve this, a hierarchical analysis approach is adopted, where each object mask is sub-divided into smaller, materially homogenous segments. This is accomplished by systematically generating a grid of point prompts within the primary mask and re-applying the MobileSAM predictor to decompose the object into its constituent parts. This secondary segmentation enables a granular and accurate material assessment.

The classification of these segments is performed using a hybrid engine that combines the semantic capabilities of the Contrastive Language-Image Pre-Training (CLIP) model (ViT-B/32) with image processing heuristics. The CLIP model evaluates each image patch against a curated list of detailed, context-aware text prompts describing various materials (e.g., "close-up of red brick texture with visible mortar lines," "reflective glass facade showing sky reflection"). Its ability to measure similarity between visual and textual concepts provides a robust primary classification. To enhance accuracy and resolve ambiguities, this score is integrated with three heuristic analyses: (1) a color analysis examining the mean RGB value of the patch to favor materials consistent with the detected color profile; (2) a texture analysis, based on the variance of the Laplacian, to quantify surface complexity and distinguish between smooth materials like glass and textured ones like brick; and (3) a positional prior considering the segment's vertical location within the image, leveraging architectural conventions as a contextual clue. The final material label for each patch is determined through a weighted fusion of the CLIP scores and these heuristics, ensuring a resilient and context-aware classification. The results from all sub-segments are aggregated to provide a comprehensive quantitative summary of the material composition for the selected objects.

Surface Area Calculation: The estimation of an architectural structure's surface area from a single 2D photograph presents a significant challenge due to the inherent ambiguity of scale in monocular vision. To address this, a hybrid computational pipeline is employed, synergizing deep learning with 3D geometric reconstruction to transform a qualitative RGB image into a quantitative, metrically accurate 3D model. The process begins with depth inference using the apple/DepthPro-hf model, a Vision Transformer-based architecture that generates a high-resolution depth map, where each pixel's value corresponds to its estimated distance from the camera. A user-provided binary segmentation mask is applied to isolate the building's geometry from the surrounding environment, ensuring that only the target structure is analyzed.

The critical challenge of calibrating the geometrically plausible but metrically arbitrary point cloud is addressed through a statistically robust scaling technique. Instead of relying on the unstable absolute minimum depth value, which is susceptible to outliers, the 1st percentile of all positive depth values within the masked region is computed to provide a stable representative value for the building's closest surface. This value is calibrated against a user-provided ground-truth distance (in meters) to the closest point on the object, yielding a scaling factor that anchors the entire point cloud in metric space.

The scaled point cloud is then converted into a continuous, measurable surface using the Poisson Surface Reconstruction algorithm, chosen for its robustness to noise and ability to generate high-quality, watertight 3D meshes. The algorithm leverages estimated normals to compute a continuous surface, with low-density vertices removed to eliminate artifacts. The total visible surface area is calculated by summing the areas of all constituent triangular facets in the final mesh, providing a precise quantitative output for the pipeline.

As illustrated in Fig. 1, the architecture of our segmentation pipeline integrates each of these steps into a unified workflow, ensuring accurate and interpretable material mapping for urban heat island analysis.

B. Component 2: UHI Detection and VLM-Based Mitigation Suggestion

Structured Metadata and Sampling: Each segmented scene is divided into spatial segments (e.g., roads, buildings, poles, vegetation), each treated as an independent data sample. For each segment, structured metadata—including material type, surface temperature, humidity, and surface area (in cm^2)—is extracted using a multi-stage image processing pipeline. To improve model generalizability, additional training data was incorporated from open-source datasets (e.g., Kaggle), ensuring diversity in object types, materials, and urban layouts.

UHI Classification: Each segment's metadata is numerically encoded (e.g., material mapped to an integer ID) and passed to a trained logistic regression model built using Scikit-learn and NumPy. The classifier outputs a binary prediction indicating whether the region is likely contributing to Urban Heat Island (UHI) formation. Aggregated indicators such

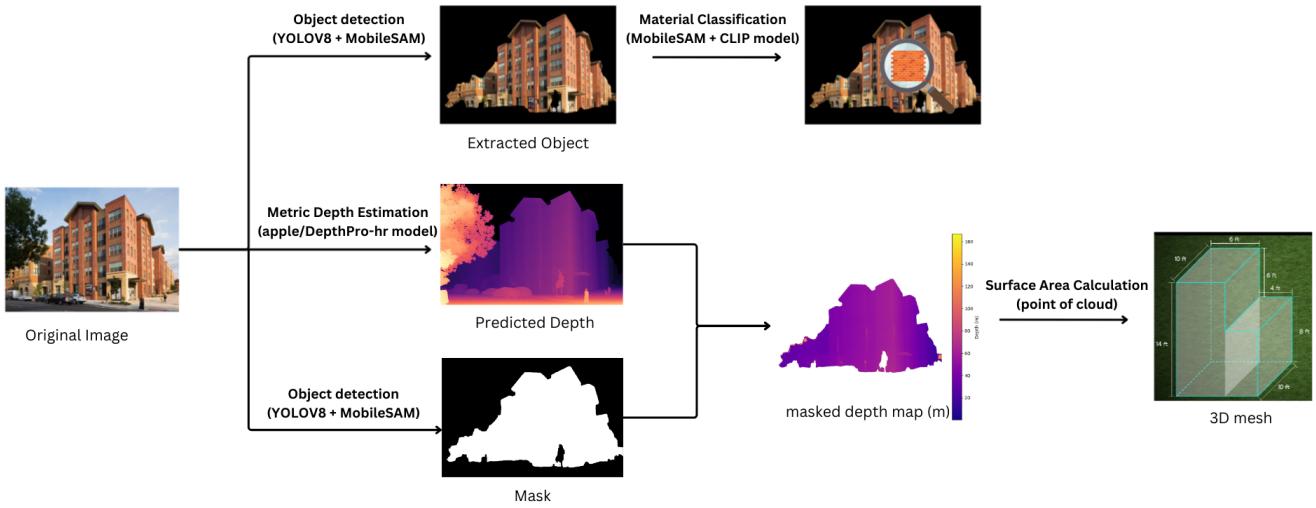


Fig. 1: Architecture of the Image Segmentation Pipeline.

as average temperature, heat-retaining surface coverage, and vegetation percentage are calculated to assess overall heat island conditions at the scene level.

Prompt Engineering and VLM Inference: If a heat island is detected in the scene, a structured natural language prompt is dynamically generated using both the aggregated metadata and the original urban image. This prompt is passed to a visual language model—specifically, `Gemini 1.5 Flash` from Google via the `google-generativeai` Python SDK. The prompt includes quantitative metrics (e.g., "Heat-retaining surface: 99.6%") and instructs the model to output three actionable and affordable mitigation strategies. Gemini's response is then post-processed to remove markdown artifacts, resulting in a clean, ready-to-display recommendation set.

Visual Heat Map Overlay: To support interpretability, the system overlays visual annotations on the original input image. Each segment is marked with a color-coded bounding box: red for predicted heat island contributors, green for non-contributing regions. Labels (e.g., "road", "building") are placed above each box to aid spatial context.

Explainable AI Integration: To enhance transparency, explainable AI (XAI) techniques such as SHAP (SHapley Additive exPlanations) are employed. For each segment, feature importance scores are computed to highlight which metadata attributes (e.g., material type or temperature) had the most influence on the classification outcome. This interpretability mechanism helps urban planners better understand the rationale behind each mitigation recommendation.

C. Validation, Reliability, and Ethical Considerations

Validation and Reliability: Model performance is evaluated using standard metrics such as Intersection over Union (IoU), accuracy, precision, recall, and F1 score. The consistency of VLM-generated recommendations is assessed through repeated trials and expert review, ensuring reliability and practical relevance.

Ethical Considerations: All images and metadata were either synthetically generated or sourced from public datasets, including Kaggle, with strict adherence to privacy and licensing requirements. No personal or sensitive information is collected or processed. The generated recommendations are intended for academic and planning research, not for direct real-world deployment without further validation.

D. Limitations and Challenges

Sample Size and Data Diversity: Some material classes are underrepresented in the training dataset, which may affect classification accuracy. Ongoing data collection and augmentation efforts aim to address this limitation.

Model Constraints: Although VLMs provide valuable recommendations, they may occasionally generate vague or less relevant suggestions. We mitigate this through prompt refinement and post-processing.

Computational Resources: Resource constraints, particularly GPU availability, are managed by optimizing batch sizes and model precision during training and inference.

As shown in Fig. 2, the architecture of the VLM component integrates metadata processing, machine learning classification, and vision-language inference into a cohesive pipeline for actionable Urban Heat Island mitigation.

This comprehensive methodology enables precise detection of UHI-prone areas and the generation of actionable, context-aware mitigation strategies, forming the core intelligence of the HeatScape system.

IV. Results and Discussion

A. Segmentation Accuracy

This study marks the initial phase of a computer vision-based framework for urban surface analysis, with a focus on segmenting key urban components using a two-stage pipeline: YOLOv8 for object detection and the Segment Anything

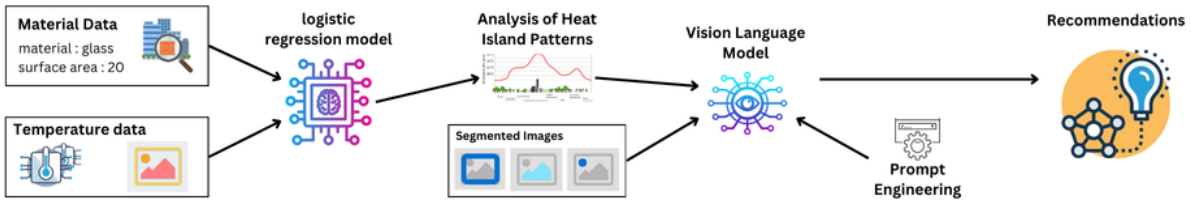


Fig. 2: Architecture of the Vision-Language Model

Model (SAM) for instance segmentation. This phase aims to evaluate the visual efficacy of the detection and segmentation processes, laying the groundwork for subsequent material classification and spatial analysis.

The YOLOv8 model, trained on a curated subset of the Cityscapes dataset, was utilized to detect fundamental urban features, including buildings, roads, sidewalks, vegetation, and fences. The resulting bounding boxes served as input prompts for SAM to generate high-resolution, object-specific segmentation masks.

While a formal quantitative evaluation (e.g., using Intersection over Union or pixel-wise accuracy) is pending, qualitative analysis reveals that the pipeline demonstrates strong visual performance. In particular, segmentation masks for prominent and structurally consistent objects such as buildings and roadways closely align with actual object boundaries in the images. SAM effectively enhanced YOLOv8's coarse detections, capturing fine contours and providing high fidelity in complex urban scenes.

However, the system exhibited reduced performance in scenarios involving occlusion, complex textural backgrounds, or adverse lighting. Under these conditions, the mask boundaries occasionally became fragmented or imprecise. These findings highlight the need for improved preprocessing techniques and more robust post-processing in future iterations.

The successful generation of visually accurate segmentation masks establishes a solid foundation for the next phase of this research, which will address material classification and surface area computation. Data collection for this upcoming phase is currently in progress, and future evaluations will incorporate quantitative metrics such as IoU to comprehensively assess system accuracy.

B. Vision-Language Model Accuracy

The proposed AI-driven framework for Urban Heat Island (UHI) analysis demonstrated high efficacy in both the detection and mitigation recommendation phases. A logistic regression classifier was trained using structured environmental and material-related metadata - including surface temperature, humidity, material type, and surface area - to identify regions likely to contribute to UHI. In a representative case study, the model accurately classified a high-risk urban environment characterized by an average surface temperature of 37.0 °C and humidity of 27.1%. Heat-retaining materials such as asphalt, concrete, and metal comprised 99.63% of the total area, while vegetation accounted for only 0.37%, clearly exceeding the defined thresholds for UHI classification.

For the mitigation phase, the framework employed multimodal reasoning through the Gemini 1.5 Pro Vision-Language Model (VLM), which processed both the urban scene image and its corresponding metadata. Gemini exhibited strong contextual awareness, effectively combining spatial patterns with quantitative metrics to generate targeted recommendations. For the aforementioned scenario, it suggested a comprehensive strategy including green roof installations, high-albedo surface treatments, and increased urban vegetation—each directly linked to the detected environmental conditions such as low vegetation coverage and extensive heat-retaining surfaces.

These findings highlight the value of integrating machine learning classification with advanced VLMs for holistic UHI mitigation. The framework's ability to synthesize multimodal inputs—image, structured metadata, and learned representations—presents a scalable and adaptive approach to urban heat analysis. Notably, the performance of Gemini underscores the importance of data-grounded prompting and context-aware reasoning in generating actionable, site-specific interventions. Such capabilities position VLMs as a transformative tool in climate resilience planning, enabling automated yet nuanced insights for sustainable urban development.

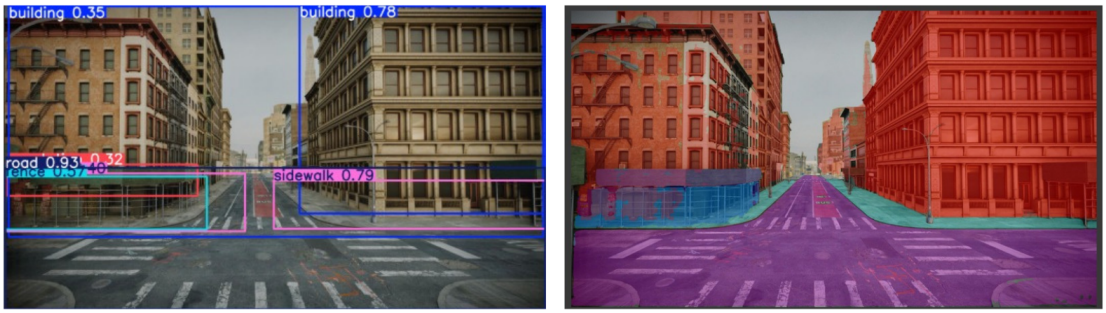
Figure 3 illustrates the outputs of both object detection and instance segmentation. The detection result (left) identifies key urban elements using coarse bounding boxes, while the segmentation output (right) provides the precise, pixel-level masks essential for subsequent material classification and UHI mitigation analysis.

TABLE I: Supervised Model Accuracy Comparison

Model Name	**Type**	**Accuracy (%)**
Logistic Regression	Linear Model	**94.25**
Random Forest	Ensemble (Bagging)	92.80
Support Vector Machine (SVM)	Margin-based Classifier	90.60
K-Nearest Neighbors (KNN)	Instance-Based Learning	88.90

V. Conclusion

This study proposed a module-based AI-based solution, HeatScape, to Urban Heat Island (UHI) effect detection and mitigation protocol by combining image segmentation and vision-language models (VLM). The system successfully fills the conceptual space between raw visual data of the urban environment and planning advice on context through the combination of hybrid semantic segmentation methods and multimodal prompt-based reasoning.



Object Detection Output Segmentation Output

Fig. 3: Object Detection and Segmentation Outputs

YOLOv8 and Segment Anything Model (SAM) have allowed identifying and classifying urban materials contributing to UHI intensity with precision and structured metadata permitted the training of supervised machine learning models specifically Logistic Regression and Random Forest, which could provide UHI classification accurately. Explanatory linear models were found to work in environmental classification tasks; the best prediction was made by logistic regression (94.25%).

The system employs Gemini 1.5 Pro to generate locally tailored mitigation plans using a prompt engineering approach that integrates environmental metadata with visual context. This fusion of structured data and imagery improves the specificity and relevance of the recommendations while supporting the principles of Explainable AI (XAI), thus improving transparency, trust, and confidence in decision making among urban stakeholders.

Although there are certain issues related to the lack of material diversity of the training data, the VLM generic outputs, and the limitations of hardware, the framework can be scaled down and up and is reliable and versatile. The protection of ethics was also treated as a priority by using open-source data and synthetic data.

In conclusion, HeatScape shows how recent AI methods can be central to the design of sustainable cities. Future studies will involve incorporation of real-time data provided by SLAM sensors, and addition of more data on materials and operation in a live urban setting with the ability to keep learning. This study has been one of the contributions to an effective, intelligent, data-based environmental resilience to cities.

REFERENCES

[1] J. A. Voogt and T. R. Oke, "Urban heat island: causes and solutions," *The Urban Climate*, vol. 29, no. 3, pp. 199–206, 2003.

[2] D. Li and E. Bou-Zeid, "Urban heat island effect simulation and mitigation strategies: A review," *Advances in Climate Change Research*, vol. 10, no. 4, pp. 203–212, 2019.

[3] F. Wang, R. Zhang, and X. Xu, "Urban planning with remote sensing and ai: Applications and challenges," *International Journal of Applied Earth Observation and Geoinformation*, vol. 70, pp. 23–34, 2018.

[4] T. Bottigheimer and H. Muller, "Envi-met simulation-based evaluation of urban microclimate and mitigation strategies," *Sustainable Cities and Society*, vol. 46, p. 101414, 2019.

[5] L. Zhang, Q. Huang, and F. Li, "Urban scene understanding with deep learning for smart cities," *IEEE Transactions on Industrial Informatics*, 2022.

[6] J. Tan, W. Zhang, and L. Zhao, "Predicting urban heat island intensity using machine learning algorithms," *Sustainable Cities and Society*, vol. 65, p. 102623, 2021.

[7] Q. Weng, "Remote sensing of impervious surfaces in the urban areas: Requirements, methods, and trends," *Remote Sensing of Environment*, vol. 85, no. 3, pp. 214–225, 2004.

[8] A. Mathew and R. Krishnan, "Thermal infrared remote sensing for urban climate research: A review," *Urban Climate*, vol. 36, p. 100789, 2021.

[9] X. Hu, X. Li, and et al., "Deepuhi-net: Deep learning-based high-resolution mapping of urban heat islands," *ISPRS Journal of Photogrammetry and Remote Sensing*, 2021.

[10] A. Arrieta, D. Garcia, and B. Qadir, "A deep learning approach for urban heat island prediction from aerial images," *Sustainable Cities and Society*, vol. 80, p. 103778, 2022.

[11] M. Alsharif and S. Kim, "A review of deep learning applications for urban heat island detection and mitigation," *Remote Sensing*, vol. 15, no. 2, p. 452, 2023.

[12] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, 2015.

[13] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Rethinking atrous convolution for semantic image segmentation," in *arXiv preprint arXiv:1706.05587*, 2017.

[14] X. Zhou, P. Krähenbühl, and V. Koltun, "Semantic segmentation with densely connected convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[16] A. Kirillov, E. Mintun, N. Ravi, and et al., "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[17] R. Fan, Y. Lin, T. Wu, and X. Zhang, "An ensemble learning approach for predicting urban heat island intensity using environmental indicators," *Sustainable Cities and Society*, vol. 92, p. 104521, 2023.

[18] H. Liu, C. Zhang, and et al., "Llava: Large language-and-vision assistant," *arXiv preprint arXiv:2304.08485*, 2023.

[19] OpenAI, "Gpt-4 with vision (gpt-4v)," https://openai.com/research/gpt-4v-system-card, 2023, accessed: 2025-06-16.

[20] J. Zhang, Y. Wu, S. Chen, Y. Wang, and E. Zhang, "A survey on vision-language models and applications," *arXiv preprint arXiv:2301.04161*, 2023.

[21] J. Kim, Y. Li, Z. Wang, and Y.-C. Lin, "Greenai: A participatory tool for neighborhood-scale urban heat mitigation using ai and urban analytics," *Computers, Environment and Urban Systems*, vol. 95, p. 101830, 2022.

[22] A. Youssef, M. Ibrahim, and A. Al-Ali, "Towards resilient cities: A smart dashboard for urban heat island monitoring and mitigation," in *Proceedings of the 18th International Conference on Smart Cities*. IEEE, 2021, pp. 144–151.