scientific reports

Check for updates

OPEN

Mapping facade materials utilizing zero-shot segmentation for applications in urban microclimate research

Nada Tarkhan[1,3]✉, Mikita Klimenka[1,3], Kelly Fang[2], Fabio Duarte[2], Carlo Ratti[2] & Christoph Reinhart[1]

To address the Urban Heat Island (UHI) effect—a significant urban climate challenge—detailed urban microclimate modeling is essential. Such modeling typically requires data on urban surface properties and morphologies from street canyons and buildings. Most urban surveying efforts have focused on morphological attributes such as sky view factor, vegetation or building surface ratio, while the mass-collection of facade materials has been hindered by the complexity of the segmentation task and the need for large and diverse labeled datasets. Recognizing the importance of mapping facade materials for urban thermal comfort, envelope heat emissions, and building energy studies, we employ computer vision-based state-of-the-art zero-shot learning paradigms for high-fidelity facade material extraction. Our approach circumvents the traditional need for extensive labeled training data, allowing for adaptation to a variety of urban contexts and material types. Tested in Dubai, Amsterdam, and Boston (three architecturally diverse cities), our algorithm successfully detects the predominant facade material in 68% of cases and identifies the top three present material classes in 85% of cases. Additionally, we show how material coverage identification is crucial for assessing outdoor thermal comfort, as evident in shifts in annual cold and heat stress hours across the climates of the three cities in a sample urban canyon.

In the context of urban climatology, quantifying urban characteristics plays an integral role in the advancement of various urban science studies. With shifts in climatic extremes, cities must analyze the effects of new and existing construction on local radiative fluxes and wind patterns at high spatial and temporal resolutions. In recent decades, the phenomenon of Urban Heat Islands (UHI) has emerged as a critical area of study within the urban planning and environmental science disciplines. The UHI phenomenon can be traced to the alteration of natural landscapes into built environments where the urbanization process replaces vegetation with asphalt, concrete, and buildings, surfaces that absorb and retain solar radiation more than natural landscapes. This retention of heat is further amplified by anthropogenic heat release from vehicles, industrial processes, and air conditioning units, contributing to temperature increases[1–5]. The UHI effect exacerbates the impacts of global warming on urban populations, making it a pivotal concern for sustainable urban development. Among the key drivers of the UHI effect are climate, anthropogenic heat, sky view factor, heat absorbing surface materials as well as building morphology[6].

The implications of the UHI effect are profound, including but not limited to, escalated energy demands for cooling, increased emissions of air pollutants and greenhouse gasses, a rise in heat-related illnesses, and the degradation of thermal comfort[7–9]. These ramifications underscore the urgency of identifying effective strategies to mitigate the UHI effect and measure it accurately, thereby enhancing urban livability and resilience to climate change. Among the set of factors influencing the UHI phenomenon, the material composition of urban facades stands out due to its direct interaction with solar radiation. Properties such as albedo (the measure of reflectivity), thermal emissivity, and specific heat capacity determine the extent to which urban surfaces absorb, retain, and emit heat[10]. Materials with high albedo values, capable of reflecting a substantial portion of incoming solar radiation, present a promising avenue for UHI mitigation[11]. However, the heterogeneity of urban environments, composed of a vast array of materials, presents a significant challenge for systematically identifying and analyzing these surfaces on a large scale.

[1]School of Architecture, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]Nada Tarkhan and Nikita Klimenko contributed equally to this work. ✉email: ntarkhan@mit.edu

To examine the implications further, especially on the nuanced assessment of urban thermal comfort, the role of albedo variations is intricately linked with Mean Radiant Temperature (MRT), a crucial determinant of surface temperature. Urban surfaces showcase significant diversity in albedo—for instance, brick could range from 0.2 to 0.4, concrete from 0.1 to 0.35 and untreated glazing from 0.07 to 0.15[12,13]. Although most building materials exhibit high long-wave emissivity, it's the variability in albedo that notably affects MRT. High-albedo materials, while mitigating UHI through their reflective properties, can also redirect shortwave radiation through windows, potentially increasing building heat loads either through diffuse or specular reflection. This underscores the importance of a balanced approach in material selection to optimize for thermal comfort, both indoor and outdoor.

Although the integration of accurate material priors in the field of urban analytics holds promise in advancing several tracks of urban and building assessment research, limited studies have explored the detection of different material classes in urban environments due to the complexity of the problem. The proliferation of computer vision models has now enabled multiple challenges in urban detection to be overcome. One of the significant challenges in this domain is the exhaustive dataset requirement inherent to urban detection tasks. Data scarcity, compounded by the difficulties in collection, annotation, and inspection, poses a substantial barrier. Specifically, in the material detection problem, labeling large datasets of every material present in a facade is extremely challenging and time consuming. In addressing the challenges of urban material accounting and advanced compositional analysis, our research focuses on two pivotal attributes: the development of cost-effective methodologies for material categorization and segmentation through state-of-the-art zero-shot learning methods[14] and scalability across diverse urban environments. Our proposed solution aims to diminish the model's dependency on labeled datasets, a strategy central to the goals of zero-shot learning. This approach enables the model to classify unseen classes, offering an invaluable solution in contexts where large, annotated datasets are unavailable. Despite the method applicability, a critical challenge remains in the resolution of the material segmentation problem, specifically in determining the optimal encoding for multiple material features, which necessitates further investigation and experimentation for fine-tuned texture detection.

To enhance urban climate adaptation strategies and facilitate more precise urban thermal performance studies, our research leverages zero-shot learning for high-precision urban material classification and segmentation. This approach yields three key outputs:

- *Material classification:* an identification of predominant material categories in a facade, providing a foundational understanding of material presence on building facades.
- *Facade composition analysis:* a quantification of the percentage distribution of materials on a facade, addressing the challenge as a multi-material segmentation issue that informs façade composition.
- *Thermal impact assessment*: a demonstration that enables the utilization of facade material coverage knowledge for thermal assessments. Here we showcase an experiment to examine how variations in material properties affect outdoor thermal comfort, highlighting the significant influence of facade materials in an urban canyon utilizing the Universal Thermal Climate Index (UTCI)[15].

This study presents several key contributions that differentiate it from existing research in urban material segmentation and climate studies. While prior works have mainly focused on urban morphology (e.g., sky view factor, vegetation, building surface ratios), our research tackles the underexplored challenge of urban facade material classification using state-of-the-art zero-shot learning methods. Unlike traditional approaches that rely on extensive labeled datasets and are restricted to specific regions or limited material classes, our method reduces the dependency on labeled data, enabling scalable segmentation across diverse urban settings such as Dubai, Amsterdam, and Boston. Furthermore, while urban heat island (UHI) research has primarily emphasized vegetation and geometry, the critical influence of facade materials always posed a data availability challenge. Our method addresses this gap by enabling large-scale mapping of material heterogeneity, which can unlock key insights into Mean Radiant Temperature (MRT) and outdoor thermal comfort. Lastly, by providing open-source workflows, our framework is adaptable to other research domains, with potential applications in building retrofits, urban energy efficiency, and environmental monitoring.

Related work: urban material segmentation

Most street view segmentation tasks have leveraged segmentation networks that are pre-trained on large urban scenery datasets, such as ADE20K[16]. Utilized as benchmarks in computer vision, they have led to the improvement of various neural network architectures. Within the context of urban scenery analysis, ADE20K-based segmentation networks have been used to detect the share of pixels belonging to sky, building facades, and vegetation in street view images. This approach is instrumental in studying the magnitude and impact of the UHI effect across urban environments. Due to its robustness for both perspective and fish-eye images of various distortions, ADE20K-based segmentation is especially applicable to this class of tasks which underscores its ubiquitous use in a variety of applications[17,18]. Some of these applications include urban sidewalk material segmentation and glazing area extraction from building facades[19,20]. However, such tasks come with several challenges, specifically related to; specialized applications centering on limited datasets that do not allow for robust training, highly specific geographic contexts or data formats and limited segmentation labels that require re-training if adapted to wider contexts and features.

Consequently, a resultant aspiration is to attain a large foundation dataset, akin to ADE20K, that could include more detailed material categories, yet a limited number of datasets exist. Some specific examples include a building facade materials dataset tailored to road views[21] and a hyperspectral facade material segmentation dataset from light industrial environments[22]. However, since material surfaces have high texture encodings, the labeling strategy would need to be guided by the application domain: for instance, to pursue the goal of

understanding the impact of urban surfaces for urban heat island assessments, the grouping of materials by classes of similar thermal properties may be most useful, while for architectural studies, identifying facade elements may take precedence. As a result, universally representative datasets may pose a challenge beyond the logistics of data accumulation, in that the representative material classes need to cater to the application domain. Previous approaches to this problem have reduced building material detection to a classification task, choosing to identify overall material presence as opposed to compositional elements. These applications have included the classification of building structural typology for seismic risk monitoring or carbon accounting and urban circularity focused applications[23,24]. While these approaches may provide sufficient insight in the respective domain of applications, they do not allow for higher precision segmentation outputs. To address this gap, we introduce a workflow that goes beyond the classification task to focus on segmentation on a patch-level scale and explore the scalability to facades in diverse urban environments.

In recent years, the domain of computer vision and machine learning has been heavily leveraging foundation models which are large neural networks trained on vast arrays of diverse data and trained to encode input information to extract general patterns and subsequently be fine-tuned to a specific application domain. Originally applied to language-related tasks, these approaches made their way into the domain of computer vision. While foundation models initially were used for subsequent training at smaller scale and cost (fine-tuning), the combination of large visual and language models created zero-shot frameworks that can perform classification or object detection without any need to train or pre-train neural networks. This provides a notable advantage to our material segmentation problem where the goal is to produce high granularity instance segmentation without training reliance. Our approach to material segmentation in building facades has thoroughly assessed current emerging models to enhance the identification and localization of different materials—these models are outlined further in the methodology.

Methods

Zero-shot models

To address the challenges of urban facade material segmentation, this study employs a series of state-of-the-art zero-shot learning and foundation models, known for their robust capabilities in computer vision tasks. These models, originally designed for vast and diverse datasets, are adapted here to encode and extract patterns specific to urban material classification without the necessity for extensive labeled training data.

- OpenAI CLIP: utilized for its zero-shot classification capabilities, leveraging text and image embeddings to match facade materials with textual descriptions[25]. Despite its proficiency in material recognition through text prompts, CLIP does not inherently localize material regions within images, necessitating supplementary mechanisms for detailed segmentation.
- CLIPSeg: deployed to enhance CLIP's capabilities, this model uses attention mechanisms to facilitate low-resolution segmentation of materials in images, highlighting specific weighted areas of the image that significantly influence classification decisions[26].
- SAM (segment anything model): acts as a foundational tool for clustering surface types within images[27]. Although it does not specify material types that belong to the same class, it effectively segregates different surface areas and bounding regions, providing a reliable segmentation base.
- Grounding DINO: this zero-shot framework complements the above models by optimizing object detection[28]. It iteratively finds the optimal bounding box containing the object of interest that is prompted through text inputs. It is capable of recognizing specific objects in an image but cannot process ubiquitous surfaces that are not explicitly concentrated in an object-like fashion. Once an object is detected, panoptic segmentation may be used to define its exact boundaries as opposed to simply drawing a bounding box[29].

Recent research efforts have focused on merging these models for unsupervised segmentation[30,31]. These innovative approaches, while promising, currently best suit simple scenes with minimal textures and are challenging for complex building facades. They tend to perform less effectively in delineating textures not confined within specific image areas. Panoptic segmentation offers advancements in delineating their image region contours, by carrying out pixel-wise segmentation while unifying the typically distinct tasks of semantic segmentation (pixel-level class labels) and instance segmentation (object instance detection and segmentation). Yet it primarily focuses on distinct well-clustered objects rather than dispersed textures, indicating a gap in addressing the nuanced requirements of building facade material segmentation. Few-shot learning (for both classification, detection, and segmentation), presents an extended approach, providing an ability to learn from a small dataset (2–30)[32]. This provides an alternative avenue, particularly when combined with fine-tuning large foundation models. However, this area of research is in active development and may exhibit catastrophic forgetting as well as domain bias[33].

While zero-shot foundation models excel at their respective tasks they require further adaptation and tuning to match respective problem domains. Recent advancements in zero-shot learning have significantly expanded its capabilities across various domains. Notable among these are the developments in attention mechanisms integrated into models like CLIPSeg, which improve segmentation by focusing on the most relevant regions of an image, particularly useful in complex urban environments with diverse materials. These mechanisms help refine localization, addressing the challenges of delineating textures that are dispersed across a facade. While zero-shot models reduce the dependency on labeled data, they may still struggle with fine-grained textures or rare materials, whereas supervised approaches might excel in these areas but at the cost of requiring extensive labeled datasets and manual intervention.

Another key advancement is the use of semantic embeddings, which allow models to link visual information with high-level semantic descriptions. This technique has been widely employed in zero-shot object detection,

where models can detect unseen object classes by leveraging relationships between known and unknown categories[34]. In remote sensing, such semantic embeddings have enabled land cover classification with minimal supervision in remote sensing applications[35,36], and in industrial material recognition, they have facilitated the accurate detection of novel materials in complex manufacturing settings as well as in textile applications[37]. The use of multimodal zero-shot learning in robotics is also noteworthy, where tactile texture recognition is performed without prior tactile training samples by combining visual and semantic information. This method has demonstrated the ability to classify previously unseen materials through tactile sensing, highlighting the versatility of zero-shot learning in recognizing complex material properties across domains beyond vision-based tasks[37]. These technical developments are particularly relevant to urban material segmentation, where the heterogeneity of facade materials and the limited availability of annotated datasets pose significant challenges. By incorporating attention-based mechanisms and semantic embeddings, we improve the detection and segmentation accuracy in diverse urban environments, enabling more scalable and flexible urban material mapping.

Hence, the application of foundation zero-shot frameworks to urban surface mapping remains nascent. The method outlined in this paper aims to combine and extend existing zero-shot approaches into a joint workflow that specifically addresses the needs of the urban scientist. Specifically, we hope to leverage the strengths of the different models and assess their applicability to the different segmentation sub-tasks. Given the ever-evolving nature of computer vision frameworks, our goal is not to offer a definitive way to segment materials but rather to demonstrate the power of zero-shot models for urban material applications and offer one potential open-source strategy.

Workflow overview: image sourcing, detection and segmentation

The first step in the workflow is image sourcing and processing. The target images are those that show direct views of the facade with minor perspectival distortions, as these are the views that typically yield the highest clarity. StreetView 360[38] was used to download high-resolution urban panoramas and subsequently warp the panorama to obtain corresponding frontal views of building facades. Panoramic images were selected for their capability to capture multiple buildings at once, increasing the efficiency of large-scale urban material analysis. It is also possible to implement multi-view integration to improve segmentation accuracy across different perspectives. We consequently use a semantic segmentation network trained on the ADE20K dataset to filter out all elements not related to urban facades such as streets and vegetation before we carry out façade segmentation. This is detailed in Fig. 1.

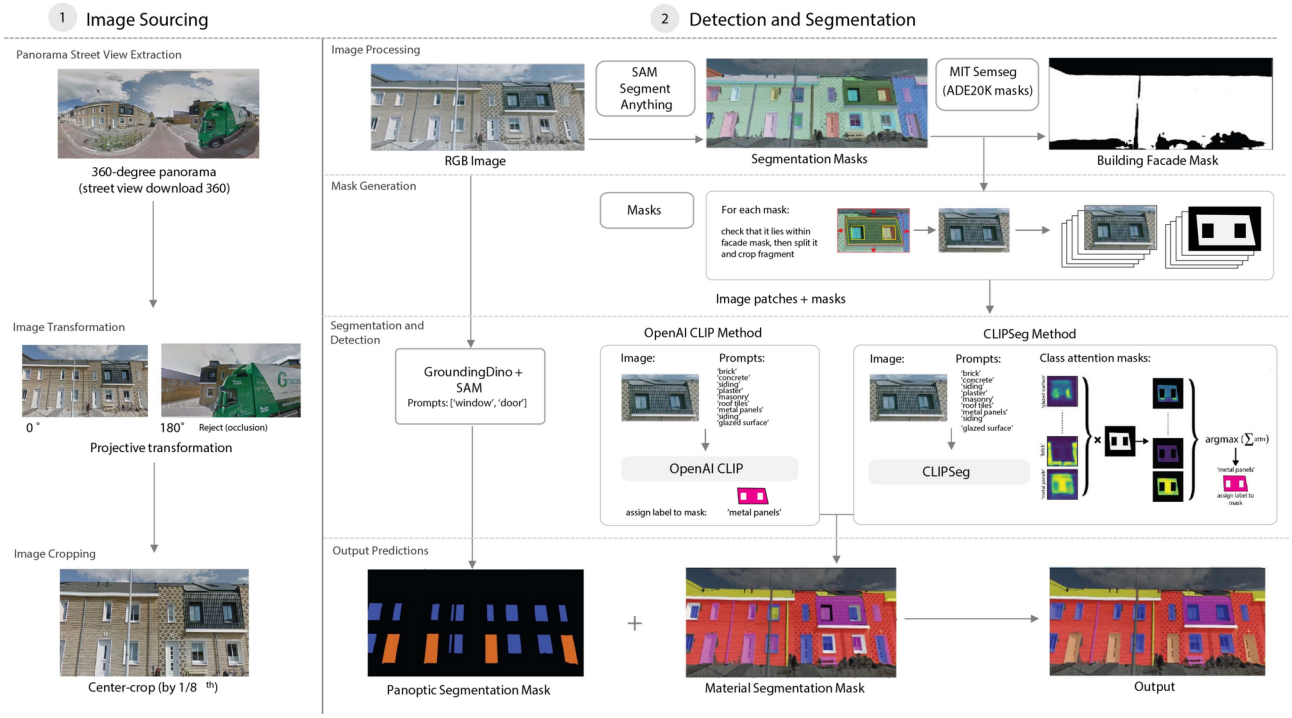


Fig. 1. Workflow overview for Image sourcing, transformation and joint detection-segmentation of facade regions (materials) and objects (elements). This has two parts: image sourcing: A schematic detailing the preparation of street view panoramas for subsequent segmentation, utilized in the preparation of the validation dataset and subsequent neighborhood-scale material mapping. Detection and Segmentation: a layout of the segmentation workflow. At the stage of image fragment classification, we explore two approaches: (1) Utilizing CLIP to classify an image patch (2) Using CLIPSeg to calculate the class triggering the most attention within the identified object mask.

In the datasets we collected, we identified the following types of materials present on facades, based on predominant construction classes, typically defined in thermal simulation studies; (1) Brick (2) Concrete (3) Glazing (4) Roof tiles (5) Metal panels (6) Wood siding (7) Plaster/stucco. It is important to equally detect façade objects, such as balconies and Air Conditioners for further analysis as they may also contribute to local microclimate conditions- for instance, balconies provide shading and ACs release heat into urban canyons contributing to anthropogenic heat. Therefore, these objects are detected initially and isolated from the regional segmentation task. The object classes are as follows; (1) AC (2) Storefront (3) Balcony (4) Door. We detect these object-based surfaces with Grounding DINO and feed them into the final panoptic segmentation prediction before we execute material segmentation. To account for facade occlusions, we rejected images with significant obstructions from our dataset and incorporated a process that automatically treats non-material classes, such as cars or lamps, as occlusions, ensuring only relevant materials are segmented. Additional automated methods, such as inpainting and object removal can also be integrated[39].

To tackle the challenge of high texture complexity in the material segmentation task, we propose to apply zero-shot classification to separate smaller patches within the facade. To do so, we first parse the facade into separate segments using SAM. Next, we crop each segment out of the image and classify it as belonging to one of the material labels using OpenAI CLIP which distinguishes between material textures by using contextual information from an image patch, padded by twice the width and height of its bounding box. To improve detection inference, a low-level segmentation is performed provided by CLIPSeg that leverages regions of attention to specific materials. As noted previously, CLIPSeg is adapted to finding regions of attention in data from a base 225×225 transformer mask which makes it unsuitable for performing segmentation on large and complex scenes. However, given that our image fragment constitutes most of the data needed, it is sufficient for distinguishing which classes draw attention within the classification mask. To allocate a material class to the target patch, we compute the normalized sum of CLIPSeg attention that only falls within the boundaries of the mask and return the label with the highest result. Figure 1 details the full pipeline. With reference to detecting material classes, we compared the detection capabilities of OpenAI CLIP and CLIPSeg. CLIPSeg showed higher accuracy in detection across the seven material classes analyzed, hence it was selected for our workflow. In the glazing condition, two distinct approaches were adopted detecting them as both objects (distinct windows) and surfaces (fully glazed facades). This dual strategy allows for a more thorough representation.

For the segmentation process, we utilized image resolutions of either 1000×1000 or 912×912 pixels, depending on the complexity of the scene, and set the classification patch size to 1.3 times the bounding box around each detected element to improve material classification accuracy. For SAM, we fine-tuned several hyperparameters- specifically, we increased the number of points per side to 64 to achieve finer segmentation, set the prediction Intersection over Union (IoU) threshold to 0.75 for boosted accuracy, and raised the stability score threshold to 0.75 for more reliable segmentation results. CLIPSeg tuning involved crafting precise text prompts to accurately capture material classes, as some materials required multiple or more specific descriptions to improve segmentation accuracy. For example, 'plaster' and 'stucco' were both used to describe stucco walls, while 'concrete' was labeled as 'exposed concrete' for clarity. These prompt refinements, referred to as prompt engineering, are detailed in Table 1 in the supplementary data and explored further in the paper for their contribution to prediction accuracy. While the pretrained segmentation models are efficient during inference, computational complexity increases with image resolution and segmentation tasks. Our segmentation runtime depends on the performance of integrated models and includes: (1) semantic segmentation with ADE20K (0.4 ± 0.6 s/image); (2) patch detection with SAM (13.4 ± 1.5 s); (3) iterative classification of all patches across image with either OpenAI CLIP (3.9 ± 2.1 s) or CLIPSEG (23.8 ± 9.8 s). The total segmentation runtime is either 17.6 s (OpenAI Clip) or 37.6 s (CLIPSeg) per image, as run on a single NVIDIA GeForce RTX 2080 Ti. The panoptic segmentation of windows and doors required us to run GroundingDINO on CPU, which added an additional 20.4 ± 0.2 s per image.

Class	IOU	Precision	Recall
Miscellaneous	0.273	0.313	0.768
Vegetation	0.724	0.894	0.743
Glass window	0.471	0.776	0.496
Brick	0.587	0.694	0.603
Concrete surface	0.525	0.706	0.565
Concrete blocks	0.131	0.293	0.136
Metal	0.062	0.500	0.062
Door	0.255	0.278	0.278
Timber	0.003	0.004	0.030
Total	0.337	0.495	0.409

Table 1. Performance by class on LIB-HSI dataset utilizing weighted IoU, Precision, and Accuracy as reported in the Methodology section. The segmentation classes were adopted from the superclass structure of the original LIB-HSI dataset.

Model assessment

The evaluation of our urban facade material segmentation model was systematically conducted in three distinct phases to ensure robustness and applicability across different urban settings and scales. To explore the performance trade-off between zero-shot and pretrained models, we have also included a comparison with SegFormer[40], a state-of-the-art segmentation network that we trained on our dataset.

Close-range image testing

The initial phase of testing involved a dataset comprising 393 close-range material segmentation images in light industrial environments (LIB-HSI)[22]. This dataset was specifically chosen to validate the model's ability to accurately recognize and classify a wide variety of building materials under controlled conditions where texture delineations are more prominent. Besides the RGB images, it provides corresponding infrared images and is primarily targeted at exploring accurate material segmentation via joint RGB and infrared inputs. Here we compare the performance of our algorithm to the segmentation network trained on the RGB portion of the dataset.

Cross-city architectural representation testing

Subsequently, the model was tested on a more diverse set of images to evaluate its performance across varied architectural styles and urban environments. To explore the applications to diverse urban contexts we collected 144 facade images (~50 from each city) from three cities: Boston (North America), Amsterdam (Europe) and Dubai (Asia). The intention was to obtain representative buildings in each material class. For instance, Amsterdam is characterized by a larger proportion of brick construction in its building stock, while some areas of Boston have higher proportions of single-family wood construction. These images were manually labeled using Segments.ai[41] and will be used as ground truth when computing the assessment scores. This selection aimed to evaluate the model's adaptability and accuracy in different global contexts. Alongside our model's evaluation, the trained SegFormer model was also tested on this dataset to benchmark the accuracy of our approach across different urban contexts. Key model parameters used during training included seven material classes, a learning rate of 0.00006, and 10 epochs with a batch size of 10. The training set consisted of 144 images, which was split into 5 folds; 4/5 folds were the training set and 1/5 was the validation set.

Large-scale urban deployment

In the final phase, the model was deployed in a sample area in our assessment cities, measuring roughly 1 km^2 in area. This deployment aimed to assess the model's effectiveness in mapping material distributions at a neighborhood scale. The focus was to showcase the spatial distribution of materials and their prevalence within typical urban blocks, providing insights into urban material distributions and potential impact on urban heat island effects. The assessment points each represent two panorama images with two viewpoints (0 and 180°). Within the spatially defined areas, about 1200 images were captured per city. The number of points per city varied slightly due to varying urban built-up densities.

Evaluation metrics

The first metric computed is a domain specific computer vision metric that assesses the extent of segmentation accuracy. The weighted Intersection over Union (IoU) is an adaptation of the standard IoU metric used extensively in image segmentation tasks to evaluate the overlap between predicted segmentation masks and ground truth masks. This modified metric is particularly relevant to material segmentation where the presence of various materials may be unevenly distributed across an image, influencing their impact on performance metrics. In the context of material segmentation, where different materials can significantly vary in their surface coverage and impact on the building's thermal properties, using a weighted IoU ensures that the evaluation metric aligns with the practical implications of correctly or incorrectly segmenting each material. The weighted IoU is reported per material class and also as an aggregate score using Eq. (1). We also document precision and recall scores to measure positive predictions[42].

$$Weighted\,IoU = \frac{\sum_{i=1}^{n} w_i x\,(IoU_i)}{\sum_{i=1}^{n} w_i} \tag{1}$$

where IoUi is the IoU for each material class and wi is the weight based on computed pixel ratio of the class.

Additionally, we implement a material presence threshold to check predicted material classes that occupy only a few pixels of the image ensuring that minor false positives do not skew the evaluation.

To advance the interpretability further of the detected facade materials, we propose two material-specific metrics that report different granularities. Each of these metrics provides a different insight through which the effectiveness and comprehensiveness of material segmentation can be assessed, to tackle more performance domain specific applications. This approach is particularly relevant in areas where both the detection (presence) of objects or materials and the accuracy of their spatial localization are critical for the application domain- this is only relevant to the cross-city facade validation. This allows researchers and urban planners to focus on precision where it matters most, ensuring that critical material knowledge is derived. They are as follows:

Predominant material class
This metric identifies the material that covers the largest area of the building's facade, indicating the primary material. This could potentially be used to infer construction and surface coverage for higher level studies.

$$\Pr edo\min antMaterial = m \in \arg\max_{m} A_m \tag{2}$$

where *Am* is the area covered by the material *m* and *M* is the set of all detected materials on the facade.

Material presence
This metric measures the completeness of material detection by calculating the percentage of correctly identified materials present on the facade. We utilize this metric to check the presence of the top three material classes present in the facade with the largest area coverage.

$$Material\Pr esence = \left(\frac{N\det ected}{Ntotal}\right) \times 100\% \tag{3}$$

where *Ndetected* is the number of materials correctly detected on the facade, and *Ntotal* is the total number of different materials actually present on the facade.

UTCI impact measurement
To demonstrate the impact of material coverage in our studied cities, the Universal Thermal Climate Index (UTCI) was calculated to compute the annual thermal comfort, heat stress and cold stress hours in a test urban canyon. The UTCI is a measure that integrates the effects of air temperature, wind speed, humidity, and MRT to evaluate human thermal exposure in outdoor environments. MRT was sourced from the building surfaces, representing the average temperature of all surfaces surrounding the midpoint in the canyon area, weighted by the angle of exposure and emissivity of the assigned material. MRT and UTCI were calculated utilizing ClimateStudio[43], a plugin for environmental studies which provided hourly surface temperatures based on physical models of solar radiation, surface properties, and environmental conditions, utilizing the EnergyPlus solver[44]. The canyon geometry measured 5 m in width and 9 m average building height. The adjacent enclosing walls were North and South facing. Nine scenarios were simulated to study the effect of three material classes (brick, glass, and wood) in the three selected cities (Dubai, Amsterdam, and Boston). We assume an 80% facade glazing coverage in the glass setup and model the wood scenario as light-colored siding. The material properties utilized are provided in Fig. 5 along with the results. All other geometric and simulation-based settings were kept constant across the cases. The outcomes helped in understanding the impact of material properties on the surface temperatures and subsequent human thermal comfort levels.

Results
As detailed in the methodology, we showcase the model's performance capabilities in both close-range and context-based images from urban settings. For each subsection, we discuss optimizing prompt input labels to enhance performance- the material classes present in both validation outputs are representative of the labels in the annotated dataset.

Close-range material facade segmentation (LIB-HSI)
The results of this dataset are represented in reference to computed weighted IoU per material class in Table 1. We report that brick, glass, concrete blocks and vegetation have the highest detection accuracies as indicated by the higher scores, within the range of 0.525 to 0.724. At close range, material textures of concrete and brick are clearer, contributing to higher detection accuracy. The lowest scores are found in the metal and timber classes. A possible explanation is that the algorithm is not sensitive enough to make fine-tuned distinctions between concrete surfaces and concrete blocks, resulting in the lower scores. Inspecting outputs further, we observe effective material segmentation mask outputs as shown in Fig. 2.

Method comparison across material classes
To evaluate our zero-shot pipeline, we compared the IoU scores across four different methods, including our full pipeline with CLIPSeg (Method 1) and OpenAI CLIP (Method 2), a direct application of CLIPSeg without the full pipeline (Method 3), and the SegFormer model trained on a small labeled dataset (Method 4). This comparison illustrates how each method performs in the contexts of Amsterdam, Boston, and Dubai across all material classes. Results are shown in aggregate in Table 2.

The comparison of methods highlights key performance differences across the evaluated segmentation approaches. Method 1 (Pipeline + CLIPSeg) achieved the highest overall IoU score (0.366), demonstrating the benefit of incorporating the full pipeline, which includes SAM and Grounding DINO for enhanced segmentation accuracy. Method 2 (Pipeline + OpenAI CLIP) yielded a slightly lower IoU score (0.312), indicating that while OpenAI CLIP provides strong classification capabilities, its performance is slightly below CLIPSeg in segmentation tasks when applied within the pipeline. Method 3 (Only CLIPSeg) performed the weakest (0.232 IoU), confirming that applying CLIPSeg directly without the support of SAM and Grounding DINO leads to suboptimal results, particularly for complex facade scenes. Method 4 (SegFormer) achieved a total IoU score

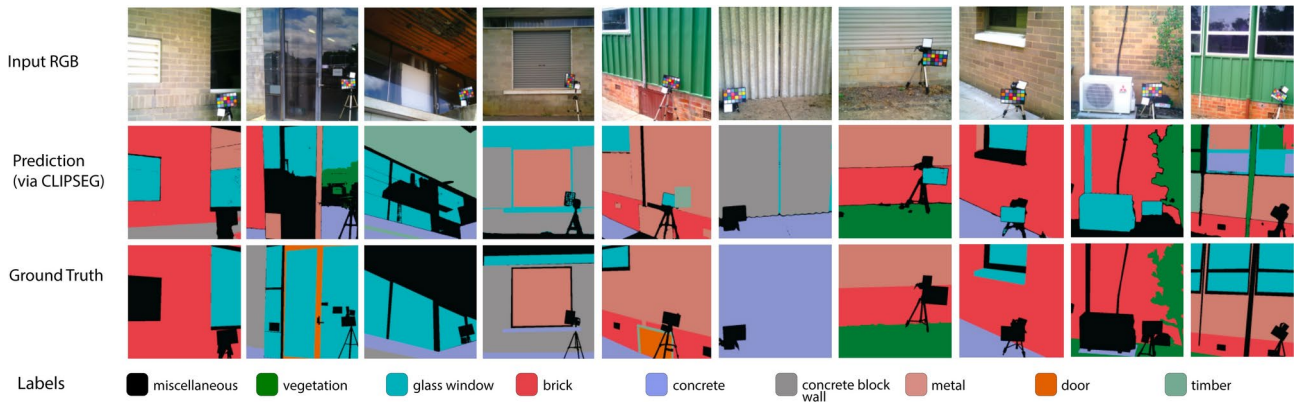


Fig. 2. Example of performance on the LIB-HSI dataset against the superclasses. The examples reveal that some low scores are due to the inconsistency between the class morphology and material surface properties. The ceiling in column 3 is identified as timber, but belongs to the 'miscellaneous' class in the LIB-HSI superclasses. Despite the inconsistencies between Door, Timber, and Metal classes, the network succeeds at differentiating between brick, concrete, vegetation, and metal surfaces.

Class	Method 1: Pipeline + CLIPSeg	Method 2: Pipeline + OpenAI CLIP	Method 3: Only CLIPSeg	Method 4: SegFormer
Glass	0.610	0.601	0.530	0.444
Concrete	0.165	0.121	0.043	0.074
Brick	0.479	0.436	0.379	0.437
Siding	0.448	0.196	0.100	0.274
Plaster	0.304	0.234	0.252	0.243
Metal	0.092	0.132	0.086	0.215
Door	0.463	0.463	NaN	NaN
Total (weighted)	0.366	0.312	0.232	0.281

Table 2. Aggregate performance across different detection methods. Table 2 shows the aggregate performance (measured in IoU) of various detection methods (Pipeline + CLIPSeg, Pipeline + OpenAI CLIP, Only CLIPSeg, and SegFormer) across multiple material classes, highlighting the overall performance against Method 1: Pipeline + CLIPSeg which is the one implemented.

of 0.281. While better than Method 3, it falls short of the full zero-shot pipeline, underscoring the advantage of zero-shot learning in material segmentation across diverse urban settings without the need for extensive labeled data. Method 1 achieved the highest IoU for glass, siding and plaster, demonstrating its strength in detecting diverse materials. Doors were omitted in the SegFormer method because they are treated as objects in the framework utilizing Grounding DINO.

We conducted three sensitivity analyses to investigate the performance of our CLIPSeg model under different conditions: (1) altering the patch factor, (2) excluding prompt engineering, and (3) reducing image resolution. Altering the patch factor from 1.3 to 2.0 resulted in a performance degradation, with a 9.6% decrease in IoU, indicating that larger patches reduce the model's ability to capture fine-grained material information. Reducing the patch factor to 1.0 also led to a decrease in IoU, although less severe than the patch factor of 2.0. Excluding prompt engineering reduced the IoU by 8.2%, highlighting its importance for better material classification and segmentation. Finally, lowering image resolution to 75% led to a significant drop in performance, with a 14.9% decrease in IoU, demonstrating the model's sensitivity to image quality for accurate segmentation. These results emphasize the need for careful optimization of model parameters to achieve robust material segmentation performance.

Cross-city facade segmentation dataset

Using the formulated zero-shot pipeline with CLIPSeg (Method 1), we investigate the performance across different cities by computing the weighted Intersection-over-Union (IoU) metric and aggregating it for each material class within the studied cities. The results are presented in Table 3.

When looking at the city-specific metrics, a clear ordinal trend is evident where the glass, brick and door classes consistently display the highest IoU scores, with lower scores seen for materials like metal panels, siding, and plaster/stucco. The lower IoUs for certain materials are largely due to their scarce presence in the dataset. In 68% of the cases, the algorithm accurately detects the predominant material on the façade, and for the top three material classes, an 85% detection accuracy is observed, highlighting the strong alignment between ground truth and predictions. Among these, brick stands out with a 92% accuracy in identifying it as the predominant

City	Amsterdam			Boston			Dubai		
Class	IOU	Precision	Recall	IOU	Precision	Recall	IOU	Precision	Recall
Glass	0.711	0.824	0.831	0.550	0.651	0.796	0.564	0.703	0.758
Concrete	0.157	0.253	0.231	0.080	0.200	0.100	0.015	0.140	0.097
Brick	0.578	0.629	0.807	0.450	0.552	0.591	0.092	0.138	0.245
Siding	–	–	–	0.172	0.254	0.313	–	–	–
Plaster	0.217	0.321	0.326	0.342	0.608	0.424	0.273	0.418	0.327
Metal	0.252	0.295	0.446	0.044	0.050	0.104	0.184	0.308	0.333
Door	0.571	0.734	0.656	0.527	0.670	0.580	0.409	0.476	0.461
Total (weighted)	0.414	0.509	0.550	0.310	0.427	0.415	0.256	0.364	0.370

Table 3. Aggregate performance across material categories and target cities utilizing weighted IoU, Precision, and Recall. Glass and brick surfaces have the highest IOU.

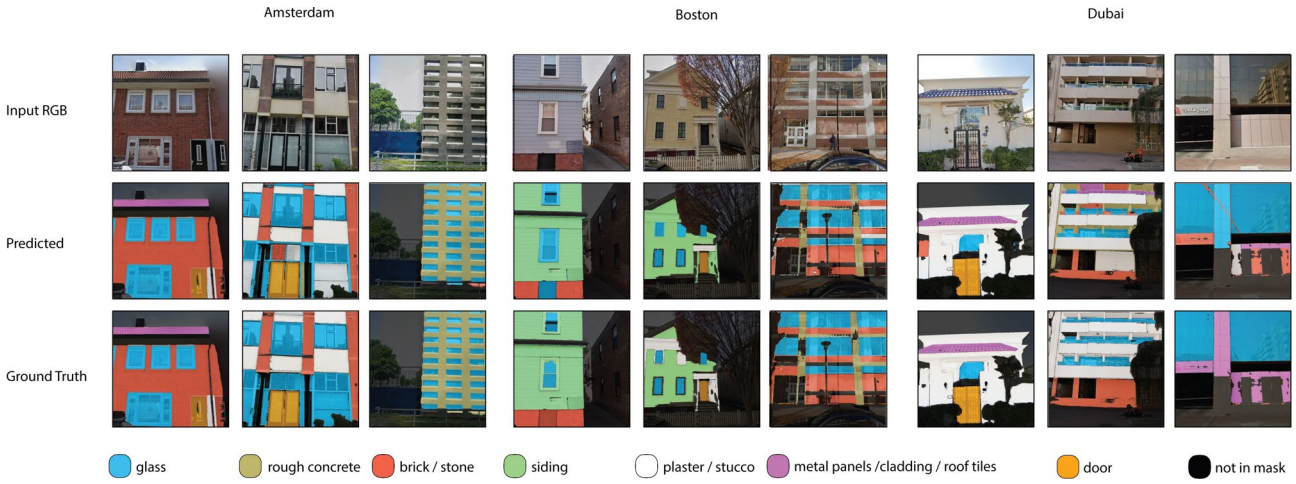


Fig. 3. Example output segmentations on panoramas in Amsterdam, Boston, and Dubai across diverse material types.

material across the analyzed conditions, indicating the robustness of the model for well-represented materials. This indicates a high alignment in composition between ground truth and predictions. The conditions with the lowest scores are images where classes of materials may not be distinct enough to differentiate, often found in the case of plaster and concrete. To quantify these differences, we conducted an analysis of variance (ANOVA) followed by a post-hoc Tukey test, which revealed statistically significant variations ($p < 0.05$) between material types. Glass (IoU = 0.711) and brick (IoU = 0.578) showed significantly higher IoU scores compared to concrete (IoU = 0.157) and metal panels (IoU = 0.252). Significant differences were also observed between cities, with Boston outperforming Dubai, particularly for materials like plaster and stucco, where Dubai's scores were significantly lower ($p < 0.05$). The detailed calculations are provided in the supplementary data.

We display the performance across representative material classes for a sample subset of the results in Fig. 3. In addition to this, urban obstructions and distortions in viewing angles may also contribute to lower scores. Within diverse urban contexts we also note that resolution of images plays a large role specifically in larger street widths such as those present in Dubai. This is one possible explanation behind the lower aggregate IoU score present there. Another explanation for lower detection accuracies in Dubai can be attributed to the presence of visually challenging construction typologies, dominated by a wide variation in plaster and stucco coatings. In contrast, Amsterdam exhibits the highest scores, given the large representation of glazing and brick, two of the model's highest-performing classes.

To investigate the correlations with morphological attributes in the analyzed images and their impact on the model's performance, we explored the relationship between scene morphological elements and IoU scores. The analysis shows that building fraction has a positive correlation with IoU ($r = 0.33$), where r represents Pearson's correlation score, indicating better model performance when a larger portion of the image consists of building facades, particularly when the building fraction exceeds 0.8.In contrast, sky view factor ($r = -0.29$) and scene elements fraction ($r = -0.22$) both have negative correlations with IoU, suggesting that higher content of sky or scene elements (e.g., cars, trees) reduces the model's performance. This is likely due to the decreased relevance of non-building features for facade material segmentation. These results emphasize the importance of clear building visibility for accurate model performance and highlight how morphological features can influence

model accuracy, with a general correlation observed between larger street widths and an increase in scene elements and sky view factor.

Mapping material distribution in the assessment cities

To demonstrate the applicability and scalability of the workflow on the city-level scale, we show the material distributions in a sample neighborhood in the three selected cities (Boston, Amsterdam, and Dubai) and spatially document the distribution of materials across urban panoramas. In addition to material classes, the visuals indicate percent coverage of the facade per material class. In Boston, brick construction is the most prominent material, with an average coverage of approximately 24%, followed by glazing at 16%. The presence of stucco and siding is minimal, averaging 8% and 6.7%, respectively. These values are consistent with the architectural style of the area, which emphasizes brick-based structures. In Dubai, the distribution is quite different, as stucco dominates with an average coverage of 34%, reflecting the region's architectural style. Glazing follows with an average of 16%, while brick and concrete are less present. These findings confirm general architectural observations in each city and demonstrate the utility of this method in capturing both local material knowledge and detailed distributions at a granular level (shown in Fig. 4).

Use case: impact on outdoor thermal comfort in an urban canyon

This experiment is designed to showcase the impact of material choices on outdoor thermal comfort across the material variations, noting that generally an inverse relationship exists between albedo and surface temperature (Surfaces with high albedo, such as white roofs, reflect a larger portion of solar radiation, absorbing less heat and hence remaining cooler). The results show that for the hottest assessed climate in Dubai, the glazing material shows the highest heat stress in the canyon due to the absorbed and re-emitted solar radiation. The annual heat stress hours are about 5% higher (220 additional hours) than in the lowest stress hours, wood surface condition. No cold stress is experienced in Dubai. In Amsterdam where the climate is temperate, we see less variation across the material classes, yet we note that the glazing coverage exacerbates both hot and cold stress hours annually when compared to other material coverages. The wood surface has the lowest heat and cold stress annual hours. Finally, in Boston we see a similar trend where the glazing coverage results in the highest annual discomfort hours. However, when it comes to brick and wood, we note that while the wood coverage reduces the heat stress hours, the cold stress hours are higher than in the brick condition by about 17.5% (174 additional hours). This

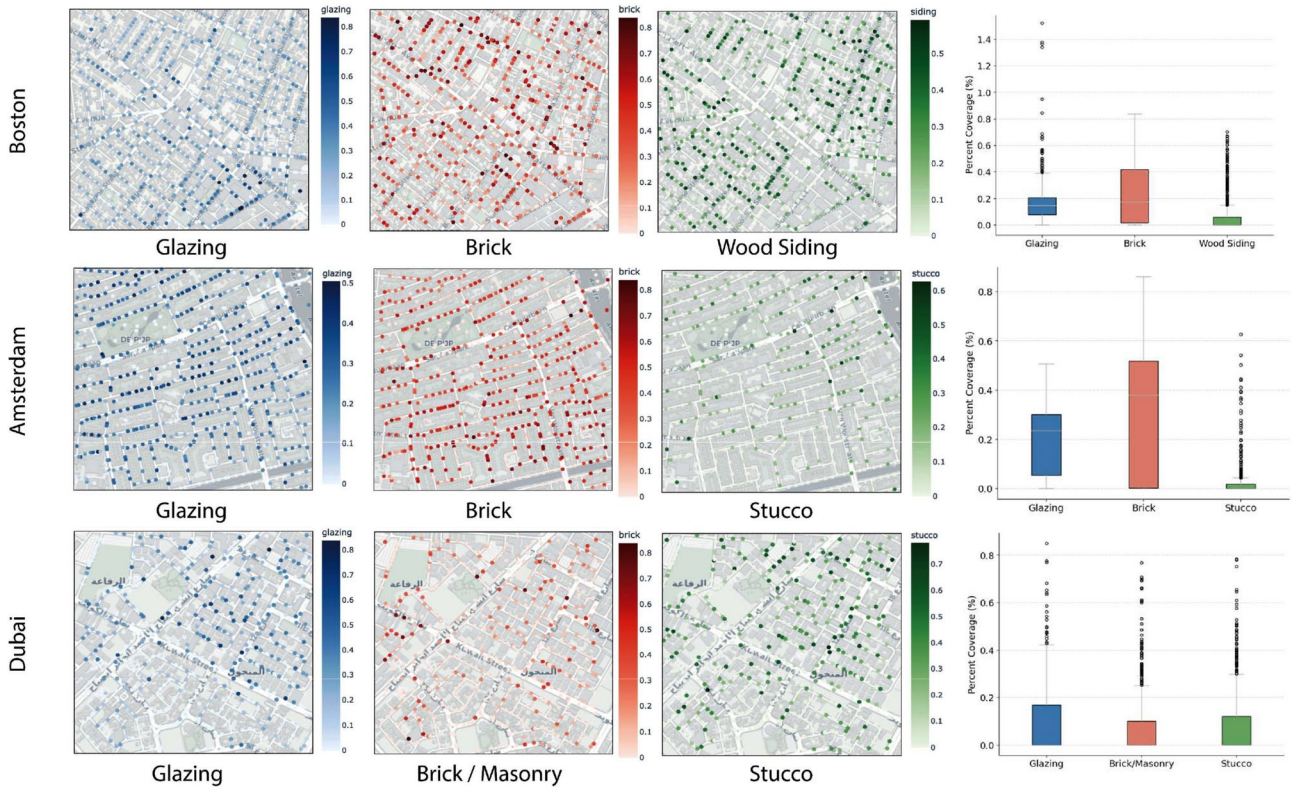
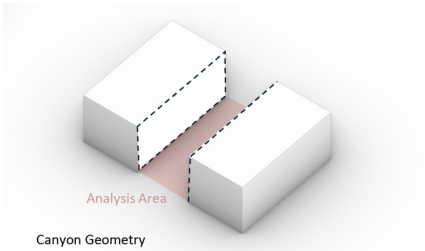


Fig. 4. Visualization of material distributions in Boston, Amsterdam, and Dubai for three most characteristic materials. The distributions corroborate the general observations of the urban morphology. In Boston (Central Cambridge), for example, residential siding buildings concentrate away from the central street (Massachusetts Avenue), while brick and glazing are more predominant along the main street and closer to the MIT Campus. In Dubai, most glazing is observed in the more densely populated area in the north, while stucco and masonry are ubiquitous across the area.

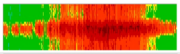
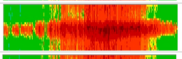
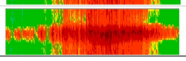
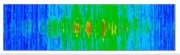
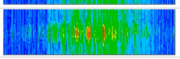
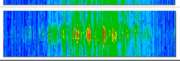
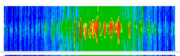
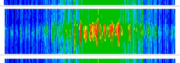
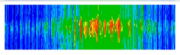


Canyon Geometry

Brick	
Property	**Value**
Solar Absorptance	0.65
Conductivity (W/m·K)	0.7
Density (kg/m³)	1900
Specific Heat (J/kg·K)	850

Glazing (untreated)	
Property	**Value**
Solar Absorptance	0.25
Conductivity (W/m·K)	0.8
Density (kg/m³)	2500
Specific Heat (J/kg·K)	750

Wood siding (light colored)	
Property	**Value**
Solar Absorptance	0.3
Conductivity (W/m·K)	0.15
Density (kg/m³)	500
Specific Heat (J/kg·K)	1800

	UTCI hourly annual Map	Annual Comfort hours	Annual Heat stress hours	Annual Cold stress hours
Dubai				
Brick		2695	4395	0
Glazing (80%)		2521	4521	0
Wood		2690	4301	0
Amsterdam				
Brick		2956	62	572
Glazing (80%)		2901	85	598
Wood		3024	28	541
Boston				
Brick		3226	269	997
Glazing (80%)		3146	275	1051
Wood		3218	150	1171

UTCI Scale

Fig. 5. Impact of Material on UTCI-based thermal comfort in urban canyon across climates. Annual results are showcased in a simulation-based canyon setup to investigate the effect of three main material coverages on outdoor thermal comfort across the three climates. The UTCI metric is utilized and the material assumptions and properties are documented.

is largely due to brick having thermal mass properties and higher conductivity that allows for the retention and release of heat during colder periods. The results are shown in Fig. 5.

Discussion

Outdoor thermal comfort performance and wider applications

While our method provides evidence of the effective use of zero-shot learning to the material segmentation task to extract dominant materials present as well as multi-class material compositions, we would like to note that utilizing this insight in urban heat island and thermal comfort research would require additional validation. With the proliferation of various composite materials, surface coatings and cool wall solutions, the extraction of advanced material properties such as thermal emissivity, roughness and conductivity becomes increasingly challenging. As a result, the inference of exact MRT and albedo values may not be directly possible. Nevertheless, as demonstrated by the surface temperature results under the different material classes, identifying the predominant material coverage can still provide substantial knowledge. On the use cases surrounding outdoor thermal comfort, extensive literature has studied the effects of material properties on UTCI metrics[45,46]- here we utilize our simulation outputs to demonstrate the importance of material selection, especially in more extreme climates as indicated by the higher stress hours in Dubai and Boston.

For extended validation, a noteworthy method involves integrating the material-specific temperature profiles, obtained through simulation software or physical testing. This approach enables a direct comparison between predicted material effects on urban thermal environments and their real-world thermal behaviors. Furthermore, envelope thermography offers a robust approach for validating the thermal impact of detected material classes[47]. By overlaying thermal images with segmented material maps, researchers can visually and quantitatively assess how different materials contribute to surface temperatures within urban landscapes. Such validation not only reinforces the reliability of the segmentation results but also provides tangible evidence of the materials' roles in urban heat dynamics.

By integrating this method into urban planning tools and energy modeling software connected to platforms such as Rhino 3D, stakeholders can evaluate how material choices affect thermal comfort and surface temperatures, and even indoor conditions through construction typology inference, supporting decisions around retrofitting initiatives, and climate adaptation strategies. This approach can also be applied in regulatory frameworks to enforce building codes related to energy efficiency and urban resilience. For example, planners could use material segmentation results to prioritize the use of high-albedo surfaces in existing developments, helping to mitigate the Urban Heat Island (UHI) effect. The ability to perform large-scale, automated assessments could help governments and municipalities prioritize interventions based on material composition where it's most critical.

Limitations and improvements

One major limitation of the method is its reliance on patch-based image processing, which is dependent on high-resolution panoramas. Obtaining high resolution street view imagery globally remains a challenge, yet crowd sourcing avenues in some contexts provide promise[48]. Another challenge lies in erroneous data labeling that may arise from unclear material patches, difficult to discern visually. Additionally, despite the fact that our workflow is based on zero-shot algorithms, some preliminary tuning and decision-making is still required to adapt it to new contexts. This primarily relates to segmentation intensity within SAM and the choice of class labels for materials. Finally, as our workflow is based on processing individual patches, it requires invoking a zero-shot classifier for every fragment of the image, which may take significant time for very complex façade scenes.

Another inherent limitation of zero-shot learning is its reliance on generalized representations, which may not always adapt well to specific urban contexts. These models, trained on broad datasets, may struggle with local variations in urban environments such as lighting conditions, weather patterns, and facade degradation. For instance, shadows cast by tall buildings or reflective surfaces can skew material detection results, while weather events like rain or fog may obscure material textures. Furthermore, facade wear and degradation over time (e.g., surface peeling, staining) can alter the visual appearance of materials, introducing additional challenges to accurate classification. Pretrained models may also exhibit biases due to geographical limitations and privacy restrictions in cities with stricter street view imagery protocols, which can result in uneven training data availability and raise ethical concerns about fairness in model performance across different cities.

Zero-shot models may exhibit domain bias, where well-represented materials from the pre-trained data are favored, potentially leading to misclassification of less common/unseen materials. This can reduce detection accuracy in cities with unique architectural styles or underrepresented materials. For instance, in Dubai, detection accuracy is lower due to wider streets leading to lower resolution images and a building stock dominated by facade coatings such as plaster and stucco, which present challenges due to their variation in texture and color. Future work could focus on fine-tuning the model by incorporating few-shot learning to improve detection of local materials, especially in cities with distinct architectural styles. However, even with additional training, certain material classes may continue to pose detection challenges. These namely include metal panels that do not have apparent reflection, or difficult distinctions between concrete and plaster. While our results show that the zero-shot framework outperforms SegFormer on the same small dataset, this could change with a larger training dataset, though the extensive time required for labeling such a dataset underscores the advantage of zero-shot frameworks.

Future work

The limitations discussed above are inherently connected to the nature of the material segmentation problem in street view-based facades. In this paper, we aimed to demonstrate the high potential of zero-shot computer vision models to advance the collection of urban data and the study of building surface properties, a challenge to urban thermal assessments. We anticipate that the workflow shared will be further improved, adapted, and tailored to various applications to the urban science community as capabilities of computer vision tools advance further. While we performed material distribution mapping throughout the urban areas and highlighted pathways for implementation, we see broader potential in affordable applications of ubiquitous material mapping to urban simulation tools. One possible avenue is inputting this information into the Urban Weather Generator (UWG)[49] for higher precision micro-climate approximations. The method can also be extended to derive solar orientations of detected surfaces and couple that with the knowledge on urban surfaces to identify potential hotpots (such as in South facing, low albedo facades).

Conclusion

This paper presents a comprehensive zero-shot learning pipeline that segments material classes and facade regions in urban buildings, advancing our understanding of urban texture detection and its implications for UHI approximations and outdoor comfort studies. Our results demonstrate the effectiveness of zero-shot learning in this domain, with the model accurately detecting the predominant facade material in 68% of cases and identifying the top three material classes in 85% of the dataset. These results highlight the model's capacity to capture material compositions at a high level of detail. One of the key contributions of this study is the ability to scale the model across diverse urban contexts, as tested in cities with varying architectural styles and climate conditions—Boston, Amsterdam, and Dubai. By leveraging material segmentation for thermal comfort studies, our method illustrates the significant impact of material choices on outdoor thermal comfort, particularly in extreme climates like Dubai. We also tested our approach against a state-of-the-art detection model, SegFormer, and demonstrated improved performance with a 32% increase in aggregate IoU. These insights demonstrate the practical utility of material segmentation in informing urban design and climate adaptation strategies.

In addition, the scalability of this approach provides a powerful tool for large-scale urban material mapping, offering granular insights into local material distributions and facilitating evidence-based urban planning and policy interventions. While the results are promising, this study also acknowledges limitations related to image resolution, lighting conditions, and material variations, especially in cities with more complex facade textures. Future work will focus on improving the adaptability of the model by fine-tuning it for specific urban contexts and exploring the integration of additional material properties, such as emissivity and thermal conductivity, to enhance predictions in thermal comfort assessments. Additionally, efforts to optimize computational efficiency and expand the model's applicability to broader datasets will further support its integration into real-world urban applications.

Overall, this study illustrates the potential of zero-shot learning in addressing the challenges of urban material detection, paving the way for more precise urban climate simulations and policy-driven interventions.

Data availability

The full segmentation pipeline code, along with the dataset, can be accessed at the following link: https://github.com/Nadatarkhan/Zero-shot-Facade-Material-Segmentation.git.

Received: 21 April 2024; Accepted: 8 January 2025
Published online: 14 February 2025

References

1. Zhou, D., Zhao, S., Liu, S., Zhang, L. & Zhu, C. Surface urban heat island in China's 32 major cities: Spatial patterns and drivers. *Remote Sens. Environ.* **152**, 51–61 (2014).
2. Taleb, D. & Abu-Hijleh, B. Urban heat islands: Potential effect of organic and structured urban configurations on temperature variations in Dubai, UAE. *Renew. Energy* **50**, 747–762 (2013).
3. Oke, T. R. City size and the urban heat island. *Atmos. Environ.* **1967**(7), 769–779 (1973).
4. Stewart, I. D. & Oke, T. R. Local climate zones for urban temperature studies. *Bull. Am. Meteorol. Soc.* **93**, 1879–1900 (2012).
5. Kamoutsis, A. P., Matsoukis, A. S. & Chronopoulos, K. I. Air temperature estimation by using artificial neural network models in the Greater Athens area, Greece. *ISRN Meteorol.* **2013**, 1–7 (2013).
6. Ward, K., Lauf, S., Kleinschmit, B. & Endlicher, W. Heat waves and urban heat islands in Europe: A review of relevant drivers. *Sci. Total Environ.* **569–570**, 527–539 (2016).
7. Vanos, J. K., Kalkstein, L. S. & Sanford, T. J. Detecting synoptic warming trends across the US Midwest and implications to human health and heat-related mortality. *Int. J. Climatol.* **35**, 85–96 (2015).
8. Kalkstein, L. S., Greene, S., Mills, D. M. & Samenow, J. An evaluation of the progress in reducing heat-related human mortality in major U. S. cities. *Nat. Hazards* **56**, 113–129 (2011).
9. Sangiorgio, V., Fiorito, F. & Santamouris, M. Development of a holistic urban heat island evaluation methodology. *Sci. Rep.* https://doi.org/10.1038/s41598-020-75018-4 (2020).
10. Doulos, L., Santamouris, M. & Livada, I. Passive cooling of outdoor urban spaces. The role of materials. *Solar Energy* **77**, 231–249 (2004).
11. Nazarian, N., Dumas, N., Kleissl, J. & Norford, L. Effectiveness of cool walls on cooling load and urban temperature in a tropical climate. *Energy Build.* **187**, 144–162 (2019).
12. Bradley, A. V., Thornes, J. E., Chapman, L., Unwin, D. & Roy, M. Modelling spatial and temporal road thermal climatology in rural and urban areas using a GIS. *Clim. Res.* **22**, 41–55 (2002).
13. Morini, E. et al. Experimental analysis of the effect of geometry and façade materials on urban district's equivalent albedo. *Sustainability* **9**, 1245 (2017).
14. Pourpanah, F. et al. A Review of generalized zero-shot learning methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 1–20 (2023).
15. Jendritzky, G., de Dear, R. & Havenith, G. UTCI-Why another thermal index?. *Int. J. Biometeorol.* **56**, 421–428 (2012).
16. Zhou B et al. Scene parsing through ADE20K dataset. In Proceedings—30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. (2017).
17. Li, B., Xing, H., Cao, D., Yang, G. & Zhang, H. Exploring the effects of roadside vegetation on the urban thermal environment using street view images. *Int. J. Environ. Res. Public Health* **19**, 1272 (2022).
18. Gong, F. Y. et al. Mapping sky, tree, and building view factors of street canyons in a high-density urban environment. *Build. Environ.* **134**, 155–167 (2018).
19. Tarkhan, N., Szcześniak, J. T. & Reinhart, C. Façade feature extraction for urban performance assessments: Evaluating algorithm applicability across diverse building morphologies. *Sustain. Cities Soc.* **105**, 105280 (2024).
20. Hosseini, M., Sevtsuk, A., Miranda, F., Cesar, R. M. & Silva, C. T. Mapping the walk: A scalable computer vision approach for generating sidewalk network datasets from aerial imagery. *Comput. Environ. Urban Syst.* **101**, 101950 (2023).
21. Cai, S., Wakaki, R., Nobuhara, S. & Nishino, K. 2023 RGB road scene material segmentation. In *Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (eds Wang, L. et al.) 256–272 (Springer, 2023).
22. Habili, N. et al. A hyperspectral and RGB dataset for building façade segmentation. In *Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (eds Karlinsky, L. et al.) 258–267 (Springer, 2023).
23. Raghu, D., Bucher, M. J. J. & De Wolf, C. Towards a 'resource cadastre' for a circular economy—urban-scale building material detection using street view imagery and computer vision. *Resour. Conserv. Recycl.* **198**, 107140 (2023).
24. Xu, Z. et al. Prediction of structural type for city-scale seismic damage simulation based on machine learning. *Appl. Sci.* **10**, 1795 (2020).
25. Radford, A. et al. Learning transferable visual models from natural language supervision. *Proc. Mach. Learn. Res.* **139**, 8748–8763 (2021).
26. Luddecke T, Ecker A. Image Segmentation Using Text and Image Prompts. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. (2022).
27. Kirillov A, Mintun E, Nikhila R. Segment Anything. In Computer Science—Computer Vision and Pattern Recognition. (2023).
28. Liu S et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In European Conference on Computer Vision (pp. 38–55). (2025).
29. Mohan, R. & Valada, A. EfficientPS: Efficient panoptic segmentation. *Int. J. Comput. Vis.* **129**, 1551–1579 (2021).
30. Ren Tet al. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. arXiv preprint arXiv:2401.14159. (2024).
31. Wang H et al. Sam-clip: Merging Vision Foundation Models towards Semantic and Spatial Understanding. In: Proceedings of the IEEE/CVFConference on Computer Vision and Pattern Recognition (pp. 3635–3647) (2024).
32. Wang, Y., Yao, Q., Kwok, J. T. & Ni, L. M. Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv.* **53**, 1–34 (2020).
33. Ramasesh VV, Lewkowycz A, Dyer E. Effect of scale on catastrophic forgetting in neural networks. International Conference on Learning Representations. (2022).
34. Rahman, S., Khan, S. & Porikli, F. Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. In *Lecture Notes in Computer Science* (eds Jawahar, C. V. et al.) 547–563 (Springer International Publishing, 2019).
35. Xu, W., Wang, J., Wei, Z., Peng, M. & Wu, Y. Deep semantic-visual alignment for zero-shot remote sensing image scene classification. *ISPRS J. Photogramm. Remote Sens.* **198**, 140–152 (2023).
36. Pradhan, B., Al-Najjar, H. A. H., Sameen, M. I., Tsang, I. & Alamri, A. M. Unseen land cover classification fromhigh-resolution orthophotos using integration of zero-shot learning and convolutional neural networks. *Remote Sens.* **12**, 1676 (2020).
37. Cao, G., Jiang, J., Bollegala, D., Li, M. & Luo, S. Multimodal zero-shot learning for tactile texture recognition. *Rob. Auton. Syst.* **176**, 104688 (2024).
38. Orlita T. Street View Download 360 (Version 4.0.18) [Software]. (2024).
39. Zhang, J., Fukuda, T. & Yabuki, N. Automatic object removal with obstructed façades completion using semantic segmentation and generative adversarial inpainting. *IEEE Access* **9**, 117486–117495 (2021).

40. Xie, E. et al. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Adv. Neural Inf. Process. Syst.* **34**, 12077–12090 (2021).
41. Segments.ai. Segments.ai [Software]. (2024).
42. Padilla R, Netto SL, Da Silva EAB. A Survey on Performance Metrics for Object-Detection Algorithms. In International Conference on Systems, Signals, and Image Processing. (2020).
43. Solemma. ClimateStudio. (Version 2.0 Release Candidate) [Software]. Retrieved from https://www.solemma.com/climatestudio (2024).
44. U.S. Department of Energy. EnergyPlus (Version 24.2.0) [Software]. Retrieved from https://energyplus.net/(2024).
45. Lobaccaro, G. & Acero, J. A. Comparative analysis of green actions to improve outdoor thermal comfort inside typical urban street canyons. *Urban Clim.* **14**, 251–267 (2015).
46. Schrijvers, P. J. C., Jonker, H. J. J., de Roode, S. R. & Kenjereš, S. The effect of using a high-albedo material on the universal temperature climate index within a street canyon. *Urban Clim.* **17**, 284–303 (2016).
47. Martin, M., Chong, A., Biljecki, F. & Miller, C. Infrared thermography in the built environment: A multi-scale review. *Renew. Sustain. Energy Rev.* https://doi.org/10.1016/j.rser.2022.112540 (2022).
48. Neuhold G, Ollmann T, Bulo SR, Kontschieder P. The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes. In Proceedings of the IEEE International Conference on Computer Vision. (2017).
49. Bueno, B., Norford, L., Hidalgo, J. & Pigeon, G. The urban weather generator. *J. Build. Perform. Simul.* **6**, 269–281 (2013).

Author contributions

Nada Tarkhan and Nikita Klimenko have equal contributions to the work. This is detailed further below along with the contributions of all other authors. Conceptualization-Nada Tarkhan and Nikita Klimenko. Methodology-Nada Tarkhan and Nikita Klimenko both set the methodology framework. Nada contributed with the Grounding DINO integration. Nikita Klimenko contributed with the CLIPSeg, OpenAI CLIP integration and testing. Nikita Klimenko synthesized all the model components and carried out tuning for the final code. Datasets-Kelly Fang, Nada Tarkhan and Nikita Klimenko all contributed to the data labelling and annotation. Results-Nikita Klimenko produced the IoU results for the analyzed datasets as well as the city-based material distributions. Nada Tarkhan extracted the material coverage metrics and conducted the outdoor thermal comfort analysis. Writing-introduction, methodology (joint Nada Tarkhan and Nikita Klimenko). Results, discussion, conclusion (Nada Tarkhan). Supervision-Carlo Ratti, Christoph Reinhart, Fabio Duarte.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-86307-1.

Correspondence and requests for materials should be addressed to N.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.