Working Paper: Predicting Urban Heat Island Intensity via Computer Vision and Spatial Feature Distribution

Ayush Lahiri[a,c]
[a] Georgetown University; [c] Ashoka University

Abstract

Abstract
The Urban Heat Island (UHI) effect poses significant risks to public health, environmental equity, and biodiversity, particularly in densely populated metropolitan areas. Current methodologies for estimating UHI intensity are predominantly ex-post, relying on deterministic algorithmic processing of temperature readings or aggregated scalar metrics—such as average tree canopy coverage—which fail to capture the granular spatial arrangement of urban features. This study investigates whether the visual spatial distribution of buildings, street networks, and vegetative cover can be utilized to accurately predict UHI effects using computer vision architectures. We introduce a novel modeling approach utilizing Convolutional Neural Networks (CNNs) to analyze geospatial representations of urban forms. By moving beyond aggregated statistics to visual feature extraction, we aim to provide a preemptive tool for urban planning, enabling policymakers to estimate temperature outcomes of theoretical city layouts before physical infrastructure is developed.

1 Introduction

Urban greenspaces are a critical component of sustainable development, providing essential ecosystem services ranging from pollution mitigation to the support of community well-being [8]. However, the economic imperatives of urbanization often place these natural assets at odds with the demand for impervious infrastructure. This trade-off is most acutely visible in the phenomenon of the Urban Heat Island (UHI) effect, where the concentration of heat-absorbing structures—buildings, roads, and concrete—causes urban centers to experience temperatures significantly higher than their rural surroundings.

The Environmental Protection Agency notes that daytime urban temperatures can exceed outlying areas by 1–7°F. The consequences of this thermal disparity are profound: UHI exacerbates air and water quality degradation, disrupts native species migration [13], and directly impacts human health. Crucially, these burdens are not shared equally; recent scholarship indicates that UHI intensity is often highest in neighborhoods inhabited by people of color, compounding existing socioeconomic inequalities [7].

1.1 The Limits of Current Measurement

Despite the severity of the issue, current UHI measurement paradigms suffer from a critical temporal limitation: they are fundamentally *ex-post*. Standard detection relies on deterministic algorithmic processing of observed temperature readings [3]. Consequently, the thermal profile of a neighborhood is only reliably known after it has been planned, built, and inhabited for several seasons. Rectifying UHI effects at this stage requires costly retrofitting—a process that is often politically unfeasible or inequitably applied.

Furthermore, existing predictive models that attempt to forecast UHI often rely on aggregated geographical data, such as the percentage of vegetation types or average canopy coverage within a census tract [1, 16]. While valuable, these scalar metrics fail to capture the *spatial* distribution of urban features. Two districts may possess identical aggregate building densities, yet the specific clustering or uniform dispersal of those structures can yield vastly different surface temperature outcomes [21].

1.2 A Computer Vision Approach

To address this gap, this study proposes a shift from scalar aggregation to visual spatial analysis. We posit that the complex geometric relationships between buildings, streets, and vegetation are best analyzed through computer vision methods. To our knowledge, this is the first attempt to predict UHI intensity by applying Convolutional Neural Networks (CNNs) directly to geospatial imagery of urban infrastructure.

By validating a model that links visual spatial distribution to thermal outcomes, we aim to provide a tool for preemptive urban planning. Such a capability would allow planners to estimate the UHI footprint of theoretical developments prior to investment, minimizing environmental costs and protecting vulnerable populations in both developed and data-scarce developing regions.

2 Data

To evaluate the relationship between urban morphology and thermal outcomes, we model the spatial distribution of three primary urban characteristics: building footprints, street networks, and vegetative cover. The analysis spans three major metropolitan areas in the United States: Chicago, New York City, and Los Angeles.

We curate two distinct datasets to represent these spatial distributions. The first, detailed below, utilizes segmented vector and raster data to create noise-free composite representations.

2.1 Dataset A: Segmented Urban Spatial Features

This dataset synthesizes distinct geospatial layers into composite visual inputs, treating the Census Block Group (CBG) as the fundamental observational unit.

2.1.1 Input Features

We construct the feature space using three primary data sources:

- **Buildings:** Unique shapefiles containing the precise geographical position, shape, and dimensions of building footprints were sourced from municipal open data portals: the City of Chicago (2023), New York Office of Technology and Innovation (2024), and Los Angeles GeoHub (2017).
- **Street Networks:** Centerline data representing the geometry of major streets and roads were similarly acquired from the respective municipal transport repositories for Chicago, New York, and Los Angeles.
- **Vegetative Cover:** Tree canopy density is derived from the 2021 CONUS Tree Canopy cover dataset [**?**], provided by the Multi-Resolution Land Characteristics (MRLC) consortium. This raster data offers high-fidelity vegetation metrics at a 30-meter pixel resolution.

Figure 1: Raw input vector examples for Chicago building footprints and street networks. (a) Segmented Building Vectors; (b) Street Vector examples.

2.1.2 Input Processing and Composition

The raw geospatial data is processed to correspond to Census Block Group boundaries as defined by the US Census Bureau (2018). We select the block group as the unit of analysis to balance granularity with variance; they are sufficiently localized to capture micro-climate effects while large enough to display meaningful internal variation in urban layout.

For each unique block group, we iteratively subset the shapefiles and raster data to generate a composite visual output (Figure 2). This process eliminates

the "visual noise" often present in raw satellite photography (e.g., shadows, temporary objects, cars) by isolating only structural and vegetative features.

In the resulting image tensors:

- **Buildings** are rendered in light red.
- **Streets** are rendered in blue.
- **Vegetation** is rendered in green, where hue saturation corresponds to the density of the canopy cover from the source raster.

Areas falling outside the irregular polygon of a specific census block group are rendered as transparent layers to ensure the model focuses exclusively on the geometry within the administrative boundary.

Figure 2: Processed composite input representing a Census Block Group with Buildings (red), Streets (blue), and Canopy (green)

2.1.3 Target Variable: UHI Estimates

The ground truth labels for our model are derived from the Simplified Urban-Extent (SUE) estimates released by the Yale Center for Earth Observation [3]. These estimates quantify the temperature differential between the urban unit and its rural surroundings using MODIS imagery at a 300m spatial resolution.

Processing: We compute the mean of the annual average daytime and nighttime UHI intensity for each coordinate. To ensure data quality, we apply the following exclusion criteria:

1. Coordinates with missing UHI values are discarded.
2. Observations with values lower than -800 are removed. These anomalies typically represent large bodies of water (e.g., Lake Michigan). As our input features (buildings and roads) do not encode water body information, these instances are excluded to prevent model confusion.

Finally, pixel values representing UHI intensity are spatially joined to the census block group boundaries to create a 1:1 mapping between the composite input images and their corresponding thermal outcomes.

Table 1: **Dataset A: Descriptive Statistics**

Parameter	New York	Los Angeles	Chicago	Overall
Total Block Groups	6,623	6,589	2,328	15,540
Total N (Block Groups with UHI Data and complete features)	1,285	898	747	2,930
Tree Canopy Cover				
Mean	47.21	82.56	17.96	49.24
Max	255	269	254	269
Min	-38.38	-16.64	-20.55	-38.38
Urban Heat Island Effect				
Mean	-2.38	0.79	0.64	0.01
Max	2.80	3.49	2.93	3.49
Min	-51.90	-46.49	-32.34	-51.90
Number of Channels in Image	3			
Raw Image Size	1000×1000			

2.2 Dataset A: Limitations and Distribution

The intersection of valid UHI sensor data and complete geospatial feature sets results in a final sample size of $N = 2,930$ Census Block Groups (from a total universe of 15,540). This attrition is primarily driven by three factors: (1) the spatial coverage of UHI sensors in Chicago does not perfectly align with census boundaries; (2) the exclusion of areas where the UHI margin of error exceeds $3°C$ [3]; and (3) the removal of blocks with incomplete vector data for buildings or roads.

The resulting distribution of UHI values exhibits significant right skewness, predominantly clustering between $-1.5°C$ and $3°C$. This falls within the standard global distribution for UHI, which typically ranges from $-4°C$ to $7°C$ depending on geographical context. Notably, New York displays "thicker tails" in the negative region.

The Scale Challenge: A critical challenge inherent to Dataset A is the variance in physical area of the observational unit. Visual inspection reveals that Census Block Groups in Los Angeles are generally significantly larger than those in New York or Chicago. Consequently, the resulting composite images for LA appear more "zoomed out," rendering individual buildings and roads smaller relative to the image frame. This variance implies that the model must learn to identify UHI-predicting patterns across inconsistent scales, potentially confounding physical size with feature density.

2.3 Dataset B: Composite Aerial Imagery

To address the limitations of scale variance and to test the efficacy of raw visual signals versus segmented features, we curate a second dataset (Dataset B) utilizing high-resolution satellite photography.

2.3.1 Uniform Grid Construction

Unlike Dataset A, which relies on irregular administrative boundaries, Dataset B utilizes a standardized spatial grid. We overlay a generated grid across the municipal boundaries of Chicago, New York, and Los Angeles, where each cell represents a uniform $500\text{m} \times 500\text{m}$ geographic region.

Using the latitude and longitude coordinates of each cell's vertices, we retrieve corresponding aerial imagery via the Google Maps Static API. By explicitly defining the bounding box for a fixed area size, we ensure that every image in Dataset B maintains a consistent zoom level and spatial resolution (1m/pixel). This uniformity eliminates the scale invariance problem observed in the Census Block Group approach.

Figure 3: Dataset B input examples. Uniform $500\text{m} \times 500\text{m}$ aerial imagery tiles.

2.3.2 Signal and Noise Strategy

The output labels (UHI intensity) are extracted using the bounding box coordinates of each grid cell, following the same SUE algorithm source used in Dataset A.

However, the preprocessing strategy for Dataset B differs fundamentally in its treatment of noise:

1. **Inclusion of Outliers:** We do not exclude observations with extreme negative values (previously associated with water bodies). Since raw aerial imagery contains visual representations of water, sand, or industrial zones, the model has the theoretical capacity to learn these features.
2. **Full Spectrum Information:** While Dataset A was curated to remove "noise" (shadows, cars, temporary structures), Dataset B retains the full spectrum of visual information available in an urban environment.

This comparative approach allows us to test whether a model performs better when provided with clean, segmented structural data (Dataset A) or noisy, holistic visual data (Dataset B).

Table 2: **Dataset B: Descriptive Statistics**

Parameter	Value
Total 500m X 500m cells	6,184
Total N : (Block Groups with UHI Data)	5704
Urban Heat Island Effect	
Mean	-82.53
Max	3.53
Min	-898.49
Number of Channels in Image	3
Raw Image Size	600×300

3 Discussion of Data Characteristics

Dataset B yields a significantly larger sample size ($N = 5,704$) compared to Dataset A ($N = 2,032$ after exclusions). This disparity is driven by two factors: the uniform grid system of Dataset B creates smaller, more consistent observational units than the irregular Census Block Groups, and the inclusion of aerial imagery allows for the retention of extreme UHI values (e.g., water bodies) that were necessarily filtered out of the segmented feature set.

The distribution of UHI values in Dataset B remains right-skewed, clustering between $-1.5°$C and $3°$C. However, the uniform grid approach alters the statistical properties of the noise. Because the observational units in Dataset B cover a smaller physical area (0.25km^2) compared to the average Census Block Group, the mean UHI calculation for each cell relies on fewer coordinate points. Consequently, the averaging process is less diluted by spatial smoothing, leading to "fatter tails" in the distribution—specifically, higher maximum temperature values and lower minimums compared to Dataset A.

4 Methodology

4.1 Rationale for Classification over Regression

While UHI intensity is inherently a continuous variable, we frame this study as a classification task. This decision is grounded in three considerations regarding the utility of urban planning tools and data granularity.

First, the practical significance of UHI variation is non-linear. A shift from -800 to -700 (indicating water or non-urban land) has negligible implications for policy. Conversely, a shift from $1.1°$C to $1.5°$C represents a critical threshold in thermal comfort and energy demand. Planners prioritize identifying these broad risk categories (e.g., "High Heat" vs. "Neutral") over precise degree predictions.

Second, accurately regressing subtle intra-class variations (e.g., predicting $1.1°$C vs $1.3°$C) typically requires high-fidelity 3D LiDAR data to map building heights and material emissivity. Given that this study utilizes 2D aerial and vector data, attempting to regress these fine margins would likely induce overfitting to noise.

Third, our objective is to provide a preemptive screening tool. By categorizing continuous values into discrete buckets, we focus the model on learning broad morphological trends—such as building density and street orientation—rather than the granular texture details required for continuous regression.

4.2 Class Definition and Sampling

Binning Strategy: Initial experimentation with fine-grained binning (intervals of $1.5°$C) resulted in severe class imbalance, with the median class dominating predictions. To ensure generalizability, we consolidated the target variable into three balanced classes:

1. **Cool:** Less than $0°$C
2. **Neutral:** $0°$C to $1.5°$C
3. **Hot:** Greater than $1.5°$C

Exclusion of Los Angeles in Dataset A: A critical preprocessing step for Dataset A was the exclusion of the Los Angeles subsample. As noted in the Data section, LA Census Block Groups are significantly larger than those in Chicago or New York. This results in "zoomed out" input images where buildings appear disproportionately small relative to the frame. This scale invariance introduced high intra-class variation that prevented the model from learning consistent features.

Consequently, Dataset A is restricted to New York and Chicago. Dataset B, utilizing the fixed-scale grid system, retains all three cities as the zoom level is constant (500m $\times$ 500m) across all samples.

Table 3: **Comparison of Final Input Datasets**

Dataset A	**Dataset B**
Final N: 2,032 *Less than 0:* 331 *0 to 1.5:* 478 *More than 1.5:* 816	**Final N:** 5,704 *Less than 0:* 1,543 *0 to 1.5:* 1,384 *More than 1.5:* 3,156
Cities: New York, Chicago	**Cities:** New York, LA, Chicago
Each image only contains information about tree canopy cover, roads, and buildings.	Each image contains all information captured in a raw aerial image of a city.
Image shape is variable, following the shape of a census block group with different zoom levels.	Image shape is square representing an area of 500m $\times$ 500m, corresponding to a consistent grid with the same zoom level.
Census block group sizes are highly variable (varying zoom levels). Bigger block = lower zoom.	Grid sizes are constant (500m$\times 500$m), generally smaller than census blocks, leading to larger N.
Source: Various GIS files (canopy, roads, buildings) from county repositories.	**Source:** Google Maps Static API.

5 Methodology

5.1 Data Constraints and Partitioning

We acknowledge that the retrieval and segmentation of high-resolution geospatial data is computationally intensive. Due to the resource costs associated with the Google Maps Static API and the processing time required for vector segmentation (approx. 118 hours per city), our sample size is constrained to the three metropolitan areas described.

To mitigate the effects of these constraints and the class imbalances noted in Table 3, we employ a stratified sampling strategy for data partitioning. This ensures that the training (80%), validation (10%), and testing (10%) sets maintain the same class proportions as the original population.

- **Dataset A:** Training ($N = 1,625$), Validation ($N = 203$), Testing ($N = 204$).
- **Dataset B:** Training ($N = 4,494$), Validation ($N = 606$), Testing ($N = 604$).

5.2 Model Architectures

We approach the UHI prediction task as a computer vision problem where the objective is to extract topological features—such as the sharp edges of impervious surfaces versus the texture of vegetation—and map them to thermal classifications. We experiment with three tiers of architectures: custom convolutional baselines, standard transfer learning models, and state-of-the-art vision transformers.

5.2.1 Tier 1: Custom Convolutional Baselines

We designed two custom Convolutional Neural Network (CNN) architectures to establish a performance baseline. The inputs are rescaled tensors of shape 180×180 (Dataset A) and 300×300 (Dataset B).

1. **Shallow CNN (5-Layer):** A sequence of five convolutional layers with increasing filter sizes (32 to 256), designed to capture low-level geometric features.
2. **Stacked CNN (7-Layer):** A deeper architecture utilizing paired convolutional layers (e.g., two layers of 64 filters, followed by pooling) to extract higher-order semantic representations.

Augmentation Strategy: To prevent overfitting on the custom baselines, we apply dynamic data augmentation to the training pipeline, including random horizontal flips, rotations ($\pm 10\%$), and zooms ($\pm 20\%$). This forces the model to learn generalized spatial features rather than memorizing pixel coordinates.

5.2.2 Tier 2: Transfer Learning (CNNs)

We utilize two established architectures pre-trained on ImageNet to leverage learned feature maps:

- **VGG-16 [17]:** A 16-layer network characterized by its uniform 3×3 convolution kernels. We freeze the base weights and replace the top layers with a dense block (256 neurons) to adapt the feature extraction to our specific 3-class problem.

- **Inception V3 [18]:** This architecture employs "Inception modules" that perform convolutions with different kernel sizes simultaneously. This allows the model to capture features at multiple scales—crucial for Dataset B, where the semantic meaning of a "green patch" might vary based on its surrounding context.

5.2.3 Tier 3: Vision Transformers

To address the limitations of CNNs in capturing long-range dependencies, we employ the **Swin Transformer** (Shifted Window Transformer) [10]. Unlike CNNs, which operate on fixed local windows, Swin Transformers utilize a hierarchical attention mechanism with shifted windows. This allows the model to effectively "attend" to the relationship between distant pixels—for example, understanding how a large park in the corner of a 500m grid influences the temperature of a building in the opposite corner. This architecture represents the current state-of-the-art for dense prediction tasks.

5.3 Training Protocol

All models optimize for **Categorical Cross-Entropy Loss**. We prioritize minimizing validation loss over maximizing accuracy to ensure the model produces confident, well-calibrated probability distributions rather than lucky guesses.

We implement two critical callbacks:

1. **Model Checkpoint:** Saves the model weights only when validation loss decreases, ensuring we retain the most generalizable iteration.

2. **Early Stopping:** Halts training if validation accuracy fails to improve for 6 consecutive epochs, restoring the best weights to prevent overfitting.

For the complex Inception and Swin architectures on Dataset B, we additionally apply L2 regularization and adaptive learning rate decay to stabilize convergence.

6 Results

We evaluate the predictive performance of our models across both datasets using test accuracy as the primary metric. Table 4 summarizes the performance of the custom baselines, standard transfer learning architectures, and the Vision Transformer.

Table 4: **Test Accuracy Comparison (Dataset A vs. Dataset B)**

Model Architecture	Acc. (Dataset A)	Acc. (Dataset B)
Custom CNN (5-layer, Unstacked)	52%	68%
Custom CNN (7-layer, Stacked)	49%	67%
Custom CNN (Augmented)	55%	68%
VGG-16 (Frozen Base)	47%	66%
Google Inception V3 (Frozen Base)	**54%**	**70%**
Swin Transformer (Tiny, Pre-trained)	55%	71%

6.1 Performance Disparity

A fundamental disparity in signal-to-noise ratio is evident between the two datasets. For Dataset A (Segmented Features), no model achieved a test accuracy significantly exceeding 55%. With a baseline random guess probability of 33% (3-class problem), and a binary equivalent of 50%, results ranging from 47% to 55% indicate that the models are failing to extract generalized signal. Two architectures performed worse than random chance, suggesting they were fitting to noise.

Conversely, Dataset B (Aerial Imagery) yields consistent performance improvements, with accuracies stabilizing between 66% and 71%. The Inception V3 and Swin Transformer models achieved the highest fidelity, suggesting that the raw, holistic visual information contained in satellite photography holds far more predictive power for UHI intensity than the clean, segmented vector features of Dataset A.

7 Discussion

7.1 The Failure of Segmented Features (Dataset A)

The inability of any architecture to generalize on Dataset A points to intrinsic flaws in the data representation rather than model complexity. We identify three drivers of this failure:

1. The Artifact of Irregular Boundaries: The Census Block Group (CBG) unit introduces significant morphological noise. Because CBGs have irregular shapes, the input tensors must be padded with black pixels to form square images. Visual inspection of feature maps reveals that the CNNs frequently activate on these artificial black boundaries rather than the urban features themselves. The model essentially learns to classify the *shape of the census block* rather than the internal heat-generating infrastructure.

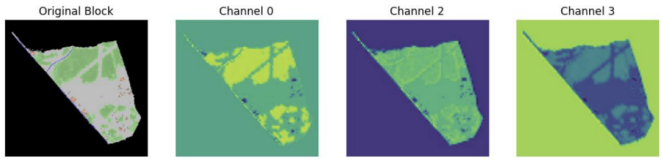

Figure 4: 5- Conv CNN with augmentation: Dataset A - Feature Map (layer 2).

2. Loss of Texture and Edge Information: By rasterizing tree canopies into green pixels, we eliminate the textural "edges" that define urban forms. In raw imagery, a tree line creates a sharp, complex boundary against concrete. In Dataset A, this is reduced to a blurred pixel blob. CNNs rely heavily on edge detection in early layers; the removal of these natural gradients likely stripped the necessary features required for thermal classification.

3. Intra-Class Inconsistency: Panel 4 illustrates the high variance within classes. We observed samples labeled as "Cool" (UHI < 0) dominated by impervious surfaces, and samples labeled "Hot" (UHI > 1.5) dominated by green space. This visual contradiction implies that the four variables selected (buildings, roads, canopy, block shape) are insufficient proxies for the complex thermodynamic processes driving UHI.

Figure 5: Visual contradiction of within class variation and across class similarity: Left: "Cool" sample with high building density vs Right: "Hot" sample with high vegetation.

7.2 Success of Aerial Imagery (Dataset B)

In contrast, Dataset B's uniform 500m $\times$ 500m grid eliminates boundary artifacts, and the raw photography retains the full spectral complexity of the urban environment.

While the Swin Transformer achieved marginally higher accuracy (71%), we posit that **Inception V3** (70%) is the superior architectural choice for this specific domain. This decision is supported by the theoretical alignment between the Inception architecture and the nature of urban spatial features.

The Multi-Scale Advantage: Urban features inherently exist at varying scales. A "green space" can be a singular street tree (5m width) or a sprawling municipal park (500m width). Standard CNNs (like VGG-16) use fixed kernel sizes (typically 3×3), forcing the model to analyze the image at a single spatial scale.

Inception V3 addresses this via "Inception Modules," which perform factorized convolutions using 1×1, 3×3, and 5×5 filters *simultaneously* within the same layer [18]. This allows the model to extract features of different sizes in parallel. In the context of UHI prediction, this means the model can simultaneously account for the micro-cooling effect of a single tree and the macro-cooling effect of a large park. Given the computational cost of Vision Transformers (which require quadratic complexity with respect to token length), Inception V3 offers an optimal balance of multi-scale feature extraction and computational efficiency for urban modeling.

7.3 Evaluation of Aerial Imagery (Dataset B)

7.3.1 Training Dynamics and Overfitting

The Inception V3 model exhibits rapid convergence on Dataset B, achieving training accuracy exceeding 95% by the 5th epoch. However, validation accuracy plateaus at approximately 70%, indicating a regime of overfitting. Despite aggressive regularization attempts—including reducing the classification head dimensionality (512 to 256 neurons), applying L2 regularization ($\lambda = 0.01 - 0.05$), and implementing exponential learning rate decay—the gap between training and validation performance persists.

This suggests that the Inception architecture, with its 24 million parameters, possesses a model capacity that exceeds the complexity of the available 5,704 samples, leading to memorization of specific training examples. However, we observe that validation *loss* continues to decrease even as accuracy fluctuates. This decoupling indicates that while the model's discrete classifications are static, its probabilistic calibration is improving—it is becoming more "confident" in its correct predictions and less confident in its errors.

7.3.2 Error Analysis and Confusion Matrix

To understand the classification boundaries, we analyze the confusion matrix (Table 5). The model demonstrates high efficacy in identifying the "Hot" class (Class 2), with a recall of 81% ($255/313$). This is likely driven by the class imbalance, as Class 2 represents the majority of the training data.

Table 5: **Confusion Matrix: Inception V3 (Dataset B)**

		Predicted Label		
		0 (Neutral)	**1** (Cool)	**2** (Hot)
True Label	**0 (Neutral)**	**50**	36	51
	1 (Cool)	16	**118**	20
	2 (Hot)	30	28	**255**

Key: 0 = 0 to 1.5°C; 1 = < 0°C; 2 = > 1.5°C

The primary source of error lies in Class 0 (Neutral: $0°C - 1.5°C$), which is frequently misclassified as Class 2 (Hot). We hypothesize that the model operates on a logic of "Dominant Feature Detection":

- **Cool (Class 1):** Dominated by vegetation.
- **Hot (Class 2):** Dominated by impervious concrete.
- **Neutral (Class 0):** A mixed distribution where neither feature dominates.

The model struggles to define the threshold for "mixed." Unlike the extreme classes, the Neutral class likely relies on non-visual latent variables—such as wind corridors, population density, or geographic orientation—which are absent from the aerial imagery.

7.3.3 Feature Map Inspection

Visual inspection of the Inception V3 feature maps confirms that errors are driven by edge-detection failures in complex transition zones.

False Positives (Predicted Hot, True Neutral): As seen in the example in Figure 6 , errors occur in neighborhoods where tree canopies are finely dispersed among dense housing. The model fails to resolve the high-frequency edges separating the trees from the buildings. It effectively "blurs" the small green patches into the surrounding concrete, interpreting the scene as a contiguous block of infrastructure (Class 2).

False Negatives (Predicted Cool, True Neutral): Conversely, in Figure 7, the model incorrectly classifies areas with tree-lined roadways as "Cool." The strong edge signal of the vegetation along the street dominates the feature map, causing the model to ignore the asphalt road itself.

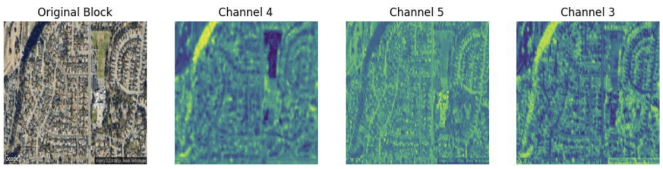

Figure 6: Predicted Class: Hot | Original Class = Neutral

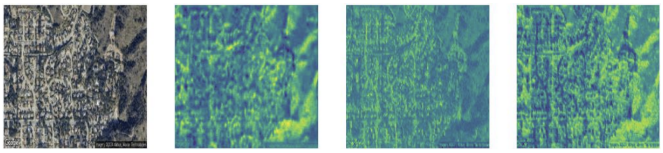

Figure 7: Predicted Class: Cool | Original Class = Neutral

Correct Classifications: In successful predictions (Figures 9 and 10), the images exhibit a distinct spatial separation between green and grey areas. The model successfully activates on the sharp boundaries between a park and a city block, correctly balancing the two signals to predict the Neutral class.

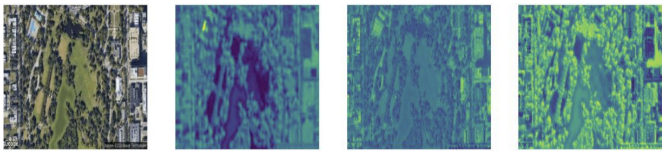

Figure 8: Predicted and Original Class = Neutral

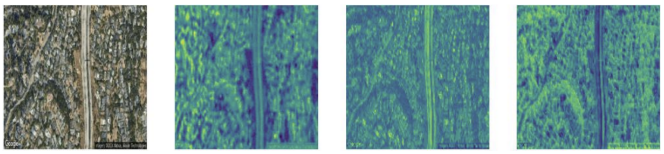

Figure 9: Predicted and Original Class = Neutral

7.3.4 Comparison with Swin Transformer

This edge-detection limitation explains the marginal performance gain of the Swin Transformer (71% vs 70%). The Swin Transformer utilizes a *shifted window* attention mechanism, which allows it to model the relationship between adjacent patches more effectively than a standard CNN.

In the "dispersed canopy" examples (Panel 7) where Inception failed, the Swin Transformer is theoretically better equipped to recognize that a repeating pattern of "house-tree-house-tree" represents a distinct texture (Neutral), rather than blurring it into a single concrete block (Hot). This ability to capture global context from local textures suggests that Transformer-based architectures may be the future standard for analyzing heterogeneous urban morphologies.

8 Conclusion and Future Work

The Urban Heat Island effect is a thermodynamic phenomenon driven by a complex interplay of variables, including building materials, anthropogenic heat emissions, wind corridors, and geographic baselines. This study sought to determine whether the visual arrangement of urban features—specifically the density and distribution of impervious surfaces versus vegetation—could serve as a sufficient proxy for these underlying physical processes.

8.1 Performance and The Visual Ceiling

Our results suggest that while visual spatial data is predictive, it has a theoretical "performance ceiling." We hypothesized that a successful model would approach 80% accuracy, capturing the nuances of spatial distribution that tabular data misses.

However, even with state-of-the-art architectures (Inception V3, Swin Transformer) and extensive hyperparameter tuning, our accuracy plateaued at 70% on Dataset B. Visual inspection of misclassified samples reveals that this remaining 30% error margin likely stems from latent non-visual variables. For instance, two neighborhoods may have identical visual layouts of trees and concrete, yet vastly different thermal profiles due to population density or coastal proximity. Therefore, we conclude that visual morphology is a *necessary but insufficient* predictor of UHI intensity.

8.2 The Planning Dilemma (Dataset A vs. B)

A primary objective of this study was to create a tool for preemptive urban planning. Dataset A (Segmented Features) was designed to mimic early-stage planning schematics—clean vector maps of proposed buildings and roads. Its

failure to yield predictive signal (Accuracy $< 55\%$) represents a significant limitation.

Because the model only succeeded on Dataset B (Aerial Imagery), its current utility is restricted to *ex-post* analysis of existing cities. Since retrofitting fully developed urban areas is slow and capital-intensive, the inability to predict thermal outcomes from the "blueprint" stage (Dataset A) highlights a gap between current computer vision capabilities and practical planning needs.

8.3 Technical Limitations and GIS Constraints

The failure of Dataset A is largely attributable to the "Scale Invariance" problem introduced by the Census Block Group unit. The irregular shapes and varying zoom levels prevented the model from learning consistent spatial features.

Ideally, Dataset A would have utilized the same uniform $500\text{m} \times 500\text{m}$ grid as Dataset B. However, this was technically unfeasible due to the conflicting Coordinate Reference Systems (CRS) of the distinct vector layers (roads vs. buildings vs. raster canopy). Re-projecting these disparate government datasets into a unified metric grid for segmentation remains a computationally expensive challenge for future research.

8.4 Pathways for Improvement

To bridge the gap between our 70% accuracy and the 80% target, future work must move beyond unimodal analysis and fixed-scale constraints.

1. Multi-Modal Fusion: We propose a **Multi-Modal Architecture** that fuses the Convolutional/Transformer image embeddings with a parallel Multi-Layer Perceptron (MLP) processing scalar data (e.g., population density, elevation, distance to water). By combining the spatial intuition of computer vision with the hard metrics of tabular data, we believe a robust, "blueprint-ready" UHI prediction tool is achievable.

2. Spatial Pyramid Pooling (SPP) for Irregular Inputs: To specifically rehabilitate the utility of Dataset A (Segmented Features), future architectures may benefit from the incorporation of Spatial Pyramid Pooling layers [6] between the convolutional and fully connected layers.

A primary failure mode in our study was the introduction of "black padding" noise required to force irregular Census Block shapes into square tensors. SPP eliminates this requirement by pooling features at multiple scales (e.g., 1×1, 2×2, 4×4), generating a fixed-length representation regardless of the input image's size or aspect ratio. This would allow the model to ingest the raw, irregular vector shapes directly, preserving spatial fidelity without the artificial boundary artifacts that confounded our current models.

References

[1] Assaf, G., Hu, X., & Assaad, R. H. (2023). Predicting Urban Heat Island severity on the census-tract level using Bayesian networks. *Sustainable Cities and Society*, 97, 104756.

[2] Bohn, K. (2023). *Strategic city planning can help reduce urban heat island effect*. Penn State University.

[3] Chakraborty, T., Hsu, A., Manya, D., & Sheriff, G. (2020). A spatially explicit surface urban heat island database for the United States: Characterization, uncertainties, and possible applications. *ISPRS Journal of Photogrammetry and Remote Sensing*, 168, 74–88.

[4] City of Chicago. (2023). *Building footprints in Chicago* [Data set]. Chicago Data Portal.

[5] Cook County GIS. (2023). *Aerial imagery reference tiles [Illinois–Cook County]* [Data set]. Cook Central.

[6] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9), 1904–1916.

[7] Hsu, A., Sheriff, G., Chakraborty, T., & Manya, D. (2021). Disproportionate exposure to urban heat island intensity across major US cities. *Nature Communications*, 12(1).

[8] Kajosaari, A. et al. (2024). Predicting context-sensitive urban green space quality to support urban green infrastructure planning. *Landscape and Urban Planning*, 242, 104952.

[9] Levy, J. (2024). *Street center lines in Chicago* [Data set]. City of Chicago.

[10] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012-10022.

[11] Los Angeles GeoHub. (2017). *Building Footprints*. City of Los Angeles.

[12] Los Angeles GeoHub. (2020). *Streets (Centerline)*. City of Los Angeles.

[13] National Geographic. (2022). *Urban Heat Island*. National Geographic Society.

[14] Office of Technology and Innovation. (2024). *Building Footprints*. NYC Open Data.

[15] Office of Technology and Innovation. (2024). *NYC Street Centerline (CSCL)*. NYC Open Data.

[16] Oh, J. W. et al. (2020). Using deep-learning to forecast the magnitude and characteristics of urban heat island in Seoul Korea. *Scientific Reports*, 10(1), 3559.

[17] Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *International Conference on Learning Representations (ICLR)*.

[18] Szegedy, C., Vanhoucke, V., Ioffe, S., & Shlens, J. (2015). Rethinking the Inception Architecture for Computer Vision. *arXiv preprint arXiv:1512.00567*.

[19] United States Census Bureau. (2022). *Cartographic boundary files*.

[20] United States Environmental Protection Agency. (2014). *Heat Island Effect*.

[21] Wang, M., & Xu, H. (2021). The impact of building height on urban thermal environment in summer: A case study of Chinese megacities. *PLoS ONE*, 16(4), e0247786.