sustainability

MDPI

Article

Deep Learning-Based Assessment of Urban Visual Perception from Street View Images: Implications for Sustainable Urban Planning in the Macau Peninsula

Chen Pan [1], Jiaming Guo [1], Jun Yang [1,2,*] and Junming Chen [3,*]

[1] Architecture and Civil Engineering Institute, Guangdong University of Petrochemical Technology, Maoming 525000, China
[2] School of Chemical Engineering, Guangdong University of Petrochemical Technology, Maoming 525000, China
[3] Faculty of Humanities and Arts, Macau University of Science and Technology, Macau 999078, China
* Correspondence: yangj529@alumni.sysu.edu.cn (J.Y.); jmchen@must.edu.mo (J.C.)

Abstract

Urban street-view imagery provides a scalable basis for examining visible environmental conditions relevant to human-centered sustainability. This study combines SegFormer-B5 semantic segmentation, six ResNet-50 classifiers trained using Place Pulse 2.0 annotations, spatial autocorrelation, XGBoost–SHAP, and Grad-CAM to analyze 20,256 images from 5064 sampling locations in the Macau Peninsula. We derived four objective visual indicators from the segmentation results. The mean classification accuracy across the six perception classifiers was 73.70% in the source-domain validation. The image-derived indicators and model-predicted scene categories exhibit significant spatial clustering, including global Moran's I values of 0.673 for the Boring score and 0.637 for the Green View Index. XGBoost–SHAP further identifies nonlinear associations between 19 semantic scene components and the six model-predicted perception scores, while Grad-CAM visualizes the image regions that contribute to individual predictions. The integrated framework links streetscape measurement, spatial diagnosis, and model interpretation to support systematic identification of streets requiring detailed urban-design assessment and locally informed renewal planning.

Keywords: street-view imagery; urban visual perception; deep learning; spatial heterogeneity; explainable artificial intelligence; sustainable urban planning; urban renewal; Macau Peninsula

Check for updates

Academic Editor: Miguel Amado

Received: 25 August 2026
Revised: 30 September 2026
Accepted: 3 November 2026
Published: 5 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Rapid urbanization has intensified spatial differences in urban environments, making access to attractive and usable public space a concern for human-centered sustainability. The visible characteristics of streets, including greenery, building interfaces, street enclosure, pedestrian facilities, and traffic, constitute an important observable dimension of urban environmental quality. Street-view imagery and computer vision provide a consistent way to quantify these characteristics and estimate scene-level perceptual attributes across large urban areas.

The spatial distribution of visible streetscape features can help identify locations for more detailed assessment of environmental quality and spatial equity. Researchers have investigated street-level greenery and pedestrian environments in relation to walking and health-related outcomes [1–5]. Image-based prediction can therefore serve as a scalable screening layer that is subsequently integrated with field observations, locally collected

opinions, and other planning evidence. In the present study, these image-derived outputs are treated as model-predicted visual attributes rather than direct reports of residents' or tourists' lived perceptions, and the relationships examined are interpreted as associations rather than causal effects.

Street-view imagery has been widely applied to urban analytics and public-health-oriented environmental assessment [6,7]. Related image-based urban research has also extended to multi-view city understanding [8] and image-based prediction of building-related attributes, such as residential electricity consumption [9]. Semantic segmentation estimates the proportions of observable objects, while models trained on human image comparisons assign scores to specific perceptual categories. Related computer-vision research includes semi-supervised object detection [10] and ConvNeXt architectures [11], which provide broader methodological context. Spatial statistics identify geographic clustering, and explainable models characterize associations between image features and predicted labels. Together, these methods provide a reproducible basis for comparing urban visual conditions across space and for identifying scene elements associated with different prediction outcomes.

This study applies an integrated analytical framework to the Macau Peninsula. Three questions guide the analysis: (1) How are six model-predicted streetscape attributes spatially distributed? (2) Which measurable image features are associated with these predictions, and how do the associations differ across categories? (3) How can these spatial and visual patterns support the assessment of streetscape-renewal needs? The contribution is a reproducible sequence linking visual feature measurement, model-predicted perception mapping, spatial diagnosis, and interpretable associations to sustainable urban planning.

2. Literature Review

Assessing streetscape appearance combines measurements of observable physical attributes with human observer evaluations. The former can be estimated from imagery and semantic segmentation, while the latter can be represented by human responses or by models trained on those responses. Biljecki and Ito [6] provide a systematic review of street-view imagery in urban analytics, and Kang et al. [7] review the use of visual measures in public-health studies. Landscape-appraisal research has also documented differences between public and professional attitudes [12] and preference heterogeneity in machine-learning-based survey analysis [13]. These studies provide complementary foundations for image-based measurement and perception modeling.

2.1. Street View Images and Urban Perception

Researchers have used street-view imagery to characterize greenery, urban form, and the visible pedestrian environment. Li et al. examined urban green space through eye-tracking responses [14], while Ki and Lee used a Green View Index to investigate walking time [1]. Li et al. described image-based approaches for mapping street landscapes [15]. Other studies have used deep learning to assess street-level greenery [16], scene elements to represent place locales [17], and metrics to examine covariation in visual and ecological landscape qualities [18]. Collectively, these studies demonstrate the value of quantitative scene descriptors for city-scale streetscape assessment.

Deep-learning approaches can also predict comparative urban scene assessments. Porzi et al. studied convolutional models of urban perception [19], while Dubey et al. introduced the Place Pulse 2.0 image-comparison dataset and associated prediction methods [20]. Helbich et al. examined associations between street-view green and blue spaces and geriatric depression using separate health-related observations [21]. Comparative and mapping studies have further examined alternative deep-learning urban-perception models [22]

and image-based mapping of human landscape perception [23]. These studies establish a methodological basis for predicting and spatially comparing perceptual attributes from street-scene imagery.

2.2. Spatial Heterogeneity of Urban Perception

Recent work has examined the relationship between landscape characteristics and urban vibrancy [24]. Researchers have also modeled urban vitality using urban morphology [25], historical preservation and multisource built-environment data [26–29], particulate matter and spatiotemporal indicators [30,31], and comparative or post-pandemic urban perspectives [32,33]. These studies demonstrate that spatial heterogeneity can be investigated through multiple forms of urban evidence. This study focuses on geographic variation in six model-predicted perception dimensions derived from street-view scenes.

2.3. Environmental Correlates of Urban Perception

Researchers have examined associations between visible environmental features and urban outcomes using multiple data sources and analytical frameworks. For example, studies have assessed street-level greenery and walking behavior [3], visual enclosure and older adults' mental health [34], street life and pedestrian form [35,36], and broader built-environment-health relationships involving population density, car ownership, and obesity [37]. Data-enabled urban-space perception frameworks [38] and scene-based perception for urban renewal [39] illustrate complementary analytical approaches. Building on this work, the present study investigates nonlinear associations between segmented image features and output scores from an image-based perception model.

2.4. Urban Visual Perception and Sustainable Urban Development

From a sustainability perspective, image-based indicators can support systematic screening of visual conditions, particularly when combined with fieldwork and community participation. Studies of greenery, visual walkability, and urban perception demonstrate the planning relevance of streetscape characteristics [1,5,39]. In this study, the mapped outputs are interpreted as model-predicted visual attributes that can identify spatial patterns and guide subsequent place-based assessment.

Three linked gaps motivate this study. First, researchers often analyze image-derived physical attributes and predicted scene assessments separately, making their relationship difficult to interpret. Second, descriptive perception maps do not by themselves reveal which observable scene features are most strongly associated with different prediction outcomes. Third, model interpretability is rarely integrated with spatial diagnosis in a form that can inform urban-renewal assessment. The proposed framework addresses these gaps by linking semantic segmentation, perception prediction, spatial autocorrelation, XGBoost–SHAP, and Grad-CAM within a unified analytical sequence.

3. Materials and Methods

3.1. Study Area

This study focuses on the Macau Peninsula to quantitatively evaluate urban visual perception (Figure 1a). The Macau Peninsula, the central and most densely urbanized part of the Macau Special Administrative Region, covers approximately 9.1 km^2 and has a population of around 388,000. Its compact urban form combines historic neighborhoods, high-density residential areas, major commercial and tourism destinations, reclaimed districts, and a fine-grained street network. These characteristics make the peninsula a suitable case for examining how visual environmental quality varies within a high-density city and

how perception-based evidence can support sustainable urban renewal under land and spatial constraints.

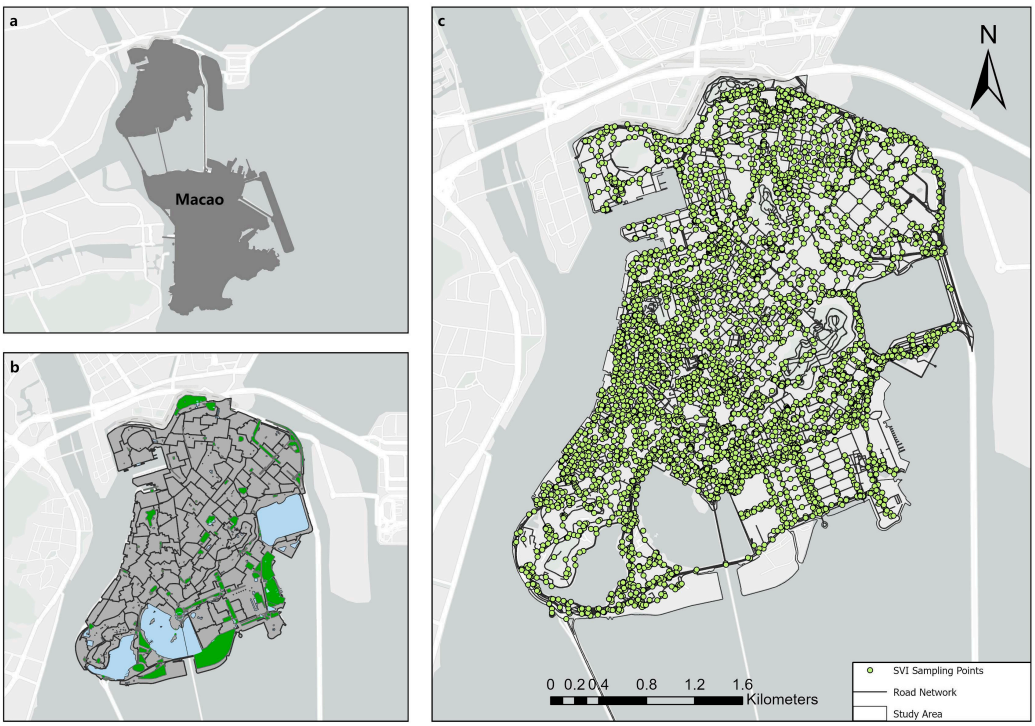


Figure 1. (**a**) Location of the Macau Special Administrative Region; (**b**) Macau Peninsula; and (**c**) street-view sampling locations ($n = 5064$).

3.2. Data Sources

3.2.1. Street View Image Sources

We established sampling points at approximately 100 m intervals along the OpenStreetMap network in the Macau Peninsula (Figure 1). After excluding locations without complete directional imagery, we retained 5064 locations. Four Baidu Street View images were obtained at headings of 0°, 90°, 180°, and 270° per location, yielding 20,256 images at a nominal resolution of 960 × 720 pixels. Geographic coordinates were retained for subsequent spatial analysis, linking the four directional images to a common sampling location.

3.2.2. Source Perception Dataset and Volunteer Annotation

The Place Pulse 2.0 dataset contains 110,988 street-scene images from 56 cities and approximately 1.17 million pairwise judgments submitted by 81,630 online volunteers [20]. During each comparison task, a volunteer viewed two street-scene photographs and selected the image that better represented one specified attribute: beautiful, boring, depressing, lively, safe, or wealthy. Repeated pairwise judgments were aggregated using the Microsoft TrueSkill algorithm to obtain relative scene scores [40]. We then used these source-domain annotations to construct category-specific training samples for six ResNet-50 classifiers. Table 1 reports the training and validation sample counts and the corresponding classification accuracies. Across the six categories, the mean source-domain classification accuracy was 73.70%. We collected the Place Pulse 2.0 annotations from online volunteers rather than a Macau-specific participant sample. Detailed demographic information sufficient to establish the volunteer pool's cultural or geographic representativeness is not available for the present analysis. Accordingly, the source annotations may contain population- and

context-specific biases, and the resulting ResNet-50 outputs are treated as source-trained model predictions rather than locally validated Macau perceptions.

Table 1. Source-domain training/validation samples and classification accuracy for the six perception categories.

Perception Category	Beautiful	Boring	Depressing	Lively	Safety	Wealthy
Training sample count	30,414	29,650	30,376	28,493	29,054	31,328
Validation sample count	6081	5929	6074	5698	5810	6264
Source-domain classification accuracy (ACC)	0.779148	0.647495	0.699045	0.771850	0.797418	0.727331

3.3. Analytical Framework

The analytical sequence consists of four related tasks (Figure 2). First, SegFormer-B5 estimates the proportions of 19 visible semantic classes and supports calculating four derived visual indicators. Second, six ResNet-50 classifiers trained on Place Pulse 2.0 labels generate model-predicted scene scores. Third, global and local Moran's I summarize spatial dependence in location-level scores and image-derived indicators. Fourth, XGBoost–SHAP quantifies how the 19 semantic fractions relate to the six ResNet-50 outputs, while Grad-CAM localizes image regions that contribute to selected ResNet-50 predictions. SHAP therefore interprets the second-stage regression, and Grad-CAM visualizes the image regions used by the perception classifier. Each component addresses a distinct analytical question: semantic segmentation quantifies observable streetscape composition, ResNet-50 estimates scene-level perceptual attributes, Moran's I identifies their spatial organization, XGBoost–SHAP examines feature–output associations, and Grad-CAM localizes classifier responses at the image level. These methods are therefore complementary rather than interchangeable components of the analytical framework.



Figure 2. Technical framework.

3.3.1. Image Semantic Segmentation

Street-view images were segmented with SegFormer-B5 using a Cityscapes-pretrained model [41]. The network assigns image pixels to 19 semantic categories, including vegetation, sky, buildings, roads, sidewalks, pedestrians, and vehicles. For image i and class c, the semantic fraction P_ic is defined as the number of pixels assigned to class c divided by the total number of valid image pixels. These semantic fractions form the basic quantitative descriptors of streetscape composition and are subsequently used to calculate composite visual indicators and as predictors in models implemented using XGBoost version 2.1.4 in Python 3.9.

$$S_n^0 = \mathrm{CNN}(i_n), 1 \tag{1}$$

$$S_n^{\acute{1}} = f_{seg}\left(\mathrm{MAS}\left(\mathrm{LN}\left(s_n^{1-1}\right)\right)\right), 1 = 1,\ 2,\ \ldots, L \tag{2}$$

$$1_n = \left[\frac{0 \text{ in } f_{seg}\left(S_n^{\acute{1}}\right)}{\text{total pixels}}, \frac{1 \text{ in } f_{seg}\left(S_n^{\acute{1}}\right)}{\text{total pixels}}, \frac{18 \text{ in } f_{seg}\left(S_n^{\acute{1}}\right)}{\text{total pixels}}\right] \tag{3}$$

We calculated the composite objective visual indicators from semantic classes extracted from Macau street-view images. Specifically, the Green View Index (GVI) represents the proportion of vegetation pixels within the valid image area, while the Sky View Factor (SVF) represents the proportion of visible sky pixels. Enclosure characterizes the degree of visual containment formed by surrounding streetscape elements, including buildings, walls, fences, and vegetation. Visual walkability reflects the relative presence of pedestrian-related streetscape elements, such as sidewalks and other pedestrian facilities. These indicators describe observable physical characteristics of the photographed streetscape and are conceptually distinct from the six model-predicted perception scores generated by ResNet-50. Table 2 summarizes each indicator's definition and interpretation.

3.3.2. Perceptual Mapping Using ResNet-50

For each perception category, we selected images at the upper and lower ends of the source-domain score distribution to construct the binary classification dataset. We defined the upper and lower thresholds using the category-specific mean score and standard deviation, as shown in Equation (4).

$$Q_{high} = Q + \sigma\ , Q_{low} = Q - \sigma \tag{4}$$

In Equation (4), Q and σ denote the mean and standard deviation, respectively, of the source-domain scores for a given perception category. Images with scores greater than or equal to Q_{high} were assigned to the high-score class, whereas images with scores less than or equal to Q_{low} were assigned to the low-score class. Images with scores between the two thresholds were excluded from classifier training. This procedure was applied separately to the six perception categories: Beautiful, Boring, Depressing, Lively, Safe, and Wealthy.

For each category, the encoded positive class corresponds to the high-score class of that category. Accordingly, higher predicted probabilities indicate stronger model-predicted Beautiful, Lively, Safe, or Wealthy attributes for the corresponding positive categories, and stronger model-predicted Boring or Depressing attributes for the corresponding negative categories. Here, "positive class" refers only to the binary label coded as 1 and does not imply a favorable perceptual meaning.

Cross-entropy was used for the binary classification task, and category-specific classification accuracy was calculated as the proportion of correctly classified evaluation images (Equations (5) and (6)). The classification accuracies reported in Table 1 summarize the source-domain validation performance for the six ResNet-50 perception classifiers.

$$L_C = -\sum_{i=1}^{n} y_i \log\left(\hat{y}_i\right) \tag{5}$$

$$A_{cc} = \frac{N_{\hat{y}_i = y_i}}{N} \quad (6)$$

Binary cross-entropy was used as the loss function for training each ResNet-50 classifier, as shown in Equation (5). In this equation, $y_i \in \{0, 1\}$ denotes the ground-truth binary label of sample i, p_i denotes the predicted probability that sample i belongs to the encoded positive class, and N is the total number of training samples. Classification accuracy was calculated using Equation (6), where $\hat{y}_i$ denotes the predicted class label, y_i denotes the corresponding ground-truth label, and $1(\hat{y}_i = y_i)$ equals 1 when the prediction is correct and 0 otherwise. The category-specific accuracies reported in Table 1 represent the source-domain validation performance of the six ResNet-50 classifiers.

3.3.3. Spatial Autocorrelation Analysis

Spatial dependence in the image-derived indicators and location-level classifier outputs was summarized using global Moran's I and local indicators of spatial association (LISA) [42,43]. A positive global Moran's I indicates that locations with similar values tend to occur near one another. Local Moran's I further distinguishes high–high and low–low clusters from high–low and low–high spatial outliers, providing a localized representation of spatial association across the Macau Peninsula.

$$I(d) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij}(q_i - \overline{q})\left(q_j - \overline{q}\right)}{S^2 \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij}} \quad (7)$$

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (q_i - \overline{q})^2 \quad (8)$$

In Equations (7) and (8), x_i and x_j denote the observed values at locations i and j, respectively, and W_{ij} denotes the spatial weight describing the spatial relationship between the two locations. In this study, the observed variable is either a location-level model-predicted perception score or an objective image-derived indicator. Global Moran's I measures the overall degree of spatial autocorrelation across the study area, whereas Local Moran's I identifies location-specific spatial associations between each observation and its neighboring observations. Positive Moran's I values indicate spatial clustering of similar values, negative values indicate spatial association among dissimilar values, and values close to zero indicate weak or no global spatial autocorrelation.

For the six model-predicted perception dimensions, the classifier output probability p was transformed to a continuous score using $S = 10_p$, and the four directional images at each sampling location were averaged to obtain the location-level scores used in the spatial autocorrelation analysis.

3.3.4. Perception Mapping and Objective Visual Indicators

We obtained six model-predicted category scores from the ResNet-50 classifiers and computed four objective indicators (GVI, SVF, enclosure, and visual walkability) from Macau image segmentation [1,39,44]. Equation (9) represents the semantic pixel proportion. The two sets of variables serve complementary purposes: the six scores describe model-predicted perceptual attributes, whereas the four indicators quantify visible physical characteristics of the streetscape.

$$P_i = \frac{C_i}{C_{sum}}, 1 \in \{\text{Road, Sidewalk, } \ldots, \text{ Bicycle}\} \quad (9)$$

For semantic category c, P_c denotes the proportion of valid image pixels assigned to category c, calculated as the number of pixels belonging to category c divided by the total number of valid pixels in the image. These semantic proportions characterize the visible

composition of each street-view image. Local Moran's I was then applied separately to each model-predicted perception score and each objective visual indicator to identify statistically significant local spatial associations. Significant locations were classified into four spatial configurations: high–high (H–H), low–low (L–L), high–low (H–L), and low–high (L–H), while locations that did not meet the adopted significance criterion were classified as non-significant.

Table 2. Segmentation-derived indicators of visible streetscape composition.

Index	Equation	Definition and Interpretation
GVI	$GVI_i = P_{Vegetation}$	GVI represents the proportion of visible vegetation pixels within the valid street-view image area and provides an image-derived measure of streetscape greenery [1].
SVF	$SVF_i = P_{sky}$	SVF represents the proportion of visible sky and characterizes image-based street openness [44].
Enclosure	$Enclosure_i = P_{Building} + P_{Wall} + P_{Fence} + P_{Pole} + P_{Light} + P_{sign} + P_{Vegetation}$	Enclosure characterizes the degree of visual containment formed by surrounding streetscape elements, including buildings, walls, fences, and vegetation [34].
Walkability	$Walkability_i = \frac{P_{Fence}+P_{Sidewalk}}{P_{road}}$	Visual walkability represents the relative presence of pedestrian-related streetscape elements, such as sidewalks and other pedestrian facilities, and is treated as an image-derived proxy rather than a direct measure of pedestrian-network accessibility [39].

Note: Green View Index is represented by GVI. Sky View Factor is represented by SVF.

3.3.5. XGBoost Regression and Model Interpretation

Six separate XGBoost regressions were used to relate visible scene composition to the six continuous ResNet-50 outputs [45]. For every regression, the predictor vector X comprised the 19 image-level semantic fractions generated by SegFormer-B5: road, sidewalk, building, wall, fence, pole, traffic light, traffic sign, vegetation, terrain, sky, person, rider, car, truck, bus, train, motorcycle, and bicycle. The four composite indicators and ResNet embeddings were not included as XGBoost predictors. The response y was the corresponding ResNet-derived scene score on the 0–10 scale ($S = 10p$); for the depression dimension, y was the Depressing score, $S(\text{depressing}) = 10 \times P(\text{depressing} \mid \text{image})$. Each Macau street-view image constituted one regression observation, whereas the four-view means at each sampling location were used for mapping and spatial analysis. Place Pulse volunteer comparisons were used only in the upstream training of the ResNet-50 classifiers and were not treated as Macau observations in the XGBoost regressions.

The model-development records indicate that the XGBoost hyperparameters were optimized using the Sparrow Search Algorithm (SSA). The selected (learning_rate, max_depth, n_estimators) settings were Beautiful (0.2957096, 9, 190), Boring (0.2858257, 9, 183), Depressing (0.2779112, 9, 188), Lively (0.2950642, 9, 195), Safety (0.2796208, 9, 184), and Wealthy (0.2765400, 9, 179). These category-specific settings were used in the six regression models reported below.

Cross-validation was used in the XGBoost model-development process. Because the response variables exhibit spatial autocorrelation, geographically grouped cross-validation provides an additional robustness strategy for evaluating geographic transferability [46]. In such a design, the four directional images from each sampling location should remain in the same fold, contiguous groups of locations should be withheld for testing, and model tuning should be performed only within the training folds. This spatial validation strategy

can be incorporated in subsequent model evaluation to complement the reported cross-validation results.

$$\text{Objective} = \sum_{i=1}^{n} L(y_i, \check{y}_i) + \sum_{k=1}^{k} \Omega(f_k), \quad (10)$$

$(y_i, \check{y}_i)$Squared error was used as the regression loss function, where y_i denotes the observed 0–10 ResNet-derived target score for sample i and $\check{y}_i$ denotes the corresponding XGBoost prediction, as shown in Equation (11).

$$L(y_i, \check{y}_i) = (y_i - \check{y}_i)^2 \quad (11)$$

The regularization term penalizes tree complexity and the magnitude of leaf weights, thereby reducing overfitting

$$\Omega(f_k) = \gamma M + \frac{1}{2}\lambda \sum_{j=1}^{M} \omega_j^2, \quad (12)$$

where $\Omega(f_k)$ denotes the regularization term for tree k, M is the number of leaves in the tree, γ is the penalty coefficient associated with the number of leaves, λ controls the L2 regularization of leaf weights, and ω_j denotes the weight assigned to leaf j.

At boosting iteration t, a new tree f_t is added to the existing ensemble, and the objective function can be expressed as

$$\text{Objective}^t = \sum_{i=1}^{n} \left\{ y_i - \left[\check{y}_i^{(t-1)} + f_t(x_i) \right] \right\}^2 + \gamma M + \frac{1}{2}\lambda \sum_{j=1}^{M} \omega_j^2, \quad (13)$$

where $\check{y}_i^{(t-1)}$ is the cumulative prediction for sample i after the first $t-1$ boosting iterations, $f_t(x_i)$ is the prediction contributed by the tree added at iteration t, and n is the number of training samples.

After T boosting iterations, the final prediction is obtained by summing the outputs of all trees in the ensemble

$$\check{y}_i = \sum_{t-1}^{T} f_t(x_i), \quad (14)$$

where $\check{y}_i$ is the final XGBoost prediction for sample i, T is the total number of trees, and $f_t(x_i)$ is the contribution of tree t to the prediction of sample i.

4. Results

Representative Macau street-view scenes and their corresponding model-predicted category scores are shown in Figure 3. SegFormer-B5 segmented the 20,256 Macau images into semantic classes, from which we calculated four objective visual indicators. Six ResNet-50 classifiers generated the model-predicted scene-category outputs. Global Moran's I statistics are positive for all ten variables shown in Table 3, with the highest values for Boring (0.673) and GVI (0.637). Reported *p*-values of 0.000 are expressed as $p < 0.001$. The results indicate clear spatial clustering in both streetscape composition and model-predicted perception across the Macau Peninsula.

Table 3. Reported global Moran's I statistics for six model-predicted scene categories and four objective visual indicators.

Index	Safety	Lively	Wealthy	Beautiful	Depressing	Boring	Walkability	Enclosure	SVF	GVI
Moran's I	0.559	0.592	0.594	0.546	0.458	0.673	0.326	0.390	0.358	0.637
Z-score	95.615	101.133	101.478	93.377	78.248	115.057	55.856	66.631	61.230	108.858
p value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001



Figure 3. Example street-view scenes from the Macau Peninsula and their model-predicted category scores.

4.1. Macau Peninsula Global Moran's I Cluster

Global Moran's I indicates that similar location-level image indicators and model-predicted category scores tend to occur near one another. Positive values for all ten variables in Table 3 indicate spatial clustering, although clustering strength differs among indicators. Boring (Moran's I = 0.673) and GVI (Moran's I = 0.637) show the strongest global spatial autocorrelation, followed by Wealthy, Lively, Safety, and Beautiful among the perception categories.

4.1.1. Spatial Characteristics of Model-Predicted Perception Scores

The six model-predicted perception scores exhibit marked spatial differentiation across the Macau Peninsula (Figure 4). Relatively high Beautiful, Safe, Lively, and Wealthy scores are concentrated in several central and historically established districts, particularly the traditional commercial core, the Nam Van area, and the vicinity of Avenida de Almeida Ribeiro (San Ma Lo). These areas feature a relatively dense mix of commercial, cultural, tourism, and public-space functions, along with comparatively continuous street frontages, pedestrian-oriented spaces, and visible greenery. The spatial coincidence of these streetscape characteristics with higher model-predicted scores suggests that the central urban area contains a distinctive combination of visible environmental features associated with several positive perception dimensions.

In contrast, relatively high Boring and Depressing scores appear more often in several peripheral or functionally specialized areas, including parts of Areia Preta, Fai Chi Kei, Toi San, Ilha Verde, Iao Hon, and adjoining residential or industrial blocks in the northern Macau Peninsula. Compared with the traditional urban core, some of these areas exhibit more homogeneous streetscape composition, less continuous pedestrian-oriented interfaces, fewer visually distinctive public-space elements, and lower visible greenery in selected locations. These characteristics correspond spatially with concentrations of relatively high Boring and Depressing scores.

The contrast between the traditional commercial and historic core and the northern and peripheral districts therefore indicates substantial intra-urban heterogeneity in the model-predicted visual attributes of the Macau Peninsula. However, these mapped values represent scene attributes predicted from street-view imagery rather than direct assessments obtained from Macau residents or visitors. The observed spatial patterns are therefore interpreted

as image-based prediction patterns and provide a basis for subsequent comparison with objective streetscape characteristics.

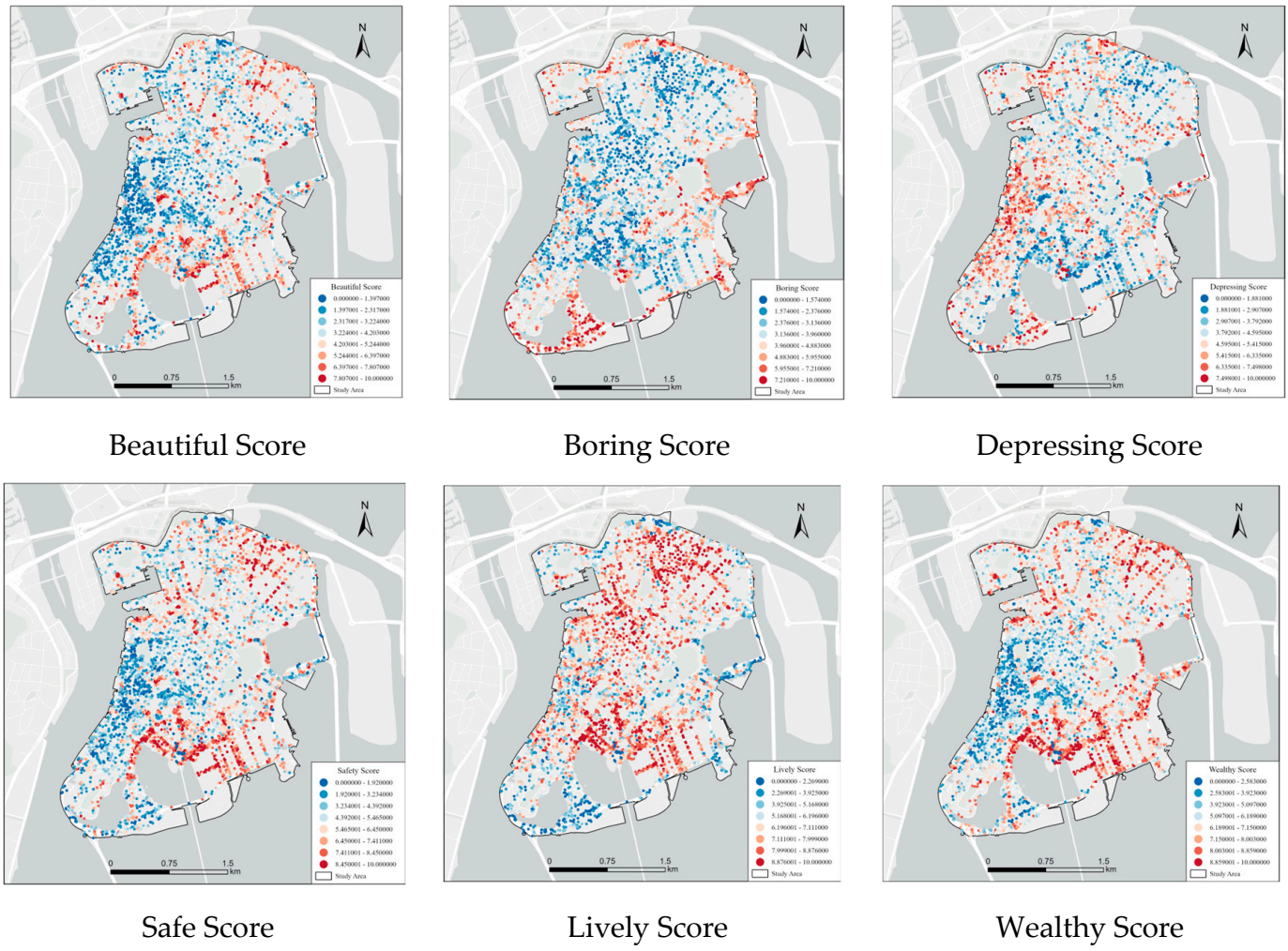


Figure 4. Spatial distribution of six model-predicted perception scores across the Macau Peninsula.

4.1.2. Spatial Characteristics of Objective Visual Indicators

The four segmentation-derived objective visual indicators—Green View Index (GVI), Sky View Factor (SVF), enclosure, and visual walkability—also exhibit clear spatial heterogeneity across the Macau Peninsula (Figure 5). Their distributions provide a physical streetscape context for interpreting the spatial patterns of the six model-predicted perception scores.

Streets with higher GVI values have more visible vegetation, including parts of the central and southern peninsula and selected streets around Nam Van. Some of these locations also correspond spatially with relatively high Beautiful, Safe, and Lively scores in Figure 4. This spatial correspondence indicates an association between visible greenery and several model-predicted positive scene attributes.

SVF displays a different spatial pattern. Relatively high values occur in several more open streets and reclaimed or peripheral areas, where wider road spaces and lower visual obstruction increase the proportion of visible sky. Parts of the northern peninsula and reclaimed districts show comparatively open visual fields, but these locations do not consistently coincide with high Beautiful, Safe, Lively, or Wealthy scores. This suggests that sky

visibility represents a distinct component of streetscape composition rather than a direct proxy for overall positive visual perception.

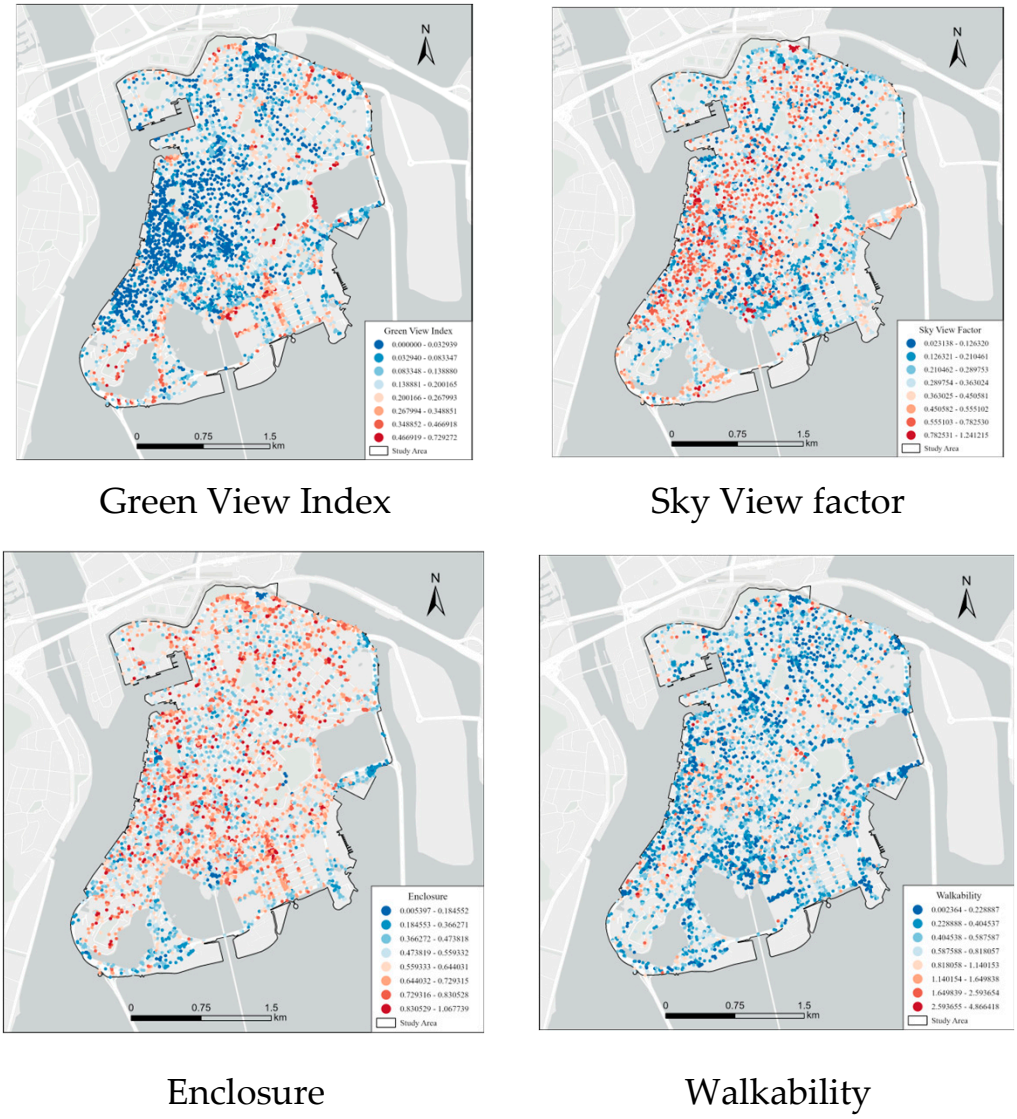


Figure 5. Spatial distribution of four segmentation-derived objective visual indicators across the Macau Peninsula.

Enclosure is comparatively pronounced in parts of the traditional urban core, including the vicinity of Avenida de Almeida Ribeiro and adjacent high-density historic and commercial streets, where continuous building frontages and relatively narrow street sections create stronger visual containment. In contrast, more open reclaimed and peripheral areas generally display lower enclosure values. The spatial distribution of enclosure partially overlaps with areas of relatively high model-predicted perception scores, although this relationship varies among perception dimensions.

Visual walkability also varies substantially across space. Relatively high values occur where sidewalks, pedestrian facilities, street interfaces, and other pedestrian-related elements are more visible, particularly in parts of the traditional commercial core and established mixed-use districts. Lower values are more frequently observed in streets dominated by vehicular infrastructure, large residential compounds, industrial functions, or discontinuous pedestrian facilities, including selected areas of Areia Preta, Fai Chi Kei, Ilha Verde,

and Iao Hon. These patterns provide an image-based representation of pedestrian-related streetscape characteristics.

Taken together, Figure 5 shows that the four objective indicators capture different dimensions of the visible streetscape. Their partial spatial correspondence with the model-predicted perception scores in Figure 4 provides a basis for examining associations between visible scene composition and perception-model outputs in the subsequent XGBoost–SHAP analysis.

The global Moran's I results further confirm significant spatial clustering across all six model-predicted perception scores and four objective visual indicators (Table 3). Among the perception dimensions, Boring exhibits the strongest global spatial autocorrelation (I = 0.673), followed by Wealthy (I = 0.594), Lively (I = 0.592), Safe (I = 0.559), Beautiful (I = 0.546), and Depressing (I = 0.458). Among the objective indicators, GVI shows the strongest clustering (I = 0.637), whereas enclosure (I = 0.390), SVF (I = 0.358), and visual walkability (I = 0.326) display comparatively weaker but still significant spatial autocorrelation. These results show that both model-predicted visual attributes and observable streetscape characteristics are spatially structured across the Macau Peninsula, although their clustering intensity differs by variable. The observed spatial correspondence between some perception-score patterns and objective indicators suggests potential associations, which the XGBoost–SHAP analysis examines further rather than interpreting as direct causal relationships.

4.2. Clustering and Outliers of Local Moran's I in Macau Peninsula

Historic streets, high-density residential neighborhoods, commercial districts, and reclaimed areas within the Macau Peninsula exhibit distinct spatial configurations of photographed streetscape characteristics. Local Moran's I was used to identify statistically significant local clusters and spatial outliers in the model-predicted perception scores and image-derived objective indicators. High–high (H–H) clusters represent locations with relatively high values surrounded by neighboring locations with similarly high values, whereas low–low (L–L) clusters indicate concentrations of relatively low values. High–low (H–L) and low–high (L–H) patterns represent local spatial outliers relative to their surrounding areas. These local patterns provide a finer-scale view of intra-urban spatial heterogeneity than the global Moran's I statistics.

4.2.1. Clustering and Outlier Patterns of Model-Predicted Perception Scores

Figure 6 shows the local spatial associations of the six model-predicted perception scores across the Macau Peninsula. The Safe score exhibits several H–H clusters in central and southern areas, including locations around the Nam Van Lake waterfront and nearby established commercial districts, whereas L–L clusters are more evident in parts of Iao Hon, Ilha Verde, and other older residential or industrial areas in the northern peninsula. Several H–L and L–H outliers occur around highly heterogeneous urban interfaces, including the vicinity of the Lisboa area, where large-scale tourism and commercial facilities are embedded within a more varied surrounding street environment.

The Lively score displays a similarly heterogeneous local pattern. H–H clusters are concentrated in parts of the traditional commercial core, including the vicinity of Avenida de Almeida Ribeiro (San Ma Lo), where continuous commercial frontages, pedestrian activity spaces, and mixed urban functions are visually prominent. In contrast, L–L clusters occur more frequently in parts of Toi San and Areia Preta, where the photographed streetscapes generally show more functionally homogeneous residential or transport-oriented characteristics. Local outliers also appear around major tourism and mixed-use areas such as the

Ruins of St. Paul's and the Horta e Costa/Shui Hang Mei vicinity, reflecting abrupt spatial transitions between visually active street segments and their immediate surroundings.

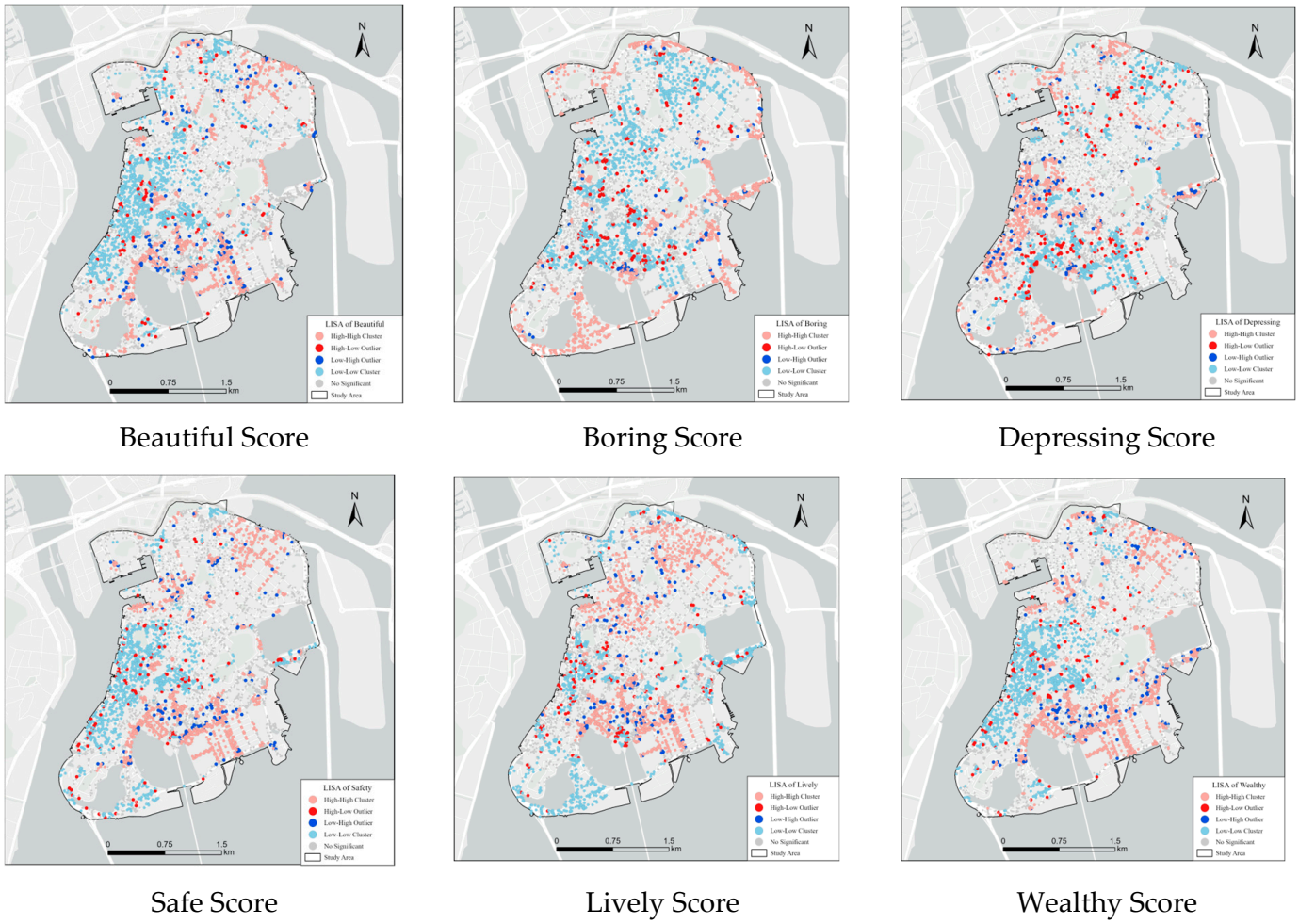


Figure 6. Local Moran's I clusters and spatial outliers for six model-predicted perception scores across the Macau Peninsula.

Beautiful and Wealthy scores also show localized clustering in several established central and southern districts. H–H clusters appear around parts of Penha Hill and adjacent lower-density residential areas, where greenery, hillside landscapes, detached buildings, and relatively open street views are more prominent. By contrast, portions of the Inner Harbour and adjoining logistics- or freight-related streets display L–L clusters, corresponding to streetscapes dominated by transport infrastructure, service functions, and comparatively limited visual diversity. Spatial outliers also occur along major urban transition corridors such as Avenida da Amizade, where modern high-rise development, tourism facilities, and older urban fabrics coexist within short distances.

For the Boring and Depressing dimensions, H–H clusters occur in several spatially concentrated locations rather than being uniformly distributed across the peninsula. Relatively high Depressing-score clusters occur in parts of older, high-density residential areas characterized by closely spaced buildings and visually constrained street sections, including selected "handshake-building" environments in the northern peninsula. Relatively high Boring-score clusters occur in some large-scale, functionally specialized spaces, including parking, transport, or infrastructure-dominated areas near the Macau Tower and selected reclaimed districts. These patterns indicate localized concentrations of model-predicted

visual attributes rather than direct evidence of boredom, depression, or other psychological states among residents or visitors.

Overall, the Local Moran's \(I\) results reveal pronounced spatial heterogeneity in the six model-predicted perception dimensions. Central commercial, tourism, and historic districts tend to contain more H–H clusters for Beautiful, Safe, Lively, and Wealthy scores, whereas selected northern residential, industrial, and peripheral districts contain more localized H–H clusters for Boring and Depressing scores. However, H–L and L–H outliers in several transition zones show that a uniform core–periphery structure does not define these patterns. Instead, the local spatial configuration reflects the fine-grained heterogeneity of Macau's urban morphology and photographed streetscape composition.

4.2.2. Clustering and Outlier Patterns of Objective Visual Indicators

The Local Moran's I results for the four segmentation-derived objective visual indicators reveal pronounced local spatial heterogeneity across the Macau Peninsula (Figure 7). High–high (H–H) clusters indicate neighboring locations with similarly high indicator values, low–low (L–L) clusters identify concentrations of relatively low values, and high–low (H–L) and low–high (L–H) classes represent local spatial outliers. The resulting patterns demonstrate that greenery, sky visibility, enclosure, and visible pedestrian-related streetscape elements are unevenly distributed across different urban contexts.

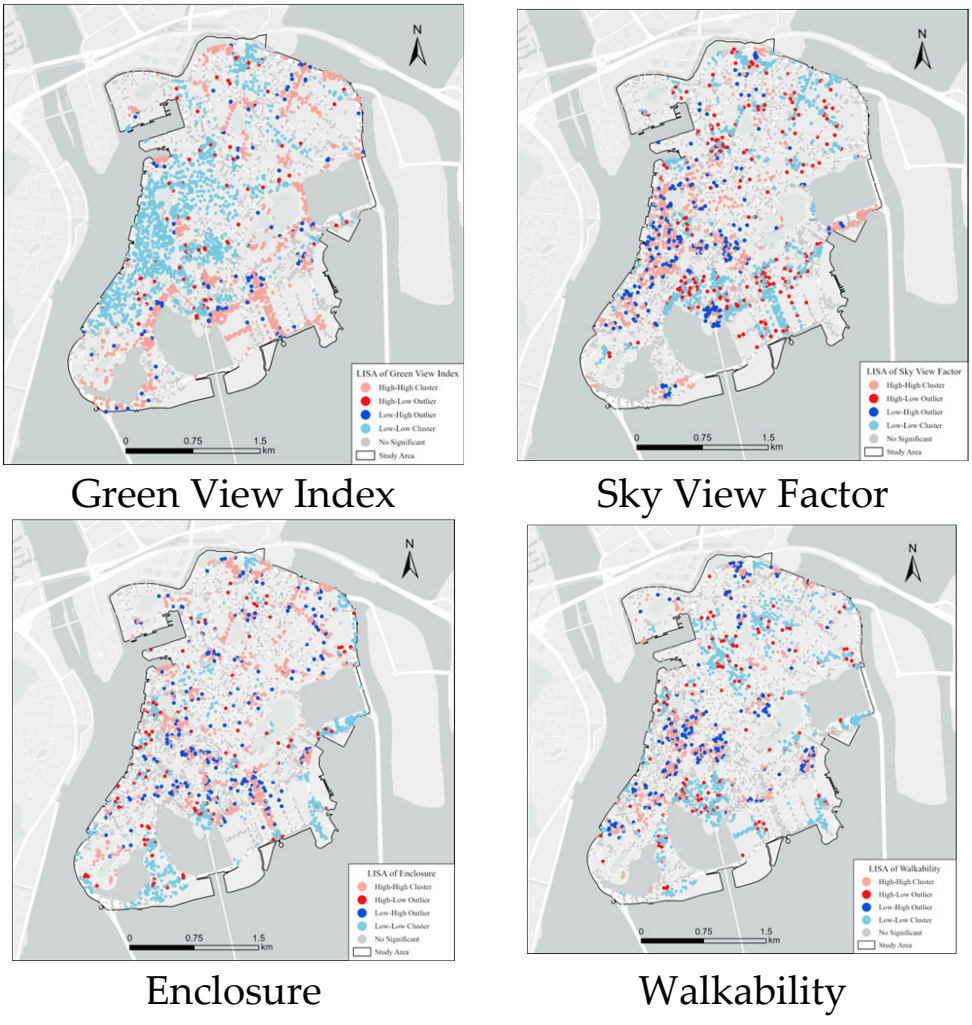


Figure 7. Local Moran's I clusters and spatial outliers for four segmentation-derived objective visual indicators across the Macau Peninsula.

For visual walkability, H–H clusters are concentrated in several established historical and mixed-use districts, where continuous sidewalks, pedestrian-oriented street interfaces, and other visible pedestrian-related facilities are comparatively prominent. In contrast, L–L clusters occur in some newly developed and reclaimed areas, including streets associated with the Hong Kong–Zhuhai–Macao Bridge area, where large-scale road infrastructure and more automobile-oriented spatial configurations are visually dominant. Several local outliers are also observed around sloping street environments, including the ramp areas near Dongwangyang, indicating that streets with distinctive topographic and morphological conditions may differ substantially from their immediate surroundings in terms of visible pedestrian-related elements. These results should be interpreted as image-based differences in streetscape composition rather than as direct measurements of actual pedestrian accessibility or walking behavior.

Sky View Factor (SVF) presents a different local spatial configuration. H–H clusters occur in relatively open streets and spaces with fewer visual obstructions, whereas L–L clusters are more evident within compact traditional street networks and areas surrounded by continuous or high-rise building frontages. In the traditional streets and alleys around Avenida de Almeida Ribeiro (San Ma Lo), local SVF values vary considerably as street width and surrounding building form change over short distances. Lower SVF clusters are also observed in several highly developed transport and commercial areas, including locations around major underground transport facilities and high-density business districts. These patterns reflect localized differences in visible sky openness rather than overall environmental quality.

Enclosure exhibits relatively strong clustering within parts of Macau's compact historical urban fabric. Streets characterized by continuous façades, narrow sections, and closely spaced buildings tend to form localized clusters of higher enclosure values. Spatial outliers are observed around transitional locations such as Shui Hang Mei, where abrupt changes in street width, building setbacks, or surrounding built form produce values that differ from those of adjacent streets. Such outliers highlight the fine-grained morphological transitions typical of Macau's high-density urban environment.

GVI also displays marked local clustering. Higher-value clusters are observed in areas with more continuous visible vegetation, including parts of the Guia Hill and related recreational or hillside environments, where roadside planting and green spaces contribute a greater proportion of vegetation pixels within street-view images. In contrast, L–L clusters are found in selected densely built traditional neighborhoods, including parts of the San Kio area, where visible vegetation is comparatively limited. Local GVI outliers are also present in locations containing highly localized greening interventions, such as podium or roof-garden environments embedded within otherwise dense built-up areas. These outliers illustrate how small-scale vertical or elevated greenery can generate localized differences in image-derived vegetation visibility relative to surrounding streets.

Overall, the Local Moran's I patterns demonstrate that the four objective indicators capture distinct aspects of Macau's streetscape structure. Historical districts tend to exhibit stronger clustering in enclosure and visible pedestrian-related elements, while reclaimed and infrastructure-dominated areas show contrasting patterns of openness and walkability-related visual features. Greenery and sky visibility display their own localized distributions that do not necessarily coincide with one another. The presence of H–L and L–H outliers further indicates that Macau's compact urban form contains abrupt transitions in streetscape composition over relatively short distances. These spatial configurations provide a basis for subsequent comparison with the six model-predicted perception scores, without implying direct causal effects on residents' or visitors' perceptions.

4.3. Regression Analysis of Macau Peninsula

We developed six separate XGBoost regression models to examine statistical associations between visible streetscape composition and the six model-predicted perception scores across the Macau Peninsula. For each regression, the predictor variables consisted of the 19 image-level semantic fractions derived from SegFormer-B5, including road, sidewalk, building, wall, fence, pole, traffic light, traffic sign, vegetation, terrain, sky, person, rider, car, truck, bus, train, motorcycle, and bicycle. The response variable was the corresponding ResNet-50-derived scene score on the 0–10 scale for Beautiful, Boring, Depressing, Lively, Safe, or Wealthy. Place Pulse 2.0 volunteer comparisons were used only in the upstream training of the ResNet-50 classifiers and were not treated as Macau observations in the XGBoost regression models.

We applied cross-validation during model development and evaluated performance using mean squared error (MSE), mean absolute error (MAE), and the coefficient of determination (R^2). Table 4 summarizes the predictive performance of the six regression models. The Beautiful model achieved an R^2 of 0.7344, followed by Wealthy ($R^2 = 0.7261$), Safe ($R^2 = 0.7098$), Boring ($R^2 = 0.7084$), Lively ($R^2 = 0.6860$), and Depressing ($R^2 = 0.6434$). These results indicate that the 19 visible semantic scene fractions account for a substantial proportion of the variation in the six ResNet-derived model outputs. However, model performance differs across perception dimensions.

Table 4. Evaluation performance of the six XGBoost regression models for predicting the 0–10 ResNet-derived perception scores.

Perception Categories	Beautiful	Boring	Depressing	Lively	Safe	Wealthy
MAE	0.7231	0.6492	0.7225	0.7925	0.8253	0.7288
RMSE	1.0709	0.9517	1.0520	1.1842	1.2138	1.1018
R^2	0.7344	0.7084	0.6434	0.6860	0.7098	0.7261

The relatively higher (R^2) values for Beautiful, Wealthy, Safe, and Boring suggest that these model-predicted scene attributes are more strongly associated with the visible semantic composition captured by the street-view images. By comparison, the lower R^2 values for Depressing and Lively indicate that these outputs are less fully explained by the selected semantic fractions alone. This difference does not necessarily reflect greater subjectivity of these dimensions; rather, it suggests that additional visual, spatial, or contextual information not represented by the 19 semantic proportions may contribute to the corresponding ResNet-50 predictions.

Overall, the regression results demonstrate that observable streetscape composition is systematically associated with the six model-predicted perception scores. However, interpret the XGBoost models as explaining variation in the ResNet-derived image-based outputs rather than directly predicting the perceptions of Macau residents or visitors. The following section further examines each semantic class's relative contribution to each perception dimension using SHAP analysis.

4.4. Visual Analysis of SHAP in Macau Peninsula

SHAP was used to interpret the six fitted XGBoost models by quantifying the contribution of each semantic scene fraction relative to the model-specific baseline [47]. For a given model, a positive SHAP value indicates that the corresponding feature increases the predicted target score, whereas a negative SHAP value decreases it. This interpretation is maintained consistently across all six regression targets. Accordingly, positive SHAP values increase the Boring score in the Boring model and increase the Depressing score in the Depressing model. The SHAP results therefore describe model-specific associations

between visible scene composition and the six ResNet-derived perception scores rather than causal effects of individual streetscape elements on human perception.

4.4.1. Relationships Between Positive Perception Scores and Objective Environmental Characteristics

Figure 8 presents the SHAP summary plots for the Beautiful, Lively, Safe, and Wealthy XGBoost models. Across these four models, vegetation, sky, buildings, pedestrian-related elements, and road-related features show heterogeneous contributions in both magnitude and direction, indicating that different combinations of visible streetscape characteristics are associated with different model-predicted perception dimensions.

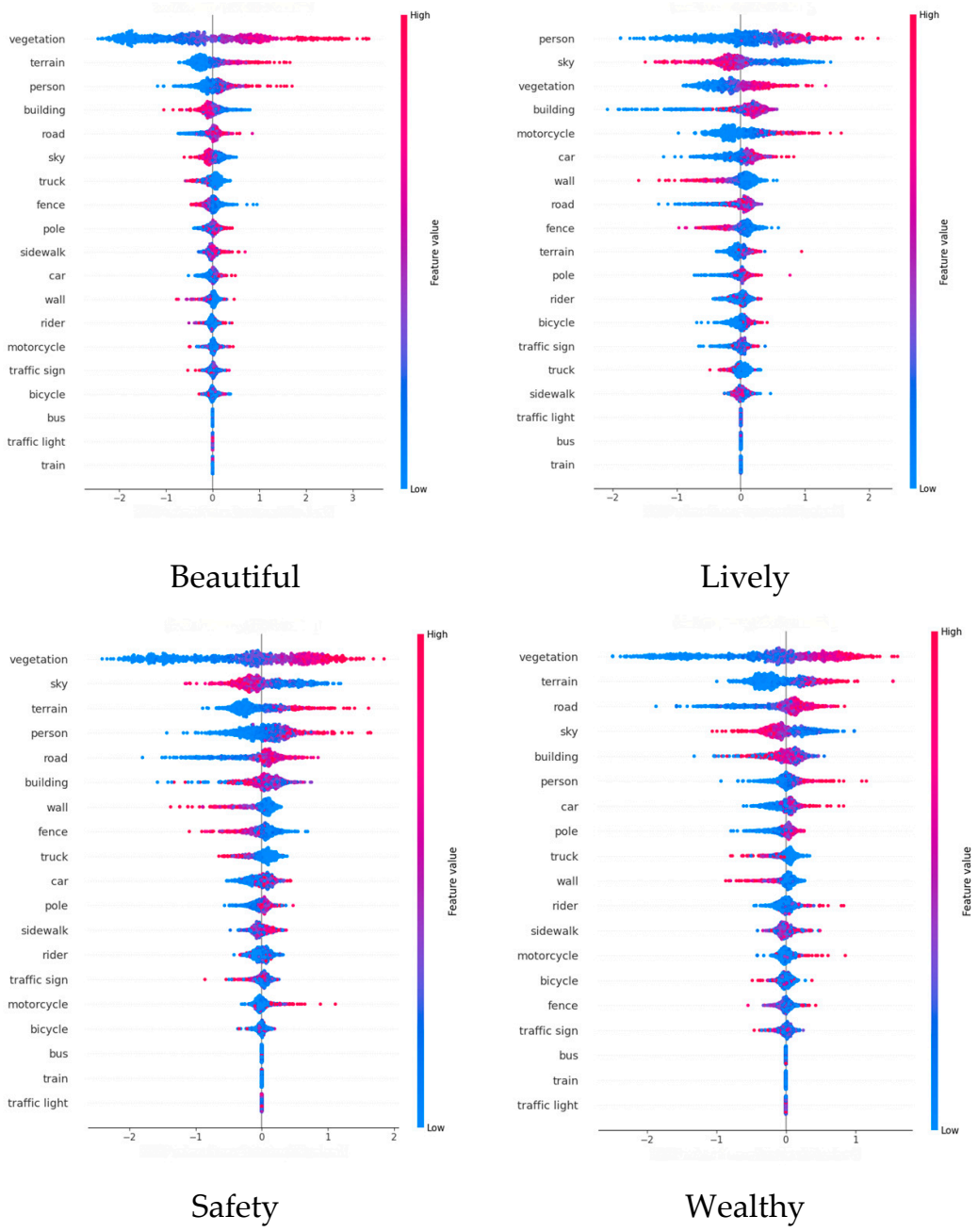


Figure 8. SHAP summary plots for the Beautiful, Lively, Safe, and Wealthy XGBoost models.

For the Beautiful model, vegetation and sky show comparatively prominent contributions, suggesting that visible greenery and openness are important components associated

with higher model-predicted Beautiful scores. This pattern is spatially consistent with several greener and more visually open areas identified in the preceding spatial analysis, including parts of the Nam Van Lake waterfront and other central or southern streets with relatively high vegetation visibility. Building- and terrain-related features also contribute to the model output, indicating that the visual composition of the built environment is considered alongside natural elements rather than through a single dominant factor.

For the Lively model, pedestrian-related elements and selected traffic- and street-related features display notable SHAP contributions. These results suggest that streets containing a greater visible presence of people, active street interfaces, and transport-related visual elements may be associated with higher model-predicted Lively scores. These relationships broadly align with the spatial concentration of relatively high Lively scores in established commercial and tourism areas such as Avenida de Almeida Ribeiro (San Ma Lo), the Ruins of St. Paul's surroundings, and other mixed-use central districts. However, the SHAP results identify statistical associations within the fitted model and do not directly measure actual pedestrian volume or observed street activity.

For the Safe model, vegetation, sky, and selected built-environment features also contribute to variation in the model-predicted scores. These features suggest that combinations of visible greenery, openness, and surrounding urban form are associated with the Safe output. This interpretation is consistent with the relatively high Safe scores observed in parts of the Nam Van area and other visually open central districts. However, it should not be interpreted as evidence that these features directly increase residents' psychological security or reduce actual safety risks.

For the Wealthy model, building-related, road-related, vegetation, and terrain features contribute differently to the predicted score. The comparatively strong role of built-form characteristics indicates that the visual appearance and composition of buildings and street infrastructure matter for the model-predicted Wealthy dimension. This pattern is broadly consistent with higher Wealthy scores observed in parts of established commercial areas, modern development corridors, and higher-end residential environments around central and southern Macau. At the same time, natural elements such as vegetation and terrain also contribute to the fitted output, suggesting that the Wealthy model responds to a combination of built and natural visual characteristics rather than to architectural form alone.

Overall, the SHAP results demonstrate that the four positively oriented model-predicted perception dimensions are associated with different configurations of visible streetscape elements. Vegetation and sky are comparatively important in the Beautiful and Safe models. Meanwhile, pedestrian- and traffic-related features matter more for the Lively model, and building- and road-related characteristics contribute substantially to the Wealthy model. These findings provide an interpretable link between semantic scene composition and model outputs. However, they should be understood as model-based associations rather than direct evidence of causal effects on residents' or visitors' perceptions.

4.4.2. Relationships of Boring and Depressing Scores with Objective Environmental Characteristics

Figure 9 presents the SHAP summary plots for the Boring and Depressing XGBoost models. In both models, positive SHAP values indicate contributions toward higher model-predicted target scores, whereas negative SHAP values indicate contributions toward lower scores. For the Depressing model, a positive SHAP value therefore corresponds to an increase in the model-predicted Depressing score, while for the Boring model it corresponds to an increase in the model-predicted Boring score. Interpret these contributions as model-specific associations between semantic scene composition and the corresponding ResNet-derived outputs, rather than as direct effects on residents' emotional or psychological states.

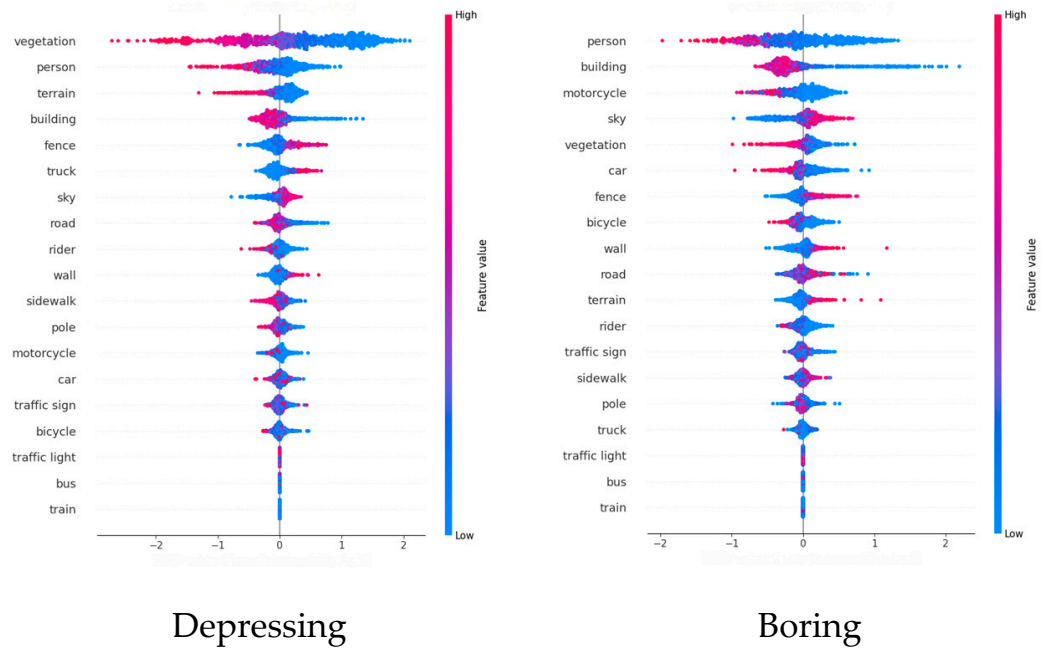


Figure 9. SHAP summary plots for the Boring and Depressing XGBoost models.

For the Depressing model, several built-environment and traffic-related semantic features, including cars, fences, and building-related elements, show positive SHAP contributions across portions of the feature distribution. This indicates that higher proportions of these visible elements are associated with higher model-predicted Depressing scores under the fitted XGBoost model. By contrast, vegetation shows a different contribution pattern, with higher vegetation-related values tending to be associated with lower Depressing scores in parts of the distribution. This result is broadly consistent with the spatial patterns observed for greener streets in the preceding analysis; however, it does not establish that vegetation directly alleviates depression or improves psychological well-being.

For the Boring model, wall- and sky-related features exhibit notable SHAP contributions, indicating that the relative presence of these scene components is associated with variation in the model-predicted Boring score. The direction of contribution varies across feature values, emphasizing that these relationships are nonlinear and model-dependent. Building, pedestrian, and traffic-related elements also contribute to the Boring output, suggesting that the model responds to combinations of streetscape composition rather than to any single environmental characteristic in isolation.

Pedestrian- and traffic-related features, including person and traffic-light fractions, also show non-negligible SHAP contributions in the two models. Their effects should not be interpreted as direct evidence that crowd activity or traffic flow reduces boredom or depression. Instead, they indicate that the visual presence of human activity and transport infrastructure forms part of the semantic scene configuration associated with the model-predicted scores.

Overall, the SHAP results indicate that the Boring and Depressing outputs are associated with distinct combinations of built, natural, pedestrian, and traffic-related visual elements. The heterogeneous direction and magnitude of SHAP values further demonstrate that the relationships between semantic scene fractions and the two model outputs are nonlinear and model-dependent. These results complement the spatial clustering analysis by identifying which visible scene components contribute most strongly to variations in the fitted Boring and Depressing scores, while avoiding causal interpretation of these associations.

4.5. Grad-CAM Visualization Results for the Macau Peninsula

Grad-CAM was applied to selected ResNet-50 predictions to identify image regions that contributed most strongly to the corresponding model outputs [48]. Unlike the XGBoost–SHAP analysis, which quantifies the contribution of semantic scene fractions to the fitted regression targets, Grad-CAM provides an image-level visualization of class-specific activation within the perception classifiers. The highlighted regions therefore indicate where the ResNet-50 model concentrated its discriminative response for a given prediction, rather than where human observers necessarily directed their visual attention.

The selected examples in Figure 10 show that building façades, vegetation, street surfaces, commercial signage, pedestrian spaces, and traffic-related elements frequently occur within regions of relatively high activation. For the Beautiful-related examples, stronger activation is often observed around visually distinctive façades, vegetation, and well-defined street interfaces, including scenes associated with historic and landscaped areas of the Macau Peninsula. These patterns suggest that the model draws on combinations of architectural and natural visual information when generating Beautiful predictions.

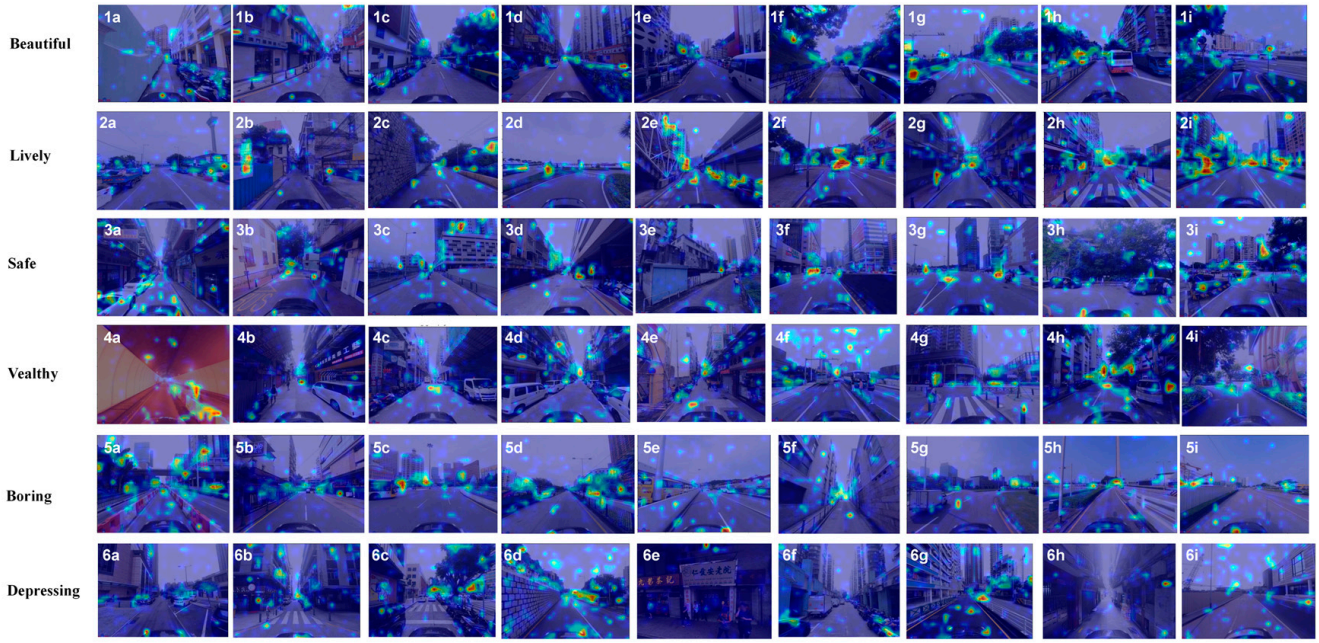


Figure 10. Grad-CAM visualizations for selected ResNet-50 predictions in the Macau Peninsula: (**1a**–**1i**) Beautiful; (**2a**–**2i**) Lively; (**3a**–**3i**) Safe; (**4a**–**4i**) Wealthy; (**5a**–**5i**) Boring; and (**6a**–**6i**) Depressing. Each group contains nine street-view examples for the corresponding perception category. The overlaid Grad-CAM heatmaps highlight image regions contributing to the respective class predictions.

For the Lively related examples, high-activation regions frequently coincide with commercial signage, public-space interfaces, pedestrian activity areas, and traffic-related elements. These patterns align with the visually complex streetscapes of major commercial and tourism districts, including areas around Avenida de Almeida Ribeiro and the Ruins of St. Paul's. However, interpret these activations as visual features used by the classifier rather than as direct measurements of actual pedestrian intensity or urban vitality.

In the Safe-related examples, activation is distributed across street edges, sidewalks, lighting-related elements, façades, and relatively unobstructed portions of the streetscape. This suggests the classifier responds to multiple visible aspects of street organization rather than a single feature. These image-level activation patterns should not be interpreted as direct evidence of actual crime risk, psychological security, or residents' perceived safety.

For the Boring and Depressing categories, several examples show relatively strong activation around extensive built surfaces, visually repetitive streetscape elements, road-dominated spaces, and areas with limited visible vegetation. In the Depressing row, higher activation corresponds specifically to the model's Depressing class, consistent with the score definition used throughout the study. These results identify image regions associated with the model classifications but do not indicate clinical depression or residents' emotional states.

The Grad-CAM results also reveal substantial variation among individual scenes within the Macau Peninsula. Central historic and commercial districts, modern tourism corridors, high-density northern neighborhoods, and peripheral or infrastructure-dominated areas exhibit different combinations of activated visual elements. Rather than supporting a simple core–periphery dichotomy, these examples illustrate the fine-grained visual heterogeneity of Macau's street environment and the model's context-dependent response.

Overall, Grad-CAM complements the XGBoost–SHAP analysis by providing a second, image-level perspective on model interpretation. SHAP identifies how semantic scene fractions relate to the fitted XGBoost outputs, whereas Grad-CAM localizes regions within individual street-view images that contribute to the corresponding ResNet-50 classifications. Combining these two interpretability approaches can help identify candidate visual elements for subsequent field-based investigation. However, it should not be treated as independent validation of local perception or as evidence of causal relationships between streetscape features and psychological or behavioral outcomes.

5. Discussion

5.1. Sustainability Implications of Spatial Variation in Model-Predicted Urban Visual Perception

The visual-category maps provide a city-scale screening of streetscape conditions and reveal systematic spatial variation in the six model-predicted perception dimensions. The concentration of relatively high and low scores in different parts of the Macau Peninsula demonstrates spatial heterogeneity in the image-derived indicators and model-predicted visual attributes across the street network. These patterns can help identify areas for more detailed field assessment and community-based evaluation of streetscape quality.

The contrast between central and peripheral streets also highlights spatial differences in visible amenities and street form. Integrating these image-derived indicators with demographic information, pedestrian audits, accessibility measures, and participatory evaluations can extend the analysis from visual diagnosis to a broader assessment of spatial equity and planning need.

The XGBoost–SHAP results further show that the associations between visible features and predicted categories vary across perception dimensions. Vegetation, sky, building surfaces, pedestrian-related elements, and traffic-related classes show different SHAP contribution patterns across the six models, indicating that the model-predicted perception dimensions are associated with different combinations of visible scene characteristics. These differentiated associations provide a practical basis for selecting site-specific elements for field assessment and design review.

5.2. Planning Implications for Sustainable Urban Renewal

The framework's planning value lies in a staged diagnostic procedure. Spatial maps identify streets with distinctive visual patterns, SHAP identifies semantic features linked to category-specific scores, and Grad-CAM highlights the image regions that contribute to individual predictions. These outputs can be combined with field surveys and stakeholder consultation to formulate place-specific streetscape renewal priorities.

Where field assessment confirms low street-level greenery and poor pedestrian conditions, consider compact interventions such as street trees, small planted areas, vertical

greening, shading, and routine maintenance. The spatial distribution of GVI and pedestrian-related semantic features can help identify locations where such interventions warrant closer design evaluation.

Where audits document fragmented sidewalks or difficult crossings, design responses may include continuous pedestrian routes, accessible crossings, legible wayfinding, and improved management of pedestrian–vehicle conflict. The visual-walkability indicator and relevant semantic classes can support the preliminary identification of streets requiring these assessments.

Interpret enclosure and frontage conditions within the local urban and heritage context. Evaluate transparent or active street interfaces, facade maintenance, and pedestrian-scale improvements in locations where enclosure and building-related features strongly influence perception scores.

Where fieldwork identifies visually monotonous streetscapes, planners may examine opportunities for public space, active frontages, compatible community uses, and greater visual diversity. In high-density and tourism-oriented areas, these design options can be considered together with land-use, mobility, noise, crowding, and resident-preference evidence.

Accordingly, the proposed planning sequence consists of computational screening, field and community assessment, feasibility evaluation, intervention design, and post-intervention monitoring. This sequence translates image-based spatial diagnosis into a structured workflow for sustainable urban renewal while retaining local planning judgment during implementation.

The analytical framework is transferable to other high-density urban contexts when it follows the same sequence of image sampling, semantic segmentation, perception prediction, spatial analysis, and interpretable modeling. Local calibration and cross-area validation can further strengthen applications in cities with different cultural, morphological, and streetscape characteristics. For application in another city, local calibration can follow four steps: (1) select a geographically stratified subset of local street-view images; (2) recruit local participants to conduct the same pairwise comparisons for the six perceptual categories; (3) compare locally elicited judgments with the source-trained model outputs and recalibrate or fine-tune the classifiers where systematic disagreement occurs; and (4) evaluate the calibrated models using geographically separated validation samples before city-wide mapping.

5.3. Future Research and Validation

Future research can extend this cross-sectional assessment with a structured longitudinal design. A practical schedule is to establish a new baseline within 6–12 months, repeat street-view acquisition during both warm/wet and cool/dry periods at comparable locations and viewing directions, and conduct at least two additional annual rounds. This design would allow evaluation of seasonal stability and year-to-year changes in the six perception scores and four objective indicators. Locally recruited participants can also assess a geographically stratified sample of Macau street-view images using the same six perceptual categories, enabling comparison between locally elicited evaluations and model-predicted scores. For the XGBoost stage, geographically grouped cross-validation can be added as a robustness analysis by assigning all four views from the same sampling location to a common fold and withholding contiguous spatial groups for testing [46].

Future analyses can further integrate image-derived results with pedestrian observations, accessibility measures, land-use information, and participatory assessment. Applying the same georeferenced sampling frame across repeated surveys would make it possible to compare temporal changes in streetscape composition, perception predictions, and spatial clustering and to evaluate how these changes correspond to urban-renewal interventions.

6. Conclusions

This study developed an integrated framework combining SegFormer-B5 semantic segmentation, ResNet-50 perception classifiers, global and local spatial autocorrelation, XGBoost–SHAP, and Grad-CAM to analyze 20,256 street-view images from 5064 locations in the Macau Peninsula. The six model-predicted perception scores and four segmentation-derived objective indicators exhibit clear spatial autocorrelation, with particularly strong clustering for the Boring score and GVI. The probability-to-score transformation is consistently defined as S = 10p, and the Depressing score retains the same direction throughout the analysis, with higher values indicating a stronger model-predicted depressing character. XGBoost uses 19 SegFormer semantic fractions as predictors of the six 0–10 ResNet-derived scores, and SHAP identifies category-specific associations between visible scene elements and these outputs. Grad-CAM provides complementary image-level localization of features contributing to individual predictions. Together, the framework provides a scalable spatial-diagnostic approach for identifying streetscape patterns, interpreting their visual composition, and supporting targeted field assessment and sustainable urban-renewal planning.

Author Contributions: C.P., conceptualization; writing—original draft; data curation; writing—review and editing. J.G., writing—review and editing; visualization. J.Y., writing—review and editing; software; supervision. J.C., writing—review; validation; resources. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by (1) Projects of Talents Recruitment of GDUPT (2023rcyj2015 and 2023rcyj2006). (2) Science and Technology Project of Maoming (2025664 and 2024011). (3) Philosophy and Social Sciences Planning Jointly Built Project of Maoming (2025GJ10). (4) Young Innovative Talents Projects in Ordinary Universities in Guangdong Province (2025KQNCX048). (5) Characteristic Innovation Projects in Ordinary Universities in Guangdong Province (2025KTSCX079).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ki, D.; Lee, S. Analyzing the effects of Green View Index of neighborhood streets on walking time using Google Street View and deep learning. *Landsc. Urban Plan.* **2021**, *205*, 103920. [CrossRef]
2. Koo, B.W.; Guhathakurta, S.; Botchwey, N. How are neighborhood and street-level walkability factors associated with walking behaviors? A big data approach using street view images. *Environ. Behav.* **2022**, *54*, 211–241. [CrossRef]
3. Lu, Y.; Sarkar, C.; Xiao, Y. The effect of street-level greenery on walking behavior: Evidence from Hong Kong. *Soc. Sci. Med.* **2018**, *208*, 41–49. [CrossRef]
4. Wang, R.; Liu, Y.; Lu, Y.; Zhang, J.; Liu, P.; Yao, Y.; Grekousis, G. Perceptions of built environment and health outcomes for older Chinese in Beijing: A big data approach with street view images and deep learning technique. *Comput. Environ. Urban Syst.* **2019**, *78*, 101386. [CrossRef]
5. Zhou, H.; He, S.; Cai, Y.; Wang, M.; Su, S. Social inequalities in neighborhood visual walkability: Using street view imagery and deep learning technologies to facilitate healthy city planning. *Sustain. Cities Soc.* **2019**, *50*, 101605. [CrossRef]
6. Biljecki, F.; Ito, K. Street view imagery in urban analytics and GIS: A review. *Landsc. Urban Plan.* **2021**, *215*, 104217. [CrossRef]
7. Kang, Y.; Zhang, F.; Gao, S.; Lin, H.; Liu, Y. A review of urban physical environment sensing using street view imagery in public health studies. *Ann. GIS* **2020**, *26*, 261–275. [CrossRef]
8. Li, W.; Lai, Y.; Xu, L.; Xiangli, Y.; Yu, J.; He, C.; Xia, G.-S.; Lin, D. OmniCity: Omnipotent city understanding with multi-level and multi-view images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 17397–17407. [CrossRef]
9. Rosenfelder, M.; Wussow, M.; Gust, G.; Cremades, R.; Neumann, D. Predicting residential electricity consumption using aerial and street view images. *Appl. Energy* **2021**, *301*, 117407. [CrossRef]

10. Liu, Y.-C.; Ma, C.-Y.; He, Z.; Kuo, C.-W.; Chen, K.; Zhang, P.; Wu, B.; Kira, Z.; Vajda, P. Unbiased teacher for semi-supervised object detection. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 3–7 May 2021.
11. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–22 June 2022; pp. 11976–11986. [CrossRef]
12. Scott, K.E.; Benson, J.F. *Public and Professional Attitudes to Landscape: Scoping Study*; Scottish Executive Social Research: Edinburgh, UK, 2002.
13. Liu, X.; Tvinnereim, E.; Grimsrud, K.M.; Lindhjem, H.; Velle, L.G.; Saure, H.I.; Lee, H. Explaining landscape preference heterogeneity using machine learning-based survey analysis. *Landsc. Res.* **2021**, *46*, 417–434. [CrossRef]
14. Li, J.; Zhang, Z.; Jing, F.; Gao, J.; Ma, J.; Shao, G.; Noel, S. An evaluation of urban green space in Shanghai, China, using eye tracking. *Urban For. Urban Green.* **2020**, *56*, 126903. [CrossRef]
15. Li, X.; Ratti, C.; Seiferling, I. Mapping urban landscapes along streets using Google Street View. In *Advances in Cartography and GIScience*; Lecture Notes in Geoinformation and Cartography; Springer: Cham, Switzerland, 2017; pp. 341–356. [CrossRef]
16. Xia, Y.; Yabuki, N.; Fukuda, T. Development of a system for assessing the quality of urban street-level greenery using street view images and deep learning. *Urban For. Urban Green.* **2021**, *59*, 126995. [CrossRef]
17. Zhang, F.; Zhang, D.; Liu, Y.; Lin, H. Representing place locales using scene elements. *Comput. Environ. Urban Syst.* **2018**, *71*, 153–164. [CrossRef]
18. Sahraoui, Y.; Clauzel, C.; Foltete, J.-C. A metrics-based approach for modeling covariation of visual and ecological landscape qualities. *Ecol. Indic.* **2021**, *123*, 107331. [CrossRef]
19. Porzi, L.; Rota Bulo, S.; Lepri, B.; Ricci, E. Predicting and understanding urban perception with convolutional neural networks. In Proceedings of the 23rd ACM International Conference on Multimedia, Brisbane, Australia, 26–30 October 2015; pp. 139–148. [CrossRef]
20. Dubey, A.; Naik, N.; Parikh, D.; Raskar, R.; Hidalgo, C.A. Deep learning the city: Quantifying urban perception at a global scale. In *Computer Vision—ECCV 2016*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2016; Volume 9905, pp. 196–212. [CrossRef]
21. Helbich, M.; Yao, Y.; Liu, Y.; Zhang, J.; Liu, P.; Wang, R. Using deep learning to examine street view green and blue spaces and their associations with geriatric depression in Beijing, China. *Environ. Int.* **2019**, *126*, 107–117. [CrossRef] [PubMed]
22. Wang, R.; Ren, S.; Zhang, J.; Yao, Y.; Wang, Y.; Guan, Q. A comparison of two deep-learning-based urban perception models: Which one is better? *Comput. Urban Sci.* **2021**, *1*, 3. [CrossRef]
23. Wei, J.; Yue, W.; Li, M.; Gao, J. Mapping human perception of urban landscape from street-view images: A deep-learning approach. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102886. [CrossRef]
24. Meng, Y.; Xing, H. Exploring the relationship between landscape characteristics and urban vibrancy: A case study using morphology and review data. *Cities* **2019**, *95*, 102389. [CrossRef]
25. Chen, W.; Wu, A.N.; Biljecki, F. Classification of urban morphology with deep learning: Application on urban vitality. *Comput. Environ. Urban Syst.* **2021**, *90*, 101706. [CrossRef]
26. Huang, X.; Gong, P.; Wang, S.; White, M.; Zhang, B. Machine Learning Modeling of Vitality Characteristics in Historical Preservation Zones with Multi-Source Data. *Buildings* **2022**, *12*, 1978. [CrossRef]
27. Li, M.; Pan, J. Assessment of influence mechanisms of built environment on street vitality using multisource spatial data: A case study in Qingdao, China. *Sustainability* **2023**, *15*, 1518. [CrossRef]
28. Li, X.; Li, Y.; Jia, T.; Zhou, L.; Hijazi, I.H. The six dimensions of built environment on urban vitality: Fusion evidence from multi-source data. *Cities* **2022**, *121*, 103482. [CrossRef]
29. Wang, Z.; Xia, N.; Zhao, X.; Gao, X.; Zhuang, S.; Li, M. Evaluating urban vitality of street blocks based on multi-source geographic big data: A case study of Shenzhen. *Int. J. Environ. Res. Public Health* **2023**, *20*, 3821. [CrossRef] [PubMed]
30. Lee, S.; Kang, J.E. Impact of particulate matter and urban spatial characteristics on urban vitality using spatiotemporal big data. *Cities* **2022**, *131*, 104030. [CrossRef]
31. Zhao, K.; Guo, J.; Ma, Z.; Wu, W. Exploring the spatiotemporal heterogeneity and stationarity in the relationship between street vitality and built environment. *SAGE Open* **2023**, *13*, 21582440231152226. [CrossRef]
32. Pakoz, M.Z.; Isik, M. Rethinking urban density, vitality and healthy environment in the post-pandemic city: The case of Istanbul. *Cities* **2022**, *124*, 103598. [CrossRef]
33. Yue, W.; Chen, Y.; Thy, P.T.M.; Fan, P.; Liu, Y.; Zhang, W. Identifying urban vitality in metropolitan areas of developing countries from a comparative perspective: Ho Chi Minh City versus Shanghai. *Sustain. Cities Soc.* **2021**, *65*, 102609. [CrossRef]
34. Wang, R.; Lu, Y.; Zhang, J.; Liu, P.; Yao, Y.; Liu, Y. The relationship between visual enclosure for neighbourhood street walkability and elders' mental health in China: Using street view images. *J. Transp. Health* **2019**, *13*, 90–102. [CrossRef]
35. Park, K.; Ewing, R.; Sabouri, S.; Larsen, J. Street life and the built environment in an auto-oriented US region. *Cities* **2019**, *88*, 243–251. [CrossRef]
36. Owens, P.M. Neighborhood form and pedestrian life: Taking a closer look. *Landsc. Urban Plan.* **1993**, *26*, 115–135. [CrossRef]

37. Yin, C.; Yao, X.; Sun, B. Population density and obesity in rural China: Mediation effects of car ownership. *Transp. Res. Part D Transp. Environ.* **2022**, *105*, 103228. [CrossRef]
38. Ji, H.; Qing, L.; Han, L.; Wang, Z.; Cheng, Y.; Peng, Y. A new data-enabled intelligence framework for evaluating urban space perception. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 400. [CrossRef]
39. Ma, X.; Ma, C.; Wu, C.; Xi, Y.; Yang, R.; Peng, N.; Zhang, C.; Ren, F. Measuring human perceptions of streetscapes to better inform urban renewal: A perspective of scene semantic parsing. *Cities* **2021**, *110*, 103086. [CrossRef]
40. Herbrich, R.; Minka, T.; Graepel, T. TrueSkill™: A Bayesian Skill Rating System. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2006; Volume 19, pp. 569–576.
41. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems 34*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 12077–12090.
42. Moran, P.A.P. Notes on continuous stochastic phenomena. *Biometrika* **1950**, *37*, 17–23. [CrossRef]
43. Anselin, L. Local indicators of spatial association—LISA. *Geogr. Anal.* **1995**, *27*, 93–115. [CrossRef]
44. Liang, J.; Gong, J.; Sun, J.; Zhou, J.; Li, W.; Li, Y.; Liu, J.; Shen, S. Automatic Sky View Factor estimation from street view photographs: A big data approach. *Remote Sens.* **2017**, *9*, 411. [CrossRef]
45. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
46. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **2017**, *40*, 913–929. [CrossRef]
47. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef] [PubMed]
48. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.