Landscape Architecture

ARTICLE

Evidence-Calibrated Urban Scene Quality Index for High-Resolution Overhead Imagery in Changsha

Kongjian Yu[1,*] and Wang Li[2]

[1] College of Architecture and Landscape Architecture, Peking University, Beijing 100871, China
[2] Turenscape Urban Planning and Design Co., Ltd., Beijing 100080, China

[*] Correspondence: kjyu@urban.pku.edu.cn

Abstract

High-resolution overhead imagery captures building arrangement, street structure, vegetation cover, industrial land use, water surfaces, and outdoor activity space in a resolution appropriate for neighborhood analysis. However, semantic classification of such imagery is reduced to an ordinal score by counting favorable criteria, despite the difference in recognition reliability between visual cues and the possibility of moving from one class to another by changing one indicator alone. An Evidence-Calibrated Urban Scene Quality Index (ECUSQI) was devised to convert five semantically identifiable visual indicators into a reliability-based five-point scale. We analyzed a test set of 3038 labeled samples of urban imagery patches extracted from central Changsha (China) among 3874 RGB patches of 250 × 250 pixels with 0.5 m ground sampling distance (GSD). Recognition reliability of 615 test scenes with respect to open building layout, grid-like street structure, vegetation coverage, lack of industrial areas, and presence of activity space was estimated in the training procedure. Jeffreys smoothing of recognition reliability normalizes each indicator increase, the posterior uncertainty component identifies scores based on less reliable semantic information, and the threshold margin term points out classes determined by a less confident inference. The test dataset includes 527 scenes with five correct decisions, 77 with four, 9 with three, and 2 with two, implying an average of 4.836 correct and 0.164 incorrect indicator interpretations per scene. The accuracies of indicators range from 93.17% for buildings to 98.86% for industrial areas. Co-attention reduces the expectation of indicator mistakes by 58.0% and decreases the multi-mistake probability by 8.48% to 1.79%. In the spatial interpretation, ECUSQI is lower in more densely populated districts of older construction and higher in green residential areas with open structure and activity space. Our index serves a concrete measuring purpose, since reliable overhead semantic information can inform fine-grained environmental evaluation, along with its threshold sensitivity.

Keywords: high-resolution remote sensing; urban environmental quality; semantic scene description; reliability calibration; Changsha; overhead imagery; neighbourhood assessment

1. Introduction

Urban environmental conditions involve observable combinations of density, connectivity, greenness, exposure, and access to outdoor activities. Built form defines openness and microclimate; street network characterizes permeability and mobility; greenness enables comfort and restoration; industry may detract from visual amenity

Received: 28 August 2022; Revised: 09 December 2022; Accepted: 30 January 2023; Published: 30 March 2023.

doi: 10.70517/la20231131-02

and functionality; and rivers, lakes, play spaces, and forests offer recreational settings. Such conditions are not distributed uniformly across cities. Changes in environmental characteristics tend to occur along parcel borders, riversides, construction sites, industrial areas, and contrasts between old and new residential neighborhoods. Detailed measurement is thus needed for local greening, priority renewal, improved accessibility, and environmental inspections [10, 11, 14, 15, 24, 26].

Existing methods for urban assessment involve field audits, surveys, administrative measures, GIS layers, and environmental monitors. Every evidence channel provides a distinct perspective on the urban environment. Perceptions and satisfaction levels can be captured through surveys, but survey administration requires substantial costs and effort. Field audits can observe quality factors that are missed by automated image-based assessments, but auditors need to be specially trained and spend considerable time in the field. Administrative measures are easy to compare over time but usually describe neighborhoods, not particular urban scenes. GIS layers enable the quantification of parcel composition, street networks, and urban form, but this requires up-to-date maps and standardized classification systems. Image-based assessment has become popular as it allows repeatable observation at the level of parcels and streets [16, 20, 21, 37, 40].

Street views have improved the ability to measure pedestrian environments, building facades, enclosure, vegetation, and other human-scale factors. In street-view images, however, visibility is limited for assessing internal features of large blocks, compound-style housing, riverside parcels, construction zones, and inter-industrial zones. The planimetric organization of these features can be efficiently observed using high-resolution overhead images in which an urban scene is captured in one image unit at sub-metre resolution.

Object detection, land use interpretation, scene classification, high-resolution image captioning, and deep feature extraction are advanced skills in remote sensing [4, 5, 13, 22, 28, 32, 35]. However, built environment quality measurement does not solely depend on the presence or absence of particular objects. Planning relevance results from a combination of objects. An open residential block with strong vegetation cover and high road permeability is a different urban environment from a compact residential block with poor vegetation cover and low road permeability. The description of an urban scene in planning language is desirable as it translates visual observations into concepts that can be measured.

A rating system is the next step after an urban scene is semantically described. An index with equal scores for every indicator and an ordinal ranking scheme is convenient as it can be intuitively interpreted. However, an equal-sum rating system assumes that each semantic recognition is equally reliable. As demonstrated by the Changsha test dataset, the reliability of the five semantic indicators varies. Recognition is more difficult for building arrangements than industrial sites, but road structure, vegetation, and activity space recognition is relatively easier. Treating every class of evidence as equally certain may overstate the precision of urban environmental conditions in some cases.

Classification thresholds constitute another issue. A five-point quality index is often evaluated using a five-level ordering scheme such as critical, constrained, transitional, adequate, and favorable. Scores such as 3.98 and 4.04 may belong to different colour classes, despite being close to each other. If the class labels are based on semantic recognition, these labels should be considered together with their associated threshold margin.

This paper addresses a practical question: how do the semantic indicators of five urban conditions observed in high-resolution overhead images of central urban scenes form an ordinal quality index that reflects their reliability and threshold sensitivity? This paper proposes the Evidence-Calibrated Urban Scene Quality Index. It retains the five categories of urban conditions in the Changsha urban corpus but uses evidence-reliability normalized increments. The index also generates posterior uncertainties and nearest threshold distances. The result is a quality score with semantic details, posterior evidence uncertainty, and threshold sensitivity for each category.

The empirical context is central Changsha, China, a city divided by the Xiang River. The eastern bank of the city comprises older and denser urban fabrics. On the western bank, newer urban fabrics with less dense housing and greater vegetation cover can be found. This contrast allows an urban scene index to be tested on various scenarios, including compact old urban districts, newly emerging urban compounds, construction lands, riverside lands, high-grade housing estates, industrial areas, rivers, lakes, playgrounds, and forested areas.

Overhead urban scene images are acquired at 0.5 m resolution. Each image patch has a 250-by-250-pixel size. A key consideration of the measurement design is image-patch size. Smaller patches, such as 125-by-125-pixel patches, cannot contain sufficient road or vegetation context. Larger patches, such as 500-by-500-pixel patches, may cover diverse environments. Patches with 250-by-250 pixels allow for the interpretation of the five semantic topics simultaneously.

At first, 3874 image patches were generated. Then, 836 patches dominated by water and woodland, or patches without relevant urban building areas, or patches with buildings outside the urban fabric were eliminated. In the end, the corpus contains 3038 labeled urban scenes. Ten seniors were recruited from a GIS program to annotate these scenes using sketch and sentence constructions. The vocabulary consisted of building area, industrial area, trees, roadways, playgrounds, rivers, lakes, and relation words like around, near, adjacent, compact, and open. The model was evaluated using an 80%-20% split and obtained 615 test scenes. The image-captioning configuration comprised ResNet152 visual features (14-by-14-by-512), 512-dimensional semantic embeddings, sentence LSTM and word LSTM hidden states (both 512), a learning rate of 0.001, the Adam optimizer, a tag loss weight of 0.5, and a word loss weight of 2 [6].

The current corpus is applicable for assessing semantic environmental conditions due to the nature of labeled patches. Each patch contains information about the physical environment of a neighborhood scene that is neither pure object-level nor district-level data.

(a) Regional setting.

(b) Urban core.

(c) Patch grid.

(d) Scene tiles.

Figure 1. Study area and scene generation.

In Figure 1, the visual flow starts from the map-to-scene generation process. Firstly, the regional locator introduces the context of Changsha in Hunan province, then urban overview presents the river-separated nature of the city, the grid panel explains how overhead images are split into grid units for local observation, while the tile panel

demonstrates the diversity of the selected urban scenes. Thus, the adoption of the 250 × 250 pixel unit makes sense because the patch captures built forms and the environmental components nearby.

1.1. Semantic indicators

Five binary indicators are selected since they are observable through high-resolution overhead imagery and also related to urban environmental perception. First, the openness of the building layout is determined by whether built forms in the scene belong to low-density open structures rather than high-density structures. Secondly, grid road structure describes whether local street patterns facilitate connectivity of the built environment in grids instead of being tree-like patterns. Thirdly, vegetation coverage measures whether vegetation covers exist in the scene, fourthly, industrial absence refers to the lack of industrial lands disturbing the residential area visually. Finally, activity-space presence reflects whether forests, lakes, rivers, or playgrounds appear in the scene.

Table 1. Urban-scene indicators.

Indicator	Favourable state coded as 1	Unfavourable state coded as 0
Building layout	Low density and open arrangement	High density and compact arrangement
Road structure	Grid-like connected pattern	Tree-like or fragmented pattern
Vegetation coverage	Visible vegetation around the building area	No visible vegetation around the building area
Industrial absence	No industrial area in the scene	Industrial area visible in the scene
Activity-space presence	Forest, lake, river, or playground visible	No activity-space element visible

(a) Building layout.

(b) Road structure.

(c) Vegetation.

(d) Industrial land.

(e) Activity space.

Figure 2. Semantic indicators of scene quality.

The coding in Table 1 gives the score a consistent positive direction. A higher value indicates a greater number of favourable visible conditions, not a judgement about the residents or social value of a neighbourhood. The industrial indicator is intentionally reversed because the environmentally favourable state is absence of industrial disturbance. This coding allows the final score to be decomposed into specific visual reasons: compactness, weak road structure, missing vegetation, industrial adjacency, or lack of activity space.

For scene s, the binary evidence vector is

$$\mathbf{x}_s = \left(x_{B,s}, x_{R,s}, x_{V,s}, x_{I,s}, x_{A,s}\right), \qquad x_{j,s} \in \{0, 1\}, \tag{1}$$

where B, R, V, I, and A denote building layout, road structure, vegetation coverage, industrial absence, and activity-space presence. This vector is intentionally simple. Its strength lies in auditability: a low value can be traced to the particular visible conditions that are missing, and a high value can be interpreted as the joint presence of open layout, connected roads, vegetation, industrial absence, and activity space.

The paired overhead examples in Figure 2 show how the binary coding in Table 1 is read visually. The favourable state is not an abstract category: it corresponds to visible open building arrangement, connected road geometry, tree or green cover, absence of industrial land, and nearby activity-space elements. The unfavourable state is similarly traceable to compact construction, fragmented access, bare surroundings, industrial adjacency, or absence of outdoor activity space.

2. Evidence-calibrated scoring

2.1. Reliability weighting of semantic decisions

Let n_j be the number of correct predictions for indicator j in the 615-scene test set. Recognition reliability is estimated with Jeffreys smoothing:

$$\widehat{r}_j = \frac{n_j + 0.5}{616}. \tag{2}$$

The 0.5 adjustment prevents the finite test-set proportion from being treated as absolute certainty. This is appropriate for urban-scene interpretation because recognition performance is an empirical estimate affected by image complexity, shadows, tree cover, mixed building types, and annotation boundaries.

The normalized reliability weight is

$$\omega_j = \frac{\widehat{r}_j^{\gamma}}{\sum_k \widehat{r}_k^{\gamma}}, \tag{3}$$

where γ controls how strongly reliability differences separate the indicators. The Changsha analysis uses $\gamma = 1$ because all five accuracies exceed 93%; stronger separation would exaggerate modest empirical differences. The scene score is then

$$S_s = 5 \sum_j \omega_j x_{j,s}, \qquad 0 \le S_s \le 5. \tag{4}$$

The factor of five preserves the familiar ordinal range. A scene with all favourable conditions scores 5, and a scene with none scores 0. Intermediate values remain close to ordinary one-point increments, but they now reflect the observed reliability of the semantic evidence.

Equations (2)–(4) calibrate evidence, not social preference. Industrial absence receives the largest one-indicator increment because it is the most reliably recognized topic in the Changsha test set, not because industrial absence is declared more important than greenery or road connectivity in every planning context. The index therefore separates visual recognition stability from policy value.

2.2. Uncertainty and threshold margin

For each indicator, reliability is treated as beta distributed with parameters

$$\alpha_j = n_j + 0.5, \qquad \beta_j = 615 - n_j + 0.5. \tag{5}$$

The reliability variance is

$$\mathrm{Var}(r_j) = \frac{\alpha_j \beta_j}{(\alpha_j + \beta_j)^2(\alpha_j + \beta_j + 1)}. \tag{6}$$

A first-order approximation to score uncertainty is

$$\sigma_s = 5\left(\sum_j \omega_j^2 x_{j,s}^2 \mathrm{Var}(r_j)\right)^{1/2}. \tag{7}$$

The uncertainty value indicates how strongly the score depends on indicators with less stable recognition. It is not a substitute for field validation, resident perception, or environmental monitoring. Its purpose is narrower: to identify scene scores that should be read with greater caution because their supporting semantic evidence is less stable.

Ordinal interpretation uses thresholds at 1, 2, 3, and 4. The nearest-threshold distance is

$$\Delta_s = \min_{\tau \in \{1,2,3,4\}} |S_s - \tau|. \tag{8}$$

The threshold caution score is

$$C_s = \exp\left(-\frac{\Delta_s}{\sigma_s + \epsilon}\right), \tag{9}$$

where ϵ prevents division by zero. A high C_s means that the score lies close to a class threshold relative to its uncertainty. This term is useful for review because a score near 4 can appear favourable on a map while remaining sensitive to a one-topic recognition error.

(a) Scene evidence.

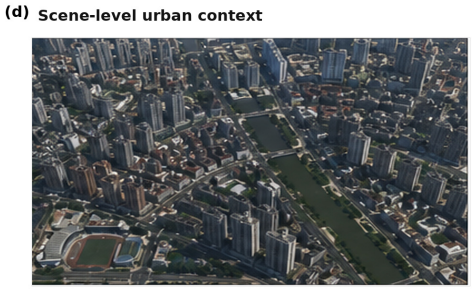
(b) Indicator reliability

Indicator	Accuracy (%)	Score increment
Building density	93.17	0.963
Road connectivity	97.72	1.010
Vegetation coverage	97.07	1.004
Industrial absence	98.86	1.022
Activity space	96.75	1.000

Test scenes: 615

(b) Reliability values.

(c) Score products.



(d) Scene output.

(e) Ordinal examples.

Figure 3. Evidence-calibrated scoring components.

2.3. Confidence-weighted spatial interpolation

Scene scores are attached to patch centres. Because a crop boundary can exclude a nearby playground, industrial building, lake edge, or green belt, isolated patch scores can be sensitive to location. A shifted grid can generate overlapping scene samples, and inverse-distance interpolation can combine nearby values. The confidence-weighted surface is

$$\widehat{S}(u) = \frac{\sum_{s \in N(u)} (d(u, c_s) + \epsilon)^{-p} (\sigma_s + \epsilon)^{-1} (1 - C_s + \epsilon) S_s}{\sum_{s \in N(u)} (d(u, c_s) + \epsilon)^{-p} (\sigma_s + \epsilon)^{-1} (1 - C_s + \epsilon)}, \tag{10}$$

where u is a map location, c_s is a patch centre, d is Euclidean distance, $N(u)$ is the neighbourhood of nearby scenes, and p is the distance-decay exponent. The search radius is 125 pixels, equal to half the patch width.

The interpolation in Equation (10) gives greatest influence to nearby, low-uncertainty, threshold-stable scenes. It does not claim exact environmental measurement at every point. Instead, it summarises overlapping semantic evidence in a way that reduces crop-boundary sensitivity while preserving caution around uncertain class assignments.

The components in Figure 3 connect the equations to the observable image material. The index begins with five scene topics, converts their test-set reliability into calibrated increments, and then produces score, uncertainty, and smoothed spatial outputs. The ordinal examples also show that the numerical classes are grounded in recognizable urban conditions rather than in an uninterpretable model output.

3. Results

3.1. Test-set semantic consistency

The 615-scene test set shows that semantic descriptions are sufficiently stable for ordinal assessment, while still requiring reliability reporting. The distribution of correct topics is given in Table 2. Five correct descriptions occur in 527 scenes, four in 77 scenes, three in 9 scenes, and two in 2 scenes. No scene has only one correct topic or zero correct topics. The implied error total is 101 indicator errors, giving 0.164 errors per scene and 4.836 correct semantic decisions per scene.

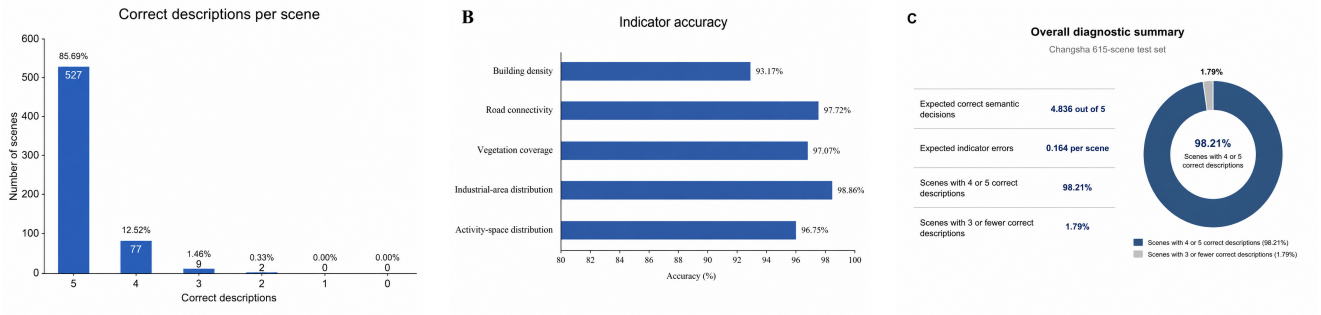
Table 2. Semantic correctness.

Correct topics	Scenes	Proportion (%)	Errors per scene	Error total
5	527	85.69	0	0
4	77	12.52	1	77
3	9	1.46	2	18
2	2	0.33	3	6
1	0	0.00	4	0
0	0	0.00	5	0
Total	615	100.00	–	101

The values in Table 2 indicate that complete semantic failure is absent in the test set. Most scenes have either no error or a single-topic error. This supports the use of scene-level descriptions for assessment because the score is not dominated by frequent multi-topic misreading. The same distribution also explains why caution remains necessary. A single error may still move a score across a class boundary when only one favourable indicator separates two ordinal categories.

The reliability results provided by the performance panels in Figure 4 are twofold. On the scene level, stability concerns five fully correct topics, hence, the consistency of completely described scenes for most of the test scenes. At the indicator level, stability coefficients of all topics are quite different, from the lowest one for building layout

and the highest one for industrial area identification. The reliability diagnostic summary is then translated into the error burden used in score analysis.



(a) Scene-level correctness. **(b)** Indicator accuracy. **(c)** Diagnostic summary.

Figure 4. Semantic recognition performance.

The profile plot in Figure 4 indicates a highly right-skewed distribution of error counts in favor of fully correct descriptions. The low-count end has only 1.79% of scenes with three or fewer correct topics. This finding is critical in developing interpretation plans, since multiclass mistakes are the situations most prone to lead to false class assignments. Therefore, examination effort can be focused on those rare tails and boundary scores in the middle of the distribution rather than on the whole population of locations analyzed.

3.2. Indicator-level reliability

Reliability of recognition differs for all semantic indicators included in the test scenes. Building layout has been correctly identified 573 times out of 615 total decisions, hence 93.17% accuracy. Road structure has been correctly recognized 601 times, hence, 97.72% accuracy. Vegetation cover has been correctly recognized 597 times, hence, 97.07% accuracy. Industrial area has been identified in its absence correctly 608 times, hence, 98.86% accuracy. Presence of activity space has been recognized correctly 595 times, hence, 96.75% accuracy.

Table 3. Calibrated increments.

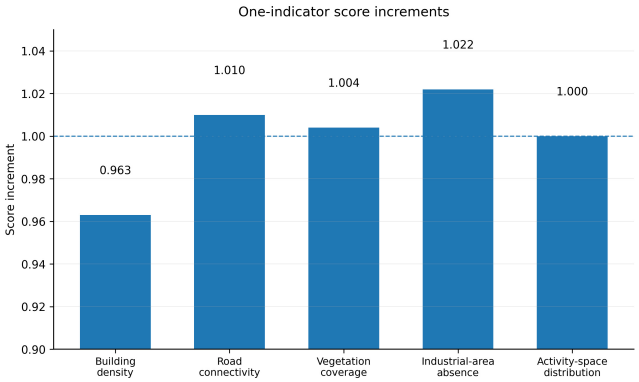
Indicator	Correct decisions	Accuracy (%)	Weight	Increment
Building layout	573	93.17	0.1927	0.963
Road structure	601	97.72	0.2021	1.010
Vegetation coverage	597	97.07	0.2007	1.004
Industrial absence	608	98.86	0.2044	1.022
Activity-space presence	595	96.75	0.2001	1.000

The increments in Table 3 show that the calibration is conservative. Equal counting would assign one point to every favourable condition. The calibrated index assigns slightly less than one point to building layout because that topic is the least reliable in the test set, and slightly more than one point to industrial absence because it is the most reliable. The largest difference between increments is 0.059, so the index preserves the familiar meaning of a five-condition score while correcting boundary cases.

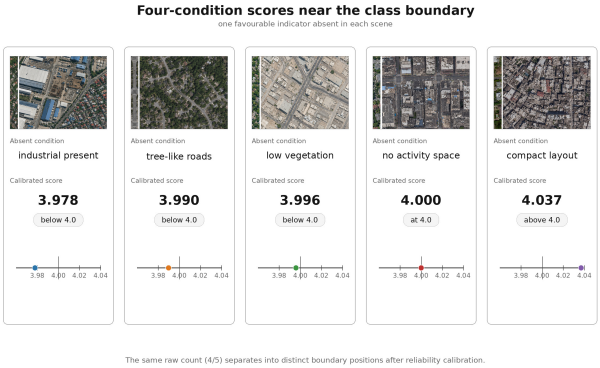
The calibrated cases illustrated in Figure 5 show why even a slight difference between increments should be taken into account. Four conditions can have the same amount of favourable indicators without exceeding the class-4 threshold since the missing topic has a different level of empirical reliability. Thus, the illustration shows that the score is not just a measure of the number of favourable elements; rather, it is a measure of the number adjusted for their stability.

As can be seen in Figure 5, the calibrated increments have actual meaning only around the thresholds. Four-condition scenes missing industrial absence are awarded 3.978, while four-condition scenes missing building openness get

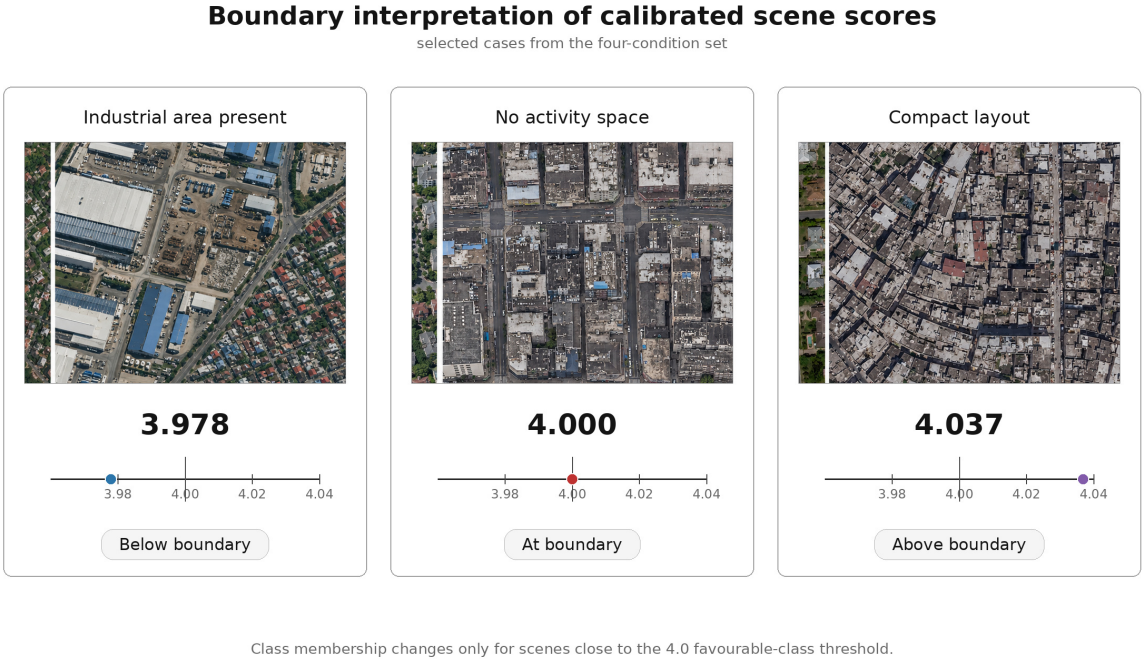
4.037. According to the non-calibrated increments, both cases would receive 4. Calibrated increments show that in the first case, a less stable favourable indicator is missing, while in the second case, a more stable one is missed. The difference is slight, but it keeps the evidence quality visible even within the class label.



(a) Indicator increments.



(b) Boundary cases.



(c) Threshold examples.

Figure 5. Ordinal effect of reliability calibration.

3.3. Effect of co-attention on stability of semantic description

Co-attention contributes to the stability of the process of describing a scene semantically by associating visual features with semantic cues prior to generating sentences. The proportion of scenes that get all five descriptions correctly with the use of co-attention is 85.69%; without co-attention, this number is only 74.32%. The expected number of correct topics decreases from 4.836 to 4.609.

The diagnostic measures in Table 4 are relevant due to non-proportionality of class error to average accuracy. A few scenes with multiple mistakes can affect the result more strongly than many scenes with single mistakes. By decreasing the high-error tail proportion from 8.48% to 1.79%, the co-attention scheme prevents that effect. From the map user's perspective, the mapped pattern will be more accurate with lower probability of being misshaped by severely misdescribed scenes.

The analysis in Table 4 confirms that decreasing the error tail proportion makes sense even with high average accuracy. The latter can generate misleading scene score assessments, if there is a relatively small number of scenes with multiple mistakes. The co-attention scheme exhibits a higher proportion of error-free and lower proportion of high-error scenes, compared with equal counting. It thus justifies itself for environmental mapping as the latter depends on the correctness of the five topics jointly.

Table 4. Co-attention stability.

Diagnostic	Co-attention	No co-attention	Difference
Five correct topics (%)	85.69	74.32	11.37 percentage points
Expected correct topics	4.836	4.609	0.227 topics
Expected indicator errors	0.164	0.391	58.0% lower with co-attention
High-error tail (%)	1.79	8.48	78.9% lower with co-attention
Error-free to high-error ratio	47.87	8.76	5.46 times larger

3.4. Score behavior under various combinations of the five topics

In the calibrated score function, each scene retains its endpoint values. The highest is 5 for a scene with open building layout, grid roads, vegetation, no industrial disturbance, and an activity space. The lowest is 0 for a scene with compact building layout, tree-like roads, no vegetation, industrial disturbance, and no activity space. The meaningful variations are those of the intermediate score values.

A scene with grid roads, vegetation, industrial absence, and activity space but compact building layout is scored as 4.037. A scene with open building layout, grid roads, vegetation, and activity space but industrial disturbance is scored as 3.978. These scenes would be equal under equal-counting approach, since only one topic varies here. However, since these topics' recognition reliability differs, the two scenes cannot be considered identical in terms of calibration. At the other extreme, the scene with industrial absence favorable gets 1.022, while that with open layout favorable gets 0.963. Neither scene can be regarded adequate from environmental perspective, yet their supporting evidence differs.

Special attention should be paid to four-condition scenes, where the difference between unfavorable and favorable classes is small. The first scene's score of 3.978 is merely 0.022 away from the favorable threshold of 4, while the second scene's score of 4.037 is only 0.037 above it. Thus, the class labels of both should be understood ambiguously, rather than unambiguously. For mapping purposes, the score layer should be complemented by a caution layer that highlights the threshold vicinity.

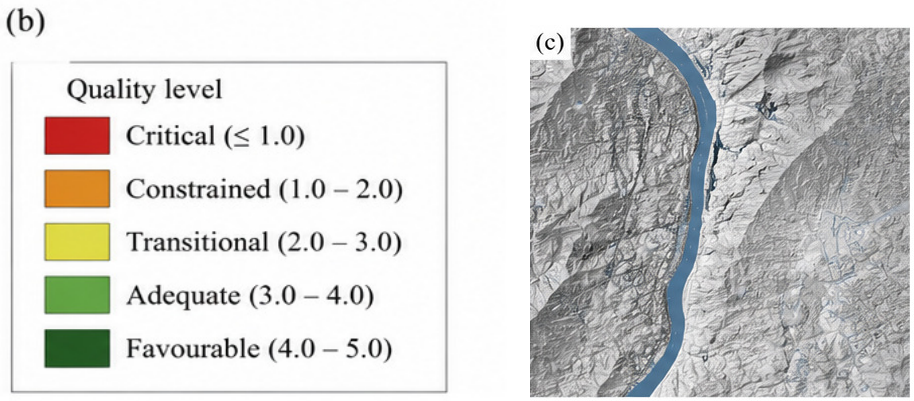
3.5. Distribution of scores in Changsha

The spatial distribution of the assessed score surface demonstrates a stark contrast between compact and green districts in Changsha. Lower values occur in the eastern part, where the scenes have compact building layout, poor road structure, lack of vegetation, and insufficient presence of activity space. Higher values appear in the western part of the city, where the scenes are characterized by open layout, good vegetation cover, good road structure, and sufficient activity space.

The city-scale quality surface in Figure 6 illustrates the east-west difference reported above. Low scores tend to locate in areas characterized by compact building layout, sparse vegetation, and lower permeability, while high scores tend to emerge in relatively green and open residential environments, particularly with identifiable activity spaces and internal road network connectivity. The supplementary images illustrate each score class without substituting the numerical map.

From the local scale perspective, the score proves sensitive to variation within a short distance. In cases where vegetation and activity spaces exist, scores will be high, while in neighboring areas with dense buildings and sparse vegetation, scores will be low. The new residential compound performs well if greenness, openness, and internal

(a) Quality surface.



(b) Legend and river context.

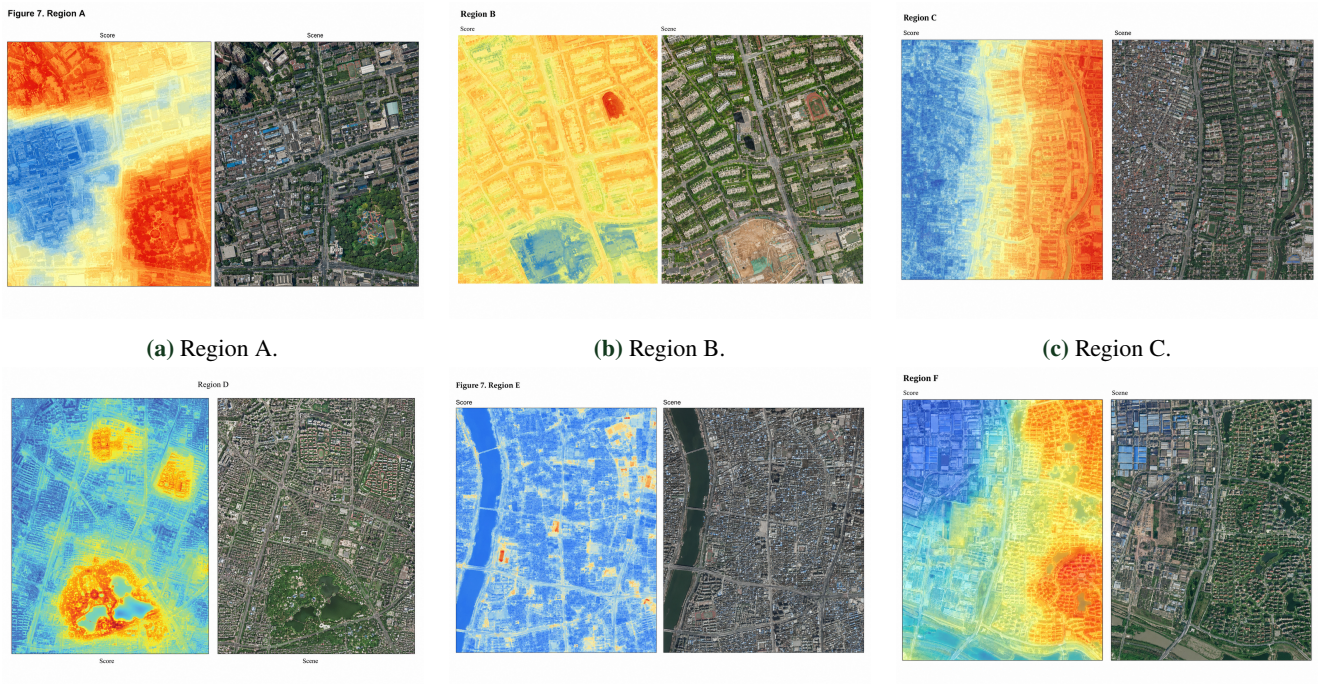
(c) Dense bank view.

(d) Green residential fabric.

(e) Urban greenway.

Figure 6. Changsha quality surface.

road network connectivity coexist. Cool patches can be found at construction sites with open surface and sparse vegetation. Local variations are also found at areas surrounding lakes, industries, and residential compounds of high grade.



(a) Region A. **(b)** Region B. **(c)** Region C.

(d) Region D. **(e)** Region E. **(f)** Region F.

Figure 7. Local diagnostic examples.

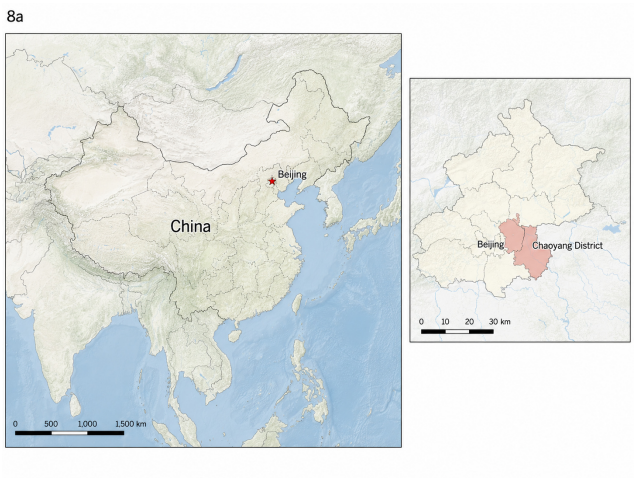
The local windows in Figure 7 prove that variability in scores is not just an effect of districts in general. Warm regions include areas with visible open residential plan, tree canopy cover, playgrounds, lakes, and rivers, while cold regions feature high-density building block, barren construction soil, industrial areas, and poor quality of activity spaces. This paired arrangement of scenes and scores makes it easy to see the class label in action at the local level.

3.6. Semantic Portability from Changsha to Chaoyang

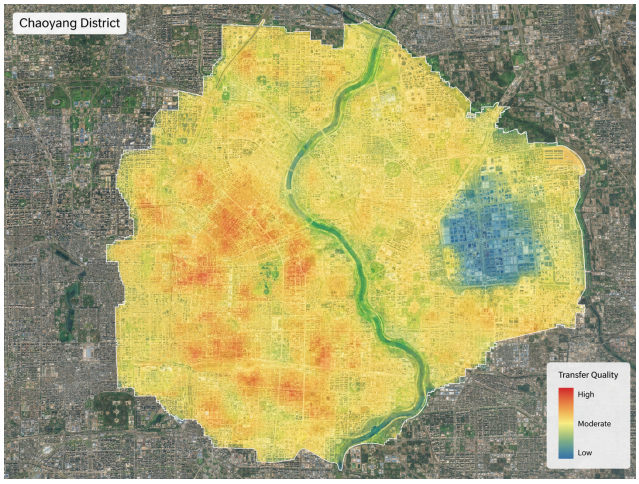
The same vocabulary of semantics is used to interpret the urban streetscape of Chaoyang District, an area in Beijing of 92 km^2. The transfer results suggest that the five indicators can still be interpreted in Chaoyang, outside of Changsha. In one local area, yellow values prevail, while orange and red values aggregate around areas with good vegetation cover and open residential plan form. In another part of the scene, warm values coincide with rivers, good vegetation cover, open residential plan form, and a playground, while the cold region corresponds to an industrial area on the right-hand side of the image.

This result demonstrates semantic portability rather than universal calibration. Buildings, roads, vegetation, industrial sites, water bodies, playgrounds, forest patches can be interpreted in both cities, but different materials for roofs, types of trees, architectural designs, and industrial patterns might change recognition accuracy. The five indicators are therefore portable across cities, while the weights have to be recomputed for each new location.

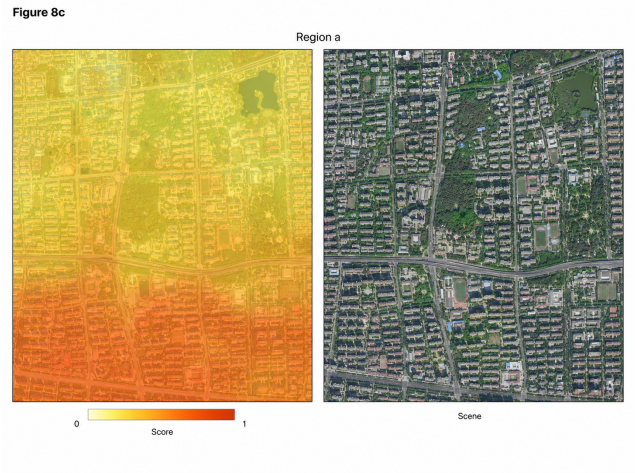
The transfer panels in Figure 8 illustrate how the same visible topics are read within an alternative urban fabric. Open residential space, waterways, vegetation, and playgrounds are associated with positive values, while industrial land use and grey structures are negatively weighted. The result confirms successful transfer of the semantic vocabulary while reinforcing the need to recalibrate reliability weights.



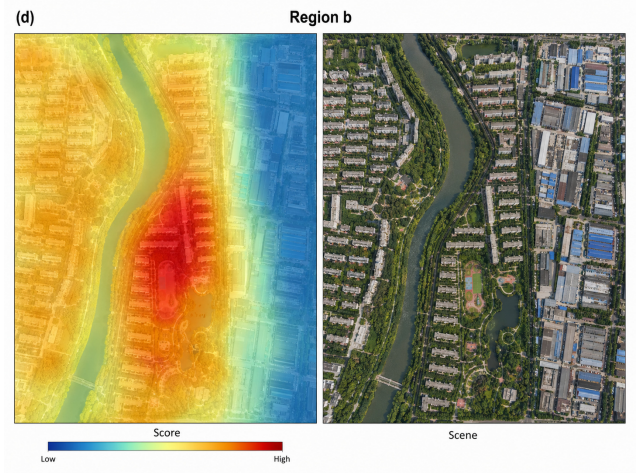
(a) Beijing locator.



(b) Chaoyang surface.



(c) Local region a.



(d) Local region b.

Figure 8. Transfer test in Chaoyang.

4. Discussion

4.1. Interpretation of the calibrated score

The Evidence-Calibrated Urban Scene Quality Index translates overhead semantic categories into scores that remain understandable at the planning scale. The key feature of the index is not numerical complexity. It is rather that every score can be broken down into five visible conditions: building layout, road pattern, vegetation, absence of industry, and presence of activity space. This characteristic makes the index easier to audit than a latent environmental value calculated using uninterpretable features.

Reliability calibration modifies the interpretation of the score without altering its practical scale. Every binary indicator contributes around one point, but there is no presumption of equal reliability among all topics. Building layout gets the lowest increment because of difficulty in identifying this factor in the test set, while industrial absence gets the highest increment due to stability. This kind of tuning works well for practical use because it affects borderline cases and preserves the communicability of the five-point scale.

The index should be interpreted as overhead visible evidence of the environment rather than comprehensive measures of urban quality. In contrast to income, housing affordability, crime rate, social integration, noise level, air quality, accessibility of services, sense of belonging, or subjective satisfaction, it evaluates observable physical aspects that contribute to urban environmental interpretation. A comprehensive analysis of urban quality will have to integrate visual evidence with data collected through field observation and survey of residents, street-level

imagery, sensors, and administration.

4.2. Importance of uncertainty for urban planning

Uncertainty is usually left out of image-based assessments, even when such assessments are based on an imperfect recognition task. The results of this study confirm that uncertainty is essential in assessing urban environments. The overall accuracy of semantic profile is high, but the building layout topic is less reliable than other indicators, and scores near the threshold are prone to switching. Communicating uncertainty avoids overconfidence of an ordinal map.

Threshold caution has direct implications for practical urban planning. Locations that lie on the boundary between adequate and favourable urban classes deserve to be distinguished from locations that lie clearly within any class. For example, a location with a 3.98 score is not automatically a scene with poor quality, nor is 4.04 a safe scene. In either case, it is important to take posterior uncertainty into account to interpret the scores carefully. This step is crucial whenever a mapping exercise influences decision making, site inspection, and urban investment.

4.3. Spatial diagnosis across urban fabrics

The Changsha findings reveal both broad and local spatial patterns. On the one hand, the old east bank neighbourhoods score lower than the newer west bank residential areas. Their poorer performance is linked to denser built-up environment, less greenspace, and greater impermeability. On the other hand, the presence of construction land, lakes, playgrounds, industrial peripheries, and high-grade residential enclaves produces clear differences in spatial score distribution.

Spatial dynamics of this sort are valuable because urban interventions tend to be locally defined. Planting trees requires the knowledge of areas lacking vegetation. Improving a network of streets depends on the ability to detect poorly connected urban fabrics. Inspection and regulation of industrial land use need the knowledge of nearby residential scenes. Planning activities spaces requires detecting areas missing river banks, lakes, forests, playgrounds, and similar sites. The index provides diagnosis for all these cases rather than just a ranking.

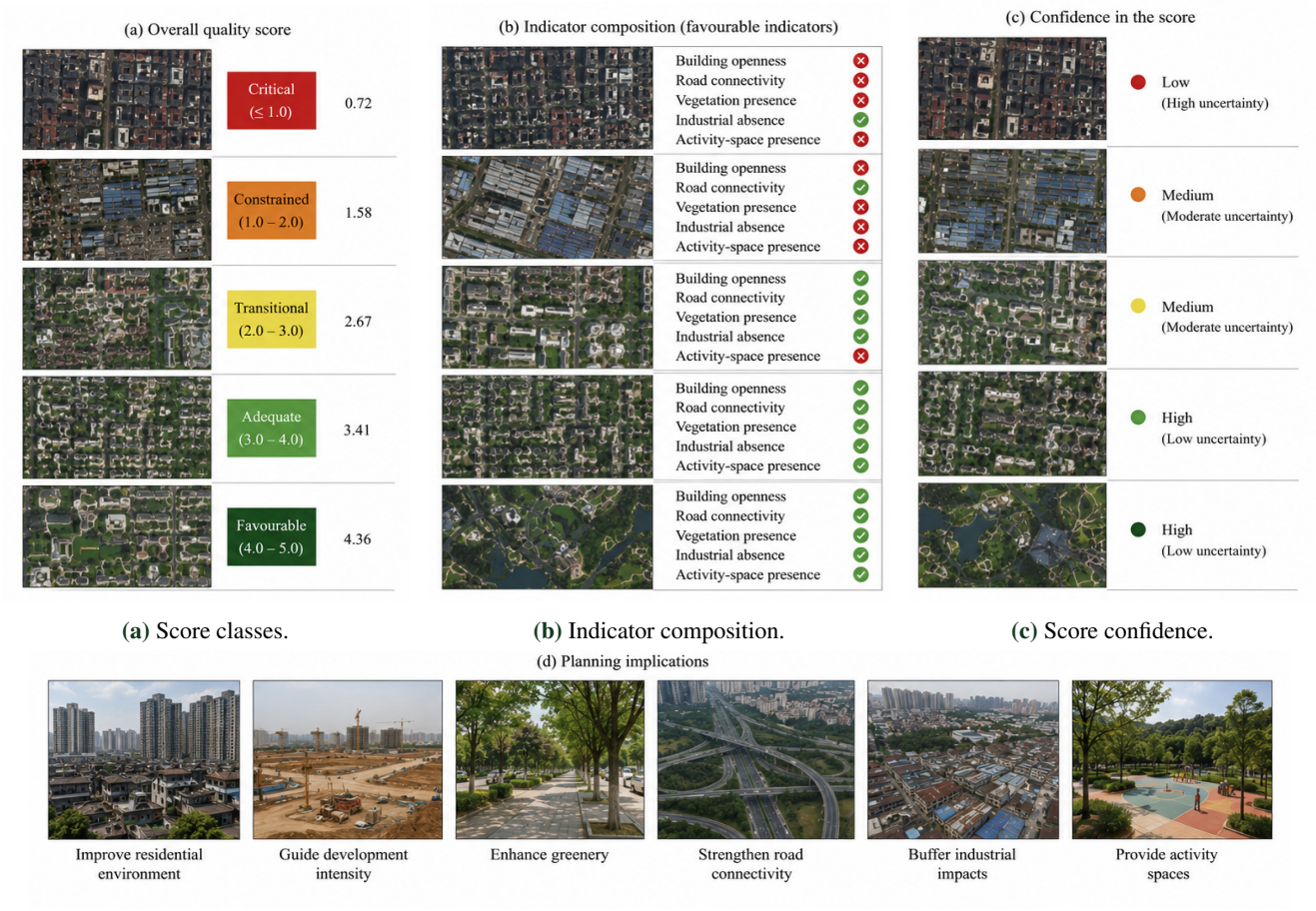
4.4. Limitations and future validation

This work uses just five binary indicators to define urban scenes. Binary coding makes results easier to read, but it eliminates variation in the continuous characteristics of the environment. Sparse trees and dense forest may be equally classified as presence of vegetation, and the mixed road network might be forced to fall into a single binary class. Future annotation will allow the introduction of ordinal and probabilistic states for canopy density, road permeability, industrial proximity, and activity spaces.

Annotation relies on training of GIS students. This approach helps maintain consistent object recognition but excludes all other perceptions from consideration. People's experience of environmental quality is known to depend on age group, physical condition, everyday route, familiarity with the area, and cultural background. Thus, overhead semantic scores must be verified in practice with field observations and residents' views in order to make proper policy decisions.

Operational use of the index requires the storage of patch centres, binary indicator vectors, calibrated scores, posterior uncertainty, and caution thresholds. In this case, an algorithm will be able to produce three kinds of output layers: a score map, a missing-indicator map, and a threshold caution map. These outputs will help planners distinguish low scores due to missing vegetation from those related to industrial proximity and poor road structure.

The decision-oriented reading in Figure 9 explains how the mapped score can be employed. A low score reveals the presence of the need for inspection or intervention, the composition panel determines the urban conditions that generate that score, and the confidence panel informs whether the assigned class can be considered stable. Planning activities can be matched to the recognized deficit, like greening, road-structure improvement, industrial buffer, and activity space creation.



(a) Score classes. **(b)** Indicator composition. **(c)** Score confidence.

(d) Planning interpretation.

Figure 9. Decision-oriented reading of score, composition, and confidence.

5. Conclusion

This paper explored whether five observable semantic indicators generated by high-resolution overhead imagery could yield a meaningful urban environmental score without masking the problem of recognition uncertainty. The findings in Changsha demonstrate an affirmative response to the question. The ECUSQI combines the interpretability of a score that considers five urban conditions with the reliability normalization of its favourable indicators. It also quantifies posterior uncertainty and threshold caution, so the map can be interpreted as an assessment of the observed scene based on evidence.

The numerical outcomes of the exercise are definite and consistent. Out of 615 test samples, 527 include five valid semantic descriptions, 77 have four, 9 scenes include three, and only 2 scenes have two correct indicators. The expected semantic correctness amounts to 4.836 topics per scene, while the error burden totals 0.164 indicator errors per scene. Indicator accuracy varies between 93.17% of building layout and 98.86% of industrial absence, providing normalized increments of 0.963 and 1.022, respectively. The effect of co-attention amounts to 58.0% of error reduction and shifts the distribution of the latter from 8.48% of the samples having high error burden to 1.79%, ensuring the stability of ordinal maps.

Spatial interpretation offers a practical response to the problem statement of section **??**. Low values are observed in the old dense districts of Changsha characterized by compact building layout, poor vegetation, thin or less connected roads, and few activity space elements. High values are associated with newer residential patterns with open layout, vegetation growth, internally connected roads, and presence of nearby lakes, playgrounds, rivers, and forests. Local analysis demonstrates that changes in the visible scores occur as a result of construction land use, industrial land use, presence of lakes, playgrounds, and residential compound conditions. ECUSQI in Chaoyang demonstrates that

a vocabulary of semantic indicators can be generalized for other cities, although local re-calibration is still required.

The index can thus serve as an auditible diagnostic layer for urban planning. It is able to pinpoint what causes urban conditions: compactness, lack of vegetation, poor connectivity, industrial intrusion, or lack of activity space. The most important feature of ECUSQI lies in making the urban environmental score interpretable by combining the numerical value with the number of favorable indicators, the uncertainty value, and threshold caution. As a result, it renders high-resolution overhead imagery suitable for neighbourhood-scale urban environmental diagnosis, inspection prioritization, greening, road-structure improvement, industrial edge management, and activity space design.

References

[1] Ahn, S., Chung, S. R., Oh, H. J., & Chung, C. Y. (2021). Composite Aerosol Optical Depth Mapping over Northeast Asia from GEO-LEO Satellite Observations. *Remote Sensing, 13*(6), 1096.

[2] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6077-6086).

[3] Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical Analysis, 27*(2), 93-115.

[4] Blaschke, T. (2010). Object based image analysis for remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing, 65*(1), 2-16.

[5] Chen, J., Han, Y., Wan, L., Zhou, X., & Deng, M. (2019). Geospatial relation captioning for high-spatial-resolution images by using an attention-based neural network. *International Journal of Remote Sensing, 40*(16), 6482-6498.

[6] Chen, J., Dai, X., Guo, Y., Zhu, J., Mei, X., Deng, M., & Sun, G. (2023). Urban built environment assessment based on scene understanding of high-resolution remote sensing imagery. *Remote Sensing, 15*(5), 1436.

[7] Cressie, N. (2015). Statistics for spatial data. John Wiley & Sons.

[8] Douglas, O., Russell, P., & Scott, M. (2019). Positive perceptions of green and open space as predictors of neighbourhood quality of life: implications for urban planning across the city region. *Journal of Environmental Planning and Management, 62*(4), 626-646.

[9] Duan, Y., Lei, K., Tong, H., Li, B., Wang, W., & Hou, Q. (2021). Land use characteristics of Xi'an residential blocks based on pedestrian traffic system. *Alexandria Engineering Journal, 60*(1), 15-24.

[10] Ewing, R., & Cervero, R. (2001). Travel and the built environment: a synthesis. *Transportation Research Record, 1780*(1), 87-114.

[11] Ewing, R., & Handy, S. (2009). Measuring the unmeasurable: Urban design qualities related to walkability. *Journal of Urban Design, 14*(1), 65-84.

[12] Ewing, R., & Handy, S. (2009). Measuring the unmeasurable: Urban design qualities related to walkability. *Journal of Urban Design, 14*(1), 65-84.

[13] Gajbhiye, G. O., & Nandedkar, A. V. (2022). Generating the captions for remote sensing images: A spatial-channel attention based memory-guided transformer approach. *Engineering Applications of Artificial Intelligence, 114*, 105076.

[14] Giles-Corti, B., Vernez-Moudon, A., Reis, R., Turrell, G., Dannenberg, A. L., Badland, H., ... & Owen, N. (2016). City planning and population health: a global challenge. *The Lancet, 388*(10062), 2912-2924.

[15] Handy, S. L., Boarnet, M. G., Ewing, R., & Killingsworth, R. E. (2002). How the built environment affects physical activity: views from urban planning. *American Journal of Preventive Medicine, 23*(2), 64-73.

[16] Herold, M., Goldstein, N. C., & Clarke, K. C. (2003). The spatiotemporal form of urban growth: measurement, analysis and modeling. *Remote Sensing of Environment, 86*(3), 286-302.

[17] Hur, M., Nasar, J. L., & Chun, B. (2010). Neighborhood satisfaction, physical and perceived naturalness and openness. *Journal of Environmental Psychology, 30*(1), 52-59.

[18] Jacobs, J. (1992). The death and life of great American cities. Vintage.

[19] Kent, J. L., & Thompson, S. (2014). The three domains of urban planning for health and well-being. *Journal of Planning Literature, 29*(3), 239-256.

[20] Larkin, A., Gu, X., Chen, L., & Hystad, P. (2021). Predicting perceptions of the built environment using GIS, satellite and street view image approaches. *Landscape and Urban Planning, 216*, 104257.

[21] Li, X., Zhang, C., Li, W., Ricard, R., Meng, Q., & Zhang, W. (2015). Assessing street-level urban greenery using Google Street View and a modified green view index. *Urban Forestry & Urban Greening, 14*(3), 675-685.

[22] Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., & Johnson, B. A. (2019). Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing, 152*, 166-177.

[23] Marans, R. W., & Stimson, R. J. (Eds.). (2011). Investigating quality of urban life: Theory, methods, and empirical research (Vol. 45). Springer Science & Business Media.

[24] Mouratidis, K. (2021). Urban planning and quality of life: A review of pathways linking the built environment to subjective well-being. *Cities, 115*, 103229.

[25] Oke, T. R. (1982). The energetic basis of the urban heat island. *Quarterly Journal of the Royal Meteorological Society, 108*(455), 1-24.

[26] Pacione, M. (2003). Urban environmental quality and human wellbeing—a social geographical perspective. *Landscape and Urban Planning, 65*(1-2), 19-30.

[27] Pfeiffer, D., & Cloutier, S. (2016). Planning for happy neighborhoods. *Journal of the American Planning Association, 82*(3), 267-279.

[28] Qu, B., Li, X., Tao, D., & Lu, X. (2016, July). Deep semantic understanding of high resolution remote sensing image. In 2016 International conference on computer, information and telecommunication systems (Cits) (pp. 1-5). IEEE.

[29] Roy, S., Bose, A., Majumder, S., Roy Chowdhury, I., Abdo, H. G., Almohamad, H., & Abdullah Al Dughairi, A. (2022). Evaluating urban environment quality (UEQ) for Class-I Indian city: an integrated RS-GIS based exploratory spatial analysis. *Geocarto International*, 2153932.

[30] Sharifi, A. (2019). Resilient urban forms: A review of literature on streets and street networks. *Building and Environment, 147*, 171-187.

[31] Shepard, D. (1968, January). A two-dimensional interpolation function for irregularly-spaced data. In Proceedings of the 1968 23rd ACM national conference (pp. 517-524).

[32] Shi, Z., & Zou, Z. (2017). Can a machine generate humanlike language descriptions for a remote sensing image?. *IEEE Transactions on Geoscience and Remote Sensing, 55*(6), 3623-3634.

[33] Stewart, I. D., & Oke, T. R. (2012). Local climate zones for urban temperature studies. *Bulletin of the American Meteorological Society, 93*(12), 1879-1900.

[34] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.

[35] Wang, B., Lu, X., Zheng, X., & Li, X. (2019). Semantic descriptions of high-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters, 16*(8), 1274-1278.

[36] Wei, Y. D., Xiao, W., Wen, M., & Wei, R. (2016). Walkability, land use and physical activity. *Sustainability, 8*(1), 65.

[37] Weng, Q. Thermal infrared remote sensing for urban climate and environmental studies: Applications and trends. *ISPRS Journal of Photogrammetry and Remote Sensing, 64* (2009): 335–344.

[38] Wolch, J. R., Byrne, J., & Newell, J. P. (2014). Urban green space, public health, and environmental justice: The challenge of making cities 'just green enough'. *Landscape and Urban Planning, 125*, 234-244.

[39] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057). PMLR.

[40] Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H. H., Lin, H., & Ratti, C. (2018). Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning, 180*, 148-160.