Explainable Artificial Intelligence for Interpreting Human Perception of Urban Streetscapes

Drew Mayo
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
drew1983@colostate.edu

Andrej Langern
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
andrejlangern01@oregonstate.edu

Serge Binner
School of Computing, Clemson University, Clemson, SC, USA.
contactserge@clemson.edu

Aoibhin A. Shiels
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
aoibhin.a.shiels@unr.edu

Abstract

Machine learning models now predict perceived safety, beauty, and walkability from street-level imagery, but their explanations often remain pixel-level and poorly aligned with the concepts urban designers and residents use. This paper develops a concept-mediated, spatially aware framework for explainable artificial intelligence in streetscape perception research. The framework treats perception ratings as ordinal, rater-dependent measurements of latent constructs; constrains visual models through interpretable concepts such as greenery, enclosure, maintenance, and human presence; decomposes predictions into concept contributions; and propagates uncertainty and spatial dependence into explanation. It also specifies evaluation logic for fidelity, stability, human interpretability, and spatial residual diagnostics. The contribution is methodological rather than empirical: it links perceptual measurement theory, concept-based vision models, attribution analysis, and spatial statistics into a single inferential workflow. The framework clarifies why objective scene descriptors and subjective ratings may cohere or diverge, how correlated urban features can produce misleading attributions, and which validation claims are defensible at the image, rater, and neighborhood scales. Its significance lies in making streetscape AI explanations auditable for design review, public deliberation, and urban policy while avoiding the false precision of unvalidated saliency maps.

Keywords

streetscape perception; explainable artificial intelligence; concept bottleneck models; urban analytics; ordinal regression; spatial uncertainty

1. Introduction

Urban streetscapes are evaluated not only by physical inventories but by how people perceive safety, comfort, beauty, and walkability. Urban design research has long treated such qualities as measurable constructs that mediate walking behavior and public life [1]. Computer vision has turned these constructs into predictive tasks: scene classifiers and regression models estimate perceived safety from street-level imagery [2], [3], and large-scale studies map human perceptions across cities [4]. The resulting maps reveal inequalities in perceived urban quality [5] and support environment-quality assessment [6]. Yet the explanatory layer remains underdeveloped. Comparative analyses of objective and subjective street perception show that physical indicators and reported perceptions can converge in some places and diverge in others [7]. An explanation system that reports only which pixels raised a safety score cannot by itself resolve that divergence, because the divergence is partly conceptual, partly rater-dependent, and partly spatial.

Most current explanations are post hoc feature attributions or saliency maps [8], with additive game-theoretic attributions [9] and local surrogate explanations [10] widely used. Such tools are useful diagnostics, but they do not guarantee that an explanation corresponds to an urban design concept that a planner, resident, or reviewer can interpret. Concept-based approaches [11] and concept bottleneck models [12] offer a more structured route by forcing predictions through intermediate semantic variables. Interpretable-by-design arguments go further and question whether high-stakes decisions should rely on post hoc explanations of black boxes at all [13]. The evaluation of such explanations also remains unsettled: fidelity to the model, stability under plausible perturbations, human comprehensibility, and alignment with domain constructs are distinct criteria that can conflict [14].

This paper addresses the gap between streetscape perception modeling and explanation validity. Its central contribution is a concept-mediated, spatially aware explanatory framework that specifies how subjective perception targets should be modeled, how visual concept representations should constrain prediction, how explanations should be decomposed and tested, and how uncertainty and spatial dependence should qualify interpretation. The framework is not presented as an evaluated deployment; it is a methodological synthesis grounded in existing work on urban perception, concept models, attribution methods, and spatial statistics. The argument proceeds by formulating the perceptual prediction problem, developing the concept-mediated explanation architecture, specifying evaluation logic, and discussing technical trade-offs and boundary conditions.

2. Problem Formulation and Perceptual Constructs

Streetscape perception is not a direct property of an image. It is a judgment made by individuals with different histories, expectations, and embodied experiences. A model trained on image-label pairs estimates a statistical relationship between visual form and reported perception, not a physical law. This distinction matters for explanation. If two images have similar greenery but different perceived safety, the difference may arise from enclosure, maintenance, lighting, visible activity, or the rater population. A model that concentrates on vegetation pixels may be faithful to its own decision rule yet misleading as an account of human perception.

The measurement model should therefore represent ordinal ratings, rater heterogeneity, and scene context. Let y_{ij} be the ordinal rating supplied by rater i for street image j, let x_j denote the image and its metadata, let z_j denote contextual covariates such as neighborhood

density or time of day, and let i index raters. A latent-response formulation separates the scene-level perception signal from rater-specific bias:

$$\mathrm{P}\left(y_{ij} \le k | x_j, i\right) = \Phi\left(\theta_k - \left(\mu\left(x_j\right) + \alpha_i + \beta^{T} z_j\right)\right)$$

Here Φ is the standard normal cumulative distribution, θ_k is the cutoff for ordinal category k, μ(x_j) is the latent scene perception estimated from imagery, α_i is a rater effect, and β captures contextual shifts. The expression is a modeling template rather than a fitted result. Its purpose is to make explicit that an explanation of μ(x_j) is not automatically an explanation of y_{ij}. Rater effects and context can alter observed ratings without changing the visual scene, so explanation claims should specify their target: model prediction, population-average perception, or individual judgment.

A second distinction concerns the construct being predicted. Perceived safety, beauty, comfort, and walkability are correlated but not interchangeable. A model trained on one construct should not be interpreted as explaining another merely because the visual predictors overlap. Construct validity requires that the rating instrument, sampling frame, and population be stated. It also requires caution about objective-subjective coherence: measurable physical attributes can support or contradict perceptual reports depending on local context and measurement scale [7]. The framework therefore treats subjective ratings as the primary target and objective scene descriptors as explanatory covariates or validation checks, not as ground truth.

3. A Concept-Mediated Explanation Framework

3.1 Model structure and concept bottleneck

The proposed architecture predicts a latent perception score through a small set of interpretable visual concepts. Let c_q(x) be the q-th concept score for image x, with q = 1, …, Q, and let ψ_q be the concept extractor. The perception model is

where h is an aggregation function, such as a sparse linear model, generalized additive model, or small monotonic network. The concept set should be specified before model fitting and justified by urban design theory and prior perception research. Candidate concepts include visible greenery, sky openness, enclosure by buildings, visual complexity, maintenance and decay, sidewalk presence, street width, and human activity. The extractor ψ_q may be a supervised segmentation or detection model, a pretrained embedding projected onto a concept direction, or a human-audited feature extractor. The aggregation h should be simple enough that the contribution of each concept can be inspected without another black box.

This concept bottleneck has an important consequence: it separates representation learning from explanatory aggregation. Deep networks can still encode rich visual information, including ResNet-style convolutional backbones [15] or transformer architectures [16], and vision-language pretraining can supply auxiliary semantic embeddings [17]. But the final prediction is routed through variables that have names and measurement protocols. The trade-off is potential loss of predictive flexibility. A concept set that omits a relevant cue, such as curb cuts or temporary construction, will force the model to explain through proxies. The framework should therefore report concept coverage and residual predictive error rather than assume that a neat concept vocabulary exhausts streetscape perception.

3.2 Attribution and decomposition

Even with a concept bottleneck, nonlinear interactions among concepts can make interpretation difficult. A local additive approximation provides a transparent summary around a specific image:

$$\mu(x) \approx \mu_0 + \sum_{q=1}^{Q} \lambda_q c_q(x) + \sum_{m=1}^{M} \gamma_m z_m(x)$$

In this expression, μ_0 is a baseline perception level, λ_q is the local contribution of concept q, z_m are contextual covariates, and γ_m their local contributions. The approximation is not asserted to be globally exact. It is a diagnostic device: if the local fit is poor, the additive story is unreliable and should not be presented as a complete explanation. Nonlinearity can be accommodated by fitting the local surrogate in a neighborhood of x or by reporting interaction terms when domain theory justifies them.

For a more principled allocation of concept influence, Shapley values can be computed over the concept set. Let Q be the set of concepts and let μ_S(x) denote the model prediction when only concepts in subset S are available. The Shapley contribution of concept q is

$$\vartheta_q(x) = \sum_{S \subseteq Q \setminus \{q\}} |S|!\,(Q - |S| - 1)!/Q!\ [\mu_{S \cup \{q\}}(x) - \mu_S(x)]$$

where the sum runs over all subsets of concepts excluding q, and |S| is the number of concepts in S. The expression distributes the prediction among concepts according to their marginal contribution across coalitions. It is attractive because it satisfies efficiency and symmetry properties, but it inherits the model's dependence structure. If greenery and enclosure are highly correlated in the training imagery, their Shapley values may split or shift influence in ways that do not match causal effects. Explanations should therefore be reported with concept correlation diagnostics and, where possible, with intervention-based checks rather than relying on attribution magnitudes alone.

3.3 Uncertainty and stability

A point explanation can overstate certainty. Perception models face aleatoric uncertainty from ambiguous scenes and rater disagreement, and epistemic uncertainty from limited coverage of visual contexts. For an ordinal prediction with category probabilities p_k(x), predictive entropy provides one summary:

$$H(x) = -\sum_{k=1}^{K} p_k(x) \log p_k(x)$$

High entropy indicates that the model's ordinal distribution is diffuse; it does not by itself distinguish rater disagreement from model ignorance. A decomposition into aleatoric and epistemic components is preferable when the architecture supports it, using Monte Carlo dropout [18] or heteroscedastic output heads [19]. In explanatory terms, an image with high uncertainty should not receive a confident concept-level narrative. The explanation should be accompanied by uncertainty intervals, and spatial summaries should separate places where the model is uncertain from places where the model is confident but wrong relative to validation ratings.

Explanation stability is a second requirement. Let a(x) be an attribution vector, such as the concept contributions λ_q or Shapley values ϑ_q, and let T_m(x) be the m-th plausible transformation of the image, such as mild cropping, exposure change, or occlusion of a nonessential region. Stability can be summarized as

$$S_a = (1/M)\sum_\{m=1\}\^M (a(x)\cdot a(T_m(x)))/(\|a(x)\|\|a(T_m(x))\|)$$

where M is the number of transformations and $\|\cdot\|$ is the Euclidean norm. A high average cosine similarity suggests that the explanation is robust to transformations that should not alter the underlying perception. A low value signals that the explanation is sensitive to nuisance variation. Stability is not equivalent to correctness, but it is a necessary safeguard against explanations that change wildly under irrelevant image operations.

4. Evaluation Logic and Spatial Diagnostics

4.1 Fidelity, comprehensibility, and human grounding

Explanation evaluation should distinguish three claims. Fidelity asks whether the explanation reflects the model's actual decision rule. Comprehensibility asks whether intended users can understand and use the explanation. Perceptual validity asks whether the explanation aligns with human judgments of the same constructs. These claims require different evidence. Fidelity can be tested by deleting or perturbing high-attribution concepts and measuring prediction change. Comprehensibility can be tested with structured judgment tasks in which planners, residents, or designers compare explanations and identify actionable scene features. Perceptual validity can be tested by correlating model concept contributions with independent human annotations of the same concepts, while recognizing that such annotations are themselves measurements with error.

Because the present framework is methodological, these are evaluation specifications rather than reported results. A defensible study would preregister the target construct, rater population, image sampling frame, concept definitions, and the criterion for a concept being present or absent. It would separate training and validation cities or neighborhoods where possible, because visual style and street typology can shift across regions. It would also report explanation fidelity for both correct and incorrect predictions. An explanation that is faithful only when the model is right is of limited diagnostic value; errors are precisely where explanation is most needed.

4.2 Spatial dependence

Streetscape perceptions are spatially structured. Adjacent images share land use, building stock, street furniture, maintenance regimes, and observer expectations. A model that ignores this dependence can produce residual patterns that look like substantive findings. For spatial units i and j, with residuals e_i from a perception model and spatial weights w_{ij}, Moran's I can diagnose residual autocorrelation:

$$I = (N/\sum_\{i,j\}w_\{ij\})\ (\sum_\{i,j\}w_\{ij\}e_i\ e_j)/(\sum_i\ e_i^2)$$

Here N is the number of spatial units, w_{ij} encodes adjacency or distance-based interaction, and e_i is the residual for unit i. Positive autocorrelation indicates that nearby units share unexplained perception variation. Such dependence can arise from unmeasured contextual variables, image capture artifacts, or neighborhood-level norms. Explanation maps should therefore report residual autocorrelation and, where appropriate, use spatial random effects or local regression rather than treating each street segment as independent. Local indicators of spatial association can identify clusters of over- or under-prediction [20], but they should be interpreted as diagnostics, not causal evidence.

5. Technical Discussion

The main advantage of the concept-mediated framework is that it makes the explanatory vocabulary contestable. A saliency map can always be produced, but a concept list can be audited: Are the concepts defined consistently? Do they cover the constructs that residents mention? Are some concepts acting as proxies for unmeasured conditions? This matters because urban design interventions operate on concepts such as shade, enclosure, and maintenance, not on abstract pixel gradients. If the model explains perceived safety through "tree canopy," a planner can ask whether the effect is really canopy, lighting, or the correlation between canopy and neighborhood income. The framework does not answer that causal question, but it makes the question visible.

The cost is that concept bottlenecks may underperform unconstrained models. A concept set chosen for interpretability can omit subtle cues, and concept extraction can introduce measurement error. The appropriate response is not to abandon concepts but to report the predictive gap between the constrained and unconstrained models, the concept extraction uncertainty, and the residual error structure. If the constrained model loses substantial accuracy, explanations may describe a simplified surrogate rather than the deployed predictor. In that case, the system should be transparent about which model is being explained and why.

Attribution methods require similar caution. Shapley values and local surrogates explain a model's behavior under specified coalitions or perturbations, not the causal effect of changing a street. Correlated concepts, unobserved context, and rater heterogeneity can all distort explanatory rankings. The framework therefore treats attributions as diagnostic summaries that must be paired with uncertainty, stability, and spatial residual checks. It also favors sparse, auditable aggregation functions over deeply interactive black boxes when the explanatory goal is design review. This is consistent with interpretable-by-design arguments [13], though it stops short of claiming that all streetscape models must be fully transparent; some exploratory mapping tasks may justify richer models if their explanations are carefully bounded.

A further boundary condition concerns population validity. A model trained on one city's imagery and rater pool may encode local visual conventions. Explanations that appear coherent in one context can fail elsewhere. Cross-city validation, rater metadata, and stratified error analysis are therefore not optional refinements; they are part of the explanation's meaning. The same caution applies to objective-subjective coherence: when objective indicators and subjective ratings diverge, the explanation should identify whether the divergence is due to concept measurement, rater composition, spatial scale, or model misspecification. These are different problems with different remedies.

6. Conclusion

Explainable artificial intelligence for urban streetscapes should not be reduced to prettier heatmaps. The central problem is to connect model behavior to the perceptual constructs that people actually use and to the spatial units at which urban decisions are made. This paper has developed a concept-mediated, spatially aware framework for that task. It treats perception as an ordinal and rater-dependent measurement, routes prediction through auditable visual concepts, decomposes predictions with local and game-theoretic attributions, and qualifies explanations with uncertainty, stability, and spatial residual diagnostics. The framework is methodological: it specifies what must be measured and reported before an explanation can support design review or policy deliberation. Its practical value lies in making streetscape AI

explanations contestable at the level of concepts, stable under irrelevant variation, honest about uncertainty, and sensitive to spatial dependence. When these conditions are met, explanations can help planners and residents interrogate why a model judges a street as safe, beautiful, or walkable, rather than merely accepting or rejecting an opaque score.

References

1. Ewing, R., & Handy, S. (2009). Measuring the unmeasurable: Urban design qualities related to walkability. Journal of Urban Design, 14(1), 65–84. https://doi.org/10.1080/13574800802451155

2. Naik, N., Philipoom, J., Raskar, R., & Hidalgo, C. (2014). Streetscore: Predicting the perceived safety of one million streetscapes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 779–785). IEEE.

3. Dubey, A., Naik, N., Parikh, D., Raskar, R., & Hidalgo, C. A. (2016). Deep learning the city: Quantifying urban perception at a global scale. In Computer Vision – ECCV 2016 (pp. 196–212). Springer.

4. 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. https://doi.org/10.1016/j.landurbplan.2018.08.020

5. Salesses, P., Schechtner, K., & Hidalgo, C. A. (2013). The collaborative image of the city: Mapping the inequality of urban perception. PLoS ONE, 8(7), e68400. https://doi.org/10.1371/journal.pone.0068400

6. Liu, L., Silva, E. A., Wu, C., & Wang, H. (2017). A machine learning-based method for the large-scale evaluation of the qualities of the urban environment. Computers, Environment and Urban Systems, 65, 113–125. https://doi.org/10.1016/j.compenvurbsys.2017.06.003

7. Song, Q., Li, M., Qiu, W., Li, W., & Luo, D. (2022, November). The coherence and divergence between the objective and subjective measurement of street perceptions for Shanghai. In International Conference on Advanced Data Mining and Applications (pp. 244-256). Cham: Springer Nature Switzerland.

8. Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (Eds.). (2019). Explainable AI: Interpreting, explaining and visualizing deep learning. Springer.

9. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30 (pp. 4765–4774). Curran Associates.

10. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). ACM.

11. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In Proceedings of the 35th International Conference on Machine Learning (pp. 2668–2677). PMLR.

12. Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020). Concept bottleneck models. In Proceedings of the 37th International Conference on Machine Learning (pp. 5338–5348). PMLR.

13. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

14. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.

15. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770–778). IEEE.

16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations.

17. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (pp. 8748–8763). PMLR.

18. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning (pp. 1050–1059). PMLR.

19. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In Advances in Neural Information Processing Systems 30 (pp. 5574–5584). Curran Associates.

20. Anselin, L. (1995). Local indicators of spatial association—LISA. Geographical Analysis, 27(2), 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x