

Article

Towards Transparent Urban Perception: A Concept-Driven Framework with Visual Foundation Models

Yixin Yu ¹, Zepeng Yu ², Xuhua Shi ¹, Ran Wan ², Bowen Wang ³ and Jiaxin Zhang ^{2,4,*} 

¹ Faculty of Humanities and Arts, Macau University of Science and Technology, Macao SAR, China; 3240001741@student.must.edu.mo (Y.Y.); 3240006050@student.must.edu.mo (X.S.)

² Architecture and Design College, Nanchang University, No. 999 Xuefu Avenue, Nanchang 330031, China; 410546420010@email.ncu.edu.cn (Z.Y.); wanran@ncu.edu.cn (R.W.)

³ D3 Center, Osaka University, 2-8 Yamadaoka, Suita, Osaka 565-0871, Japan; bowen.wang@is.ids.osaka-u.ac.jp

⁴ Environmental Design and Information Technology Laboratory, Division of Sustainable Energy and Environmental Engineering, Graduate School of Engineering, Osaka University, 2-1 Yamadaoka, Suita, Osaka 565-0871, Japan

* Correspondence: jiaxin.arch@ncu.edu.cn

Abstract

Understanding urban visual perception is crucial for modeling how individuals cognitively and emotionally interact with the built environment. However, traditional survey-based approaches are limited in scalability and often fail to generalize across diverse urban contexts. In this study, we introduce the UP-CBM, a transparent framework that leverages visual foundation models (VFMs) and concept-based reasoning to address these challenges. The UP-CBM automatically constructs a task-specific vocabulary of perceptual concepts using GPT-4o and processes urban scene images through a multi-scale visual prompting pipeline. This pipeline generates CLIP-based similarity maps that facilitate the learning of an interpretable bottleneck layer, effectively linking visual features with human perceptual judgments. Our framework not only achieves higher predictive accuracy but also offers enhanced interpretability, enabling transparent reasoning about urban perception. Experiments on two benchmark datasets—Place Pulse 2.0 (achieving improvements of +0.041 in comparison accuracy and +0.029 in R^2) and VRVWPR (+0.018 in classification accuracy)—demonstrate the effectiveness and generalizability of our approach. These results underscore the potential of integrating VFMs with structured concept-driven pipelines for more explainable urban visual analytics.

Keywords: deep learning; urban perception; visual concept; explainable AI; visual prompt



Academic Editor: Wolfgang Kainz

Received: 12 June 2025

Revised: 11 August 2025

Accepted: 14 August 2025

Published: 18 August 2025

Citation: Yu, Y.; Yu, Z.; Shi, X.; Wan, R.; Wang, B.; Zhang, J. Towards Transparent Urban Perception: A Concept-Driven Framework with Visual Foundation Models. *ISPRS Int. J. Geo-Inf.* **2025**, *14*, 315. <https://doi.org/10.3390/ijgi14080315>

Copyright: © 2025 by the authors. Published by MDPI on behalf of the International Society for Photogrammetry and Remote Sensing. 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 (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Understanding cognitive reactions to the built environment depends on insights from urban visual perception. Rather than relying solely on objective indicators, it emphasizes the subjective impressions people form when observing urban spaces [1]. Traditional efforts to measure such perception have often involved interviews, on-site surveys, and structured questionnaires [2–4]. While these methods provide depth, they tend to be labor-intensive, time-consuming, and geographically constrained.

Recent advances in computer vision have opened new avenues for large-scale cost-efficient analysis of urban perception [5]. Automated image processing allows researchers to extract perceptual signals from vast collections of street-view images (SVIs), enabling assessments across different temporal and environmental contexts [6,7]. Compared to the

conventional approaches, these technologies reduce reliance on manual labor while maintaining spatial and temporal comprehensiveness. Recent research has focused on mining hidden attributes of cities through SVIs, with particular attention to quantifying urban visual perception on a city-wide scale [8–10]. However, given the inherently subjective nature of visual perception [11], generating reliable ground-truth annotations for training machine learning models is challenging. The MIT Place Pulse project [12] has addressed this challenge by introducing an innovative online crowd-sourcing approach: participants evaluated image pairs across perceptual dimensions such as wealth, vibrancy, and visual appeal. The resulting dataset, Place Pulse 2.0 [13], has significantly enhanced the scalability of urban perception research.

In addition to these developments, visual foundation models (VFM) [14–16] have emerged as powerful tools for learning visual representations across diverse domains. Pretrained on massive image–text pairs, models such as CLIP [15] and DINOv2 [16] encode semantically rich features that can be readily adapted to downstream urban perception tasks. They have the ability to capture high-level visual concepts, which makes them especially suitable for modeling the nuanced subjective dimensions of how people perceive the urban environment. More recently, visual prompt engineering has been shown to enhance the visual concept alignment capability of CLIP models. For example, incorporating a simple red circle as a visual cue can significantly improve CLIP’s ability to recognize target concepts [17,18]. Building on this idea, we evaluate CLIP’s alignment performance for urban-related concepts in SVIs. As illustrated in Figure 1 (we use different-colored circles for better demonstration of concepts), our results show that circle-based visual prompts enable the correct alignment of CLIP with urban visual concepts.

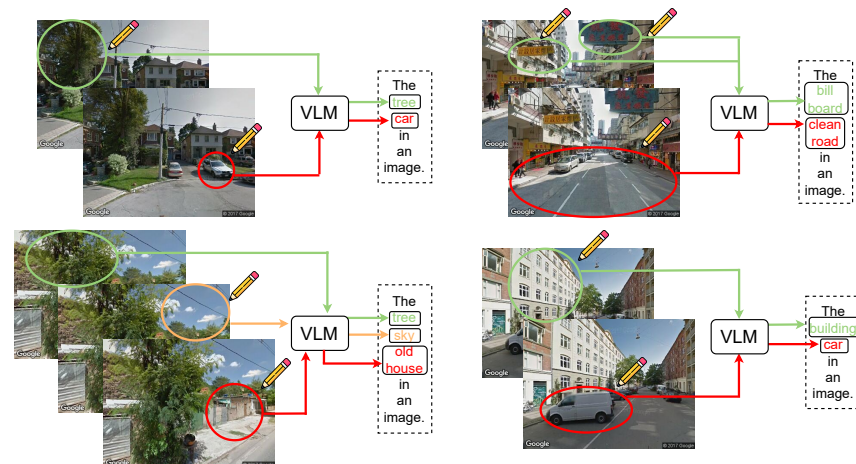


Figure 1. Using circles to pin-point urban concepts in SVIs via a VLM model, CLIP.

However, the existing VFM-based interpretability methods typically rely on global similarity maps or single visual cues to provide post hoc explanations, which often fall short of offering a complete and traceable reasoning path across the concept, spatial, and decision levels. For instance, the use of CLIP similarity maps alone can highlight relevant regions but cannot reveal how individual concepts cumulatively influence the final perception score; likewise, the direct mapping of VFM features to labels lacks any visualization of “which concepts” are involved and “how they interact”. In contrast, our proposed concept bottleneck model (UP-CBM) delivers three key explainability advances over these prior approaches: (1) **explicit concept vocabulary**: by automatically generating and then expert-filtering a set of human-readable urban concepts, our model elevates explanations from unstructured feature vectors to semantically meaningful concepts [19,20]; (2) **weakly supervised spatial pseudo-labels**: through multi-scale visual prompting, we obtain fine-

grained CLIP similarity maps that not only localize concept activations but also quantify activation strengths, enabling spatially grounded explanations; and (3) **end-to-end concept bottleneck layer (CBL)**: we train a concept bottleneck layer atop frozen VFM backbones with an MSE alignment loss, enforcing a clear “concept–activation–output” pathway that is fully traceable and verifiable.

In this paper, we introduce a novel concept bottleneck model, the UP-CBM, tailored for urban perception tasks. It begins by generating a compact human-understandable urban visual concept vocabulary using GPT-4o [21]. Through carefully designed prompts, the language model [22,23] identifies semantically meaningful visual elements that influence perception—such as “clean road,” “cars,” or “colorful facade”—across different perceptual dimensions. These concepts serve as an interpretable intermediate space between raw pixels and perceptual scores. We then design a multi-scale visual prompting strategy [18] to probe spatial concept activations at different image resolutions. By applying synthetic perturbations such as red circles across the image grid, we use CLIP to compute fine-grained similarity maps between image regions and textual concepts. These maps act as weak pseudo-labels to guide the learning of a concept bottleneck layer (CBL), which distills interpretable concept activations from the frozen backbone features of a CNN or ViT. Finally, we aggregate the spatial concept maps into a global concept vector, which is passed to a shallow predictor for either regression or classification depending on the dataset.

The overall pipeline provides transparency at multiple levels: it grounds predictions in visual concepts, localizes those concepts spatially, and supports quantitative evaluation using both pairwise comparison accuracy and regression R^2 . We also validate our method on two perception datasets: Place Pulse 2.0 [12] and VRVWPR [24]. The experimental results demonstrate that our interpretable model achieves better perception performance (0.041 accuracy and 0.029 R^2 improvement on average for Place Pulse 2.0, and 0.018 accuracy improvement for VRVWPR) while offering significant advantages in transparency and human-centered understanding of urban visual perception.

Our contributions are as follows:

- An interpretable urban perception pipeline that explicitly models perceptual reasoning through human-understandable visual concepts is proposed. Our approach introduces a concept bottleneck layer aligned with CLIP-derived similarity maps, enabling transparent and controllable perception prediction.
- We design a class-free concept discovery mechanism using GPT-4o, generating task-specific visual concepts without relying on predefined categories. This allows the model to flexibly adapt to perception tasks with subjective or continuous labels, such as safety or walkability.
- We design a multi-scale prompting strategy to spatially probe and localize perceptual concepts in street-view images for better concept extraction, achieving both accurate and interpretable predictions across multiple datasets.

2. Related Work

2.1. Visual Foundation Models

The success of large language models (LLMs) [25–27] in the field of natural language processing has significantly influenced recent advances in visual foundation models (VFMs) [28,29]. Applications in these domains have demonstrated superior generalization capabilities across a variety of tasks. Inspired by this, models such as CLIP [15] have emerged, bridging the gap between vision and language by aligning their respective feature spaces through contrastive learning on massive image–text pair datasets. This alignment strategy has enabled CLIP to achieve superior zero-shot performance in downstream visual understanding tasks [30–32].

A particularly notable advancement in vision is the use of promptable models for segmentation. A breakthrough in this direction is represented by the Segment Anything Model (SAM), which has been trained on over a billion masks with both sparse (e.g., point, bounding box, and text cues) and dense (e.g., full mask) annotations. This flexible prompting mechanism allows SAM to adapt to a wide range of downstream tasks, such as visual object tracking [33,34], 2D image segmentation [7,35], and even 3D scene reconstruction [36]. In addition to promptable segmentation, recent VFMs such as DINOv2 [16], masked autoencoder (MAE) [37], and CLIP are capable of learning robust object-centric and dense semantic features. These models capture rich visual cues, enabling detailed characterization of object properties in complex real-world environments.

Recent work has shown that visual prompt engineering [38] can enhance the visual alignment capabilities of CLIP models. For example, adding a simple red circle as a visual cue significantly improves CLIP's ability to recognize target concepts [17]. By deliberately designing visual elements such as background, positioning, and composition, images can be tailored to better “prompt” the model. Building on this idea, we assess CLIP's alignment performance for urban-related concepts in street-view imagery. As shown in Figure 1, our findings reveal that circle-based visual prompts markedly improve CLIP's alignment with urban visual concepts. For visualization purposes, we use different circle colors to distinguish between concepts, although all circles are rendered in red during actual implementation.

2.2. Quantifying Urban Perception via Foundation Models

In the domain of urban image analysis, the increasing availability of remote-sensing and street-view imagery has facilitated the development of multimodal models for urban perception. UrbanCLIP [39], for example, employs cross-modal alignment techniques to infer urban indicators from images using textual guidance. UrbanVLP [40] further integrates automatically generated textual descriptions with macro-level (satellite) and micro-level (street-view) imagery, enhancing semantic understanding and descriptive capability. Meanwhile, benchmarks like V-IRL [41] have provided standardized evaluations for LLMs on tasks such as localization and scene recognition in urban imagery. Urban planners and sociologists have long emphasized the influence of visual and physical urban characteristics on human cognition and behavior. Nasar et al. argue that a city's appearance significantly impacts residents' emotional responses [9], while scholars such as Keizer and Kelling highlight the associations between visual disorder and social issues such as crime and declining educational outcomes [3].

To quantify subjective perceptions of the urban visual environment, researchers have increasingly turned to street-view imagery (SVI) as an objective medium to reflect urban appearance [42]. Projects like Place Pulse gathered public ratings on attributes such as “safety” or “cleanliness” through large-scale image pair comparisons, resulting in the Place Pulse 1.0 dataset [13]. Building on this, Naik et al. developed the Streetscore algorithm using image features and support vector regression to predict perceived street safety across 21 U.S. cities [43,44]. However, due to its training bias toward New York and Boston images, its global generalizability remains limited.

In addition to crowdsourcing-based approaches, Griew et al. introduced the FASTVIEW tool, which combines expert audits and crowd ratings based on Google SVI to assess factors such as pavement quality, lighting, and safety in relation to physical activity [45]. Despite these innovations, traditional data collection methods—such as interviews and surveys—still suffer from small sample sizes and subjectivity, limiting their scalability and reliability [46]. Overall, the integration of street-level visual features with semantically rich LLM outputs offers a promising new paradigm for urban perception

research. This fusion of visual and semantic cues not only enhances the accuracy of city-scale variable predictions but also enables more interpretable, scalable, and fine-grained modeling of how urban environments are perceived and experienced.

2.3. Concept-Based Explanation

Recent advancements in concept-based interpretability frameworks have highlighted their importance in improving the transparency and human-understandability of deep learning models. These frameworks map the internal representations of models to mid-level semantically meaningful units known as concepts, which are nameable and perceptually accessible to humans [19,47]. Inspired by cognitive science, this approach emphasizes that human decision-making often relies more on abstract semantic concepts than on raw inputs or pixel-level features, offering interpretability that is both more intuitive and more transferable across domains.

A core component of these frameworks lies in the construction of a high-quality concept set that captures essential structures or semantic patterns in the data. Concepts can either be manually defined using expert knowledge—for instance, disease types in medical imaging or syntactic tags in language [48]—or automatically discovered through data-driven techniques such as clustering, sparse factorization, or contrastive learning [49]. Once established, model predictions can be re-expressed as concept activation patterns, allowing users to trace the decision path through semantically meaningful nodes. One of the most prominent structures embodying this idea is the concept bottleneck model (CBM) [19], which inserts an intermediate bottleneck layer representing concepts and forces the model to make predictions solely based on them. The architecture is typically split into two stages: a concept predictor and a classifier operating on predicted concepts. This modular design enhances traceability and interventionability: users can tweak concept activations to directly observe changes in predictions. Variants of the CBM have been successfully applied across tasks in vision for classification or captioning.

Notably, the SALF-CBM [18] transforms deep networks into spatially and semantically interpretable models without sacrificing performance. It excels in zero-shot segmentation, offering interactive functionalities beneficial for high-stakes domains such as medical imaging and autonomous driving. Despite the promise of this paradigm, one of its central challenges remains the construction of concept sets that are not only discriminative but also semantically coherent and nameable. Automatically discovered concepts may lack semantic clarity or deviate from human understanding [50]. To bridge this semantic gap, recent works propose hybrid approaches that combine a small number of expert-defined concepts with a larger set of machine-discovered ones, aiming for a better balance between performance and interpretability [20]. In this paper, we adopt the concept-based method for transparent urban perception assessment.

3. Method

As shown in Figure 2, we propose an interpretable framework for urban perception modeling (UP-CBM). First, task-specific visual concepts are generated by prompting a large language model and subsequently refined through expert filtering. Then, multi-scale visual prompting and CLIP embeddings are employed to compute fine-grained image–concept similarities. A concept bottleneck layer (CBL) aligns backbone features with these concept cues, enabling interpretable activations. Finally, global concept vectors are aggregated for perception prediction, with the model jointly optimized using a concept alignment loss to enhance both transparency and accuracy.

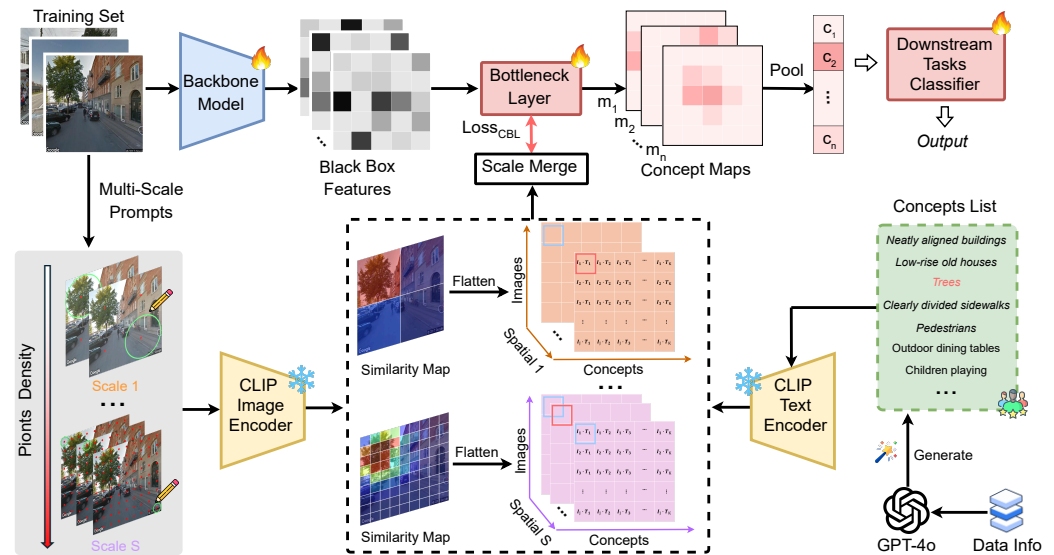


Figure 2. The overall pipeline of our proposed concept-based interpretable urban perception model.

3.1. Concept Generation

Unlike general classification tasks where discrete class labels are available for specific categories (e.g., cat or dog), the dataset we use consists of urban scene images annotated with continuous perception scores (e.g., safety and vibrancy), making the task more akin to regression or soft classification rather than strict categorization. This inherent difference motivates the need for a concept generation process that does not rely on predefined class labels but instead dynamically constructs semantically meaningful and perceptually relevant concepts that reflect the subjective nature of human urban perception.

To address this, we propose a class-free context-driven concept generation strategy that leverages the power of large-scale language models to bootstrap a task-specific visual concept space. The goal is to identify a comprehensive yet compact set of visual concepts $K = \{k_1, k_2, \dots, k_N\}$ that can be used for both interpretability and downstream reasoning.

We begin by designing a set of task-level prompts that encapsulate the core objective of perception modeling. Specifically, we employ a state-of-the-art language model, GPT-4o, to respond to the following queries (three designed prompts), as shown in Table 1.

Note that “{requirement}” can be replaced with different terms according to the specific dataset requirements. The model’s responses are concatenated and aggregated to form a raw concept pool:

$$K_{\text{raw}} = \text{GPT4o}(\text{Prompt}_1, \text{Prompt}_2, \text{Prompt}_3). \quad (1)$$

Since the raw list often contains noisy or redundant terms, we apply a filtering step to ensure interpretability and visual relevance. Specifically, we remove (by city experts):

- Overly long or ambiguous descriptions (e.g., “a sense of openness and connectedness”);
- Semantically overlapping entries (e.g., “car” and “vehicle”);
- Items not visually grounded in the dataset (e.g., abstract concepts like “justice”).

The resulting refined list $K = \text{Filter}(K_{\text{raw}})$ represents the task-specific concept vocabulary, which is then used to anchor our interpretability framework. These concepts serve as an intermediate bottleneck representation between the raw visual features and the high-level perception prediction, enabling more transparent and controllable decision-making. In Table 2, we also show the number of concepts and some examples for the two datasets used in this paper.

Concept refinement was carried out by three experts. Two independently screened the GPT-4o-generated concept pools (452 for Place Pulse 2.0 and 344 for VRVWPR), and a third adjudicated any disagreements. Terms exhibiting “semantic overlap” (cosine similarity > 0.80 in CLIP text embeddings) were jointly removed by the first two experts, while “visual grounding” (concepts unlikely to appear or be visually identifiable, e.g., “justice”) was assessed by having the two experts inspect 1000 randomly sampled street-view images from the dataset and eliminate items rarely or never observed. Post-filtering, the concept counts decreased to 317 and 195, respectively, demonstrating the necessity of expert intervention. Cohen’s κ for the initial two experts’ labels was 0.825, indicating strong agreement.

Table 1. Prompts for visual concept generation.

Prompt	Content
Prompt 1	List the most important visual features that influence a person’s perception of urban {requirement} in street-view images.
Prompt 2	What visual elements in a city scene make people feel {requirement}?
Prompt 3	List common objects, layouts, or attributes visible in urban street-view images that could impact how people feel about the place.

Table 2. Statistics of visual concepts generated for each dataset.

Dataset	Number of Concepts	Example Concepts
Place Pulse 2.0 [12]	317	Trees, Sidewalks, Pedestrians, Streetlights, Clean roads, Car, Trash, Construction site, Graffiti, Old houses, Road, etc.
VRVWPR [24]	195	People, Bicycles, Crosswalk, Bike Lane, Traffic Light, Shop, Empty roads, Dirt land, Wall, Billboard, Bus, etc.

3.2. Fine-Grained Image–Concept Similarities

To interpret perception-relevant visual concepts at fine-grained spatial levels, we adopt a multi-scale visual prompting strategy. This enables us to probe and localize semantically meaningful regions within an image that align with human-understandable concepts.

Given an input image $x \in \mathbb{R}^{H \times W \times 3}$, we generate multiple variants of the image by applying synthetic prompts at different spatial resolutions (shown in Figure 2). These prompts function as perturbations (e.g., circles or masks) that softly highlight potential regions of interest. Formally, for each scale $s \in \{1, \dots, S\}$, we define a prompting function $Q^{(s)}$ such that

$$x^{(s)} = Q^{(s)}(x, P^{(s)}), \quad (2)$$

where $P^{(s)}$ denotes a grid of positions over the image space, determining where prompts are applied. The grid is generated such that the strides between positions ensure full spatial coverage, computed as

$$\Delta_h^{(s)} = \left\lfloor \frac{H}{H^{(s)} - 1} \right\rfloor, \quad \Delta_w^{(s)} = \left\lfloor \frac{W}{W^{(s)} - 1} \right\rfloor,$$

where $H^{(s)} \times W^{(s)}$ denotes the number of prompt positions at scale s . At each grid location (h_i, w_j) , we apply a visual prompt (e.g., a red circle of radius $r^{(s)}$), resulting in a perturbed version of the image:

$$x_{i,j}^{(s)} = Q^{(s)}(x, (h_i, w_j), r^{(s)}). \quad (3)$$

These prompted images are fed into a CLIP image encoder E_I to obtain local visual embeddings $v_{i,j}^{(s)} = E_I(x_{i,j}^{(s)})$. In parallel, each concept k_n from the generated concept list

is embedded using the CLIP text encoder E_T to obtain $t_n = E_T(k_n)$. We then compute the cosine similarity between visual and textual embeddings to quantify alignment at each position:

$$O^{(s)}[n, i, j] = \frac{\langle v_{i,j}^{(s)}, t_n \rangle}{\|v_{i,j}^{(s)}\| \cdot \|t_n\|}. \quad (4)$$

This yields a stack of fine-grained spatial similarity maps, indexed by concept n , location (i, j) , and scale s . These maps serve as weak pseudo-labels for guiding interpretable concept extraction.

3.3. Bottleneck Layer for Concept Alignment

To embed these spatial concept cues into the visual backbone, we propose a concept bottleneck layer (CBL), a simple yet effective mechanism for distilling interpretable concept activations from latent feature representations.

Let $f = \text{Backbone}(x)$ be the frozen deep feature extracted from a backbone model (CNN or ViT), where $f \in \mathbb{R}^{C \times H' \times W'}$ (C represents the channel dimension; H' and W' are the spatial dimensions after feature extraction). These features capture rich semantic information but are typically entangled and not easily interpretable. To bridge this gap, we use a single 1×1 convolutional layer Conv_{CBL} to project f onto a concept space:

$$m_n = \text{Conv}_{\text{CBL}}^n(f), \quad M_n \in \mathbb{R}^{H' \times W'}. \quad (5)$$

Stacking the N concept channels yields the bottleneck output tensor:

$$M = [m_1, m_2, \dots, m_N] \in \mathbb{R}^{N \times H' \times W'}. \quad (6)$$

Each m_n is intended to reflect the spatial activation of concept k_n across the image. To ensure semantic fidelity, we align these activations with the fine-grained CLIP-based similarity maps. First, we rescale and merge the multi-scale maps $\{O^{(s)}\}$ to match the resolution of the concept bottleneck outputs:

$$\tilde{O} = \text{ScaleMerge}(\{O^{(s)}\}_{s=1}^S) \in \mathbb{R}^{N \times H' \times W'}. \quad (7)$$

We then introduce an alignment loss to encourage consistency between the predicted concept activations and the CLIP-derived pseudo-ground truth:

$$L_{\text{CBL}} = \frac{1}{NH'W'} \sum_{n=1}^N \sum_{i=1}^{H'} \sum_{j=1}^{W'} (M_n[i, j] - \tilde{O}_n[i, j])^2. \quad (8)$$

This Mean Squared Error (MSE) loss promotes the emergence of dedicated concept channels, each responsible for detecting a human-understandable visual factor, thereby enabling semantic interpretability throughout the network.

3.4. Downstream Prediction Using Concepts

Once the spatial concept maps M are learned and aligned, we aggregate them into a global concept vector via spatial average pooling:

$$c_n = \text{Pooling}(m_n), \quad c = [c_1, c_2, \dots, c_N]^\top \in \mathbb{R}^N. \quad (9)$$

Here, c_n denotes the spatially averaged activation score of concept k_n , and $c = [c_1, c_2, \dots, c_N]^\top$ is the aggregated concept vector summarizing the image-level concept activations. This vector is passed into a shallow classifier to produce the final perception prediction $\hat{y}$:

$$\hat{y} = \text{Classifier}(c). \quad (10)$$

Depending on the nature of the dataset, this classifier can be either a regression model or a discrete classification model. A corresponding loss term, denoted as L_{label} , is computed with respect to the ground-truth label y . Accordingly, the overall training objective is defined as

$$L = \alpha L_{\text{CBL}} + L_{\text{label}}, \quad (11)$$

where α controls the weight of the CBL loss.

In this way, the entire pipeline—from multi-scale spatial prompting, to concept-aligned bottleneck representation, to final prediction—forms a transparent interpretable framework grounded in human-recognizable visual elements. It not only improves model interpretability but also enhances trust and controllability in perception modeling.

3.5. Evaluation Metrics

We evaluate our model using two widely used complementary metrics: accuracy and the coefficient of determination (R^2).

Accuracy measures the proportion of correctly predicted pairwise comparisons between images. Given a set of image pairs (x_i, x_j) with ground-truth comparison labels $y_{ij} \in \{+1, -1\}$ and model-predicted comparison $\hat{y}_{ij} \in \{+1, -1\}$, accuracy is defined as

$$\text{Accuracy} = \frac{1}{N} \sum_{(i,j)} \mathbf{1}[\hat{y}_{ij} = y_{ij}], \quad (12)$$

where N is the total number of image pairs, and $\mathbf{1}[\cdot]$ is the indicator function that returns 1 if the condition is true, and 0 otherwise. We randomly sampled images from test set 10,000 times for this calculation.

Coefficient of Determination evaluates how well the predicted perceptual scores match the ground-truth ratings for individual images. Let y_i denote the ground-truth score for image i , and $\hat{y}_i$ the predicted score. Then, R^2 is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (13)$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ is the mean of the ground-truth scores. An R^2 value of 1 indicates perfect prediction, 0 indicates that the model performs no better than predicting the mean, and negative values indicate worse-than-baseline performance.

4. Experiment

4.1. Datasets

We use two city perception datasets in our experiments for both regression and classification tasks.

Place Pulse 2.0 [12] is a dataset developed by the MIT Media Lab to collect human perceptions of urban environments through visual imagery. Using an online platform, participants compare pairs of street-level images and rate them based on six perceptual dimensions: safety, beauty, depression, liveliness, wealth, and boredom. The dataset contains 110,988 street-view images from 56 cities worldwide, with approximately 1.16 million pairwise comparisons contributed by 81,630 participants. Each image is assigned perception scores derived from these comparisons, and the dataset has been rigorously evaluated for consistency and bias. Place Pulse 2.0 provides a large-scale structured foundation for studying the relationship between urban appearance, human perception, and urban design.

VRVWPR [24] is a panoramic street-view image dataset designed to quantify Visual Walkability Perception (VWP). The dataset contains 2642 panoramic images collected from seven major cities in the United States, the United Kingdom, France, Germany, and Japan. Images were evaluated through immersive VR-based pairwise comparisons across six dimensions: walkability, feasibility, accessibility, safety, comfort, and pleasurability. Each image received perception scores ranging from 0 to 10 for each of the six dimensions and was further categorized into high, medium, or low levels for training deep learning models. Thus, different to Place Pulse 2.0, we applied this dataset for classification tasks only. Accuracy for classification is applied for quantifying the performance.

4.2. Implementation Details

We conducted experiments on the Place Pulse 2.0 and VRVWPR datasets. For Place Pulse 2.0, 80% of the data was used for training, 10% for validation, and 10% for testing. For VRVWPR, 60% of the data was used for training, 10% for validation, and 30% for testing. All backbone networks, including ResNet and ViT variants, were initialized with CLIP-pretrained weights. We employed UrbanCLIP [39] for concept alignment and adopted the AdamW optimizer [51] for model training. In the first stage, we froze the backbone parameters and trained the additional layers for 10 epochs with a learning rate of 0.0001. In the second stage, we unfroze the backbone and continued training the entire model for another 15 epochs with the same learning rate. For the CBL loss, the weight α was set to 1 by default. Regarding the scale configurations, we used two scales: 2×2 and 7×7 . All experiments were conducted using four NVIDIA GeForce RTX A6000 GPUs.

4.3. Comparison to Existing Methods

We compare the performance of our approach against five recent regression methods, each representing a distinct modeling paradigm. Streetscore [43] is a classification-based approach that integrates multiple visual features. RESupCon [52] applies supervised contrastive learning to enhance regression accuracy. Adaptive Contrast [53] is a contrastive learning technique tailored for medical image regression tasks, while UCVME [54] employs a semi-supervised learning framework to improve performance with limited labeled data. CLIP + KNN [55] is a clustering method that uses CLIP features for image score aggregation. We evaluate these methods using metrics of R^2 and comparison accuracy.

Table 3 presents a detailed comparison across six perceptual dimensions from Place Pulse 2.0, and several trends can be observed. Among the baseline methods, CLIP + KNN achieves the best performance, consistently outperforming traditional CNN-based methods (Streetscore, Adaptive Contrast, and RESupCon) in both accuracy and R^2 across all dimensions. This highlights the strong generalization ability of CLIP features for visual perception tasks. Additionally, compared to all baselines, our proposed UP-CBM approach achieves new state-of-the-art results. The UP-CBM with an RN-101 backbone already surpasses CLIP + KNN by notable margins in most categories. The UP-CBM with a ViT-B backbone further improves the performance, achieving the best results across all six dimensions. Specifically, it improves the accuracy in the Safe dimension to 0.9352 and the R^2 in the Beautiful dimension to 0.6174, which represents substantial improvements.

A clear trend is also observed that ViT-based models (ViT-S and ViT-B) generally outperform ResNet-based models (RN-50 and RN-101), reflecting the advantages of transformer architectures in modeling complex visual semantics for perceptual regression. The most important discovery is that our method not only improves the comparison accuracy but also significantly raises the R^2 scores, indicating better regression quality and consistency with human perceptual judgments. These results demonstrate the effectiveness of

our UP-CBM framework in capturing perceptual cues and improving generalization across diverse urban perception dimensions.

Table 3. Performance comparison of different regression-based methods. We evaluate six perceptual dimensions from Place Pulse 2.0 using comparison accuracy (Acc.) and R^2 scores. Results with underline and bold font highlight the best performance.

Method	Safe		Lively		Beautiful		Wealthy		Boring		Depressing	
	Acc.	R^2	Acc.	R^2	Acc.	R^2	Acc.	R^2	Acc.	R^2	Acc.	R^2
RN-50 [56]	0.8120	0.4551	0.7289	0.3522	0.8443	0.5090	0.6937	0.3227	0.6451	0.2789	0.6720	0.3136
RN-101 [56]	0.8605	0.5182	0.7778	0.4174	0.8912	0.5783	0.7482	0.3921	0.6958	0.3386	0.7225	0.3784
ViT-S [57]	0.8473	0.4989	0.7831	0.4292	0.8576	0.5374	0.7344	0.3740	0.7042	0.3417	0.7103	0.3599
ViT-B [57]	0.8695	0.5381	0.7967	0.4352	0.8732	0.5479	0.7521	0.3996	0.7603	0.4113	0.7294	0.3832
Streetscore [43]	0.8120	0.4510	0.7290	0.3510	0.8430	0.5010	0.6920	0.3240	0.6440	0.2770	0.6700	0.3120
UCVME [54]	0.8450	0.4920	0.7750	0.4210	0.8590	0.5390	0.7370	0.3760	0.7030	0.3390	0.7090	0.3580
Adaptive Contrast [53]	0.7980	0.4420	0.7150	0.3390	0.8320	0.4950	0.6810	0.3080	0.6330	0.2650	0.6600	0.3020
CLIP + KNN [55]	0.9060	0.5791	0.8415	0.4796	0.9012	0.5741	0.7867	0.4412	0.7951	0.4456	0.7475	0.4008
RESupCon [52]	0.8842	0.5583	0.8155	0.4546	0.8831	0.5580	0.7649	0.4180	0.7724	0.4193	0.7381	0.3939
UP-CBM RN-101 (Ours)	0.9150	0.5852	0.8480	0.5012	0.9154	0.5823	0.7845	0.4401	0.8052	0.4587	0.7550	0.4170
UP-CBM ViT-B (Ours)	0.9352	0.6038	0.8661	0.5239	0.9487	0.6174	0.7893	0.4483	0.8294	0.4827	0.7718	0.4385

We further evaluate the performance of different concept-based classification methods on the VRVWPR dataset, which contains six perceptual dimensions: Walkability, Feasibility, Accessibility, Safety, Comfort, and Pleasurability. For concept-based methods, we briefly introduce them in this section. SENN [58] and ProtoPNet [59] are pioneering works that introduce interpretable prototypes or relevance scores to enhance model explainability. BotCL [20] improves concept learning by combining bottleneck constraints with contrastive learning, whereas P-CBM [60] proposes a post hoc concept bottleneck model for flexible concept supervision. LF-CBM [61] refines concept bottlenecks by leveraging label factorization to disentangle concept dependencies. We also include standard deep models (ResNet50 and ViT-Base) as baselines for comparison.

As shown in Table 4, the UP-CBM consistently achieves superior performance across all dimensions. Our UP-CBM based on RN50 already matches or exceeds the strongest prior methods, whereas the UP-CBM with a ViT-B backbone further boosts the accuracy to new state-of-the-art levels. Specifically, it achieves an overall accuracy of 0.8835, outperforming the previous best method, LF-CBM (0.8647), by a notable margin. Across individual dimensions, the UP-CBM ViT-B attains the highest accuracy in all six perceptual categories. For example, it improves Walkability to 0.9021, Feasibility to 0.9405, and Comfort to 0.9253, demonstrating that our framework can more effectively capture nuanced perceptual concepts compared with previous approaches. In addition, transformer-based models (ViT-B) consistently outperform convolutional networks (ResNet50) in both baseline and UP-CBM settings, highlighting the advantage of transformer architectures for concept-aware visual perception modeling. These results validate the effectiveness of our UP-CBM framework in enhancing both interpretability and prediction accuracy for perceptual classification tasks.

Table 4. Performance comparison of different concept-based classification methods. We evaluate six perceptual dimensions from the VRVWPR dataset using classification accuracy. Results with underline and bold font highlight the best performance.

Models	Walkability	Feasibility	Accessibility	Safety	Comfort	Pleasurability	Overall
ResNet50 [56]	0.8652	0.9084	0.7561	0.7984	0.8950	0.8549	0.8497
ViT-Base [57]	0.8915	0.9321	0.7894	0.8361	0.9173	0.8805	0.8728

Table 4. Cont.

Models	Walkability	Feasibility	Accessibility	Safety	Comfort	Pleasurability	Overall
ProtoPNet [59]	0.8601	0.9001	0.7550	0.8002	0.8925	0.8508	0.8440
SENN [58]	0.8530	0.9019	0.7423	0.7820	0.8861	0.8427	0.8346
P-CBM [60]	0.8794	0.9206	0.7722	0.8189	0.9080	0.8655	0.8608
BotCL [20]	0.8743	0.9198	0.7705	0.8127	0.9042	0.8630	0.8574
LF-CBM [61]	0.8880	0.9297	0.7846	0.8295	0.9150	0.8780	0.8647
UP-CBM RN50 (Ours)	0.8825	0.9260	0.7790	0.8242	0.9192	0.8701	0.8655
UP-CBM ViT-B (Ours)	0.9021	0.9405	0.7980	0.8452	0.9253	0.8899	0.8835

4.4. Concept Analysis

The objective of this experiment is to investigate the influences of various visual concepts on the prediction of urban perception dimensions. For each concept (In Figure 3), we compute the product of its activation value and the corresponding weight from the fully connected classifier and normalize the result to the range $[-1, 1]$ (computed across the whole test set). A higher normalized value (closer to 1) indicates a positive contribution to the perception prediction, whereas a lower value (closer to -1) indicates a negative impact. To intuitively illustrate the role of different concepts, we plot bar charts for each perception dimension. Positive concepts are shown in red, negative concepts in blue, and a bold ellipsis visually separates positive and negative concepts (the top 6 positive and negative concepts are demonstrated).

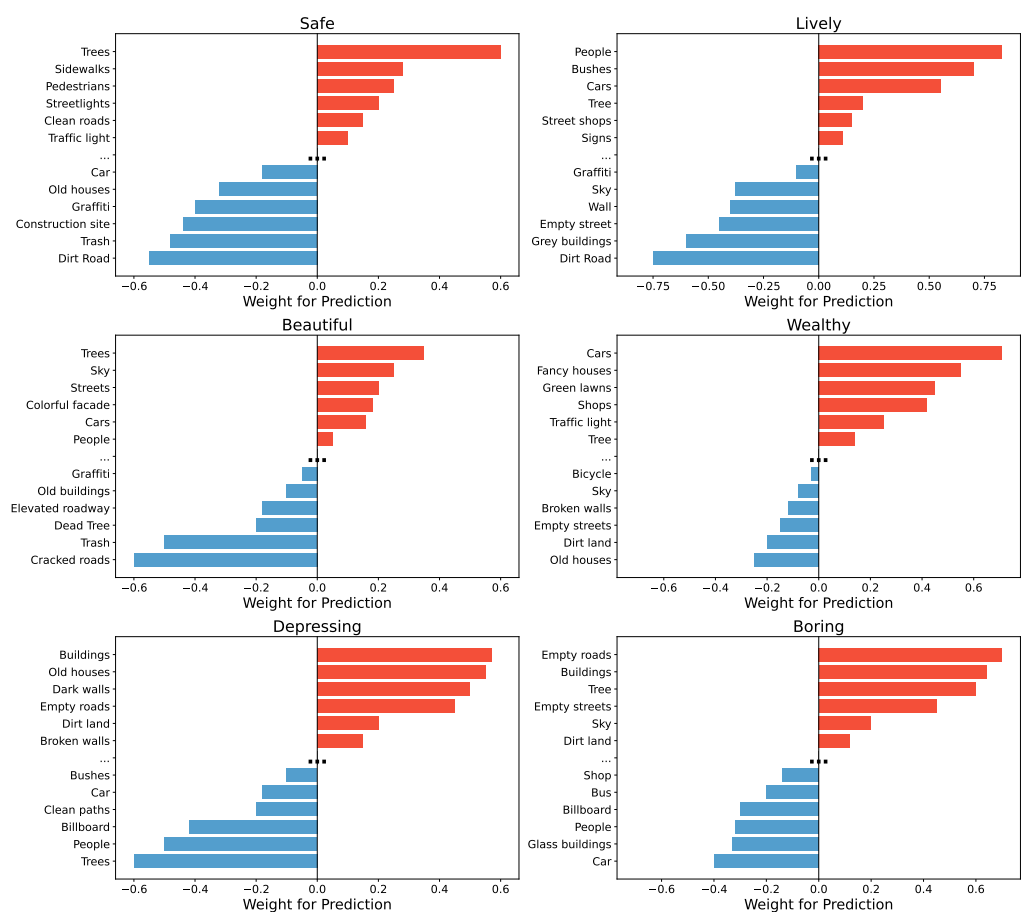


Figure 3. The analysis of concept contribution for six dimensions. We show top 6 positive and negative concepts for the final prediction via classifier weights and concept activation statistics. Calculation is implemented across the whole test set.

Across the different perception dimensions, we observe distinct patterns: For *Safe*, natural and pedestrian-friendly elements such as “Trees”, “Sidewalks”, and “Pedestrians” make strong positive contributions, highlighting the importance of greenery and pedestrian infrastructure for perceived safety. Conversely, elements indicative of environmental deterioration, such as “Trash”, “Construction site”, and “Graffiti” negatively affect safety perception. For *Lively*, the presence of “People”, “Bushes”, and “Cars” strongly enhances the sense of liveliness, emphasizing the role of human activity and transportation in shaping urban vibrancy. In contrast, empty streets (“Streets”) and monotonous structures (“Grey buildings”) significantly diminish the feeling of liveliness. For *Beautiful*, natural scenery elements such as “Trees” and “Sky”, as well as colorful architectural features (“Colorful facade”), positively influence beauty perception. On the other hand, environmental damage such as “Cracked roads” and “Trash” substantially lowers perceived beauty, demonstrating the critical role of cleanliness and natural elements. For *Wealthy*, indicators of affluence, including “Fancy houses”, “Shops”, and “Green lawns”, contribute strongly to the sense of wealth. Meanwhile, dilapidated elements like “Old houses” and “Cracked roads” exert a negative impact, suggesting that visual cues of maintenance and prosperity are closely tied to wealth perception. For *Depressing*, old and deteriorated structures like “Old houses”, “Dark walls”, and “Empty roads” are the major contributors, while the presence of “Trees” and “People” alleviates the depressing feeling. For *Boring*, monotony-related features such as “Empty roads”, “Buildings”, and “Empty streets” increase boredom, whereas “People” and “Shops” mitigate it by introducing diversity and activity. Overall, these results demonstrate that specific visual elements have consistent and interpretable influences on different dimensions of urban perception, providing valuable insights for urban design and planning.

Figure 4 illustrates the concept-based interpretation process for the “Safe” score prediction during inference. The input street-view image is first processed by a backbone model to extract high-level visual features. These features are then passed through a bottleneck layer that generates a set of semantically meaningful concept maps, each highlighting spatial regions associated with specific urban concepts. For each concept, a scalar contribution score is calculated, reflecting its influence on the final “Safe” score. Positive contributions indicate concepts that enhance perceived safety, while negative contributions represent factors that detract from it.

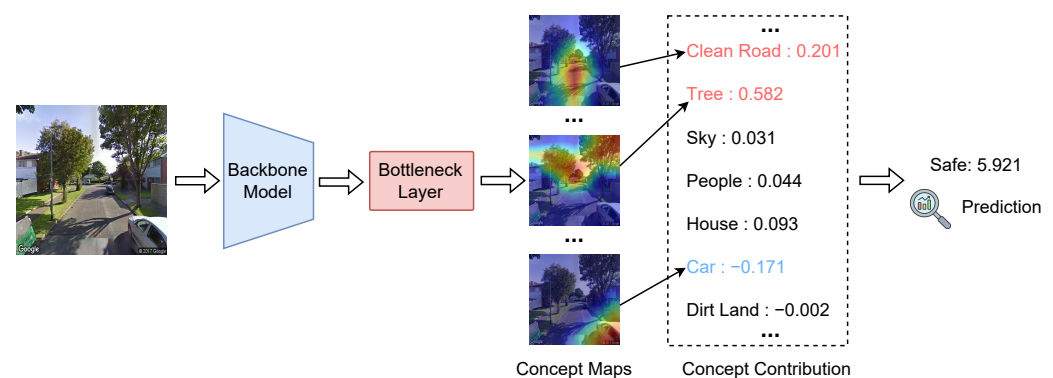


Figure 4. Concept analysis for “Safe” score prediction during inference phase.

In this example, “Tree” exhibits the highest positive contribution (0.582), suggesting that the presence of trees strongly promotes the perception of safety. “Clean Road” also contributes positively (0.201), reinforcing the intuition that well-maintained infrastructure improves safety perceptions. Other positive but smaller contributions come from “House” (0.093), “People” (0.044), and “Sky” (0.031). Conversely, the presence of a “Car” negatively impacts the safety score (−0.171), implying that visible vehicles in this context are associated

with lower perceived safety. “Dirt Land” shows a minor negative effect (-0.002). By aggregating these weighted contributions across concepts, the model produces a final “Safe” prediction score of 5.921. This decomposition provides an interpretable pathway from low-level visual features to high-level safety assessments, offering transparency and explainability for the model’s decision-making process.

Similarly, the concept-based analysis for the “Wealthy” score prediction (Figure 5) shows that “Dirt Land” makes the largest negative contribution (-0.601). “Tree” (0.034) and “Shop” (0.013) have minor positive effects, while “Sky” (-0.058) and “Old House” (-0.031) negatively impact the wealth prediction. “People” (0.007) and “Car” (0.001) contribute minimally. These aggregated concept contributions result in a final “Wealthy” score of 2.417.

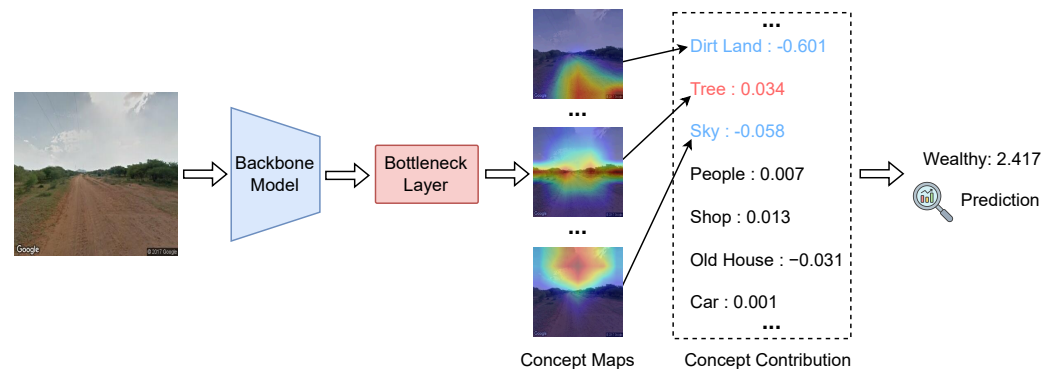


Figure 5. Concept analysis for “Wealthy” score prediction during inference phase. Samples are randomly selected from Place Pulse 2.0.

In addition, we conduct a qualitative assessment to confirm that each concept in the concept bottleneck layer is effectively linked to its designated semantic meaning (as illustrated in Figure 6). After training a UP-CBM on Places365, we sample five concepts from the concept bottleneck layer. For each selected concept, we extract the top six validation images with the highest activations for the corresponding concept. For most concepts, the images with the highest activations generally align well with the intended concept semantics. However, in cases such as “Construction,” some images containing unrelated visual elements (e.g., “old buildings”) are also retrieved. This suggests that the semantic consistency for each concept is not flawless and still exhibits certain limitations.

To further evaluate semantic consistency, we conducted a user study involving 20 participants. Each participant was shown all concepts in the format presented in Figure 6 and asked to rate the semantic consistency of each concept on a scale from 1 to 5. The results are illustrated in Figure 7. Based on the average scores, we categorized the concepts into three groups: the top 30%, middle 40%, and bottom 30%. Overall, the results indicate that the concepts learned by the UP-CBM are generally interpretable to human evaluators. This user study involved 20 participants from Nanchang University, comprising 16 undergraduate students and 4 master’s students. While the relatively small sample size provides valuable preliminary insights into concept semantic consistency, it may limit the generalizability of the findings. Future work will expand the participant pool and include individuals from more diverse backgrounds to enhance the robustness and applicability of the conclusions.

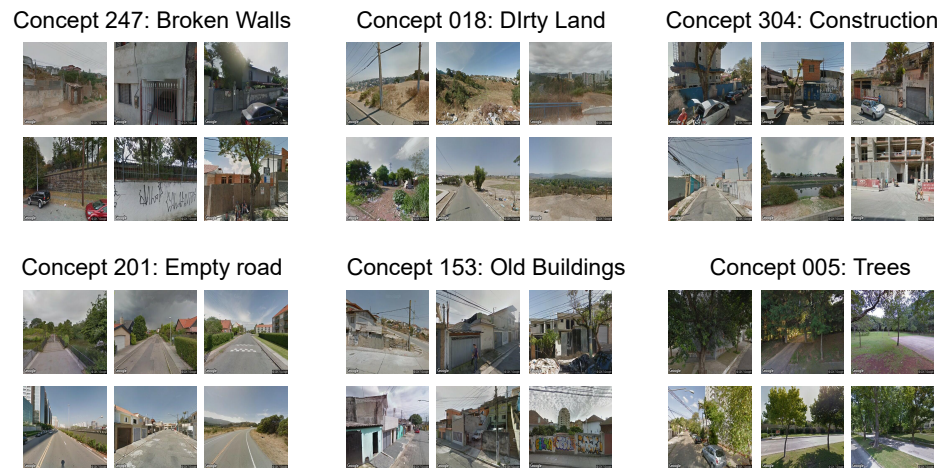


Figure 6. Visualization of some concepts with most activated samples in validation set. Model is trained with Place Pulse 2.0.

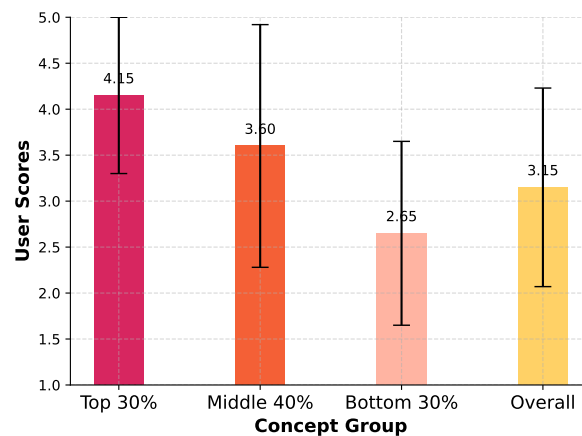


Figure 7. User study results for concept consistency evaluation.

4.5. Ablation Analysis

In this section, we analyze the impact of different module and hyperparameter settings.

As shown in Table 5, the results demonstrate the effectiveness of employing multi-scale visual prompts. When both 2×2 and 7×7 prompts are used together, the model achieves the best performance across both datasets, with an accuracy of 0.8567 and an R^2 of 0.5191 on Place Pulse 2.0, and a classification accuracy of 0.8835 on VRVWPR. Using only the 2×2 prompts leads to a notable performance drop of approximately 7–8 percentage points, indicating that coarse-scale features alone are insufficient for optimal perception modeling. In contrast, using only the 7×7 prompts results in a much smaller decrease (around 1–2 points), suggesting that fine-grained visual information contributes more significantly to performance. These results highlight the importance of integrating multiple spatial resolutions to capture complementary information for perception tasks.

Table 5. Ablation study for using multi-scale visual prompt.

Multi-Scale Settings		Place Pulse 2.0		VRVWPR
2×2	7×7	Acc. (Comparison)	R^2	Acc. (Classification)
✓		0.7802	0.4405	0.8054
	✓	0.8421	0.5032	0.8712
✓	✓	0.8567	0.5191	0.8835

In Figure 8, we analyze the impact of the CBL coefficient (α) on model performance across Place Pulse 2.0 and VRVWPR. When $\alpha = 0$, the absence of CBL leads to notable performance degradation: the accuracy of Place Pulse 2.0 drops to 0.7367 (−12%) and R^2 falls to 0.3691 (−15%), while VRVWPR accuracy decreases to 0.7635 (−10%). As α increases to 0.1 and 0.2, the performance improves but remains below baseline, with Place Pulse 2.0 accuracy reaching 0.7867 and 0.8167, respectively. At $\alpha = 0.5$, the results become much closer to baseline levels, with Place Pulse 2.0 accuracy at 0.8467, R^2 at 0.4791, and VRVWPR accuracy at 0.8785. Further increasing α to 0.8 and 1.0 leads to stabilization, where both accuracy and R^2 nearly match or slightly exceed baseline values. Beyond $\alpha = 1.0$, although minor fluctuations occur (e.g., slight gains in accuracy and small decreases in R^2), no substantial improvement is observed. These results indicate that applying a moderate CBL coefficient (α between 0.5 and 1.0) effectively enhances model generalization, while larger α values yield diminishing returns.

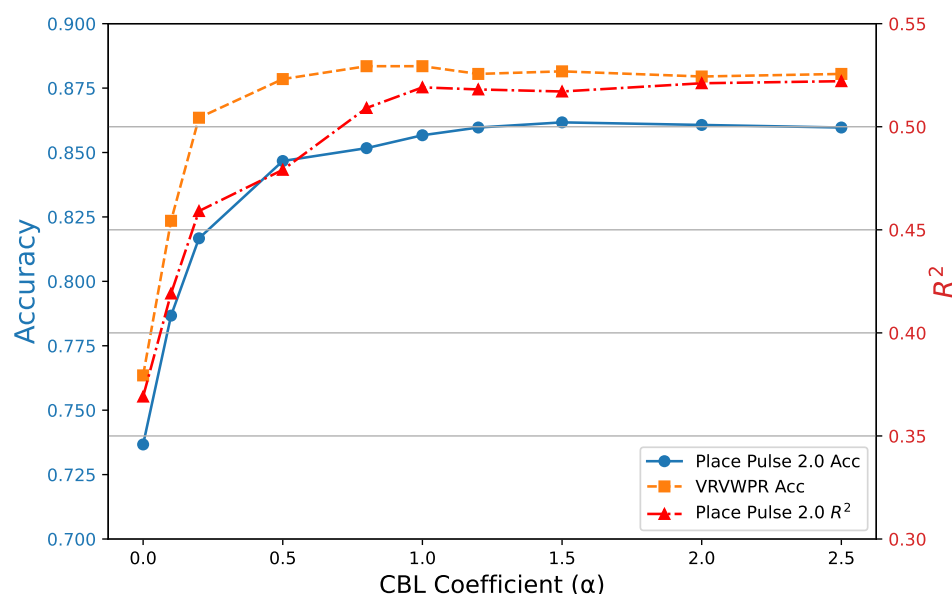


Figure 8. Effect of CBL coefficient (α) on performance. All experiments use ViT-B as the backbone model.

We adopted a single convolutional CBL for two principal reasons. First, channel independence: the 1×1 convolution ensures that each concept channel attends exclusively to its corresponding semantic feature, avoiding any cross-channel interference. Second, effective alignment: despite its simplicity, this design—when paired with an MSE alignment loss—consistently guides each concept activation to match its CLIP-derived pseudo-label. In preliminary experiments, we also evaluated a two-layer CBL variant (Table 6) but observed no significant gains in overall accuracy. However, slight performance difference happens between Place Pulse 2.0 and VRVWPR dataset (the former performs worse with the two-layer setting, whereas the latter benefits from it). This may be because, in classification tasks with discrete labels, the extra non-linear transformation enhances feature separability and captures subtle concept interactions, offering a small performance gain over the simpler design. As a result, we retained the single-layer configuration to maximize interpretability of the model’s reasoning process. Nevertheless, we acknowledge that incorporating more sophisticated multi-scale feature fusion within the CBL could further enhance the network’s capacity to discover and represent complex urban perception concepts. For instance, integrating deformable or dilated convolutions would allow the CBL to adaptively aggregate spatial information at varying receptive fields, capturing both fine-grained details and

broader contextual cues. Future work will explore these architectures and quantify their impact on both predictive performance and explanatory fidelity.

Table 6. Complexity settings of CBL structure.

Settings	Place Pulse 2.0		VRVWPR
	Acc. (Comparison)	R^2	Acc. (Classification)
One layer	0.8567	0.5191	0.8835
Two layers	0.8502	0.5083	0.8919

5. Discussion

5.1. Advantages and Potential Applications of the Proposed Method

The UP-CBM presents several significant advantages for interpretable urban perception modeling. First, by constructing a task-specific visual concept vocabulary through GPT-4o and expert filtering, we address the challenge of subjective and continuous urban perception tasks without relying on rigid class labels. Second, the multi-scale visual prompting strategy enables fine-grained spatial probing of concept activations, bridging the gap between semantic-level concepts and spatial image features. Third, the integration of a CBL aligns model predictions with human-understandable visual elements, substantially enhancing interpretability without sacrificing predictive performance. Extensive experiments demonstrate that the UP-CBM achieves state-of-the-art results on both the Place Pulse 2.0 and VRVWPR datasets, validating the effectiveness of our transparent and scalable approach.

The transparent and interpretable nature of the UP-CBM provides opportunities for several practical applications. In urban planning, our framework can help policymakers and designers to identify specific visual factors—such as greenery, street cleanliness, or architectural style—that positively or negatively influence public perception. This enables targeted interventions to enhance city livability and safety. Furthermore, in real estate, our method can be used to quantitatively assess and visualize neighborhood attractiveness, supporting value prediction and marketing strategies. Another promising application is in autonomous navigation and robotic urban exploration, where understanding human-centered perception dimensions could contribute to safer and more socially aware behavior planning in urban environments.

To assess the computational implications of our approach, we measured the time consumption of the UP-CBM and the ViT-S baseline during both training and inference on a single NVIDIA A6000 GPU (Table 7). During training, the UP-CBM incurred a higher computational cost (156.3 ms vs. 45.8 ms) because multi-scale prompting was employed to supervise the bottleneck layer, involving additional perturbation generation and similarity computations for concept alignment. Importantly, the textual features for the concept list are pre-extracted and cached offline, and the bottleneck layer itself is a single layer with minimal parameters, ensuring that the added complexity remains contained. Once training is complete, multi-scale prompting is no longer required, and the model only computes concept activations through the bottleneck layer to produce predictions. Consequently, the UP-CBM's inference time (23.3 ms) is nearly identical to that of the baseline ViT-S (22.5 ms), demonstrating that the additional training overhead does not translate into runtime inefficiency, and that the method remains highly practical for deployment.

To further assess the stability of the proposed UP-CBM, we conducted five independent training runs on the Place Pulse 2.0 dataset and evaluated each run on the test set. The aggregated results are presented in Figure 9 as the mean $\pm$ standard deviation for both comparison accuracy and R^2 across the five trials. As shown in the figure, the accuracy exhibits minimal fluctuation, with the largest standard deviation below 0.0122, while R^2

shows slightly greater variability but remains stable, with the largest standard deviation still below 0.0205. Table 8 also presents 95% confidence intervals (CIs) for all the metrics using the t-distribution. The inclusion of CIs provides a more rigorous statistical characterization of the results. These observations confirm the robustness and reliability of our method's performance under repeated training conditions.

Table 7. Computation time (ms) comparison between baseline classification and UP-CBM.

Model	Training	Inference
Baseline (ViT-S)	45.8	22.5
UP-CBM	156.3	23.3

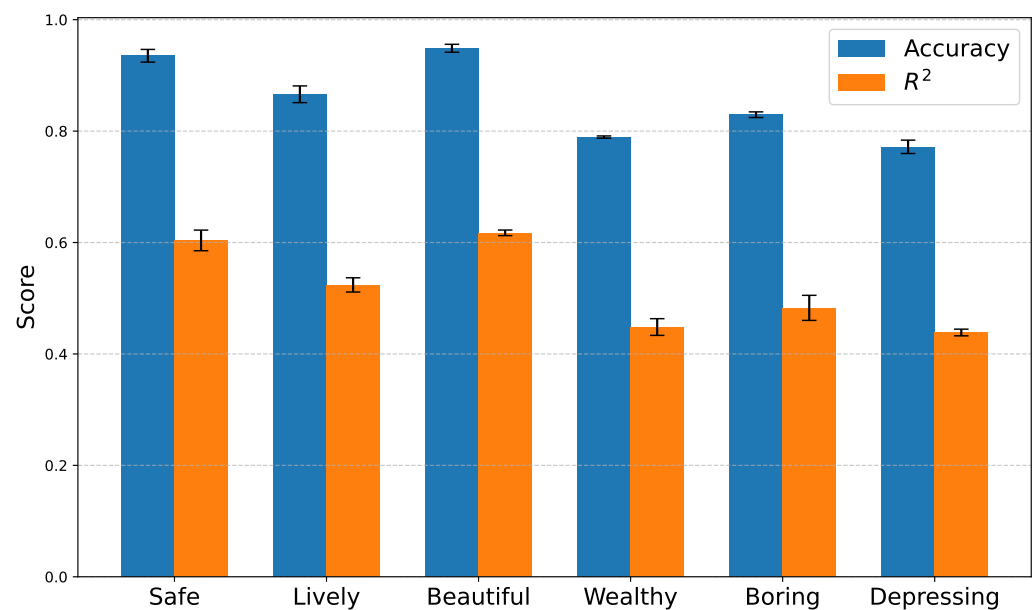


Figure 9. Robustness experiments for UP-CBM on five different training datasets.

Table 8. Robustness experiments for UP-CBM, shown with 95% confidence intervals (CIs).

Dimension	Accuracy (95% CIs)	R ² (95% CIs)
Safe	[0.9210, 0.9494]	[0.5808, 0.6268]
Lively	[0.8475, 0.8847]	[0.5080, 0.5398]
Beautiful	[0.9399, 0.9575]	[0.6112, 0.6236]
Wealthy	[0.7868, 0.7918]	[0.4297, 0.4669]
Boring	[0.8231, 0.8357]	[0.4548, 0.5106]
Depressing	[0.7569, 0.7867]	[0.4311, 0.4459]

5.2. Limitations and Future Work

Despite its advantages, our framework still has several limitations. The concept generation process, although aided by large language models and expert filtering, remains somewhat dependent on the initial prompt design and may introduce biases toward certain visual features. In addition, our reliance on pretrained vision–language models such as CLIP could limit the adaptability of the UP-CBM to highly localized cultural contexts, where visual semantics differ across regions. While multi-scale prompting improves spatial grounding, the synthetic nature of the prompts (e.g., circles) might introduce artifacts or subtle domain shifts. Future work could explore dynamic learned prompting strategies and adapt the concept vocabulary more flexibly across different cultural or environmental settings. Moreover, extending the method to support multimodal inputs (e.g.,

textual or geographic data) could further enhance both perception modeling performance and interpretability.

Beyond methodological considerations, several practical technical bottlenecks could hinder the real-world deployment of the UP-CBM. One major issue is the processing throughput required for large-scale imagery: the framework's multi-scale prompting generates multiple perturbed versions of each image, each requiring a separate forward pass through the CLIP backbone. While acceptable for controlled experiments, this approach becomes computationally intensive when scaled to millions of images, resulting in long inference times and high GPU resource demands. Another challenge lies in data management and interoperability. Urban imagery is drawn from diverse sources—street-view platforms, municipal archives, drone captures, and crowd-sourced uploads—with wide variations in resolution, metadata quality, and licensing restrictions. Building a unified pipeline that normalizes, validates, and maintains these heterogeneous data streams requires considerable engineering effort. Model updating and concept drift also present significant long-term obstacles. Cities evolve rapidly: streets are redesigned, signage is changed, and new structures appear. If the model continues to rely on a fixed concept vocabulary and static weights, its predictions and interpretability may degrade over time, underscoring the need for mechanisms to enable incremental retraining and dynamic concept updates.

6. Conclusions

This study introduces the UP-CBM, a concept-based interpretable framework for urban perception modeling that connects visual features to human-understandable concepts, delivering both transparent reasoning and state-of-the-art performance on the Place Pulse 2.0 and VRVWPR datasets. The innovation of the UP-CBM lies in its integration of three key ideas: a concept bottleneck architecture that grounds predictions in explicit interpretable concepts; a class-free concept discovery process powered by GPT-4o that flexibly generates task-specific visual vocabularies without predefined labels; and a multi-scale visual prompting strategy that probes and localizes concepts across different spatial resolutions, enhancing both interpretability and predictive accuracy. By uniting these components, the UP-CBM not only improves model performance but also provides an interpretable decision pathway that helps urban planners, designers, and researchers to understand how specific visual elements shape human perception of cities. These contributions demonstrate how VFMs and concept-based reasoning can be combined to create a truly transparent human-centered framework for urban perception analysis, and they yield promising directions for future research on cultural adaptability, dynamic concept refinement, and multimodal extensions that further bridge the gap between AI-driven analysis and real-world urban decision-making.

Author Contributions: Conceptualization, Yixin Yu, Bowen Wang, and Jiaxin Zhang; methodology, Yixin Yu, Bowen Wang, and Jiaxin Zhang; software, Zepeng Yu and Bowen Wang; validation, Jiaxin Zhang and Ran Wan; formal analysis, Yixin Yu, Xuhua Shi and Bowen Wang; investigation, Yixin Yu; resources, Jiaxin Zhang; data curation, Zepeng Yu; writing—original draft preparation, Yixin Yu, Xuhua Shi, Jiaxin Zhang, and Zepeng Yu; writing—review and editing, Yixin Yu, Bowen Wang, and Jiaxin Zhang; visualization, Bowen Wang, Xuhua Shi and Zepeng Yu; supervision, Jiaxin Zhang; project administration, Jiaxin Zhang; funding acquisition, Jiaxin Zhang. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Jiangxi Provincial Department of Science and Technology Natural Science Foundation, Grant No. 20242BAB20223.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data and source code are available by contacting corresponding author.

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

References

1. Qiu, W.; Li, W.; Liu, X.; Zhang, Z.; Li, X.; Huang, X. Subjective and objective measures of streetscape perceptions: Relationships with property value in Shanghai. *Cities* **2023**, *132*, 104037. [\[CrossRef\]](#)
2. Wang, L.; Han, X.; He, J.; Jung, T. Measuring residents' perceptions of city streets to inform better street planning through deep learning and space syntax. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 215–230. [\[CrossRef\]](#)
3. Keizer, K.; Lindenberg, S.; Steg, L. The spreading of disorder. *Science* **2008**, *322*, 1681–1685. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Kelling, G.L.; Wilson, J.Q. Broken windows. *Atl. Mon.* **1982**, *249*, 29–38.
5. Zhang, J.; Yu, Z.; Li, Y.; Wang, X. Uncovering Bias in Objective Mapping and Subjective Perception of Urban Building Functionality: A Machine Learning Approach to Urban Spatial Perception. *Land* **2023**, *12*, 1322. [\[CrossRef\]](#)
6. Xue, Y.; Li, C. Extracting Chinese geographic data from Baidu map API. *Stata J.* **2020**, *20*, 805–811. [\[CrossRef\]](#)
7. Wang, B.; Zhang, J.; Zhang, R.; Li, Y.; Li, L.; Nakashima, Y. Improving facade parsing with vision transformers and line integration. *Adv. Eng. Inform.* **2024**, *60*, 102463. [\[CrossRef\]](#)
8. Fan, Z.; Zhang, F.; Loo, B.P.; Ratti, C. Urban visual intelligence: Uncovering hidden city profiles with street view images. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2220417120. [\[CrossRef\]](#)
9. Nasar, J.L. The evaluative image of the city. *J. Am. Plan. Assoc.* **1990**, *56*, 41–53. [\[CrossRef\]](#)
10. Zhang, J.; Hu, J.; Zhang, X.; Li, Y.; Huang, J. Towards a Fairer Green city: Measuring unfairness in daily accessible greenery in Chengdu's central city. *J. Asian Archit. Build. Eng.* **2024**, *23*, 1776–1795. [\[CrossRef\]](#)
11. Yao, Y.; Liang, Z.; Yuan, Z.; Liu, P.; Bie, Y.; Zhang, J.; Wang, R.; Wang, J.; Guan, Q. A human-machine adversarial scoring framework for urban perception assessment using street-view images. *Int. J. Geogr. Inf. Sci.* **2019**, *33*, 2363–2384. [\[CrossRef\]](#)
12. Saleses, M.P. Place Pulse: Measuring the Collaborative Image of the City. Master's Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2012.
13. Saleses, P.; Schechtner, K.; Hidalgo, C.A. The collaborative image of the city: Mapping the inequality of urban perception. *PLoS ONE* **2013**, *8*, e68400. [\[CrossRef\]](#)
14. Awais, M.; Naseer, M.; Khan, S.; Anwer, R.M.; Cholakkal, H.; Shah, M.; Yang, M.H.; Khan, F.S. Foundation Models Defining a New Era in Vision: A Survey and Outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 2245–2264. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, PMLR, Shenzhen, China, 26 February–1 March 2021; pp. 8748–8763.
16. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. Dinov2: Learning robust visual features without supervision. *arXiv* **2023**, arXiv:2304.07193.
17. Shtedritski, A.; Rupprecht, C.; Vedaldi, A. What does clip know about a red circle? Visual prompt engineering for vlms. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 11987–11997.
18. Benou, N.; Chen, L.; Gao, X. SALF-CBM: Spatially-Aware and Label-Free Concept Bottleneck Models. *ICLR* **2025**.
19. Koh, P.W.; Nguyen, T.; Tang, Y.S.; Mussmann, S.; Pierson, E.; Kim, B.; Liang, P. Concept Bottleneck Models. In Proceedings of the 37th International Conference on Machine Learning, PMLR, Online, 13–18 July 2020; Volume 119, pp. 5338–5348.
20. Wang, B.; Li, L.; Nakashima, Y.; Nagahara, H. Learning Bottleneck Concepts in Image Classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023.
21. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. GPT-4 technical report. *arXiv* **2023**, arXiv:2303.08774. [\[CrossRef\]](#)
22. Sun, W.J.; Liu, X.F. Learning Temporal User Features for Repost Prediction with Large Language Models. *Comput. Mater. Contin.* **2025**, *82*, 4117–4136. [\[CrossRef\]](#)
23. Wan, R.; Zhang, J.; Huang, Y.; Li, Y.; Hu, B.; Wang, B. Leveraging diffusion modeling for remote sensing change detection in built-up urban areas. *IEEE Access* **2024**, *12*, 7028–7039. [\[CrossRef\]](#)
24. Li, Y.; Yabuki, N.; Fukuda, T. Measuring visual walkability perception using panoramic street view images, virtual reality, and deep learning. *Sustain. Cities Soc.* **2022**, *86*, 104140. [\[CrossRef\]](#)
25. Liu, J.; Li, L.; Xiang, T.; Wang, B.; Qian, Y. Tera-llm: Token compression retrieval augmented large language model for inference cost reduction. *arXiv* **2023**, arXiv:2310.15556. [\[CrossRef\]](#)
26. Brown, T.B. Language models are few-shot learners. *arXiv* **2020**, arXiv:2005.14165. [\[CrossRef\]](#)
27. Wang, B.; Chang, J.; Qian, Y.; Chen, G.; Chen, J.; Jiang, Z.; Zhang, J.; Nakashima, Y.; Nagahara, H. Direct: Diagnostic reasoning for clinical notes via large language models. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 74999–75011.

28. Han, Y.; Liu, J.; Luo, A.; Wang, Y.; Bao, S. Fine-Tuning LLM-Assisted Chinese Disaster Geospatial Intelligence Extraction and Case Studies. *ISPRS Int. J. Geo-Inf.* **2025**, *14*, 79. [[CrossRef](#)]
29. de Moraes Vestena, K.; Phillipi Camboim, S.; Brovelli, M.A.; Rodrigues dos Santos, D. Investigating the Performance of Open-Vocabulary Classification Algorithms for Pathway and Surface Material Detection in Urban Environments. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 422. [[CrossRef](#)]
30. Cheng, Y.; Li, L.; Xu, Y.; Li, X.; Yang, Z.; Wang, W.; Yang, Y. Segment and track anything. *arXiv* **2023**, arXiv:2305.06558.
31. Andriashen, V.; van Liere, R.; van Leeuwen, T.; Batenburg, K.J. Unsupervised foreign object detection based on dual-energy absorptiometry in the food industry. *J. Imaging* **2021**, *7*, 104. [[CrossRef](#)]
32. Xu, S.; Zhang, J.; Li, Y. Knowledge-driven and diffusion model-based methods for generating historical building facades: A case study of traditional Minnan residences in China. *Information* **2024**, *15*, 344. [[CrossRef](#)]
33. Park, W.; Choi, Y.; Mekala, M.S.; SangChoi, G.; Yoo, K.Y.; Jung, H.y. A latency-efficient integration of channel attention for ConvNets. *Comput. Mater. Contin.* **2025**, *82*, 3965–3981. [[CrossRef](#)]
34. Wang, H.; Zhang, Y.; Zhu, C. YOLO-LFD: A Lightweight and Fast Model for Forest Fire Detection. *Comput. Mater. Contin.* **2025**, *82*, 3399–3417. [[CrossRef](#)]
35. Zhang, K.; Liu, D. Customized segment anything model for medical image segmentation. *arXiv* **2023**, arXiv:2304.13785. [[CrossRef](#)]
36. Shen, Q.; Yang, X.; Wang, X. Anything-3d: Towards single-view anything reconstruction in the wild. *arXiv* **2023**, arXiv:2304.10261.
37. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–23 June 2022; pp. 16000–16009.
38. Zhang, J.; Wang, B.; Li, L.; Nakashima, Y.; Nagahara, H. Instruct me more! random prompting for visual in-context learning. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2024; pp. 2597–2606.
39. Yan, Y.; Wen, H.; Zhong, S.; Chen, W.; Chen, H.; Wen, Q.; Zimmermann, R.; Liang, Y. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In Proceedings of the ACM Web Conference 2024, Singapore, 13–17 May 2024; pp. 4006–4017.
40. Hao, X.; Chen, W.; Yan, Y.; Zhong, S.; Wang, K.; Wen, Q.; Liang, Y. UrbanVLP: Multi-granularity vision-language pretraining for urban socioeconomic indicator prediction. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 27 February–4 March 2025; Volume 39, pp. 28061–28069.
41. Yang, J.; Ding, R.; Brown, E.; Qi, X.; Xie, S. V-irl: Grounding virtual intelligence in real life. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; Springer: Cham, Switzerland, 2024; pp. 36–55.
42. Zhang, J.; Fukuda, T.; Yabuki, N. Development of a city-scale approach for façade color measurement with building functional classification using deep learning and street view images. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 551. [[CrossRef](#)]
43. Naik, N.; Philipoom, J.; Raskar, R.; Hidalgo, C. Streetscore-predicting the perceived safety of one million streetscapes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 23–28 June 2014; pp. 779–785.
44. Naik, N.; Raskar, R.; Hidalgo, C.A. Cities are physical too: Using computer vision to measure the quality and impact of urban appearance. *Am. Econ. Rev.* **2016**, *106*, 128–132. [[CrossRef](#)]
45. Griew, P.; Hillsdon, M.; Foster, C.; Coombes, E.; Jones, A.; Wilkinson, P. Developing and testing a street audit tool using Google Street View to measure environmental supportiveness for physical activity. *Int. J. Behav. Nutr. Phys. Act.* **2013**, *10*, 103. [[CrossRef](#)]
46. Halpern, D. *Mental Health and the Built Environment: More than Bricks and Mortar?* Routledge: Abingdon-on-Thames, UK, 2014.
47. Wang, B.; Li, L.; Verma, M.; Nakashima, Y.; Kawasaki, R.; Nagahara, H. Match them up: Visually explainable few-shot image classification. *Appl. Intell.* **2023**, *53*, 10956–10977. [[CrossRef](#)]
48. Ghorbani, A.; Wexler, J.; Zou, J.; Kim, B. Towards Automatic Concept-based Explanations. In Proceedings of the Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019.
49. Ge, S.; Zhang, L.; Liu, Q. Robust Concept-based Interpretability with Variational Concept Embedding. In Proceedings of the Conference on Neural Information Processing Systems (NeurIPS), Online, 6–14 December 2021.
50. Laugel, T.; Lesot, M.J.; Marsala, C.; Renard, X.; Detyniecki, M. The Dangers of Post-hoc Interpretability: Unjustified Counterfactual Explanations. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Macao, China, 10–16 August 2019.
51. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
52. Zhou, Z.; Zhao, Y.; Zuo, H.; Chen, W. Ranking Enhanced Supervised Contrastive Learning for Regression. In Proceedings of the Advances in Knowledge Discovery and Data Mining, Taipei, Taiwan, 7–10 May 2024; pp. 15–27.
53. Dai, W.; Li, X.; Chiu, W.H.K.; Kuo, M.D.; Cheng, K.T. Adaptive contrast for image regression in computer-aided disease assessment. *IEEE Trans. Med Imaging* **2021**, *41*, 1255–1268. [[CrossRef](#)] [[PubMed](#)]

54. Dai, W.; Li, X.; Cheng, K.T. Semi-supervised deep regression with uncertainty consistency and variational model ensembling via bayesian neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 7304–7313.
55. Zhang, J.; Li, Y.; Fukuda, T.; Wang, B. Urban safety perception assessments via integrating multimodal large language models with street view images. *Cities* **2025**, *165*, 106122. [[CrossRef](#)]
56. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
57. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16×16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4–8 May 2021.
58. Alvarez Melis, D.; Jaakkola, T. Towards robust interpretability with self-explaining neural networks. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 7786–7795.
59. Chen, C.; Li, O.; Tao, D.; Barnett, A.; Rudin, C.; Su, J.K. This looks like that: Deep learning for interpretable image recognition. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 8930–8941.
60. Yuksekgonul, M.; Wang, M.; Zou, J. Post-hoc concept bottleneck models. *arXiv* **2022**, arXiv:2205.15480.
61. Oikarinen, T.; Das, S.; Nguyen, L.M.; Weng, T.W. Label-free concept bottleneck models. *arXiv* **2023**, arXiv:2304.06129. [[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.