Urban Perception Analysis via Transformer-based Image Segmentation: A Deep Dive into City Understanding

Nam-Hai Hoang Le[1] and Thanh-Danh Nguyen[†2,3]

[1]University of Economics Ho Chi Minh City, Vietnam
[2]University of Information Technology, Ho Chi Minh City, Vietnam
[3]Vietnam National University, Ho Chi Minh City, Vietnam
haihln@ueh.edu.vn, danhnt@uit.edu.vn, [†]*corresponding author*

***Abstract*—Urban analytics recently relies on visual data to understand the complex relationships between physical environments and socioeconomic conditions. The key challenge is how to leverage such abundant visual data to investigate the urban perception effectively. We propose a novel dual-track framework, dubbed UPFormer, that employs the diffusion-based image segmentation architectures alongside urban perception classification models to analyze $2,680$ Street View Images (SVIs) from Nha Trang, Vietnam. By exploring SVIs from diverse regions, UPFormer exploits the urban perception classification via the semantic relationship among the components captured in the images automatically. The integration of these two streams enables a robust statistical analysis linking visual urban elements to perceived socioeconomic traits, providing interpretable insights into how physical features influence urban perception. To this end, evaluated on the Place Pulse v2 benchmark and a custom Nha Trang dataset, our approach achieves statistically significant correlations ($p < 0.05$) between visual elements and socioeconomic perceptions, offering a scalable, data-driven tool for human-centric urban governance in the Global South. Code can be found at https://github.com/danhntd/UPFormer.**

***Index Terms*—Urban Perception Analysis, Image Segmentation, Transformer-based Architecture.**

I. Introduction

By 2050, 68% of the world's population will live in cities requiring urban analytics to capture the human experience of city. Urban perception - emotions like safety, lively, or wealth - offers a vital insights to have resilient cities [1]–[3]. Unlike coarse satellite imagery, Street View Imagery (SVI) captures street-level details, such as sidewalks, vegetation, and buildings. In Vietnam, rapid urbanization amplifies socioeconomic disparities, but time-consuming manual surveys and coarse satellite data miss cultural nuances, while fragmented analytics prioritize either infrastructure detection (e.g., potholes) or socioeconomic estimation (e.g., property values), rarely integrating both [4]. Cultural biases in perception data further limit generalizability in non-Western contexts. We focus on Nha Trang, a city reflecting Southeast Asia's diverse urban fabric.

Research question

- How do urban features (e.g., roads, vegetation) influence socioeconomic perceptions (e.g., "wealthy") in Nha Trang?
- How can transformer-based frameworks scalably unify objective and subjective urban analysis?

We propose UPFormer, a dual-track framework that processes SVI to extract objective urban features and subjective perceptions, offering a comprehensive view of urban environments. The task is defined as follows:

- **Input:** We use $2,680$ high-resolution SVI images (from Mapillary and Google Street View (GSV)) from Nha Trang (2019–2024), covering diverse zones.
- **Output:** The framework delivers (1) semantic segmentation maps at the pixel level identifying urban features (e.g. roads, buildings, vegetation) with mean Intersection over Union (mIoU) > 0.75; (2) perception classification scores for categories like "wealthy" or "lively" with $>$ 70 percent accuracy; (3) statistical correlations ($p < 0.05$) linking segmented features to perception labels.
- **Processing:** A transformer-based pipeline using OneFormer for semantic segmentation (pre-trained on Cityscapes/ADE20K) and a Vision Transformer (ViT) fine-tuned on Place Pulse v2 for perception classification, integrated via OLS regression and SHAP values

To address cultural biases, we supplement Place Pulse v2 with a localized Nha Trang dataset and use SHAP values for model interpretability, ensuring insights are context-specific and reducing the risk of perpetuating stereotypes.

Recent studies highlight SVI-based urban analytics, but remain fragmented. [5] used CNN-based feature extraction from GSV to identify urban elements with approx. 80 percent accuracy, yet omitted perception analysis. [6] employed manual SVI audits for accessibility planning, improving infrastructure design, but lacking scalability. Transformer-based models like Segmenter (mIoU $\approx$ 0.78) and SegFormer (mIoU $\approx$ 0.80) excel in urban scene segmentation but are not tailored for perceptual tasks. Mask2Former (mIoU ≈ 0.82) unifies segmentation tasks but remains limited to instance-level analysis without socioeconomic integration.

To bridge this gap, we propose a unified approach using transformers - in which the self-attention mechanism is used. This AI-driven approach addresses Vietnam's scalability challenges, using Nha Trang as a case study to pioneer data-driven urban analytics in the Global South.

To summarize, our contributions are:

- First - UPFormer: Transformer-based urban perception analysis for decision-making.
- Second - Localized dataset: 2,680 Nha Trang SVIs.

II. Related Work

A. Urban Perception Analysis

Overview of Urban Analytics and Perception. Urban analytics has become an essential discipline for understanding and managing the complexities of urban environments, driven by the need for detailed, scalable, and actionable data to support urban planning and policy [6], [7]. Traditional satellite imagery, while useful for land use and sprawl analysis, lacks the granularity to capture human-scale experiences critical for holistic urban studies [2], [8].

Urban perception research, focusing on emotions like safety, liveliness, or wealth, has shifted from labor-intensive street audits to AI-driven frameworks. Street View Imagery (SVI) has emerged as a vital tool, offering rich, human-scale data via platforms like Google Street View (GSV) and Mapillary. Recent scholarship has harnessed SVI to map these subjective experiences over time. For example, [8] used SVI in Singapore to uncover temporal shifts in perceptions, linking revitalized districts' "liveliness" to urban renewal, while noting persistent "depressing" sentiments in neglected areas. Similarly, [3], synthesized AI-SVI applications to chart emotional landscapes, scaling beyond traditional surveys. [6] comprehensive review of more than 100 studies underscores the potential of AI to model complex emotions like stress, but notes that single-dimensional analyses (e.g., safety alone) fail to capture the multifaceted nature of urban experiences. [4] revealed demographic influences, with women and introverts perceiving streets as less "safe", highlighting the need for inclusive design. These studies underscore SVI's potential but lack integration of objective feature extraction with subjective perception, particularly in the Global South.

Datasets. In the evolving field of urban analytics, the Place Pulse v2 dataset [1] stands as a pivotal resource for quantifying subjective urban perceptions through visual data. This dataset comprises over 1.17 million pairwise comparisons of approximately 110,988 street-level images sourced from GSV across cities worldwide, annotated with perception scores across six dimensions: safety, liveliness, boredom, wealth, depression, and beauty. The dataset's utility lies in its ability to support deep learning models that predict urban perceptions based on visual cues [9], [10]. However, the dataset has limitations that contextualize its use in this study. Its reliance on pairwise comparisons can introduce subjective biases, as perceptions may vary as a challenge noted in [4] for global-scale analyses. Additionally, while Place Pulse v2 [1] covers diverse cities, its coverage in non-Western contexts like Vietnam is sparse, necessitating supplementary local datasets for comprehensive analyses.

B. Image Segmentation for Urban Analysis.

The increasing reliance on technologies such as autonomous driving [11], intelligent surveillance [12], and urban analytics [9] underscores the need for precise scene understanding. Urban environments, characterized by complex and dynamic elements such as vehicles, pedestrians, and infrastructure [13], demand segmentation-based approaches for effective analysis. Semantic segmentation provides high-level comprehension by assigning each pixel to meaningful categories [14], and has traditionally been modeled using CNNs. More recently, transformer-based architectures [15], [16], building on advances in computer vision and natural language processing [17], [18], have achieved strong performance. However, semantic segmentation struggles in urban contexts due to difficulties in separating individual objects of the same class. Instance segmentation addresses this limitation by delineating precise object boundaries, with methods ranging from two-stage approaches [19]–[21] to efficient one-stage variants [5], [22], [23], [23], [24]. Recent unified frameworks such as OneFormer [25] combine semantic, instance, and panoptic segmentation into a single model, while FastInst [5] improves real-time performance. As cities grow denser, advancing such segmentation methods remains vital for building robust, real-time urban perception systems [26]. In this work, we adopt OneFormer [25] as the baseline for urban scene analysis.

Datasets. Dataset like Cityscapes [13] provides high-resolution urban street scenes with annotations. And ADE20K [27] covering diverse scenes, serves as a standard benchmark for scene parsing. These datasets collectively drive advances in segmentation models but are Western-centric, underscoring the need for localized datasets like ours. [2] applied SVI for urban feature extraction but lacked perception integration, reinforcing our study's focus on unified analytics.

C. AI-based Urban Perception Analysis

AI-driven urban perception leverages segmentation to quantify subjective experiences from SVI. Early

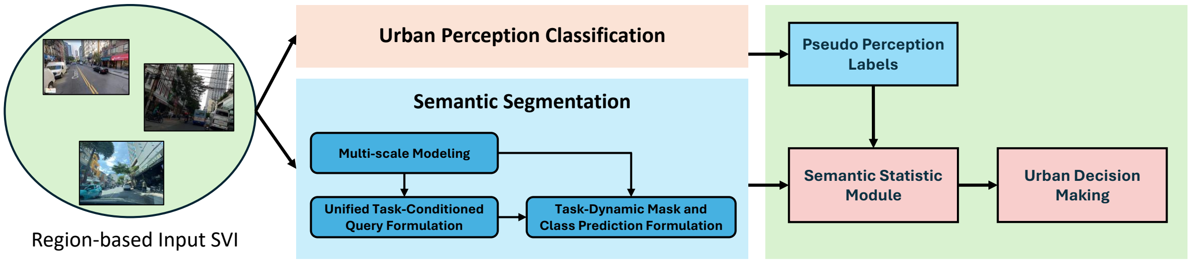
Urban Perception Classification
Semantic Segmentation
Multi-scale Modeling
Unified Task-Conditioned Query Formulation
Task-Dynamic Mask and Class Prediction Formulation
Pseudo Perception Labels
Semantic Statistic Module
Urban Decision Making
Region-based Input SVI

Fig. 1. Overview of our proposed UPFormer, an urban perception analysis framework via Transformer-based image segmentation.

works used crowdsourced ratings to predict attributes like safety or beauty [28], [29]. Recent studies employ deep learning, with Place Pulse [28], [30] enabling scalable perception modeling, though limited in non-Western contexts [31]. Our framework advances this by integrating OneFormer's pixel-level feature extraction with ViT-based perception classification, tailored to Nha Trang's 2,680 SVIs, to support equitable smart city planning, public policy, and urban design.

III. Proposed Framework

A. Overview

Figure 1 illustrates the pipeline for our urban analysis framework, named after UPFormer, beginning with inputting SVIs from multiple locations. Each geometric region contributes around 100 SVIs, forming the foundational input for the pipeline. From this point, the images are processed along two parallel tracks. The first track is responsible for predicting high-level urban perception categories (i.e., "Lively", "Wealthy", "Beautiful", "Boring", "Safety", or "Depressing") using a pre-trained classification model inherited from [2] built on top of MaxViT [32]. The second track sends the same images through a semantic segmentation model [25] to extract semantic pixel-level labels across various visual categories, followed by the definition of the Cityscapes benchmark [13]. Ultimately, these two tracks are aligned by analyzing the statistical distributions of segmented objects in relation to the predicted urban perception classes, enabling a deeper understanding of how visual urban elements contribute to perceived socioeconomic traits.

B. Urban Perception Classification

This urban perception classification module is designed to leverage the pretrained models [2] built on top of MaxViT [32]. The classification is well trained on the Place Pulse v2 [1], [33] and fine-tuned on the dataset of Nha Trang. This module is responsible for predicting the pseudo perception labels for the region-based input SVI, which are in six aforementioned types. This model learns to associate input images with one subjective urban perception with $> 70\%$ accuracy. The output from this track not only labels each image with a high-level urban concept but also serves as a reference for aligning and interpreting the statistical outputs of the segmentation model in the dual stage.

C. Urban Scene Understanding

In parallel with the classification task, each SVI undergoes processing through a semantic segmentation model, which decomposes the image into a predefined set of visual categories such as roads, vehicles, buildings, pedestrians, vegetation, and sky. Such categories follow the definition in the annotation of Cityscapes [13]. The segmentation model assigns each pixel to one of these categories, resulting in a dense, pixel-wise annotation map (mIoU > 0.75). Pixel counts for classes like roads or buildings are then aggregated and stored as structured numerical data. This second track utilizes OneFormer [25] to extract semantic pixel-level labels across various visual categories, which is a state-of-the-art model based on the current Transformer architecture.

D. Urban Statistics

By aligning these semantic statistics with the urban perception labels predicted from the classification track, the pipeline enables a correlation-based analysis that can uncover meaningful patterns and relationships. For example, one might observe that a higher proportion of vehicles or expansive road surfaces is commonly associated with areas labeled as "Wealthy" or "Luxury," whereas a cluttered or vehicle-scarce scene might be more indicative of "Poor" or "Unsafe" labels. This statistical alignment not only provides a quantitative basis for interpreting subjective urban qualities but also supports exploratory research and model explainability. Importantly, this approach allows for nuanced urban understanding even in the absence of explicit geographic metadata, making it scalable and adaptable to datasets without location-based annotations. To this end, our UPFormer support the downstream urban decision making tasks in urban studies.

E. Statistical Analysis and Correlation of Urban Perception

For each segmentation mask, we summarize the semantic composition of the image by counting the number of pixels belonging to each semantic category

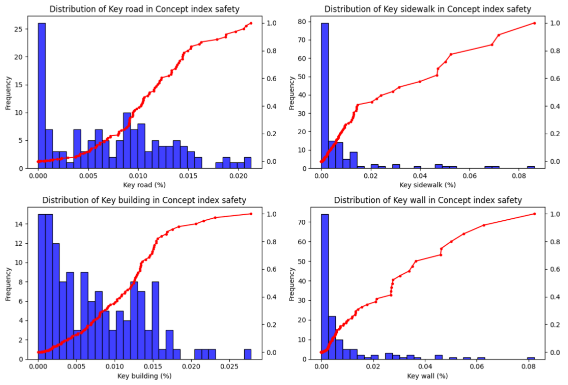

Fig. 2. Statistics on the effect of four exemplary selected key factors, including "Road", "Sidewalk", "Building", and "Wall", on the key concept "Safety".

$c \in \mathcal{C}$, where $\mathcal{C}$ follows the Cityscapes [13] annotation set. Let $M(i,j)$ denote the predicted semantic label at pixel (i,j). The total count of pixels for category c is computed as in Equation 1:

$$N_c = \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbb{1}\big(M(i,j) = c\big), \tag{1}$$

where H and W are the height and width of the image, and $\mathbb{1}(\cdot)$ is the indicator function.

The resulting vector $\mathbf{N} = (N_1, N_2, \ldots, N_{|\mathcal{C}|})$ directly represents the semantic pixel statistics of each image. These statistics are then aggregated across all images belonging to the same predicted perception category (e.g., "Safety", "Wealthy"). The aggregation can be summarized in tabular form to report average counts per category, or visualized using confusion-matrix-style representations to illustrate how certain semantic categories are distributed with respect to different perception labels. This provides an interpretable link between the pixel-level physical structure of urban scenes and their subjective perception.

IV. Experimental Results

A. Configurations Setup

Our UPFormer uses a ViT model pre-trained [2], [32] on Place Pulse v2 [1], [33] dataset to infer subjective urban perception categories and fine-tuned for Nha Trang, and a OneFormer model pre-trained on Cityscapes [13], [25], run on a single NVIDIA RTX 3090 Ti GPU. The configuration parameters follow the original publications.

B. SVIs Data Collection

To investigate urban perception in Nha Trang, Vietnam, this study employs a dual-track framework integrating semantic segmentation and perception classification, leveraging a custom dataset of $2,680$ SVIs and the Place Pulse v2 dataset [1]. The data collection process is tailored to Nha Trang's diverse urban contexts, covering 2019–2024 to ensure recent, high-quality imagery. This supports the study's objectives of extracting urban features and quantifying socioeconomic perceptions, addressing the research gap in unified objective-subjective urban analytics.

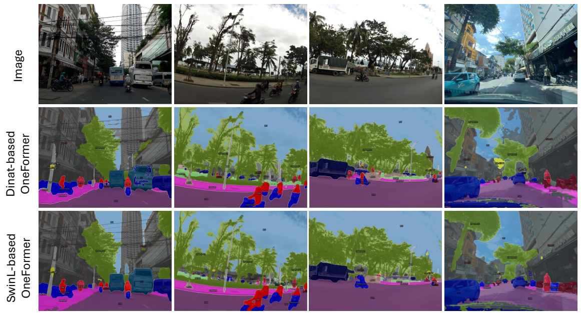

Fig. 3. Exemplary visualization on different segmentation backbones of OneFormer [25] on the collected SVIs.

The Place Pulse v2 dataset, developed by [1], is used to train the ViT for urban perception classification, ensuring generalizability before local adaptation.

A custom dataset of $2,680$ SVIs was collected, including $2,391$ images from Mapillary and 289 from GSV, covering 2019–2024 to reflect current urban conditions in Nha Trang. Sample points were generated every 300 meters along OSM-derived residential, tertiary, and service roads, ensuring coverage of diverse districts. Mapillary images, retrieved via API, emphasize pedestrian zones, while GSV images provide high-quality panoramas. Images were standardized to 1280 × 720 pixels, split into train (70%, $1,876$ images), validate (15%, 402 images), and test (15%, 402 images) sets, with augmentation to enhance robustness. Mapillary's dominance (89%) ensures pedestrian-centric coverage, mitigated by stratified sampling to balance district representation.

C. Evaluation Metrics and Validation Strategy

The dual tracks of UPFormer are built on top of a classification model and a semantic segmentation model. Thus, we straightforwardly follow the evaluation metrics of each model. For the urban perception classification module, we adopt the standard *accuracy* metric, which measures the proportion of correctly predicted perception labels among the six predefined categories. For the semantic segmentation track, we follow the common practice in semantic scene understanding and report both the *mIoU* and the overall *pixel accuracy*. To validate the effectiveness of UPFormer, we conduct experiments on the collected real-life SVIs datasets and conduct urban analysis on the real data.

D. Quantitative Evaluation and Interpretation

Our research establishes a quantitative foundation for urban perception analysis by providing an under-

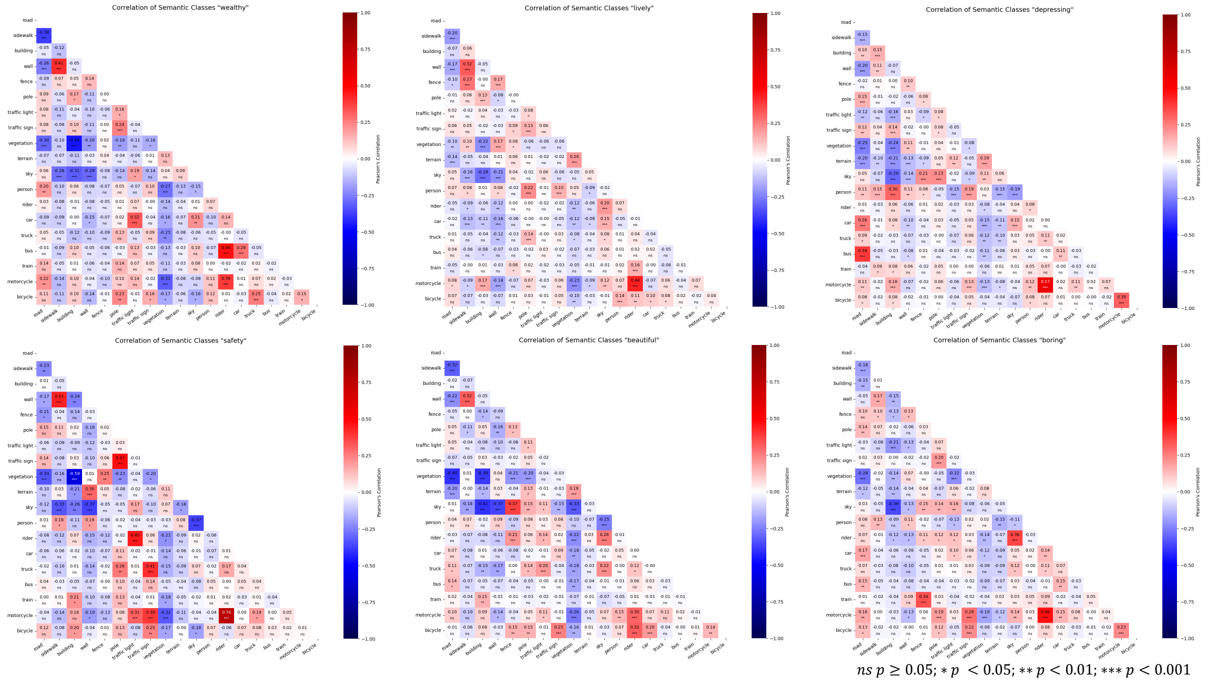

Fig. 4. Correlation matrix showing the relationships between perceived urban characteristics and urban elements across 2, 680 street-level images. The matrix includes six perception categories - "Lively", "Wealthy", "Beautiful", "Boring", "Safe", and "Depressing" - each evaluated in relation to the presence of urban elements. Cell values represent Pearson correlation coefficients, with corresponding significance levels ($p - values$). Only statistically significant correlations ($p < 0.05$) are shown/highlighted. Please zoom in for details.

standing of how specific visual features shape human perspectives. The UPFormer framework, which merges semantic segmentation with a perception classification pipeline, has been assessed using the Place Pulse v2 benchmark [1]. The core findings are articulated through three distinct figures, each offering a unique layer of insight.

The correlation matrices (Figure 4, $n = 2680$) provide a direct, interpretable link between semantic classes and urban perceptions. For instance, the "wealthy" perceptions show buses and riders correlate positively ($r = +0.46$, $p < 0.001$), indicating provided transport hubs enhance perceived affluence, while walls and sidewalks ($r = +0.41$, $p < 0.001$) suggest exclusive spaces. Conversely, vegetation and buildings correlate negatively ($r = -0.45$, $p < 0.001$), implying green, open areas or dense urban structures drive wealth perceptions. Similarly, "lively" perception class also shows postive correlation between wall and sidewalks ($r = +0.32$, $p < 0.001$). People and poles correlate positively ($r = +0.22$, $p < 0.001$), suggesting human activity paired with urban infrastructure enhances vibrancy, while motorcycles and riders ($r = +0.44$, $p < 0.001$) indicate dynamic mobility contributes to lively perceptions. Conversely, "boring" perceptions correlate negatively with sky relative to building ($r = -0.36$, $p < 0.001$), indicating sparse or overly built-up scenes foster monotony, while motorcycle and riders show a postitive correlation ($r = +0.46$, $p < 0.001$) indicating routine clusters contribute to dullness. Regarding "Beautiful" scenes, they correlate strongly with vegetation ($r = +0.28$, $p < 0.001$) and fences ($r = +0.13$, $p < 0.01$), suggesting green, orderly spaces enhance aesthetics. "Safe" perceptions align with sidewalks ($r = +0.43$, $p < 0.001$) and terrain ($r = +0.30$, $p < 0.001$).

These numerical values, which are normalized to highlight key relationships while maintain The consistency in linking between object presence and perceptual outcome.

The integrity of our quantitative analysis replies on the precision of our semantic segmentation. Figure 3 showcases the segmentation results from two different transformer-based models: the SwinL-based OneFormer and our proposed Dinat-based OneFormer. The side-by-side view demonstrates a clear difference in output quality. Our Dinat-based model consistently produces more accurate and cleaner segmentation masks, with sharper boundaries between objects like roads, buildings, and vegetation. This high-fidelity output is indispensable, as the accuracy of our semantic maps directly governs the reliability of the subsequent correlation analysis. The visual evidence confirms that our framework is built on a solid foundation of precise data extraction.

To further investigate our findings, Figure 2 illustrates the frequency distribution of four key semantic classes—road, sidewalk, building, and wall—in the context of the urban perception of "safety." The histograms and cumulative distribution functions show how the prevalence of these features varies across the dataset. The analysis reveals that the majority of images contain a low percentage of these features, a crucial insight for understanding the typical composition of urban environments in our data. This statistical approach can be extended to other perception categories, allowing us to extract more profound insights into how feature prevalence contributes to different urban perceptions.

E. Limitations

The reliance of the dataset on Mapillary introduces variable image quality, while the limited Vietnamese coverage in Place Pulse v2 necessitates fine-tuning. Cultural biases in perception labels may affect generalizability. Stratified sampling and augmentation mitigate these, but future work should explore cross-city validation and cultural adaptation.

V. Conclusion

In this work, we introduce *UPFormer*, a dual-track framework that integrates diffusion-based image segmentation with urban perception classification to advance image-driven urban analytics. By jointly analyzing street view imagery at both the pixel level and the perceptual level, the framework establishes meaningful links between physical urban elements and subjective socioeconomic perceptions. Experiments on the Place Pulse v2 benchmark confirm the effectiveness of our approach, demonstrating that UPFormer can provide interpretable and scalable insights into how the built environment shapes human perception of cities.

VI. Acknowledgements

This paper is a product of a university-level research project code CTD-CS-2025-05 funded by the University of Economics Ho Chi Minh City.

References

[1] A. Dubey, N. Naik, D. Parikh, R. Raskar, and C. A. Hidalgo, “Deep learning the city: Quantifying urban perception at a global scale,” in *IEEE/CVF ECCV*, pp. 196–212, Springer, 2016.

[2] Y. Hou, M. Quintana, *et al.*, “Global streetscapes – a comprehensive dataset of 10 million street-level images across 688 cities for urban science and analytics,” *Journal of Photogrammetry and Remote Sensing*, vol. 215, pp. 216–238, 2024.

[3] Y. Kang, S. Gao, and M. Helbich, “Urban visual intelligence: Studying cities with artificial intelligence and street-level imagery,” *Annals of the American Association of Geographers*, vol. 114, no. 5, pp. 876–897, 2023.

[4] M. Quintana *et al.*, “It's not you, it's me – global urban visual perception varies across demographics and personalities,” *Cities*, vol. forthcoming, p. arXiv:2505.12758, 2025.

[5] J. He, P. Li, Y. Geng, and X. Xie, “Fastinst: A simple query-based model for real-time instance segmentation,” in *IEEE/CVF CVPR*, pp. 23663–23672, 2023.

[6] K. Ito, “Understanding urban perception with visual data: A systematic review,” *Cities*, vol. 152, p. 105140, 2024.

[7] Z. Wang, K. Ito, and F. Biljecki, “Assessing the equity and evolution of urban visual perceptual quality with time series street view imagery,” *Cities*, vol. 145, p. 104704, 2024.

[8] X. Liang, T. Zhao, and F. Biljecki, “Revealing spatio-temporal evolution of urban visual environments with street view imagery,” *Lands. and Urban Planning*, vol. 238, p. 104812, 2023.

[9] D. Zhang, Y. Zheng, S. Qi, B. Li, and L. Ma, “Urban computing: concepts, methodologies, and applications,” *ACM TISIT*, vol. 8, no. 5, pp. 1–19, 2017.

[10] T. Zhao and F. Biljecki, “Physical urban change and its socio-environmental impact: Insights from street view imagery,” *Computers, Environment and Urban Systems*, vol. forthcoming, p. 102037, 2025. Forthcoming.

[11] C. Badue, R. Guidolini, *et al.*, “Self-driving cars: A survey,” *Expert Systems with Applications*, vol. 165, p. 113816, 2021.

[12] C. Chen, Y. Li, and L. Chen, “Intelligent video surveillance: A review through deep learning techniques for crowd analysis,” *MTAP Journal*, vol. 79, no. 33-34, pp. 21753–21780, 2020.

[13] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, *et al.*, “The cityscapes dataset for semantic urban scene understanding,” in *IEEE/CVF CVPR*, pp. 3213–3223, 2016.

[14] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *IEEE/CVF CVPR*, pp. 3431–3440, 2015.

[15] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, “Segmenter: Transformer for semantic segmentation,” in *IEEE/CVF ICCV*, pp. 7262–7272, 2021.

[16] E. Xie, W. Wang, Z. Yu, *et al.*, “Segformer: Simple and efficient design for semantic segmentation with transformers,” *NeurIPS*, vol. 34, pp. 12077–12090, 2021.

[17] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *IEEE/CVF ECCV*, pp. 213–229, Springer, 2020.

[18] A. Vaswani, N. Shazeer, *et al.*, “Attention is all you need,” *NeurIPS*, vol. 30, 2017.

[19] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *IEEE/CVF ICCV*, pp. 2961–2969, 2017.

[20] Z. Cai *et al.*, “Cascade r-cnn: Delving into high quality object detection,” in *IEEE/CVF CVPR*, pp. 6154–6162, 2018.

[21] S. Liu, L. Qi, *et al.*, “Path aggregation network for instance segmentation,” in *IEEE/CVF CVPR*, pp. 8759–8768, 2018.

[22] D. Bolya, C. Zhou, *et al.*, “Yolact: Real-time instance segmentation,” in *IEEE/CVF ICCV*, pp. 9157–9166, 2019.

[23] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, “Masked-attention mask transformer for universal image segmentation,” in *IEEE/CVF CVPR*, pp. 1290–1299, 2022.

[24] T.-D. Nguyen, D.-T. Luu, V.-T. Nguyen, and T. D. Ngo, “Ce-ost: Contour emphasis for one-stage transformer-based camouflage instance segmentation,” in *IEEE MAPR*, pp. 1–6, IEEE, 2023.

[25] J. Jain, J. Li, M. T. Chiu, A. Hassani, N. Orlov, and H. Shi, “Oneformer: One transformer to rule universal image segmentation,” in *IEEE/CVF CVPR*, pp. 2989–2998, 2023.

[26] D. Feng, C. Haase-Schuetz, L. Rosenbaum, *et al.*, “Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges,” *IEEE Trans. on ITS*, vol. 22, no. 3, pp. 1341–1360, 2020.

[27] B. Zhou *et al.*, “Scene parsing through ade20k dataset,” in *IEEE/CVF CVPR*, pp. 633–641, 2017.

[28] P. Salesses, K. Schechtner, and C. A. Hidalgo, “The collaborative image of the city: Mapping the inequality of urban perception,” *PloS one*, vol. 8, no. 7, p. e68400, 2013.

[29] N. Naik, J. Philipoom, R. Raskar, and C. A. Hidalgo, “Streetscore—predicting the perceived safety of one million streetscapes,” in *IEEE/CVF CVPRW*, pp. 779–785, 2014.

[30] N. Naik, S. D. Kominers, R. Raskar, E. L. Glaeser, and C. A. Hidalgo, “Computer vision uncovers predictors of physical urban change,” *PNAS*, vol. 114, no. 29, pp. 7571–7576, 2017.

[31] N. Yang *et al.*, “Urban perception assessment from street view images based on a multifeature integration encompassing human visual attention,” *Annals of the American Association of Geographers*, vol. 114, no. 7, pp. 1424–1442, 2024.

[32] Z. Tu, H. Talebi, *et al.*, “Maxvit: Multi-axis vision transformer,” in *IEEE/CVF ECCV*, pp. 459–479, Springer, 2022.

[33] P. Salesses, K. Schechtner, and C. A. Hidalgo, “The collaborative image of the city: Mapping the inequality of urban perception,” *PloS one*, vol. 8, no. 7, p. e68400, 2013.