IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 March 2024, accepted 30 May 2024, date of publication 3 June 2024, date of current version 28 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3408843

RESEARCH ARTICLE

Automatic Estimation for Visual Quality Changes of Street Space via Street-View Images and Multimodal Large Language Models

HAO LIANG[1], JIAXIN ZHANG[2,3], YUNQIN LI[2,3], BOWEN WANG[4], (Member, IEEE), AND JINGYONG HUANG[2]

[1]College of Landscape Architecture, Nanjing Forestry University, Nanjing 210037, China
[2]Architecture and Design College, Nanchang University, Nanchang 330031, China
[3]Division of Sustainable Energy and Environmental Engineering, Graduate School of Engineering, Osaka University, Osaka 565-0871, Japan
[4]Graduate School of Information Science and Technology, Osaka University, Osaka 565-0871, Japan

Corresponding author: Jingyong Huang (huangjingyong@ncu.edu.cn)

This work was supported in part by the Key Research Base of Humanities and Social Sciences in Jiangxi Universities, in 2023, under Project JD23003.

ABSTRACT Estimating Visual Quality of Street Space (VQoSS) is pivotal for urban design, environmental sustainability, civic engagement, etc. Recent advancements, notably in deep learning, have enabled large-scale analysis. However, traditional deep learning approaches are hampered by extensive data annotation requirements and limited adaptability across diverse VQoSS tasks. Multimodal Large Language Models (MLLMs) have recently demonstrated proficiency in various computer vision tasks, positioning them as promising tools for automated VQoSS assessment. In this paper, we pioneer the application of MLLMs to VQoSS change estimation, with our empirical findings affirming their effectiveness. In addition, we introduce Street Quality Generative Pre-trained Transformer (SQ-GPT), a model that distills knowledge from the current most powerful but inaccessible (not free) GPT-4V, requiring no human efforts. SQ-GPT approaches GPT-4V's performance and is viable for large-scale VQoSS change estimation. In a case study of Nanjing, we showcase the practicality of SQ-GPT and knowledge distillation pipeline. Our work promises to be a valuable asset for future urban studies research.

INDEX TERMS Smart city, visual quality, deep learning, multimodal large language models.

I. INTRODUCTION

Urban landscapes, encompassing a tapestry of streets, parks, and public squares, are more than mere channels of movement; they are the lifeblood of city vibrancy and communal interaction. Within this urban fabric, the aesthetic, functional, and social dimensions of street spaces emerge as pivotal to public well-being and community cohesion [1], [2]. They are not just static structures but vibrant stages where the urban narrative unfolds over time. The temporal evolution of street spaces is critical to understanding and fostering urban vitality, as they are the sites where the past converses with the present and shapes the future [3]. This recognition has sparked a series of urban design initiatives [4], underscoring the global movement towards enhancing the Visual Quality of Street Space (VQoSS) [5], [6].

The concept of VQoSS is not a contemporary invention. Its origin can be traced back to the advocacy for community-centric urban design in the mid-20th century [7]. Despite their contributions, historical limitations in data collection confined such research to the realm of small-scale studies, impeding the development of empirically grounded urban design principles that could be applied broadly. Urban planners face the complex challenge of estimating VQoSS changes as they endeavor to shape environments that align with human experiences amid the continuous progression of urban development [8].

In an era where technological advancements have revolutionized data accessibility and analytical capabilities, tools like Geographic Information Systems (GIS) for street-view

The associate editor coordinating the review of this manuscript and approving it for publication was Liangxiu Han.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

images [9], [10], remote sensing technologies [11], [12], and deep learning based methods [13], [14] present new opportunities for comprehensive urban analysis. They introduce a refined methodology for evaluating the visual quality of street spaces, both in their current state and across temporal shifts, quantifying and analyzing the temporal changes of VQoSS with unprecedented precision. By leveraging image processing and machine learning, we can dissect the physical and perceived visual qualities of urban thoroughfares, even applying our techniques to the storied alleyways of historic districts. People can examine the past, present, and potential futures of urban streets, laying the groundwork for methodologies that blend tradition with innovation in pursuit of urban spaces that are not only functional but also inviting.

Previous works [5], [15], [16] incorporate the ability of deep learning in computer vision, offering solutions for a variety task of VQoSS. However, current deep learning technologies face several limitations. One of the primary challenges is the inherent difficulty in capturing the subjective elements of aesthetics and social vibrancy that constitute the essence of street quality. They also need multiple modules to compute segmentation, detection, depth, and such for the estimation [5], which makes the pipeline very complex. Moreover, deep learning models require vast amounts of labeled data to learn effectively [17], [18], [19]. The labor-intensive process of gathering and annotating temporal visual data from urban environments can be prohibitive, often resulting in models that may not generalize well across different cities or cultural contexts [20]. Lastly, environmental and seasonal variations, along with the dynamic nature of urban life, introduce a level of complexity that deep learning models, with their need for stable and consistent patterns, struggle to accommodate. It seems that only using vision-related technology can not satisfy the real-world demands of VQoSS tasks [21]. This leads to a question: *Can we design a method that can efficiently and accurately realize the large-scale VQoSS change estimation tasks without any manual annotation?*

Recently, Large Language Models (LLMs) [22], [23] have shown their power in a variety of complex language understanding and generation tasks, fundamentally altering the landscape of natural language processing. These models, trained on expansive corpora of text, excel at grasping the nuances of human language, from writing and summarizing articles to generating coherent and contextually relevant text. This ability has been further extended to Multimodal Large Language Models (MLLMs) [24], [25], achieving outstanding accomplishments. Models like GPT-4V [26] can now not only understand and generate text but also interpret and analyze visual data. This integration of vision and language processing enables them to provide answers to questions about the content of images, effectively bridging the gap between visual perception and linguistic expression [27], [28]. By processing and understanding complex visual scenes, these models can identify objects, recognize patterns, and even infer emotions or actions depicted in images [28], [29], [30]. We can directly use them for the purpose of VQoSS estimation. Although GPT-4V is powerful, using its API is expensive [31], and its structure is still unknown, which makes it impossible for large-scale VQoSS change estimation.

In this study, we investigate the potential for automating the estimation of VQoSS changes using street-view imagery and MLLMs. Our analysis is centered on the urban center of Nanjing, China. We collected a dataset of 10,598 data points along main thoroughfares, capturing images from four cardinal directions (0°, 90°, 180°, and 270°) at each location for two different periods. We also set 9 different VQoSS change evaluation tasks to quantify the performance of MLLMs. Our findings indicate that models such as GPT-4V exhibit remarkable proficiency in discerning temporal street variations for VQoSS evaluation. Moreover, as GPT-4V is not free in use, in order to reduce the cost for large-scale estimation, we introduce the Street Quality GPT (SQ-GPT) for VQoSS change estimation, which can efficiently distill [32], [33] GPT-4V's knowledge. Through rigorous experimentation, we demonstrate that the performance of SQ-GPT is close to GPT-4V, validating its utility as a cost-effective instrument for large-scale urban street quality surveillance.

We summarized our contributions as follows: (1) Our study pioneers the integration of MLLMs into the automation process for estimating changes in the VQoSS, offering a significant tool for urban research advancements. (2) We have identified 9 critical VQoSS tasks and have rigorously evaluated the effectiveness of MLLMs through both quantitative measures and qualitative analysis. (3) We introduce SQ-GPT, an innovative model that demonstrates performance on par with GPT-4V, providing an alternative for large-scale VQoSS estimation tasks.

II. RELATED WORKS

A. VISUAL QUALITY OF STREET SPACE

The requirements to quantify the VQoSS changes in urban spaces date back to the latter half of the 20th century. Pioneering works [7] emphasized the importance of street life and social interactions in urban spaces. They underscored the need for empirical studies to connect physical appearance with social attributes. Early efforts were largely qualitative due to the challenges of quantitative data collection over large areas [34].

In the end of last century, there seems to be a shift towards more systematic methodologies for assessing the visual quality of street spaces [35]. Researchers began employing a mix of subjective assessments, systematic observations, and technological advancements such as GIS [36] for more objective measurements. Studies offered structured approaches to evaluate the physical and perceptual dimensions of urban spaces. Recent years have marked the introduction of advanced technologies into the realm of urban space assessment. The utilization of Street View Imagery (SVI) [9], [37], [38] and deep learning algorithms [5], [15], [16], [39], [40], for instance, has allowed for large-scale quantitative analyses [41]. This integration has propelled the field from qualitative descriptions to data-driven, automated evaluations, as evidenced in works [42] that leveraged SVI for extensive urban analysis.

Tang and Long [5] built upon this progression by offering a comprehensive methodology that combines physical and perceived evaluations of visual quality using machine learning (SegNet [43]), GIS analysis, and systematic user-based assessments. Their work stands at the intersection of urban planning, computer science, and psychology, showcasing a multifaceted approach to understanding the temporal dynamics of street space quality in historical areas like Beijing's Hutongs. The current trajectory of research suggests a continuous refinement of methodologies with an increasing emphasis on integrating user perceptions and technological sophistication [1], [9]. Upcoming studies will likely focus on enhancing the objectivity in measuring visual quality [44], [45], fine-tuning algorithmic approaches [46], and incorporating real-time data to capture the ever-evolving nature of urban spaces [47]. This evolution reflects a broader trend in urban planning towards smart, data-informed decision-making that prioritizes both the physical attributes and the experiential qualities of urban environments. Although previous studies have achieved promising results, large-scale estimation of VQoSS changes is still labor-intensive. Especially methods based on deep learning, their accuracy requires a large amount of data annotation.

B. MULTIMODAL LARGE LANGUAGE MODELS

In recent years, the field of multimodal learning [24], [48] has experienced rapid development, with researchers striving to construct artificial intelligence models capable of comprehending and generating multiple types of data, including text, images, videos, and more. In this section, we provide a concise overview of the relevant literature that informs our work.

To empower LLMs with the ability to interpret images and engage in user interactions based on visual inputs, a multitude of methods have emerged. These methods aim to establish connections between frozen visual encoders and LLMs while keeping the number of trainable parameters minimal. Prominent approaches in this category encompass BLIP-2 [27], InstructBLIP [49], LLaMA-Adapter [50], LLaMA-Adapter V2 [51], LLaVA [29], MiniGPT-4 [30], MultiModal-GPT [52], and PandaGPT [53]. These models introduce additional trainable parameters between the frozen image encoder (or multimodal encoder in the case of PandaGPT) and the LLMs. For instance, BLIP-2 [27] utilizes a lightweight Transformer (Q-Former) as a bridge between the image encoder and the LLMs. LLaMA-Adapter and LLaMA-Adapter V2 introduce adaptation prompts that are pre-added to higher converter layer inputs, along with new teaching cues injected into LLaMA via a zero-gated zero-initial attention mechanism. MultiModal-GPT leverages low-rank adapters (LoRA) [54] for efficient fine-tuning of OpenFlamingo [55]. MiniGPT-4 and LLaVA simply add Linear layers between the frozen visual encoder and frozen LLMs. There are also some works introduce LLMs for urban analysis. Wu et al. [56] employed a model based on CLIP [57] to facilitate vision-language multimodal learning, specifically targeting street view images and their spatial context. This approach highlights the intricate relationship between visual data and its corresponding linguistic descriptors in an urban environment. Verma et al. [58] utilized LLMs for a comprehensive evaluation of environmental surroundings. Their methodology focuses on interpreting context, framing perspectives, and predicting behavioral responses, thereby showcasing the versatility of LLMs in understanding and anticipating complex human-environment interactions.

A critical objective in developing LLMs is to ensure that they produce coherent and extended natural language responses, enabling more human-like conversations. Multimodal instruction tuning [59] is a widely adopted technique in this regard. LLaVA [29] is the first attempt to use pure language GPT-4 [26] to generate multimodal language-image instructions following data. They are based on the captions and bounding boxes on the COCO dataset [60]. InstructBLIP [49] and MultiModal GPT [52] also proposed to match the corresponding image descriptions to create instruction fine-tuning data. Using these data, LLMs are trained, resulting in a significant improvement in dialogue performance. However, LLaMA-Adapter v2 [50] and MiniGPT4 [30] only use ready-made image captioning datasets for multimodal instruction tuning, potentially resulting in relatively weak image understanding capabilities. Therefore, it is crucial to curate diverse datasets that cover various tasks and scenarios, providing rich contextual information to support the model in deep reasoning. If the examples in the dataset are too single or simple, the model may learn to produce short and uninformative responses. In summary, high-quality instruction following data is pivotal for enabling effective conversations [59], [61].

MLLMs have demonstrated a formidable capacity in bridging various modes of human communication and understanding. These models are adept at processing multiple data types, paving the way for innovative applications across diverse fields. This is particularly evident in tasks involving the VQoSS, where the application of MLLMs is especially critical. The visual quality of street spaces not only impacts the quality of life for residents but is also intricately linked to urban planning and sustainable development. Traditionally, assessing the visual quality of street spaces relies on manual annotations and subjective evaluations, a process that is both time-consuming and inefficient. Against this backdrop, the introduction of MLLMs, such as GPT-4V [26], to automate the assessment of street space visual quality is revolutionary. Firstly, by analyzing street view images from different time periods, MLLMs can identify and track urban features that change over time, offering insights into the evolving visual

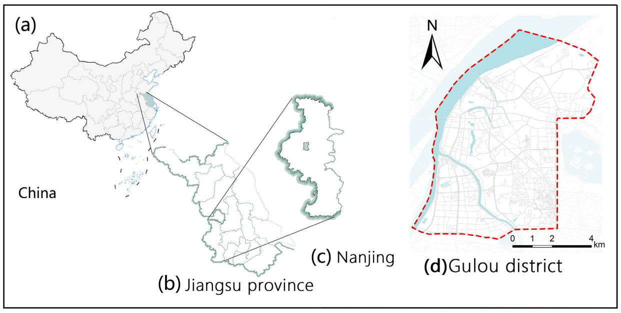


FIGURE 1. Research area of this study. All data points are collected from Gulou district (d), Nanjing (c), Jiangsu province (b), China (a).

quality of street spaces. Secondly, MLLMs are capable of processing and integrating data from various sources, such as GIS data, social media inputs, and urban planning records, to provide a more comprehensive assessment of street space visual quality. Additionally, the integration of MLLMs helps to overcome the limitations of subjective evaluations, yielding more objective and consistent assessment results.

Currently, GPT-4V stands as one of the most powerful MLLMs, albeit accessible only through its official API and associated with high costs in large-scale applications. Therefore, we have introduced the SQ-GPT model, which achieves near GPT-4 performance through efficient knowledge distillation. The development of SQ-GPT not only signifies wider accessibility and reduced costs but also represents a more effective use of advanced MLLMs technology in the VQoSS domain. It can support urban planning and community development, such as aid in designing efficient public transportation systems, optimizing energy usage in buildings, and creating green spaces in cities, paving the way for a new era of sustainability in urban development.

III. DATA AND METHOD

A. DATA

1) DATA COLLECTION

Our dataset is acquired from Baidu Street View and encompasses a total of 10,598 data points from the center of Nanjing, Gulou district, with each point paired with corresponding geographical coordinates. Figure 1 depicts our study area, the Gulou District of Nanjing. The Gulou District of Nanjing is a vital administrative region situated in the center of Nanjing City, Jiangsu Province. It covers an area of approximately 56.57 square kilometers and has a permanent resident population of about 940,400. Geographically, the district is bordered by the Yangtze River to the north, Qinhuai River to the south, Qinhuai and Pukou districts to the east, Xuanwu and Jianye districts to the south, and is adjacent to Yuhuatai District to the west and Qixia District to the north. As the core area of the city, Gulou District boasts a wealth of historical and cultural heritage, such as the Drum Tower and the Confucius Temple, and serves as Nanjing's center for commerce, culture, and tourism, with multiple commercial districts and cultural institutions. It offers convenient transportation with several subway lines and major roads running through it. In terms of education, the district houses numerous universities and research institutions, making it one of the most concentrated areas of scientific and technological resources in China. The Gulou District also enjoys Nanjing's 11-kilometer prime Yangtze River coastline, features diverse terrain, and has a distinct four-season climate, marking it as a region with a long history and profound cultural heritage.

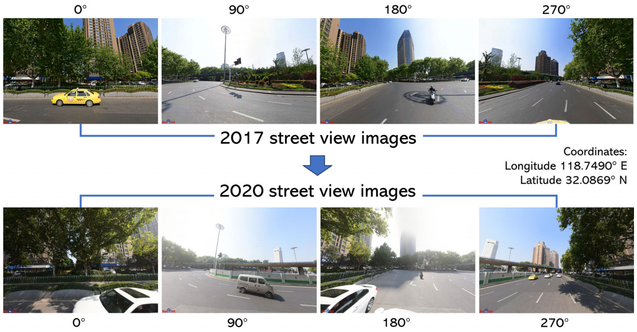


FIGURE 2. The study area's specific location at Longitude 118.7490° E, Latitude 32.0869° N has street view images from 2017 and 2020 in all four cardinal directions.

This dataset captures images from two different periods: the previous period, denoted as p, is from September 2017, and the latter period, denoted as l, is from September 2020. In both periods, we collected four images per data point, corresponding to the four cardinal directions (0°, 90°, 180°, and 270°). Therefore, we refer to our dataset as $\mathcal{X}$, and an individual image within this dataset is represented by x_i^{sd}, where d stands for the direction, i is the data point index, and s indicates the time period (either p or l). We show a sample of our collection in Figure 2.

To progress with our method and evaluation pipeline, we randomly selected approximately 10% (1,060) of the data points (each of them is a group of 8 images from different directions and periods) from our dataset to distill knowledge via GPT-4V. We denote these split data as $\bar{\mathcal{X}}$.For the purpose of user evaluation, we further randomly chose 200 data points from the remaining dataset (the remaining 90% of the data points) to quantify and compare the performance between GPT-4V and our developed model, SQ-GPT.

2) ESTIMATION SETTINGS

In Figure 3, it can be observed that our pipeline comprises 4 distinct categories, encompassing a total of 9 individual tasks ($T = 9$) designed for estimating VQoSS changes. These tasks serve as pivotal criteria for quantifying these changes, providing a comprehensive assessment. For each task, we set three levels of evaluation results: A: Positive Changes, B: Negative Changes, and C: No Changes. Our definitions for these tasks draw inspiration from the work of [5] (some samples are showing in Figure 4), and we elaborate on their specific definitions as follows (We make a concise definition

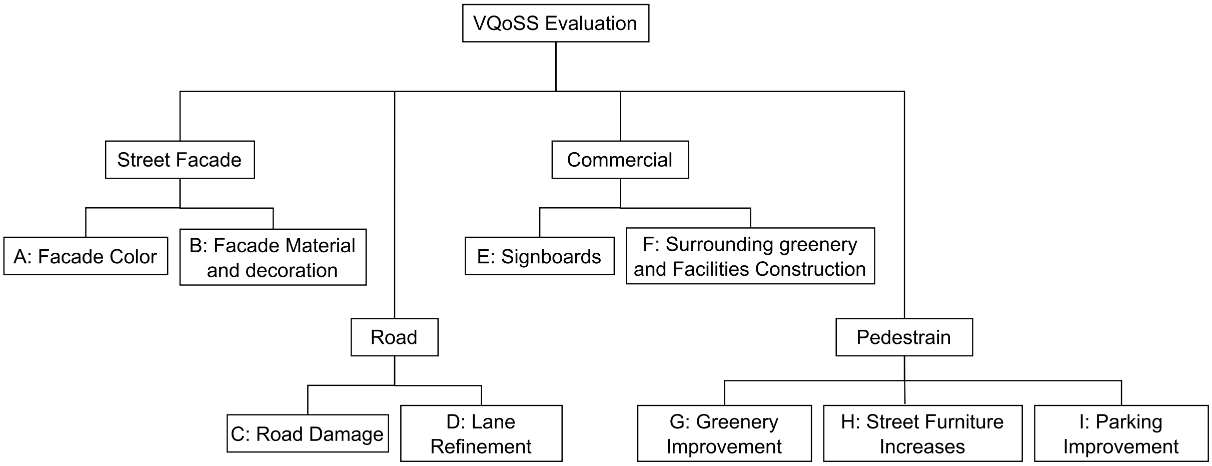


FIGURE 3. The settings for VQoSS evaluation. We totally defined 4 categories and 9 tasks for the VQoSS evaluation. These tasks can comprehensively evaluate street visual quality.

so that models can use them as prompt efficiently during questioning. Note that the ability to recognize each level is the common sense reasoning ability by MLLMs [62]):

Task A: Facade Color. In this task, the focus is on assessing the color schemes used on building facades along a street. It involves evaluating the aesthetic and visual appeal of the colors employed and considering any changes or improvements to enhance the overall quality of the street space.

Task B: Facade Material or Decoration. Task B revolves around evaluating and potentially modifying the materials or decorative elements used in building facades within the street environment. The aim is to enhance the visual and tactile qualities of facades, potentially by introducing new materials or decorations that contribute to the overall street space quality.

Task C: Road Damage. This task refers to any defects or deterioration in the street surface, such as potholes, cracks, rutting, weathering, surface defects, and debris, which impair both the functionality and the appearance of the roadway.

Task D: Lane Refinement. Task G involves assessing and refining the layout and design of lanes within the street. This may include modifications to lane width, markings, or traffic flow to enhance safety, traffic management, and overall street quality.

Task E: Signboards. In this task, the emphasis is on evaluating and optimizing the presence and design of signboards within the street space. This may involve assessing the placement, size, and aesthetics of signboards to improve the overall quality and functionality of the street.

Task F: Surrounding greenery and Facilities Construction. In Task I, the scope extends beyond the street itself to encompass the construction and improvement of greenery and facilities in the surrounding area. This can include parks, plazas, or other amenities that contribute to the quality and attractiveness of the street space and its immediate surroundings.

Task G: Parking Improvement. Task E focuses on evaluating and implementing improvements to parking facilities within the street area. This includes considerations for parking space design, accessibility, and any enhancements that can contribute to the overall quality of the street space.

Task H: Street Furniture Increases. In Task F, the objective is to check the quantity and quality of street furniture. This may involve adding benches, lighting fixtures, waste bins, or other elements that improve the comfort and usability of the street space.

Task I: Greenery Improvement. Task H focuses on evaluating and enhancing the presence of greenery, such as trees, shrubs, or plantings, within the street environment. The goal is to improve aesthetics, air quality, and the overall ambiance of the street.

Each of these tasks is associated with a distinctive question that necessitates a MLLM to provide an answer. Our objective is for the MLLM to effectively discern the differences between images captured during two periods p and l. The question format for each task adheres to the following structure:

$$\begin{aligned} Q_T = \{D\} &+ \text{“Analyze”} + \{T\} \\ &+ \text{“changes between different periods,”} + \{C\}, \quad (1) \end{aligned}$$

where D is the description of input images as “The first four images are street view images taken from four different directions of the street. The last four images are taken from the same location and the same four directions but at a different period.” T is the task defined above and C is a multiple choice question as “and finally you must give me a choice from: A: Positive B: Negative, C: No Changes. Make your statement concise”. We also demonstrate a sample of question and answer in Table 1.

B. OVERVIEW OF OUR PIPELINE

The architecture of our pipeline is illustrated in Figure 5. Initially, we distill knowledge from GPT-4 by inputting

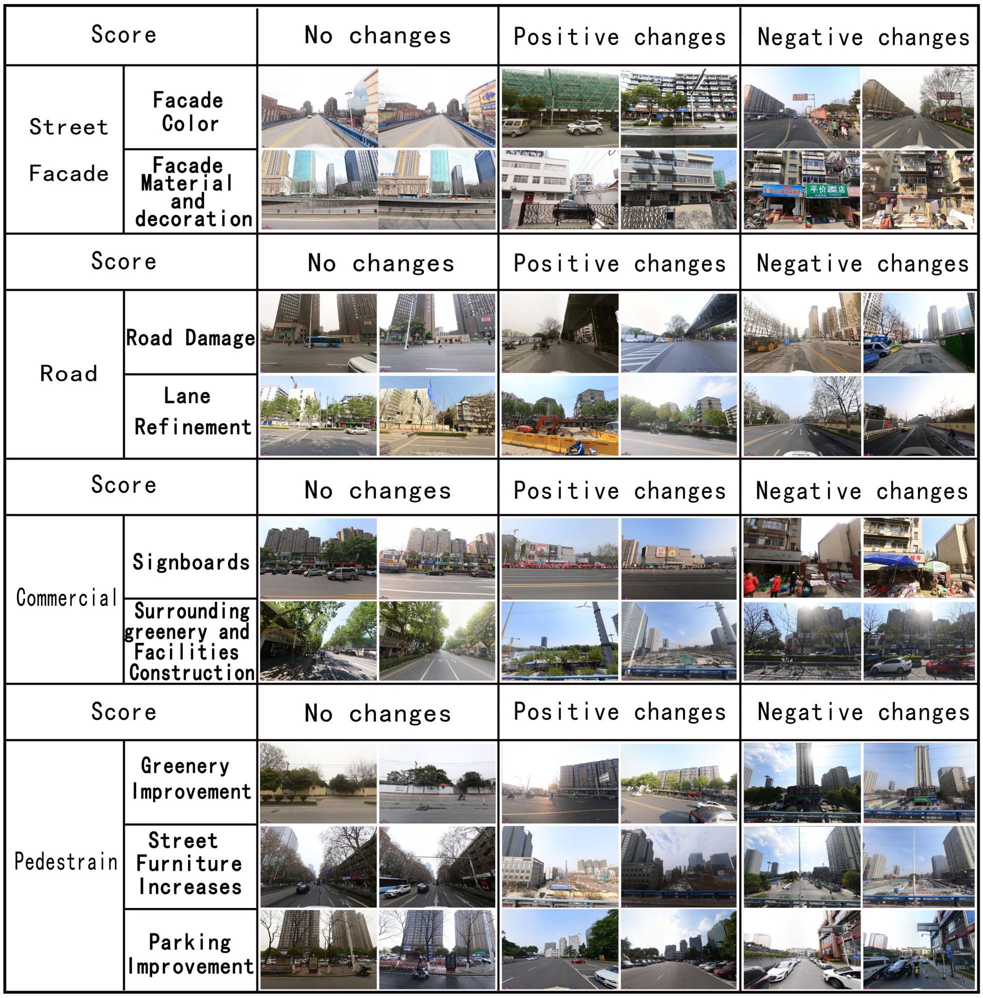


FIGURE 4. Samples for 9 tasks of VQoSS changes evaluation.

street-view images from a data point along with corresponding prompts and questions to elicit high-quality answers. This procedure is applied across all 1,060 split data points, generating a new dataset to fine-tune our own model. SQ-GPT comprises a frozen image encoder, a frozen QFormer [27] for image-to-text alignment, a Linear plus Reduction stage for additional refinement, and a frozen LLMs framework for the final answer output. The fine-tuning process is informed by insights drawn from GPT-4, ensuring that the output of SQ-GPT is improved by the distilled knowledge. We provide a detailed description of each component in the subsequent sections.

C. KNOWLEDGE DISTILLATION FROM GPT-4

We utilize the GPT-4V API (gpt-4-vision-preview) for distillation processes to generate a new dataset for the fine-tuning of our model. Table 1 presents a sample that illustrates the generation pipeline. It can be observed that GPT-4V can describe the details of input images and compare two periods from multiple aspects.

We sequentially feed GPT-4V four images from a data point's period p, aligned at 0°, 90°, 180°, and 270°. This is followed by four images from period l in the same input order. Subsequently, for a given task, the pertinent question formulated in Equation 1 is inputted (Note that we also set the definition of the corresponding task as the input prompt). To augment the dataset with a diverse set of question-answer pairs, we perform $k = 5$ iterations of answer generation. To guarantee uniformity in the responses for a specific data point within a given task, we mandate that the outcomes produced for the same question on five separate generations must be consistent. Any generation where the generated answers vary will be excluded and subjected to the regeneration of a new point. The pricing structure is as follows: \$ 0.01 per 1K tokens for input and \$ 0.03 per 1K tokens for output. Our calculations indicate that generating a single question-answer pair costs approximately \$ 0.07. This cost becomes prohibitive for large-scale operations, underlining the importance and economic value of our research. The analysis of the whole city generally involves

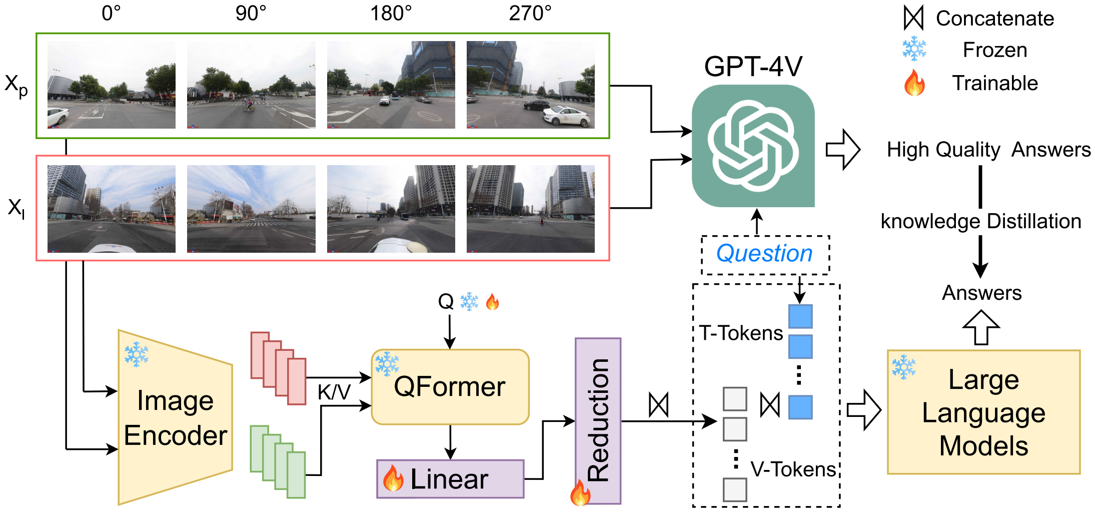


FIGURE 5. Overview of our proposed SQ-GPT. In the upper part, we first distill the knowledge in GPT-4 through split data points $\mathcal{X}$ and generate a new dataset $\mathcal{L}$. Then, in the lower part, we fine-tune our SQ-GPT by the supervision of high-quality answers from GPT-4.

TABLE 1. To show how we distill knowledge from GPT-4, we demonstrate an answer sample from GPT-4 in discriminating the color of the facade. We input four direction images of a data point for previous and later periods, respectively. Then, we ask GPT-4 to describe the color difference between two periods and select a choice from the given options.

Visual Input Example, Question for "Facade Color"	
Question	The first four images are street view images taken from four different directions of the street. The last four images are taken from the same location and the same four directions but at a different period. Analyze the facade color changes of different periods and finally you must give me a choice from: A: Positive, B: Negative, C: No Changes. Make your statement concise.
Answer	Based on the provided images, we can observe the following changes in the building facades' color over the two different time periods. Initial Period (First Four Images): The building facades are predominantly neutral-toned with a mix of whites, beiges, and light gray hues. The overall color palette is muted and uniform, with little variation in color saturation or brightness. The buildings have a clean and modern appearance, with the color of the construction materials clearly visible. Later Period (Last Four Images): The facades exhibit a similar color scheme with no significant changes in the hues. However, there may be slight variations in brightness due to differences in the time of day or weather conditions when the photos were taken. There is no evidence of major renovation or repainting that has altered the color palette of the buildings significantly. Considering these observations, the most accurate response would be: C: No Changes. .

about 100,000 data points, which may cost 7,000 $ using GPT-4V. A local server with one A40 GPU (costs about 10,000 $) is enough for the inference of a MLLM model. Therefore, for long-term planning, it is more appropriate to deploy a local server.

For a better understanding, we formulated the generation of new data set $\mathcal{L}$ as:

$$y_{i,t,k} = \int_k^5 GPT4V([\oint_s^2 \oint_d^4 x_i^{sd}] + Q_t), \tag{2}$$

$$\mathcal{L} = \{(\hat{x}_i, Q_t, y_{i,t,k}) \mid i = 1, \ldots, I; t = 1, \ldots, T; k = 1, \ldots, 5\}, \tag{3}$$

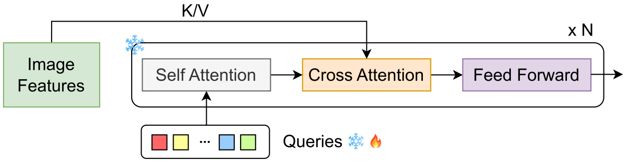


FIGURE 6. The inference pipeline of a frozen QFormer. It is a N time transformer [63] blocks and we adopted a learnable query besides the original learned query.

where $\oint$ stands the input images in order for period and direction of a data point, $\int$ is the repeat generation for k times, I is the data point number in split $\mathcal{X}$, $\hat{x}_i$ represent all images for a data point, and t is the index for all tasks T.

$\mathcal{L}$ contains 47,700 data triads, each consisting of image sequences, corresponding questions, and high quality answers. This rich dataset will be instrumental in the fine-tuning process of our model, SQ-GPT. The fine-tuning will be executed with a specific focus on improving the model's ability to understand and generate contextually relevant responses based on sequential visual inputs, to achieve knowledge distillation of GPT-4V.

D. SQ-GPT

As depicted in Figure 5's lower branch, our SQ-GPT is trained in an end-to-end manner. To commence this process, we employ a frozen image encoder E_{image}, which extracts image features from all images $\hat{x}_i$, corresponding to a given data point i. This process can be succinctly described as follows:

$$\hat{f}_i = E_{image}(\hat{x}_i), \in \mathbb{R}^{sd\times hw\times c}, \tag{4}$$

where h and w represent the spatial dimensions of the image features obtained after encoding, while c stands for the feature dimension. During this stage, we independently process the previous four images and the later four images. We stack them in the same period order and direction order as outlined in the GPT-4V distillation III-C.

Next, we leverage a frozen QFormer architecture (shown in Figure 6), as introduced [27], to facilitate image-text alignment. The choice of QFormer is crucial in our approach, as it excels in capturing intricate dependencies between textual and visual information, enabling us to establish a more robust and meaningful connection between the two modalities. This alignment step is fundamental to the overall effectiveness of our method, as it forms the basis for subsequent fusion and comprehension of image and text data. Except the learned queries Q^* in the original QFormer, we add a set of learnable queries Q' for better ability of alignment of image and text. Thus, the calculation of QFormer is as follows:

$$f_i^* = Linear(QFormer(\hat{f}_i, Q^* + Q')), \in \mathbb{R}^{sd \times hw \times c}, \tag{5}$$

where each image feature of $\hat{f}$ serves as K and V in cross-attention [63]. *Linear* is a trainable simple one-layer Fully Connection (FC) layer that can further adjust the output features from QFormer.

After the process of image-text alignment, we further reduce the volume of each image and concatenate them as a visual token entirety as follows:

$$f_i' = Cat[Reduction(f_i^*)], \in \mathbb{R}^{z \times c}, \tag{6}$$

where the reduction operation takes the form of a convolution layer with stride of 2, which effectively reduces the dimensions of the hw features to one-fourth of their original size. The concatenation of all images is represented as Cat, and the resulting total number of visual tokens is denoted as z. The inclusion of the reduction step serves a critical purpose: it shortens the length of visual tokens, mitigating the potential adverse effects of overly long visual tokens on the computations of a frozen LLM. The importance of this reduction step has been substantiated in a prior study [28].

Next we concatenate the visual tokens (as prefix) with text tokens encoded from question Q_t corresponding to a certain task. Then, a frozen LLM will take then as input and generate the answer. It can be formulated as follows:

$$y_{i,t} = LLM(Cat[f_i', Q_t]). \tag{7}$$

The output answer of $y_{i,t}$ will calculate loss with one of the high quality answer $y_{i,t,k}$. Through this knowledge distillation training, our SQ-GPT can achieve performance levels close to those of GPT-4V.

IV. RESULTS

A. EXPERIMENTAL SETTINGS

Our training pipeline is adapted from LLaVA-1.5 13B [29], with modifications that involve freezing both its image encoder and its LLM components. Additionally, we incorporate a pre-trained QFormer, sourced from BLIP-V2 [27], to enhance the capabilities of our model in handling image-text tasks. We also adopt instruction tuning as LLaMA-Adapter v2 [50] and MiniGPT4 [30] during training. Only ready-made image captioning in $\mathcal{L}$ is utilized, and no extra tuning data [29] is needed for annotation. We train all 9 tasks together to enhance the generalization of the model. The training is completed within 1 day on a GPU server with 4 Nvidia A40.

To evaluate the performance of our SQ-GPT model and benchmark it against other models, we refer to Section IV-B where a human-centric evaluation was conducted on 200 data points as mentioned in III-A1. We utilized LLaVA-1.5, SQ-GPT, and GPT-4V to generate answers for all nine tasks, employing the corresponding questions. During this process, we set the generation temperature to 0 to ensure precise and deterministic outputs. We enlisted three distinct groups of evaluators–50 laypeople, 30 architecture students, and 10 urban environment experts–to assess the quality of responses from all three models. Evaluators were instructed to cast a binary vote on whether the generated answers accurately addressed the task as defined in III-A2. Consequently, we define our metric of agreement based on these binary votes as follows:

$$Agreement_{g,m,t} = \frac{1}{|J_g|}\frac{1}{|O|}\sum_{j=1}^{J_g}\sum_{o=1}^{O} vote(A_{o,m,t}), \tag{8}$$

where g is the group, m is the model, J_g is the number of person in a group, and O is the number of generate answer for a certain model and task.

In Section IV-C, we also provide the visual analysis by mapping the answer results of all data point to the city map. We generate answers using SQ-GPT for all 9 tasks. The answer of A: Positive, B: Negative, and C: No Changes are translated to the value of 1, 2, and 3.

B. ANSWER QUALITY ANALYSIS

As shown in Figure 7, we analyze the result for human-centric evaluation in all 9 tasks. The data compares the agreement with three models: LLaVA-1.5, SQ-GPT, and GPT-4V. Across all tasks, GPT-4V consistently gains the highest agreement rates from all respondent groups, indicative of its superior performance or alignment with human judgment. Our SQ-GPT also has a performance very close to GPT-4V, which verifies the effect of our knowledge distillation. In addition, laypeople demonstrate the highest levels of agreement with the models, suggesting AI outputs are more in tune with their evaluations. Conversely, experts consistently show the lowest agreement rates, which may reflect a disparity in understanding or knowledge relative to the models and other groups. LLaVA-1.5 is consistently the model with the lowest agreement rates among all respondents, hinting at it being the least accurate or aligned with human judgment. This pattern suggests that advancements in AI, as seen with GPT-4V, although not perfect, are increasingly recognized by common sense.

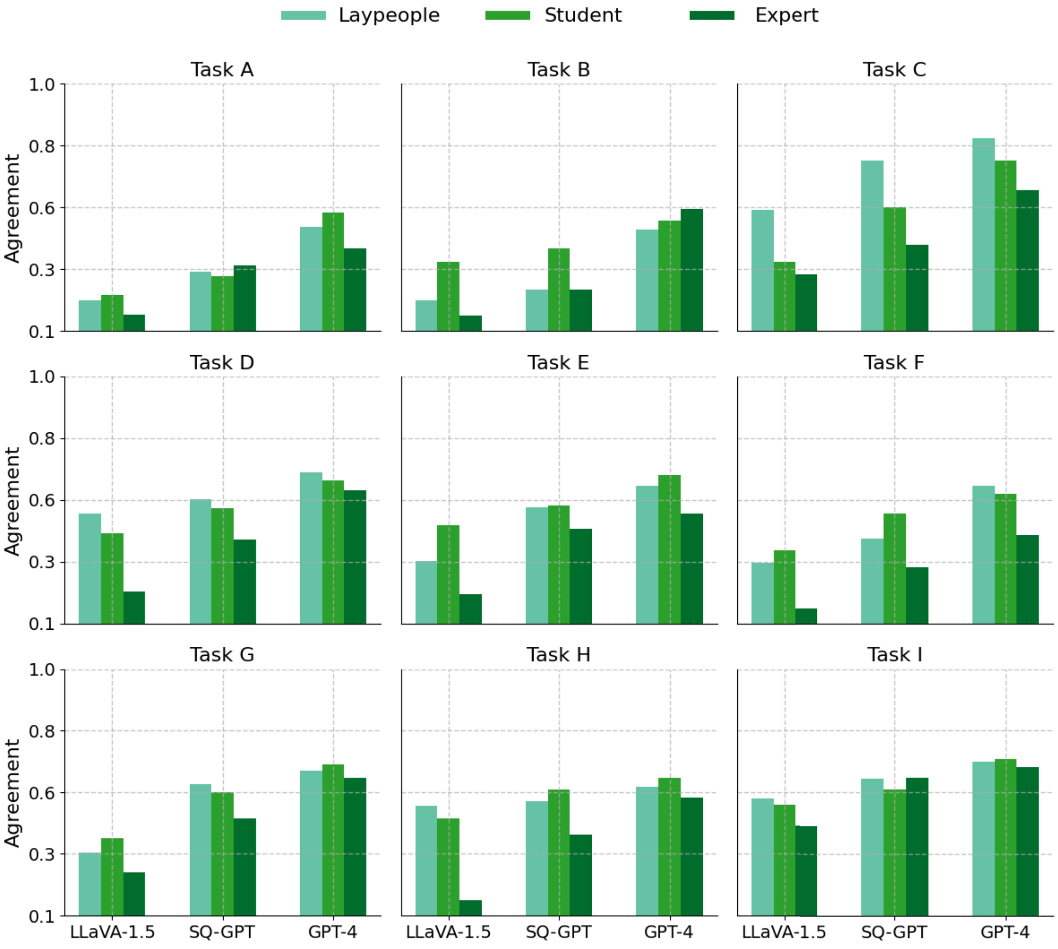


FIGURE 7. **Human-centric evaluation among LLaVA-1.5, SQ-GPT, and GPT-4V. We adopt the agreement testing for 200 data points and all 9 tasks. We employed three groups of people (50 laypeople, 30 architecture students, and 10 urban environment experts) for the quantitative evaluation.**

Among all tasks, A and B seem to be difficult. Experts give a low agreement even for the most powerful GPT-4V. We find the most possible reason is the brightness difference when capturing images from two periods. This difference may affect the analysis of models as they can capture the color changes without discriminating the condition of weather. Models perform well in tasks like C (“Road Damage”) and I (“Greenery Improvement”). These tasks pay more attention to some obvious visual forms and layout features, which are what deep models are good at. Nevertheless, tasks such as “Parking Improvement” (G) are susceptible to inaccuracies, as static images do not provide sufficient information to determine whether vehicles are stationary (parked) or in motion. In general, MLLMs can complete most VQoSS tasks well and is a good means to implement urban analysis. In particular, our SQ-GPT has obtained matching performance after distilling the knowledge of GPT-4V, and can be used as a tool for large-scale evaluation.

We also offer some samples of answer comparison between LLaVA-1.5, SQ-GPT, and GPT-4V. However, due to the length restriction, only two samples for discriminating changes of “Lane Refinement” and “Greenery Improvement” are provided in this manuscript. Referencing Table 2, we tasked the models with interpreting “Lane Refinement” based on images from varying periods. The later-period images visibly display faded road markings and numerous repair patches. All models unanimously indicated deterioration with option B: Negative. GPT-4V observed “The roads show signs of heavier use,” which corroborates our visual assessment. Interestingly, SQ-GPT’s analysis parallels GPT-4V’s but goes a step further by considering the condition of traffic signs, stating, “Despite a consistent presence of traffic signs, the overall impression hints at a deterioration in the road’s organization.” This suggests SQ-GPT’s nuanced attention to detail. LLaVA-1.5, while also noting a negative change, linked it to an uptick in transportation in the later period–an associative leap that, although not incorrect, strays from the task definition provided in III-A2. Such findings not only demonstrate each model’s unique approach to the analysis of visual changes but also underscore the

TABLE 2. Comparison of answer samples among LLaVA-1.5, SQ-GPT, and GPT-4V for the question of discriminating changes of "Lane Refinement."

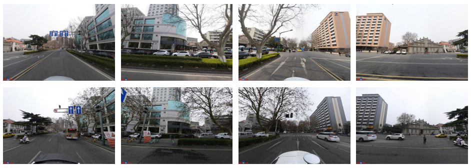
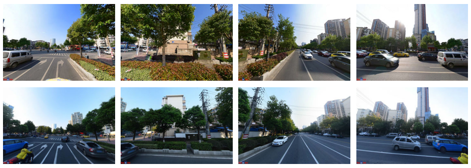
Visual Input Comparison Examples, Question for "Lane Refinement"	
Question	 {D} + "Describe changes in road refinement between two periods." + {C}
Answer GPT-4V	After reviewing the images from two different periods, it is noticeable that there has been some wear in the road markings and a general decline in lane visibility in the later images. The roads show signs of heavier use, and the increased clutter along the curbside could suggest changes in traffic or parking patterns that may not be as orderly as before. These factors contribute to a slight negative impact on lane refinement, indicating a decrease in the clarity and maintenance of traffic lanes over time. Hence, the assessment concludes with a negative change in lane refinement. Therefore, my feeling towards these changes would be option B: Negative.
Answer SQ-GPT	It's evident that lane refinement has seen a decline over time. Initially, clear lane markings in the first period have become less discernible in the second, likely as a result of wear and inadequate maintenance. This degradation in lane visibility can have serious implications, potentially compromising lane divisions' clarity and introducing confusion among drivers. Despite a consistent presence of traffic signs, the overall impression hints at a deterioration in the road's organization. Taking these factors into account, it becomes increasingly apparent that the changes in lane refinement lean towards a negative impact, aligning with Option B: Negative.
Answer LLaVA-1.5	The changes in lane refinement between the two periods are evident in the increased number of cars on the street, which suggests that the area is more congested. I feel that the changes in lane refinement are negative, as the increased number of cars on the street indicates a higher level of traffic congestion. This can lead to longer travel times, increased air pollution, and a higher risk of accidents. The increased traffic can also be a sign of urbanization and economic growth, which is a positive aspect of the city's development. However, the negative impacts on the environment and the quality of life for the residents must be considered. The final option is B.

TABLE 3. Comparison of answer samples among LLaVA-1.5, SQ-GPT, and GPT-4V for the question of discriminating changes of "Greenery Improvement."

Visual Input Comparison Examples, Question for "Greenery Improvement"	
Question	 {D} + "Describe changes in greenery improvement between two periods." + {C}
Answer GPT-4V	Upon examining the eight street view images you provided, they depict the same location from four cardinal directions, captured at two different time periods. The comparison shows visible differences in the greenery between the two periods. Initially, the area has smaller trees and less dense foliage. In the later images, the trees appear more mature and fuller, indicating growth and increased leaf density, which contributes to a lusher streetscape. Additionally, the vegetation in the median strips and along the sidewalks seems to be better maintained, suggesting an enhancement in the overall greenery management. This qualitative assessment of the vegetative growth and maintenance suggests a positive change in the greenery of the area. Therefore, based on the visual evidence from the images provided, the choice would be: A: Positive.
Answer SQ-GPT	From the images you've provided, it's evident that there have been changes in the greenery between the two periods. In the earlier images, the trees appear less mature and the foliage is sparser, while in the latter images, the trees show more growth, with denser canopies, indicating positive development in the greenery. Additionally, the bushes and the underplanting in the median strips and along the sidewalks seem to be better maintained and more abundant in the later images. This suggests a concerted effort to enhance the urban green spaces. These improvements contribute to a healthier, more aesthetically pleasing environment and potentially better air quality for the area. Thus, it is a positive change (Option A).
Answer LLaVA-1.5	Upon analyzing the images from the two different periods, the following observations can be made: The tree canopies seem consistent in size and density, indicating that the trees are mature and have been maintained. There is no significant removal or reduction of trees or plants. The color and vigor of the greenery appear similar across both periods, suggesting good health. No major landscaping additions or enhancements are visible that would suggest an intentional effort to increase green spaces. Considering these points, it can be concluded that there have been no significant changes in the greenery between the two periods depicted in the images. The urban greenery maintenance appears to be stable. Therefore, the assessment is: C: No Changes.

variability in interpreting contextual information, offering valuable insights into the current capabilities and limitations of MLLMs-based image interpretation.

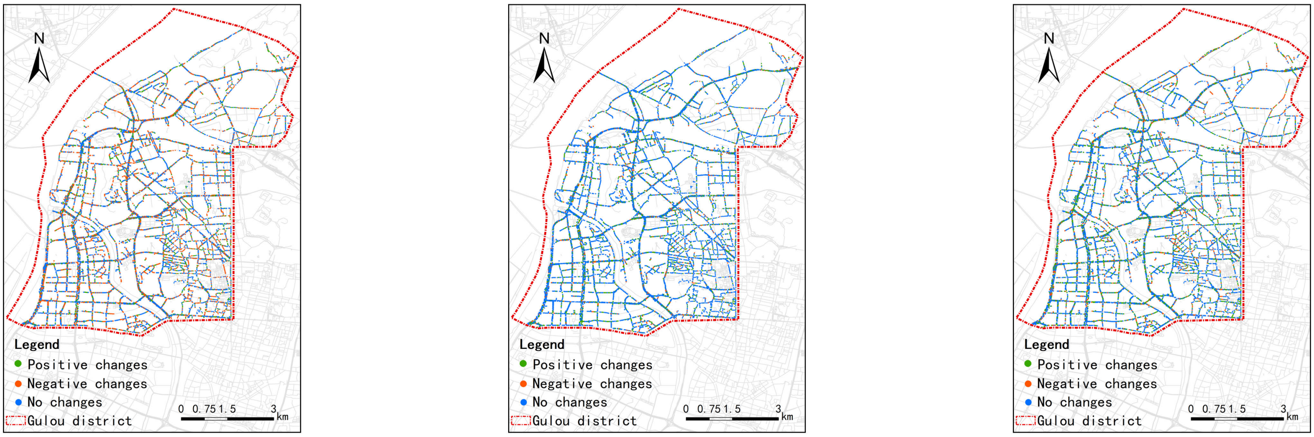
Referencing Table 3, we tasked the models with interpreting "Greenery Improvement" based on images from varying periods. The first set of images is taken on a sunny day, while the later set are take from the day with clouds. At first glance, there seems to be no obvious greenery improvement changes exist between images from two periods. LLaVA-1.5 analyse two periods in a very fairly good way and give an answer as C: No Changes. However, GPT-4V and our SQ-GPT give the answer as Positive changes. They all concern about "the later images, the trees appear more mature and fuller, indicating growth and increased leaf density, which contributes to a lusher streetscape". When we look closely at these two periods, we can find that the trees in the latter period are indeed more vigorous and the green of the bushes is more obvious. This shows that GPT-4's attention is more detailed, and our model distills the knowledge of GPT-4V and obtains similar capabilities.

C. VISUALIZATION OF VQOSS ESTIMATION

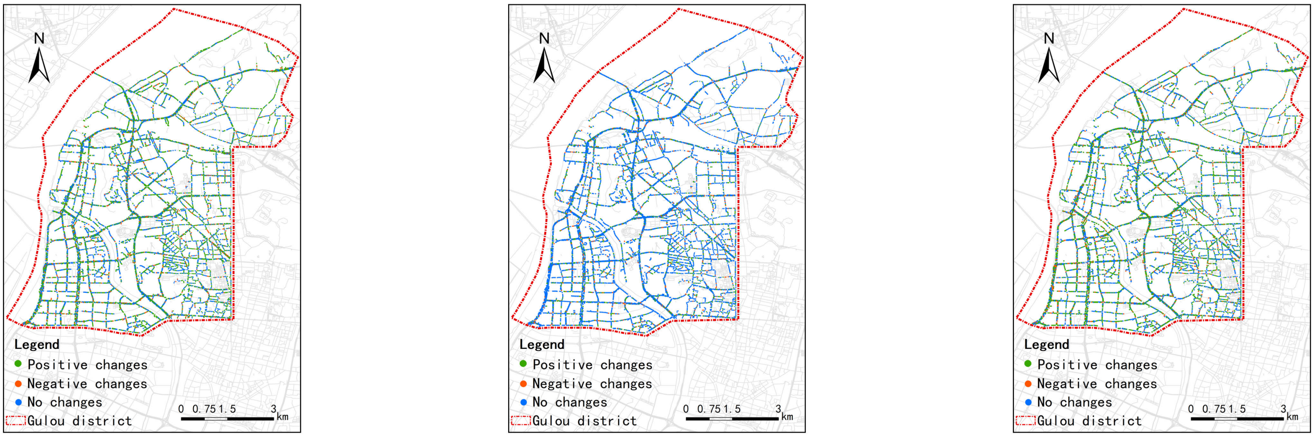
Our analysis of tasks B, C, and E reveals a persistent blue tone, suggesting that these tasks of urban development have remained relatively stable over the past three years. While stability might imply effective management in these areas, it could also signal a lack of innovation or improvement that warrants further investigation. Task G's "Parking Improvement" stands in stark contrast with a striking red spectrum dominating the visual data. This is a clear indicator of escalating urban congestion, primarily due to a surge in vehicle ownership. The density of parked cars has outpaced the availability of parking spaces, highlighting a critical area for urban policy intervention. The city must consider implementing multifaceted solutions, such as expanding public transportation, increasing parking infrastructure, and encouraging carpooling to alleviate parking pressures.

The evaluation of Task A presents an intricate scenario, with a balanced mix of colors denoting a complex and uncertain outcome. This color distribution suggests that the strategies employed are not yielding clear results. The task's low agreement, referenced in Figure 7, corroborates the ineffectiveness of current efforts. This warrants a reevaluation of the task's goals and methods to better serve the city's interests. Other tasks reflected by green shades indicate positive strides in the city's Visual Quality of Service Standards. These improvements are promising, demonstrating successful initiatives that enhance the city's visual appeal and functionality. However, it is crucial to maintain this momentum with continuous monitoring and updating of urban planning strategies to ensure sustained progress.

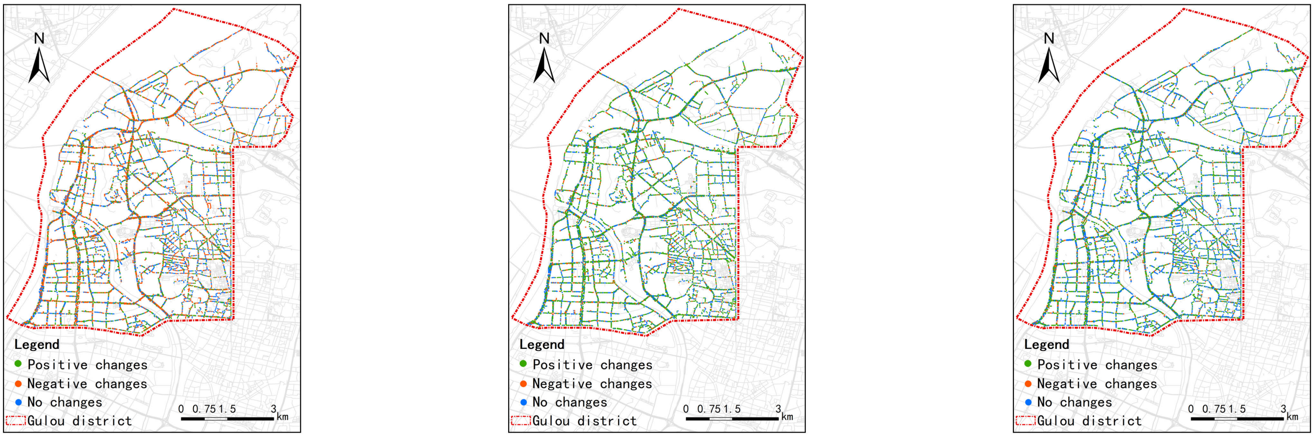
While the overall trend points towards an improvement in the city's visual quality, the looming shadow of increased congestion due to population growth cannot be ignored. This presents a paradox where the city is improving in visual standards while simultaneously grappling with the challenges of density and overuse of urban infrastructure. To address



(a) Facade Color (b) Facade Material (c) Road Damage



(d) Lane Refinement (e) Signboards (f) Surrounding Greenery



(g) Parking Improvement (h) Street Furniture (i) Greenery Improvement

FIGURE 8. Visualizations on a city map for all 9 VQoSS tasks. We use green point, red point, and blue point to represent positive changes, negative changes, and no changes, respectively. Sub-figure (a–i), (a) Facade Color; (b) Facade Material or Decoration.; (c) Road Damage; (d) Lane Refinement; (e) Signboards; (f) Surrounding greenery and Facilities Construction; (g) Parking Improvement; (h) Street Furniture Increases; (i) Greenery Improvement.

this, city planners and policymakers need to adopt a holistic approach. The above result analysis highlights the actual effect of our SQ-GPT on VQoSS tasks, indicating that our method is significant for future urban planning.

The current model evaluates 9 aspects of VQoSS. However, they may lack a comprehensive assessment, overlooking some other potential aspects and how these aspects interact to affect the overall user experience. We believe that MLLM can address this gap. It has the potential to not only quantify additional aspects of VQoSS, such as user interface fluidity and security response times, but also to integrate these factors into a cohesive overall quality assessment. By analyzing the interplay between different VQoSS components, MLLM can provide a more nuanced and realistic understanding of estimation. Our future work aims to expand MLLM's capabilities in this direction, enhancing its analytical depth while maintaining efficiency for real-time applications.

V. DISCUSSION

In this paper, we verified that When applying MLLMs to the estimation of VQoSS, these models can analyze photographic or multi-aspects street-view images to assess various quality indicators such as cleanliness, state of repair, signage clarity, and overall visual appeal. The estimation of VQoSS involves evaluating these images to rate or categorize the street space's condition and visual characteristics. This could be used for urban planning, maintenance scheduling, or to improve navigational aids by providing more detailed information about the environment. In traditional machine learning, human annotation is necessary to annotate a dataset, which the model then uses to learn and make predictions. For instance, humans might need to annotate images of streets with descriptors like "poorly maintained" or "clean" to train a model to recognize these qualities. However, with MLLMs, the AI can learn from vast amounts of unlabeled data, recognizing patterns and making inferences without explicit human-supervision. This capability is particularly advantageous because it reduces the time and resources needed to prepare data, making it easier to scale up and apply these models to large-scale tasks, such as those needed for assessing the VQoSS across different cities or regions. In addition, we can further use MLLMs with both street imagery and possibly accompanying textual data (like metadata, annotations, or descriptions) to form an assessment of the street's visual quality, which could inform various urban development and management decisions.

A distillation pipeline is also provided aimed at enhancing cost-efficiency. Given that GPT-4V is not currently free, direct API utilization for large-scale VQoSS estimation is not feasible. Our experiments with the proposed SQ-GPT demonstrated that, after distillation, our model can achieve performance comparable to GPT-4V at a reduced cost. This allows for its application in large-scale estimation tasks across various tasks for VQoSS. As the field of MLLMs evolves rapidly, GPT-4V, despite its effectiveness in numerous tasks, may be superseded in the future. Consequently, our pipeline is significant for users wishing to elevate their models to state-of-the-art MLLM technology.

In summary, we demonstrated that MLLMs can effectively interpret visual information for VQoSS estimation. This approach marks a significant departure from traditional deep learning methods, reducing the need for human-annotated data and enhancing scalability for large-scale applications. Our methodology leverages MLLMs to analyze street-view images, assessing various urban quality indicators. Furthermore, we introduced a cost-efficient distillation pipeline with our SQ-GPT model, enabling comparable performance to GPT-4V but at a reduced cost. This development is crucial for practical, large-scale VQoSS assessments and future adaptations to advancing MLLM technologies, representing a significant theoretical and practical contribution to the field.

TABLE 4. Evaluation on samples with both positive and negative changes.

	Task A	Task B	Task E	Task G	Overall
Number	8	25	44	61	138
Positive	50.0%	56.0%	52.3%	45.9%	51.1%
Negative	12.5%	12.0%	15.9%	21.3%	15.4%
No Changes	37.5%	32.0%	31.8%	32.8%	33.5%

A. LIMITATIONS AND FUTURE WORKS

The current version of SQ-GPT presents some limitations. Specifically, the integration of several frozen modules from prior research may hinder performance optimization during fine-tuning. While methodologies such as Lora [54] offer pathways for efficient training, our attempts in this direction yielded suboptimal results. Adjustments to the image encoder or LLMs adversely affected the quality of the generated answers. This could be attributed to data constraints, as the distilled dataset available for fine-tuning may be insufficient for the extensive modules in question. Addressing these issues will be a primary focus in our forthcoming research endeavors.

Furthermore, our analysis indicates that SQ-GPT's performance in tasks such as "Facade Color" recognition is not good enough, which can be primarily attributed to the challenges posed by varying light intensities and angles during image capturing. These environmental factors can significantly alter the perceived color, leading to inaccuracies in the model's answering. To counteract these issues, future version of the pipeline could incorporate advanced pre-processing algorithms that normalize lighting conditions or employ more sophisticated techniques for color constancy. Additionally, enhancing the dataset with a wider array of images captured under diverse lighting conditions may improve the model's robustness to such environmental variables. By addressing these nuances, we can refine SQ-GPT's capabilities for more accurate and reliable analysis.

We also investigated the ability of the MLLM to distinguish and detail multiple elements present in imagery, such as facades, signboards, and billboards, as well as to assess changes in scenarios like parking improvements and scenery and facilities construction. Given the complexity of such images, where both positive and negative changes might be present simultaneously, we were particularly interested in the MLLM's discriminative power and its response tendency. Upon analyzing the generated answers from all our training data, we observed instances where a single scene depicted both negative and positive changes. Specifically, shown in Table 4, we identified 8 data points for Task A, 25 for Task B, 44 for Task E, and 61 for Task G that presented such duality. In these scenarios, GPT-4V exhibited a tendency to respond with ‘No Changes’ in 33.5% of cases, ‘Positive’ changes in 51.1%, and ‘Negative’ changes in 15.4%. This indicates a propensity for GPT-4V to report positive changes more frequently. Further examination revealed that the MLLM lacks the ability to discern all details in the images precisely. However, it’s important to note that such cases were relatively rare and did not significantly impact the overall analysis. This insight into the MLLM’s capabilities and limitations is crucial for understanding its application in complex urban imagery analysis, where multiple elements and contrasting changes coexist.

Comparison to previous deep learning based works is also important. A key issue we have identified is the complexity of hyperparameter settings in previous methodologies, which not only adds to the intricacy of the models but also often obscures their replicability and comparison. Additionally, a significant barrier to progress in this field has been the lack of open-source data and code, hindering transparent and fair comparisons across different methods. Our future efforts will be directed towards creating and contributing open-source datasets and code related to our research. This initiative will not only facilitate more straightforward comparisons with existing methods but also foster a collaborative and open environment in the research community. By addressing these issues, we aim to refine the efficiency and accuracy of our models, making them more user-friendly and replicable, thus advancing the field in a meaningful and measurable way.

VI. CONCLUSION

In this paper, we proposed SQ-GPT, aimed at assessing the visual quality of street spaces using the capabilities of MLLMs like GPT-4V. We also proposed our own model SQ-GPT which comprises several key components: an image encoder, a QFormer for aligning images with text, and a refinement stage to fine-tune the process. Through efficient knowledge distillation, SQ-GPT manages to closely mirror GPT-4V’s performance while significantly cutting down on costs, thereby democratizing the technology for broader use. We utilized a dataset from Baidu Street View in Nanjing, Gulou district, capturing images from two distinct time frames to monitor changes over time.

We also assessed the model’s performance through both human-centric evaluation and visual analysis, which confirmed its effectiveness in various tasks. Nevertheless, we recognized the model’s limitations, such as the challenge of inconsistent lighting conditions affecting color recognition, and proposed future improvements, including the development of sophisticated pre-processing algorithms. In conclusion, we showed that SQ-GPT can perform on par with GPT-4V in quality assessments but at a more affordable rate, which is highly beneficial for large-scale urban planning and quality assessment projects. We also acknowledged the necessity for ongoing research to surmount the existing limitations and to boost the model’s adaptability and sturdiness across various urban settings. We believe that this paper could be a pioneer work that inspires researchers using MLLMs for the future efficient and high-quality urban planning.

REFERENCES

[1] K. Mouratidis, “Urban planning and quality of life: A review of pathways linking the built environment to subjective well-being,” *Cities*, vol. 115, Aug. 2021, Art. no. 103229.

[2] T.-J. Liu, “Study of visual quality assessment on pattern images: Subjective evaluation and visual saliency effects,” *IEEE Access*, vol. 6, pp. 61432–61444, 2018.

[3] K. Meyrick and P. Newman, “Exploring the potential connection between place capital and health capital in the post COVID-19 city,” *Npj Urban Sustainability*, vol. 3, no. 1, p. 44, Jul. 2023.

[4] A. Sharifi and A. R. Khavarian-Garmsir, “The COVID-19 pandemic: Impacts on cities and major lessons for urban planning, design, and management,” *Sci. Total Environ.*, vol. 749, Dec. 2020, Art. no. 142391.

[5] J. Tang and Y. Long, “Measuring visual quality of street space and its temporal variation: Methodology and its application in the hutong area in Beijing,” *Landscape Urban Planning*, vol. 191, Nov. 2019, Art. no. 103436.

[6] L. Song, “The fine design strategy of urban streets using deep learning with the assistance of the Internet of Things,” *IEEE Access*, vol. 11, pp. 67518–67525, 2023.

[7] W. H. Whyte, *The Social Life of Small Urban Spaces*. Washington, DC, USA: Conservation Foundation, 1980.

[8] A.-L. Balogun, D. Marks, R. Sharma, H. Shekhar, C. Balmes, D. Maheng, A. Arshad, and P. Salehi, “Assessing the potentials of digitalization as a tool for climate change adaptation and sustainable development in urban centres,” *Sustain. Cities Soc.*, vol. 53, Feb. 2020, Art. no. 101888.

[9] F. Biljecki and K. Ito, “Street view imagery in urban analytics and GIS: A review,” *Landscape Urban Planning*, vol. 215, Nov. 2021, Art. no. 104217.

[10] M. Wang, Y. He, H. Meng, Y. Zhang, B. Zhu, J. Mango, and X. Li, “Assessing street space quality using street view imagery and function-driven method: The case of Xiamen, China,” *ISPRS Int. J. Geo-Inf.*, vol. 11, no. 5, p. 282, 2022.

[11] C. Toth and G. Jóźków, “Remote sensing platforms and sensors: A survey,” *ISPRS J. Photogramm. Remote Sens.*, vol. 115, pp. 22–36, May 2016.

[12] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, “Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities,” *IEEE J. Sel. Topics Appl. Earth Observat. Remote Sens.*, vol. 13, pp. 3735–3756, 2020.

[13] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 7553.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[15] Z. Gong, Q. Ma, C. Kan, and Q. Qi, “Classifying street spaces with street view images for a spatial indicator of urban functions,” *Sustainability*, vol. 11, no. 22, p. 6424, 2019.

[16] Y. Li, N. Yabuki, and T. Fukuda, “Measuring visual walkability perception using panoramic street view images, virtual reality, and deep learning,” *Sustain. Cities Soc.*, vol. 86, Nov. 2022, Art. no. 104140.

[17] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Jan. 2009.

[18] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, “A survey on deep transfer learning,” in *Artificial Neural Networks and Machine Learning—ICANN*, Rhodes, Greece. Springer, 2018, pp. 270–279.

[19] B. Wang, L. Li, M. Verma, Y. Nakashima, R. Kawasaki, and H. Nagahara, “MTUNet: Few-shot image classification with visual explanations,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 2294–2298.

[20] A. Aurigi and N. Odendaal, “From ‘smart in the box’ to ‘smart in the city’: Rethinking the socially sustainable smart city in context,” in *Sustainable Smart City Transitions*. Evanston, IL, USA: Routledge, 2022, pp. 53–68.

[21] J. Wang, X. Wang, T. Shen, Y. Wang, L. Li, Y. Tian, H. Yu, L. Chen, J. Xin, X. Wu, N. Zheng, and F.-Y. Wang, “Parallel vision for long-tail regularization: Initial results from IVFC autonomous driving testing,” *IEEE Trans. Intell. Vehicles*, vol. 7, no. 2, pp. 286–299, Jun. 2022.

[22] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, and Y. Du, "A survey of large language models," 2023, *arXiv:2303.18223*.

[23] (2023). *OpenAI ChatGPT*. [Online]. Available: https://www.openai.com/research/chatgpt

[24] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," 2023, *arXiv:2306.13549*.

[25] P. Grifoni, A. D'ulizia, and F. Ferri, "When language evolution meets multimodality: Current status and challenges toward multimodal computational models," *IEEE Access*, vol. 9, pp. 35196–35206, 2021.

[26] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, and R. Avila, "GPT-4 technical report," 2023, *arXiv:2303.08774*.

[27] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," 2023, *arXiv:2301.12597*.

[28] K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, and Y. Qiao, "VideoChat: Chat-centric video understanding," 2023, *arXiv:2305.06355*.

[29] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proc. NIPS*, 2023, pp. 1–25.

[30] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing vision-language understanding with advanced large language models," 2023, *arXiv:2304.10592*.

[31] H. Nori, N. King, S. Mayer McKinney, D. Carignan, and E. Horvitz, "Capabilities of GPT-4 on medical challenge problems," 2023, *arXiv:2303.13375*.

[32] P. Chen, S. Liu, H. Zhao, and J. Jia, "Distilling knowledge via knowledge review," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5006–5015.

[33] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, Jun. 2021.

[34] C. Wei, J. Meng, L. Zhu, and Z. Han, "Assessing progress towards sustainable development goals for Chinese urban land use: A new cloud model approach," *J. Environ. Manag.*, vol. 326, Jan. 2023, Art. no. 116826.

[35] M. Liu, B. Zhang, T. Luo, Y. Liu, B. A. Portnov, T. Trop, W. Jiao, H. Liu, Y. Li, and Q. Liu, "Evaluating street lighting quality in residential areas by combining remote sensing tools and a survey on pedestrians' perceptions of safety and visual comfort," *Remote Sens.*, vol. 14, no. 4, p. 826, Feb. 2022.

[36] Y. Li, N. Yabuki, and T. Fukuda, "Integrating GIS, deep learning, and environmental sensors for multicriteria evaluation of urban street walkability," *Landscape Urban Planning*, vol. 230, Feb. 2023, Art. no. 104603.

[37] R. Mahabir, R. Schuchard, A. Crooks, A. Croitoru, and A. Stefanidis, "Crowdsourcing street view imagery: A comparison of mapillary and OpenStreetCam," *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 6, p. 341, May 2020.

[38] J. Zhang, Z. Yu, Y. Li, and X. Wang, "Uncovering bias in objective mapping and subjective perception of urban building functionality: A machine learning approach to urban spatial perception," *Land*, vol. 12, no. 7, p. 1322, Jun. 2023. [Online]. Available: https://www.mdpi.com/2073-445X/12/7/1322

[39] B. Wang, J. Zhang, R. Zhang, Y. Li, L. Li, and Y. Nakashima, "Improving facade parsing with vision transformers and line integration," *Adv. Eng. Informat.*, vol. 60, Apr. 2024, Art. no. 102463.

[40] B. Wang, L. Li, Y. Nakashima, and H. Nagahara, "Learning bottleneck concepts in image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023.

[41] K. Clifton, R. Ewing, G. Knaap, and Y. Song, "Quantitative analysis of urban form: A multidisciplinary review," *J. Urbanism, Int. Res. Placemaking Urban Sustainability*, vol. 1, no. 1, pp. 17–45, Mar. 2008.

[42] J. Zhang, T. Fukuda, and N. Yabuki, "Automatic generation of synthetic datasets from a city digital twin for use in the instance segmentation of building facades," *J. Comput. Des. Eng.*, vol. 9, no. 5, pp. 1737–1755, Sep. 2022.

[43] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[44] X. Li, Y. Li, T. Jia, L. Zhou, and I. H. Hijazi, "The six dimensions of built environment on urban vitality: Fusion evidence from multi-source data," *Cities*, vol. 121, Feb. 2022, Art. no. 103482.

[45] B. W. Koo, S. Guhathakurta, and N. Botchwey, "How are neighborhood and street-level walkability factors associated with walking behaviors? A big data approach using street view images," *Environ. Behav.*, vol. 54, no. 1, pp. 211–241, Jan. 2022.

[46] E. J. Hoffmann, Y. Wang, M. Werner, J. Kang, and X. X. Zhu, "Model fusion for building type classification from aerial and street view images," *Remote Sens.*, vol. 11, no. 11, p. 1259, May 2019.

[47] T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, E. L. Aiden, and L. Fei-Fei, "Using deep learning and Google street view to estimate the demographic makeup of neighborhoods across the united states," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 50, pp. 13108–13113, Dec. 2017.

[48] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[49] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "InstructBLIP: Towards general-purpose vision-language models with instruction tuning," 2023, *arXiv:2305.06500*.

[50] R. Zhang, J. Han, C. Liu, P. Gao, A. Zhou, X. Hu, S. Yan, P. Lu, H. Li, and Y. Qiao, "LLaMA-adapter: Efficient fine-tuning of language models with zero-init attention," 2023, *arXiv:2303.16199*.

[51] P. Gao, J. Han, R. Zhang, Z. Lin, S. Geng, A. Zhou, W. Zhang, P. Lu, C. He, X. Yue, H. Li, and Y. Qiao, "LLaMA-adapter v2: Parameter-efficient visual instruction model," 2023, *arXiv:2304.15010*.

[52] T. Gong, C. Lyu, S. Zhang, Y. Wang, M. Zheng, Q. Zhao, K. Liu, W. Zhang, P. Luo, and K. Chen, "Multimodal-GPT: A vision and language model for dialogue with humans," 2023, *arXiv:2305.04790*.

[53] Y. Su, T. Lan, H. Li, J. Xu, Y. Wang, and D. Cai, "PandaGPT: One model to instruction-follow them all," 2023, *arXiv:2305.16355*.

[54] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Represent.*, 2022.

[55] A. Awadalla, I. Gao, J. Gardner, J. Hessel, Y. Hanafy, W. Zhu, K. Marathe, Y. Bitton, S. Gadre, S. Sagawa, J. Jitsev, S. Kornblith, P. Wei Koh, G. Ilharco, M. Wortsman, and L. Schmidt, "OpenFlamingo: An open-source framework for training large autoregressive vision-language models," 2023, *arXiv:2308.01390*.

[56] M. Wu, Q. Huang, S. Gao, and Z. Zhang, "Mixed land use measurement and mapping with street view images and spatial context-aware prompts via zero-shot multimodal learning," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 125, Dec. 2023, Art. no. 103591.

[57] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, and J. Clark, "Learning transferable visual models from natural language supervision," in *Proc. ICML*, 2021, pp. 8748–8763.

[58] D. Verma, O. Mumm, and V. Miriam Carlow, "Generative agents in the streets: Exploring the use of large language models (LLMs) in collecting urban perceptions," 2023, *arXiv:2312.13126*.

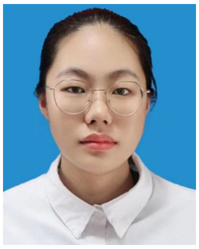
[59] Z. Xu, Y. Shen, and L. Huang, "MultiInstruct: Improving multi-modal zero-shot learning via instruction tuning," 2022, *arXiv:2212.10773*.

[60] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Computer Vision—ECCV*, Zurich, Switzerland. Springer, 2014, pp. 740–755.

[61] S. Mukherjee, A. Mitra, G. Jawahar, S. Agarwal, H. Palangi, and A. Awadallah, "Orca: Progressive learning from complex explanation traces of GPT-4," 2023, *arXiv:2306.02707*.

[62] Y. Wang, W. Chen, X. Han, X. Lin, H. Zhao, Y. Liu, B. Zhai, J. Yuan, Q. You, and H. Yang, "Exploring the reasoning abilities of multimodal large language models (MLLMs): A comprehensive survey on emerging trends in multimodal reasoning," 2024, *arXiv:2401.06805*.

[63] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.



HAO LIANG is currently pursuing the degree in landscape architecture with the College of Landscape Architecture, Nanjing Forestry University. Her academic pursuits are centered around the digitalization of urban and landscape. She has been actively involved in research projects focusing on human settlements.

JIAXIN ZHANG received the bachelor's degree in architecture from Nanchang University, in 2016, the master's degree from Southeast University, in 2019, and the Ph.D. degree from the School of Energy and Environment, Osaka University, in 2022. Since November 2022, he has been a specially-appointed Researcher with the Laboratory of Environmental Design and Information Technology, Osaka University. His research interests include city perception, integrating machine learning in urban studies, and pioneering automated measurements of architectural facades. His work reflects his architectural acumen and passion for innovative technology.

YUNQIN LI received the bachelor's degree from Nanchang University, in 2016, the master's degree from Southeast University, in 2019, and the Ph.D. degree from Osaka University, in 2022. Currently, she is a Visiting Researcher with Osaka University. Her research interests include spatial auditing, measuring, perception, understanding, and interaction, backed by new data, technologies, and methods. She has a particular interest in street perception and explainable machine learning. Despite her achievements, she remains humble in her continuous exploration of the intricate relationship between technology, space, and urban environments.

BOWEN WANG (Member, IEEE) received the B.S. degree in computer science from Anhui University, China, in 2016, and the M.S. degree in medical information and the Ph.D. degree in computer science from Osaka University, Japan, in 2020 and 2023, respectively. He is currently a specially-appointed Researcher with the Institute for Datability Science (IDS), Osaka University. His research interests include computer vision, explainable AI, city perception, and medical AI. He is a member of ACM and IPSJ. He has received the Best Paper Award at APAMI 2020.

JINGYONG HUANG is currently an Associate Professor with the Department of Architecture, School of Architecture and Design, Nanchang University. He has been appointed as a member of the Architectural Education Branch of the Architectural Society of China, the Vice Chair of the Green Building and Building Carbon Emission Professional Committee of Jiangxi Provincial Survey and Design Industry Association, and the Vice Chair of the Architectural Creation (Architect) Professional Committee. Additionally, he serves as a member for Jiangxi Provincial Urban Physical Examination and Urban Renewal Expert Database, a Reviewer for Jiangxi's Scenic Spots and Natural Heritage, and an Expert for Nanchang Municipal Architectural Art Committee.

• • •