

UrBench: A Comprehensive Benchmark for Evaluating Large Multimodal Models in Multi-View Urban Scenarios

Baichuan Zhou^{1*}, Haote Yang^{1*}, Dairong Chen^{4,2*}, Junyan Ye^{2,1*},
Tianyi Bai¹, Jinhua Yu², Songyang Zhang¹, Dahua Lin^{1,3},
Conghui He^{1,3†}, Weijia Li^{2†}

¹Shanghai AI Laboratory

²Sun Yat-Sen University

³Sensetime Research

⁴Wuhan University

{chendr7, yejy53, yujh56}@mail2.sysu.edu.cn, liweij29@mail.sysu.edu.cn,
{zhoubaichuan, yanghaote, baitianyi, zhangsongyang, lindahua, heconghui}@pjlab.org.cn

Abstract

Recent evaluations of Large Multimodal Models (LMMs) have explored their capabilities in various domains, with only few benchmarks specifically focusing on urban environments. Moreover, existing urban benchmarks have been limited to evaluating LMMs with basic region-level urban tasks under singular views, leading to incomplete evaluations of LMMs' abilities in urban environments. To address these issues, we present UrBench, a comprehensive benchmark designed for evaluating LMMs in complex multi-view urban scenarios. UrBench contains 11.6K meticulously curated questions at both region-level and role-level that cover 4 task dimensions: Geo-Localization, Scene Reasoning, Scene Understanding, and Object Understanding, totaling 14 task types. In constructing UrBench, we utilize data from existing datasets and additionally collect data from 11 cities, creating new annotations using a cross-view detection-matching method. With these images and annotations, we then integrate LMM-based, rule-based, and human-based methods to construct large-scale high-quality questions. Our evaluations on 21 LMMs show that current LMMs struggle in the urban environments in several aspects. Even the best performing GPT-4o lags behind humans in most tasks, ranging from simple tasks such as counting to complex tasks such as orientation, localization and object attribute recognition, with an average performance gap of 17.4%. Our benchmark also reveals that LMMs exhibit inconsistent behaviors with different urban views, especially with respect to understanding cross-view relations.

Project — <https://opendatalab.github.io/UrBench/>

Appendix — <https://github.com/opendatalab/UrBench/blob/master/static/appendix.pdf>

Introduction

Recently, the research community has witnessed an emergent interest in developing Large Multimodal Models (LMMs) (Achiam et al. 2023; Liu et al. 2024b; Chen et al. 2024) that have exhibited impressive abilities in a variety of

*These authors contributed equally.

†W. Li and C. He are the corresponding authors.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

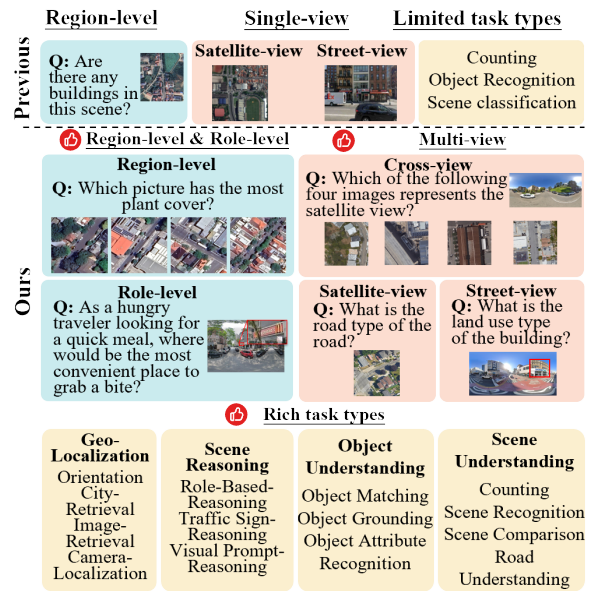


Figure 1: Comparison between UrBench and previous works. (1) UrBench contains both region-level and role-level questions, while previous benchmarks generally focus on region-level questions. (2) In addition to single-view questions in satellite or street view, UrBench also incorporates cross-view questions. (3) It evaluates LMMs on a comprehensive range of 14 diverse tasks in 4 evaluation dimensions.

benchmarks (Yue et al. 2024; Liu et al. 2023b). The central purpose behind these explorations is to build human-centric AI models that can serve as helpful assistants for everyday life. Given that over 57% of the global population resides in urban areas (World Bank 2024), it is crucial that these AI models should be capable of performing a variety of urban tasks, such as assisting government officials to manage urban development and facilitating citizens to make decisions in daily life (Zhou et al. 2024b; Feng et al. 2024a). On the other hand, urban areas are often captured from various perspectives, including the vertical view from satellite or aerial

imagery and the horizontal view from street-view imagery. To be truly effective in assisting the large urban population, AI models should also be capable of comprehensively understanding these environments from multiple perspectives.

To better evaluate and develop human-centric AI models, several works have explored the prospect of LMMs for urban environments. For example, various works evaluate the capabilities of LMMs on region-level visual recognition (Hao et al. 2024; Yan et al. 2024) and urban planning (Zhou et al. 2024b). Besides, researchers also examine the performance of LMMs with remote sensing images (Li, Ding, and El-hoseiny 2024; Kuckreja et al. 2024). However, as shown in Fig.1, while these studies primarily focus on urban understanding at a region level, they neglect human-centric tasks within the urban scenarios. A more comprehensive approach should address urban tasks across multiple levels, from region-level recognition tasks to role-level tasks such as geo-localization and scene understanding.

Another important aspect of the urban environments is that they are usually captured by multiple different perspectives. As each perspective offers unique information, it is vital for LMMs to comprehend different perspectives to complete certain tasks. For instance, geo-localization tasks require satellite-view imagery for spatial orientation and street-view imagery for detailed contexts. LMMs must utilize both perspectives to successfully perform geo-localization. Given the importance of understanding urban environments from multiple perspectives, current benchmarks that only evaluate LMMs on single-view data (Wang et al. 2024b; Feng et al. 2024b), as shown in Fig.1, are incomprehensive. Therefore, it is crucial to develop a multi-view benchmark to accurately evaluate LMMs under complex urban settings. However, one of the key challenges of curating such datasets is create annotations for cross-view scenarios (Zhu, Yang, and Chen 2021; Ye et al. 2024a). While paired street and satellite images are easy to acquire, creating questions about cross-view relations remains challenging due to lack of annotations (Li et al. 2023b).

To address these challenges, we propose UrBench, a multi-task, multi-view benchmark designed for comprehensively evaluating LMMs in urban environments. UrBench comprises over 11.6K questions across 14 tasks spanning four dimensions: Geo-Localization, Scene Understanding, Scene Reasoning, and Object Understanding. UrBench includes both region-level tasks from previous benchmarks as well as role-level tasks aimed at assisting humans in daily life. Additionally, considering the multi-view characteristics of urban environments, UrBench incorporates multiple urban perspectives to evaluate LMMs’ capabilities in understanding complex multi-view relations. In constructing UrBench, we introduce a novel cross-view detection and matching method to create multi-view annotations. We then utilized these images and annotations to construct our high-quality and diverse set of questions with various methods. As shown in Fig.2, our evaluation results indicate that LMMs lag behind human experts in most tasks, highlighting their limitations towards human-centric assistants in urban environments. Our contributions are summarized as follows:

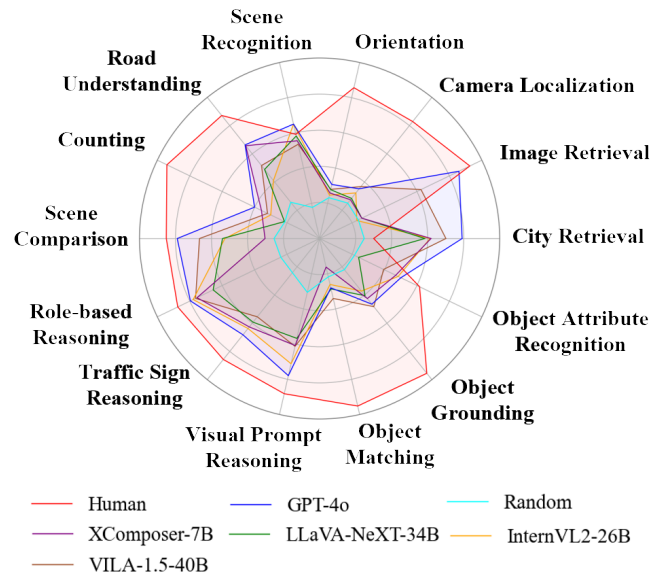


Figure 2: The performances of the 5 leading LMMs, as well as that of the human and random guess, on UrBench.

- We propose UrBench, a multi-view benchmark designed to evaluate LMMs’ performances in urban environments. Our benchmark includes 14 urban tasks that we categorize into various dimensions. These tasks encompass both region-level evaluations that assess LMMs’ capabilities in urban planning, as well as role-level evaluations that examine LMMs’ responses to daily issues.
- We introduce a novel benchmark curation pipeline that involves a cross-view detection-matching algorithm for object-level annotation generation and a question generation approach that integrates LMM-based, rule-based, and human-based methods. This pipeline ensures the creation of a large-scale and high-quality corpus of questions, significantly enhancing the diversity and depth of evaluation across multiple urban tasks.
- We evaluate 21 popular LMMs on UrBench. Our evaluation results show that current models lag behind human experts in most tasks and reveal LMMs’ inconsistent behaviors with different urban views, which demonstrates the limitations of current LMMs in urban environments.

Related Work

Large Multimodal Models

Building on the strengths of Large Language Models (LLMs) (Brown et al. 2020; Touvron et al. 2023) in complex language reasoning and understanding, Large Multimodal Models (LMMs) can process inputs from multiple modalities and accomplish sophisticated visual reasoning and understanding tasks (Yin et al. 2023; Bai et al. 2024). The rapid growth of LMMs has given rise to both closed-source models like GPT-4o (Achiam et al. 2023) and Gemini (Reid et al. 2024), as well as open-source models such as LLaVA (Liu et al. 2024b), and VILA (Lin et al. 2024), all

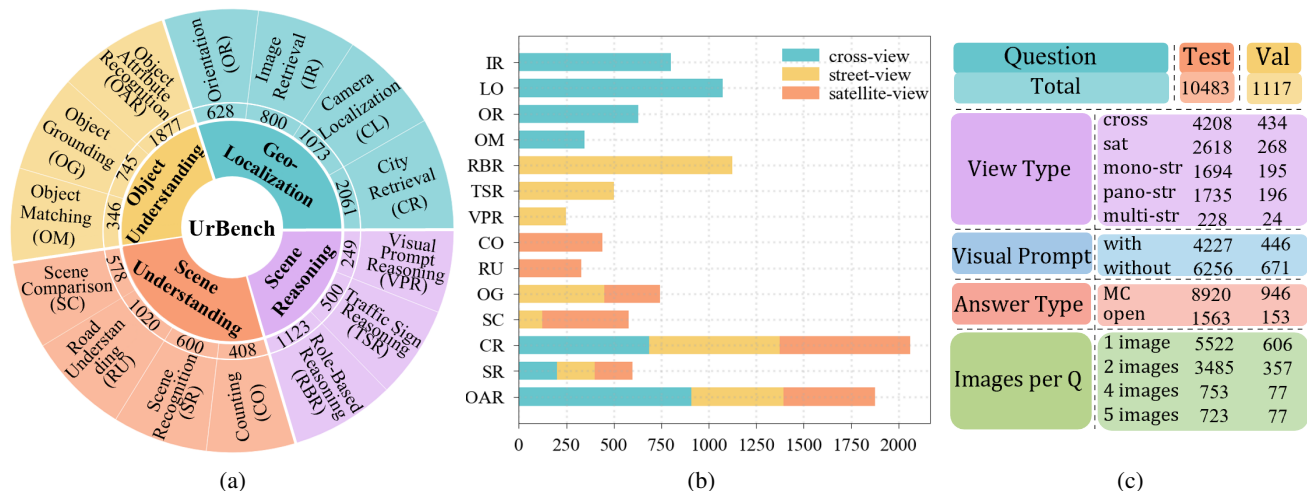


Figure 3: (a) The 14 types of tasks under 4 evaluation dimensions. (b) The view types of each task. (c) The statistics of UrBench. cross, sat, and str are the abbreviations for cross-view, satellite-view, and street-view. mono, pano, and multi are the abbreviations of monocular, panoramic, and multiple. MC and open means multiple-choice and open-ended, respectively.

of which have demonstrated significant potential for various application tasks (Cui et al. 2024; Xiao et al. 2024).

For urban-related tasks, recent developments show growing interest in utilizing LLMs and CLIP (Radford et al. 2021), covering aspects such as urban planning and vision-language navigation (Zhou et al. 2024b; Schumann et al. 2024). For instance, UrbanCLIP (Yan et al. 2024) leverages LLMs (Touvron et al. 2023) and CLIP (Radford et al. 2021) for urban region profiling using remote sensing images, while Velma (Schumann et al. 2024) combines LLMs with CLIP (Radford et al. 2021) for street view navigation. Scene-LLM leverages LLMs for multi-view 3D reasoning in in-door setups, while CityGPT (Feng et al. 2024a) studies LLMs’ performances in urban spatial understanding tasks.

Multimodal Benchmarks

With the rapid advancement of LMMs, traditional multimodal question answering benchmarks like VQA (Goyal et al. 2017) and GQA (Hudson and Manning 2019) have become inadequate for fully assessing LMM capabilities. Recently, more comprehensive benchmarks are introduced to better evaluate LMMs. For example, MME (Fu et al. 2023) is one of the first to thoroughly assess LMMs across 14 perception and reasoning tasks. MMMU (Yue et al. 2024) benchmarks expert-level knowledge using college-level questions, showing that current models still lag behind human experts. Newer benchmarks, such as those in (Jiang et al. 2024; Wang et al. 2024a), focus on multi-image reasoning, and MUIRBench (Wang et al. 2024a) includes tasks with unanswerable questions. While these works extensively evaluate LMMs in visual reasoning and multi-image understanding, few analyze performance in urban environments. Our work fills this gap by constructing a comprehensive benchmark for evaluating LMMs’ potential in urban planning, reasoning, and understanding from multi-views.

There is also a growing body of work focused on bench-

marking LMMs in the remote sensing domain. Early efforts like RSVQA (Lobry et al. 2020) comprises visual recognition tasks such as classification and detection for image sensing images. EarthVQA (Wang et al. 2024b) feature remote sensing image-question pairs centered on attributes of ground objects in urban areas. RSIEval (Hu et al. 2023) and LHRS-Bench (Muhtar et al. 2024) adapt existing remote sensing datasets to create visual reasoning benchmarks for LMMs, while Geochat (Kuckreja et al. 2024) primarily assesses regional perception capabilities. VRSBench (Li, Ding, and Elhoseiny 2024) utilizes GPT-4 (Achiam et al. 2023) to generate visual question answering data focused on object relations. More recently, CityBench (Feng et al. 2024b) evaluates LMMs in urban environments but with a limited range of tasks. Overall, these existing benchmarks are limited in task variety and lack multi-view samples. Moreover, because the image data in these works are primarily repurposed from existing datasets (Xia et al. 2018; Sun et al. 2022), their geographical diversity is constrained. In contrast, our benchmark introduces a new pipeline for collecting multi-view and multi-image data, expanding the range of tasks and incorporating images from diverse geolocations with multiple perspectives.

UrBench

Benchmark Analysis

Overview. We introduce UrBench, a novel benchmark designed for evaluating LMMs in urban scenarios. As detailed in Fig.3(c), UrBench comprises 11.6K questions, which are divided into a validation set for hyperparameter selection and a test set for evaluation. The validation set and the test set contain approximately 1.1K and 10.5K questions, respectively. Please refer to the appendix for further statistical details. The UrBench is characterized by the following features: (1) UrBench integrates street-view, satellite-view, and street-satellite cross-view images, offering a more com-

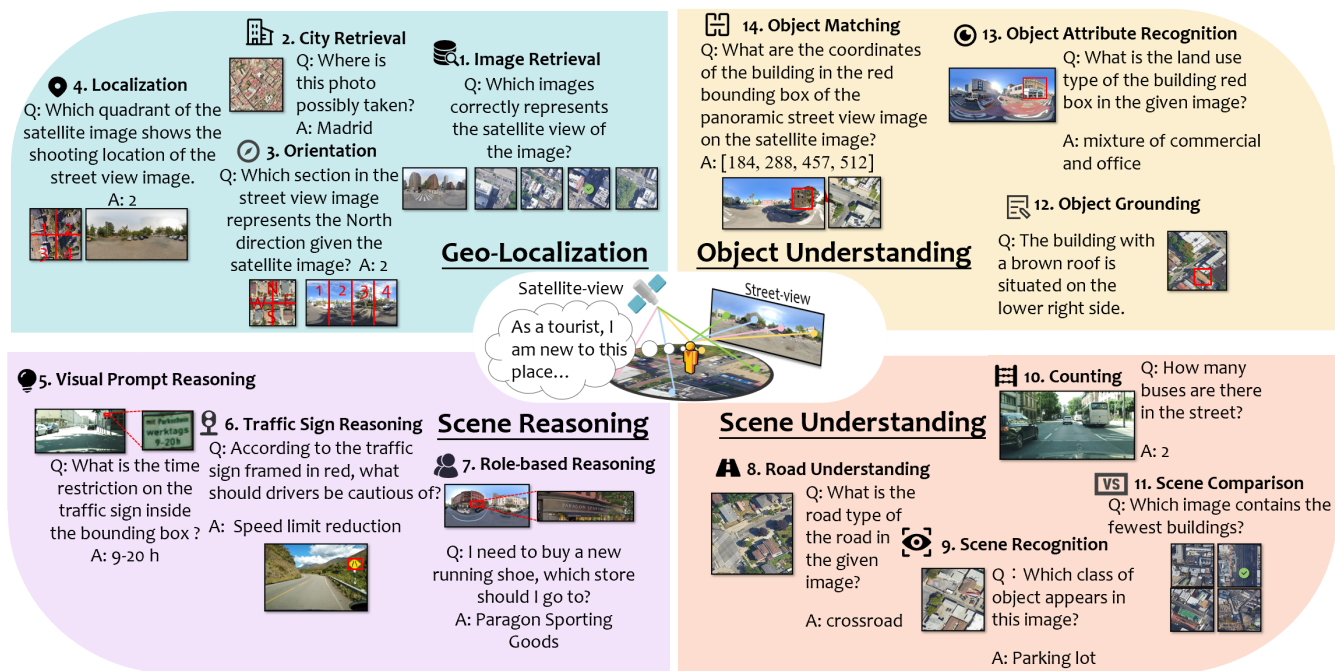


Figure 4: UrBench consists of 14 different task types, categorized into four evaluation dimensions based on the capacities and the granularity of the objects of interest assessed by the questions.

prehensive understanding of urban scenarios (Fig.3(b)). (2) UrBench evaluates the capability of LMMs focusing on urban scenarios from comprehensive dimensions, including Geo-Localization, Scene Reasoning, Scene Understanding, and Object Understanding, with a total of 14 task types (Fig.3(a)). (3) The questions of UrBench are generated by an integrated approach, encompassing model-based, rule-based, and human-based methods, which ensures the generation of a substantial and high-quality corpus of questions. **Comparison with existing benchmarks.** While general benchmarks such as MUIRBench (Wang et al. 2024a) and MMMU (Yue et al. 2024) evaluate the capacities of LMMs in general scenarios, typically from a single view, our benchmark is focused on urban scenarios from more different perspectives. On the other hand, unlike existing benchmarks in urban scenarios such as CityBench (Feng et al. 2024b) and EarthVQA (Wang et al. 2024b) that place significant emphasis on single-view images and a limited range of tasks, our benchmark incorporates questions that utilize multi-view images and cover more diverse task types.

Benchmark Tasks

UrBench comprehensively evaluates LMMs in urban scenarios from four evaluation dimensions. Fig.4 illustrates the specific task types under each evaluation dimension. Please refer to the appendix for additional task questions.

Geo-Localization. This dimension contains role-level tasks widely used in remote-sensing (Zhu, Yang, and Chen 2021), which requires LMMs to predict geographical coordinates and directions given images (Task 1-4 in Fig.4). For example, in Image Retrieval (IR), we query LMMs with satellite

or street view images to retrieve their corresponding counterpart, while City Retrieval (CR) tasks LMMs to predict the name of a city given its satellite or street view images. Orientation (OR) and Camera Localization (CR) require LMMs to pinpoint directions given cross-view information.

Scene Reasoning. Reasoning about urban scenes is crucial for assisting humans in urban environments. In this dimension, we design three role-level tasks that aim to assess the reasoning capabilities of LMMs under urban multi-view scenarios (Task 5-7 in Fig.4). In Visual Prompt Reasoning (VPR), LMMs must reason about objects framed in visual prompts, while in Traffic Sign Reasoning (TSR), LMMs predict the usage of traffic signs. To better simulate real-world human usage, we design Role-based Reasoning to incorporate questions that are posed from the perspective of urban residents, such as shoppers, visitors, and city managers.

Scene Understanding. To better evaluate LMMs’ capabilities in region-level scene understanding, we design new tasks as well as adapting existing tasks to urban environments in this dimension, as shown in (Task 8-11 in Fig.4). For example, to examine whether LMMs can detect common urban regions like crosswalks, we convert the classic Counting (CO) (Ranjan et al. 2021) task to our specific settings. We design Scene Recognition (SR) and Scene Comparison (SC) to assess LMMs’ region-level understanding of urban scenes, such as recognizing building types and plant cover. Additionally, we present Road Understanding (RU), which is aimed at examining LMMs’ ability in classifying road types and understanding traffic roads.

Object Understanding. In Object Understanding, we assess the fine-grained region-level capabilities of LMMs within

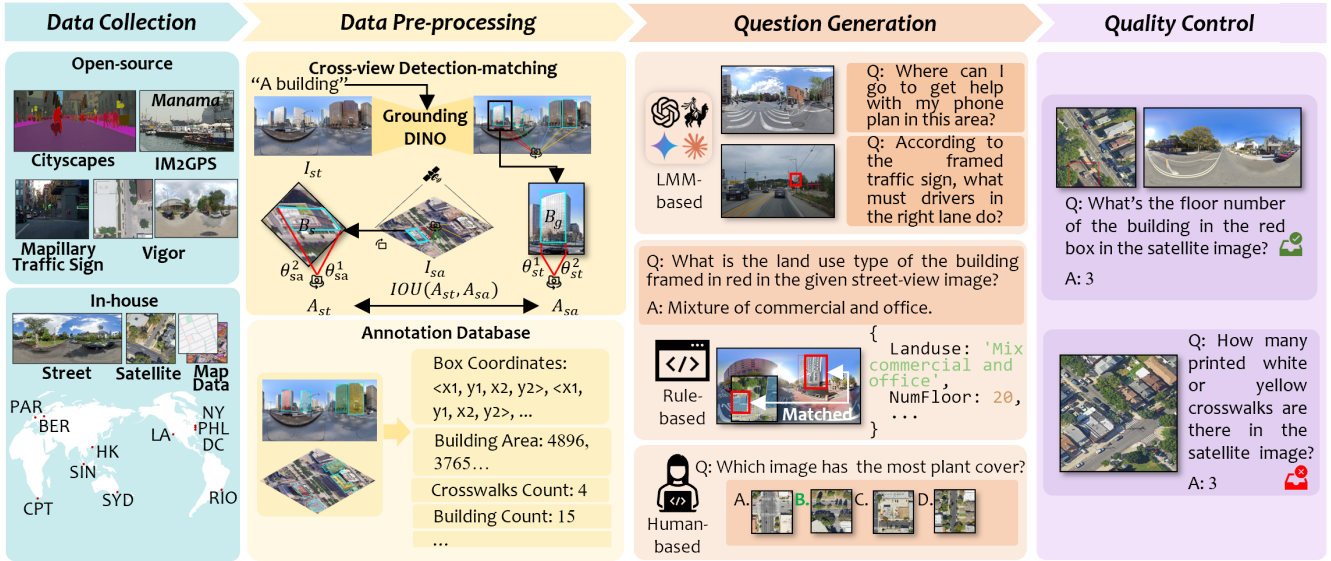


Figure 5: UrBench curation pipeline includes data collection, data pre-processing, question generation and quality control.

urban environments (Task 12-14 in Fig.4). Inspired by previous grounding tasks (Yu et al. 2016), we include Object Grounding (OG) to evaluate LMMs’ abilities to ground text phrases of objects in images across different views. To further challenge the spatial understanding of LMMs, we propose Object Matching (OM), a cross-view task where LMMs predict the locations of objects in satellite and street views based on their cross-view correspondences. Object Attribute Recognition (OAR) prompts LMMs to predict ground object attributes such as building floors and land use.

Benchmark Curation

Data Collection. As outlined in the data collection stage in Fig.5, there are two data sources of UrBench, the in-house data collected by ourselves and data from open datasets. The in-house data contains 2,604 street-view images from Google Street View and 4,239 satellite-view images from Google Earth (Level 19). Among these images, 1,965 street-satellite image pairs are fit together according to their geographical coordinates. In addition, each satellite-view image is equipped with some ground object annotation from OpenStreetMap¹. We follow previous works (Zhu, Yang, and Chen 2021) to ensure no significant temporal differences between satellite and street-view images, which were all collected in 2022-2023. To support more urban tasks, we also collect images from existing open source datasets, including Cityscapes (Cordts et al. 2016), Mapillary Traffic Sign Dataset (Ertler et al. 2020), VIGOR dataset (Zhu, Yang, and Chen 2021), and IM2GPS (Hays and Efros 2008).

Data Pre-processing. In this stage, we process our collected raw image data to produce annotations for later stages. For cross-view tasks that require object matching, we develop a cross-view detection-matching method to obtain paired-up instance level annotations. Specifically, We first use a pre-

trained Grounding DINO (Liu et al. 2023a) to obtain bounding box annotations for street-view images. Since bounding box annotations for satellite-view images are already obtained through OSM, we apply ray tracing to map the street-view boxes to the satellite view. Next, we calculate the IoUs between the mapped and original satellite-view boxes and select pairs with IoUs over 0.5 as cross-view matches. We then employ human checking to further ensure the quality of our matching algorithm. This way, we effectively align bounding boxes from different views at an instance level. Additionally, we construct a comprehensive annotation database by unifying annotations from different datasets, facilitating data generation and quality control.

Question Generation. Given the nature and requirements of different tasks, we design three methods to generate question samples for UrBench. (1) LMM-based. For Scene Reasoning tasks where questions cannot be derived from our annotations, we prompt LMMs to generate Q&A pairs based on specific task settings. To mitigate bias from LMMs generating and testing their own questions, we diversify our samples by using four different LMMs: GPT-4 (Achiam et al. 2023), Gemini-1.5-Flash (Reid et al. 2024), Claude-3.5-Sonnet (Anthropic 2024), and InternVL2-26B (Chen et al. 2024). These samples are then reviewed by humans to eliminate potential hallucinations and ensure quality. (2) Rule-based. For tasks with fixed settings, such as IR and CL, we generate questions using rule-based templates given the image annotations, which allows us to automatically convert the annotations into corresponding Q&A pairs, ensuring consistency and efficiency. (3) Human-based. For tasks like SC where answers cannot be derived from annotations, we have human annotators determine the ground truths and generate the Q&A pairs.

Quality Control. At this stage, we employ human checking to alleviate potential bias during construction of UrBench. Specifically, for cross-view images, to minimize tempo-

¹<https://www.openstreetmap.org>

Model	Geo-Localization				Scene Understanding				Scene Reasoning			Object Understanding			Overall
	CR	IR	CL	OR	SR	RU	CO	SC	RBR	TSR	VPR	OM	OG	OAR	
Human	30.0	92.6	82.9	85.7	59.2	87.2	94.1	85.1	87.4	85.7	88.2	95.2	95.5	61.6	69.9
Random	24.8	23.9	25.1	23.2	17.7	25.7	21.4	25.3	23.9	24.2	30.6	21.8	22.1	21.5	23.5
GPT-4o	79.2	85.9	35.3	30.7	65.0	66.3	40.1	79.0	79.6	68.2	77.9	28.0	46.5	50.1	61.2
Gemini-1.5-Flash	69.7	25.9	25.9	24.0	57.9	71.0	29.1	67.7	77.8	75.8	69.8	22.0	39.1	40.9	50.9
Claude-3.5-Sonnet	<u>72.3</u>	55.8	30.8	33.3	52.4	59.0	48.0	81.0	73.7	37.7	66.7	22.0	61.5	45.4	<u>55.0</u>
TinyLLaVA	51.9	23.2	24.7	27.9	8.6	9.3	9.5	27.6	40.3	32.7	48.6	22.9	41.1	18.8	29.9
InternVL2-2B	50.3	23.8	31.9	29.0	47.9	47.9	28.6	30.1	64.8	45.9	54.5	25.5	30.3	40.5	41.2
InternVL2-4B	55.0	24.2	27.4	23.1	52.3	53.1	22.1	39.4	73.0	56.6	62.6	30.6	30.2	40.2	43.5
XComposer2-4KHD	61.9	26.0	27.5	25.9	55.6	65.8	35.8	30.3	75.5	62.6	60.8	16.2	42.6	46.9	47.8
LLaVA-NeXT-7B-Mistral	51.6	25.9	24.2	24.0	55.6	47.3	34.1	28.4	59.2	42.7	45.5	43.9	33.1	30.0	39.2
LLaVA-NeXT-7B-Vicuna	51.2	24.6	27.2	23.3	56.1	20.2	34.3	25.9	49.6	51.5	54.5	31.8	27.2	31.9	37.1
InstructBLIP-Vicuna-7B	40.4	25.7	25.4	27.1	33.0	14.8	20.4	28.0	30.7	25.7	29.3	22.9	22.7	17.3	27.6
LLaVA-NeXT-Interleave-7B	57.9	41.6	27.6	25.5	52.6	50.7	37.3	48.4	65.8	55.9	63.1	37.3	37.6	41.5	40.4
Mantis-LLaMA3-SigLIP	67.0	32.4	27.0	27.2	59.2	44.5	27.4	52.4	67.6	41.6	57.7	25.2	34.2	38.6	45.3
Mantis-Idefics2	69.0	29.9	27.0	25.7	50.3	49.3	22.9	56.0	68.9	50.6	56.3	29.6	37.8	35.6	40.7
LLaVA-NeXT-8B	54.4	27.0	27.8	26.0	55.7	44.5	34.1	24.0	55.2	52.8	58.6	39.8	41.7	28.4	38.8
InternVL2-8B	50.8	26.6	31.8	25.2	53.0	52.6	43.0	51.4	74.9	54.8	62.6	30.3	32.0	41.3	48.8
Idefics-2-8b	65.5	23.8	26.0	24.1	52.3	47.9	25.9	27.8	64.7	60.4	42.8	24.8	21.1	27.4	42.7
LLaVA-NeXT-13B	52.0	24.5	27.7	26.7	53.9	50.7	33.8	25.1	54.0	52.1	52.3	31.8	34.8	26.3	46.5
VILA-1.5-13B	62.7	33.7	28.6	24.1	47.7	43.9	23.9	48.6	66.3	43.8	46.4	25.8	32.3	38.2	45.8
InternVL2-26b	61.3	23.0	32.3	24.7	65.0	41.1	30.1	52.6	77.9	63.8	71.2	26.1	37.3	48.4	46.0
LLaVA-NeXT-34B	58.4	26.0	28.5	27.8	58.3	49.0	21.6	53.9	65.6	59.3	56.8	28.7	40.5	24.1	43.7
VILA-1.5-40B	70.1	<u>62.5</u>	36.8	27.9	53.6	51.7	32.1	66.7	76.4	55.5	61.3	34.1	48.3	39.5	53.1

Table 1: The quantitative results for 3 closed-source and 18 open-source LMMs, as well as those for human and random guess across 14 tasks. The overall score is computed across all tasks. The maximum value and the next largest value of each task are indicated by the **bold** and underlined text, respectively. Task names are abbreviated for brevity.

ral inconsistencies, annotators are asked to remove images with significant temporal changes. During the preprocessing stage, to reduce potential bias introduced by our cross-view detection-matching method, human verification ensures the correctness of paired bounding boxes and filters out mismatched samples. Furthermore, for LMM-generated data, multiple annotators are engaged to eliminate hallucinations and maintain data quality. These comprehensive quality control steps ensure the robustness and accuracy of UrBench.

Experiments

In this section, we evaluate various LMMs on our proposed UrBench. We consider closed-source models, open-source single-image models and open-source multi-image models, and perform evaluations under zero-shot settings. In the following sections, we first introduce our evaluated models evaluation protocols. Then we summarize our findings of model performance with respect to different model types, view settings and tasks. Finally, we provide a detailed analysis in terms of different tasks and views.

Evaluation Setups

Evaluated Models. We evaluate 3 closed-source and 18 open-source LMMs across different model types and sizes. For closed-source models, we consider GPT-4o (Achiam et al. 2023), Gemini-1.5-Flash (Reid et al. 2024), Claude-3.5-Sonnet (Anthropic 2024). For open-sourced models, we categorize them into single-image type and multi-image type according to their training data and strategies, including LLaVA series (Liu et al. 2024a) (Zhou et al. 2024a), XComposer (Zhang et al. 2023), InstructBLIP (Li et al. 2023a) and idefics (Laurençon et al. 2024) for single-image type, and Mantis series (Jiang et al. 2024), VILA series (Lin et al. 2024), InternVL series (Chen et al. 2024) and LLaVA-NeXT-Interleave(Li et al. 2024) for multi-image type.

Evaluation Protocols. In UrBench, our questions have two response formats: multiple-choice and open-ended. We follow standard setups in MMMU (Yue et al. 2024) and MuirBench (Wang et al. 2024a) to process LMMs’ responses. To ensure reproducibility, we set the temperature to 0 and perform greedy decoding. Additionally, for models that do not support multi-image inputs, we concatenate the images as one input. More details on setups and the human evaluation protocols are provided in the Appendix.

Main Results

Overall Challenge Presented in UrBench. As indicated by Table 1, UrBench poses significant challenges to current SoTA LMMs. We find that the best performing closed-source model GPT-4o and open-source model VILA-1.5-40B only achieve a 61.2% and a 53.1% accuracy, respectively. Interestingly, our findings indicate that the primary limitation of these models lies in their ability to comprehend UrBench questions, not in their capacity to process multiple images, as the performance between multi-image and their single-image counterparts shows little difference, such as LLaVA-NeXT-8B and LLaVA-NeXT-Interleave in Table 1. Overall, the challenging nature of our benchmark indicates that current LMMs’ strong performance on the general benchmarks (Fu et al. 2023; Liu et al. 2023b) are not generalized to the multi-view urban scenarios.

LMMs’ performances across task dimensions. In Table 1, we show LMMs performances across different dimensions. We observe that most LMMs exhibit impressive capabilities in Scene Reasoning tasks such as Visual Prompt Reasoning (VPR) and Role-based Reasoning (RBR), which are greatly aligned with their SFT objectives (Liu et al. 2024b). In Scene Understanding, models perform relatively well in region-level tasks such as Scene Recognition (SR) and Scene Comparison (SC), but are bad at Counting (CO). While LMMs achieve impressive results in City Retrieval

Acknowledgments

This project was funded by National Natural Science Foundation of China (Grant No. 42201358) and Shanghai AI Laboratory.

References

- Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. Technical report, Anthropic.
- Bai, T.; Liang, H.; Wan, B.; Yang, L.; Li, B.; Wang, Y.; Cui, B.; He, C.; Yuan, B.; and Zhang, W. 2024. A Survey of Multimodal Large Language Model from A Data-centric Perspective. *arXiv preprint arXiv:2405.16640*.
- Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.
- Chen, Z.; Wang, W.; Tian, H.; Ye, S.; Gao, Z.; Cui, E.; Tong, W.; Hu, K.; Luo, J.; Ma, Z.; et al. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.
- Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Cui, C.; Ma, Y.; Cao, X.; Ye, W.; Zhou, Y.; Liang, K.; Chen, J.; Lu, J.; Yang, Z.; Liao, K.-D.; et al. 2024. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 958–979.
- Ertler, C.; Mislej, J.; Ollmann, T.; Porzi, L.; Neuhold, G.; and Kuang, Y. 2020. The mapillary traffic sign dataset for detection and classification on a global scale. In *European Conference on Computer Vision*, 68–84. Springer.
- Feng, J.; Du, Y.; Liu, T.; Guo, S.; Lin, Y.; and Li, Y. 2024a. CityGPT: Empowering Urban Spatial Cognition of Large Language Models. *arXiv preprint arXiv:2406.13948*.
- Feng, J.; Zhang, J.; Yan, J.; Zhang, X.; Ouyang, T.; Liu, T.; Du, Y.; Guo, S.; and Li, Y. 2024b. CityBench: Evaluating the Capabilities of Large Language Model as World Model. *arXiv preprint arXiv:2406.13945*.
- Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; et al. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*.
- Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6904–6913.
- Hao, X.; Chen, W.; Yan, Y.; Zhong, S.; Wang, K.; Wen, Q.; and Liang, Y. 2024. UrbanVLP: A Multi-Granularity Vision-Language Pre-Trained Foundation Model for Urban Indicator Prediction. *arXiv preprint arXiv:2403.16831*.
- Hays, J.; and Efros, A. A. 2008. IM2GPS: estimating geographic information from a single image. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8.
- Hu, Y.; Yuan, J.; Wen, C.; Lu, X.; and Li, X. 2023. Rsgpt: A remote sensing vision language model and benchmark. *arXiv preprint arXiv:2307.15266*.
- Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6700–6709.
- Jiang, D.; He, X.; Zeng, H.; Wei, C.; Ku, M.; Liu, Q.; and Chen, W. 2024. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.
- Kuckreja, K.; Danish, M. S.; Naseer, M.; Das, A.; Khan, S.; and Khan, F. S. 2024. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27831–27840.
- Laurençon, H.; Tronchon, L.; Cord, M.; and Sanh, V. 2024. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*.
- Li, F.; Zhang, R.; Zhang, H.; Zhang, Y.; Li, B.; Li, W.; Ma, Z.; and Li, C. 2024. LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models. *arXiv preprint arXiv:2407.07895*.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.
- Li, W.; Lai, Y.; Xu, L.; Xiangli, Y.; Yu, J.; He, C.; Xia, G.-S.; and Lin, D. 2023b. OmniCity: Omnipotent city understanding with multi-level and multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17397–17407.
- Li, X.; Ding, J.; and Elhoseiny, M. 2024. VRSBench: A Versatile Vision-Language Benchmark Dataset for Remote Sensing Image Understanding. *arXiv preprint arXiv:2406.12384*.
- Lin, J.; Yin, H.; Ping, W.; Molchanov, P.; Shoeybi, M.; and Han, S. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26689–26699.
- Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024a. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.
- Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023a. Grounding dino:

- Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.
- Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2023b. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.
- Lobry, S.; Marcos, D.; Murray, J.; and Tuia, D. 2020. RSVQA: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12): 8555–8566.
- Muhtar, D.; Li, Z.; Gu, F.; Zhang, X.; and Xiao, P. 2024. Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model. *arXiv preprint arXiv:2402.02544*.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.
- Ranjan, V.; Sharma, U.; Nguyen, T.; and Hoai, M. 2021. Learning to count everything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3394–3403.
- Reid, M.; Savinov, N.; Teplyashin, D.; Lepikhin, D.; Lillcrap, T.; Alayrac, J.-b.; Soricut, R.; Lazaridou, A.; Firat, O.; Schrittwieser, J.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- Schumann, R.; Zhu, W.; Feng, W.; Fu, T.-J.; Riezler, S.; and Wang, W. Y. 2024. Velma: Verbalization embodiment of llm agents for vision and language navigation in street view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18924–18933.
- Shi, Y.; and Li, H. 2022. Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17010–17020.
- Sun, X.; Wang, P.; Yan, Z.; Xu, F.; Wang, R.; Diao, W.; Chen, J.; Li, J.; Feng, Y.; Xu, T.; et al. 2022. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184: 116–130.
- Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Wang, F.; Fu, X.; Huang, J. Y.; Li, Z.; Liu, Q.; Liu, X.; Ma, M. D.; Xu, N.; Zhou, W.; Zhang, K.; et al. 2024a. Muir-Bench: A Comprehensive Benchmark for Robust Multi-image Understanding. *arXiv preprint arXiv:2406.09411*.
- Wang, J.; Zheng, Z.; Chen, Z.; Ma, A.; and Zhong, Y. 2024b. Earthvqa: Towards queryable earth via relational reasoning-based remote sensing visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5481–5489.
- World Bank. 2024. Urban Population (% of Total Population) - World. Accessed: 2024-08-07.
- Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; and Zhang, L. 2018. DOTA: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3974–3983.
- Xiao, H.; Zhou, F.; Liu, X.; Liu, T.; Li, Z.; Liu, X.; and Huang, X. 2024. A comprehensive survey of large language models and multimodal large language models in medicine. *arXiv preprint arXiv:2405.08603*.
- Yan, Y.; Wen, H.; Zhong, S.; Chen, W.; Chen, H.; Wen, Q.; Zimmermann, R.; and Liang, Y. 2024. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM on Web Conference 2024*, 4006–4017.
- Ye, J.; Luo, Q.; Yu, J.; Zhong, H.; Zheng, Z.; He, C.; and Li, W. 2024a. SG-BEV: Satellite-Guided BEV Fusion for Cross-View Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27748–27757.
- Ye, J.; Lv, Z.; Li, W.; Yu, J.; Yang, H.; Zhong, H.; and He, C. 2024b. Cross-view image geo-localization with Panorama-BEV Co-Retrieval Network. *arXiv preprint arXiv:2408.05475*.
- Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; and Chen, E. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.
- Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*, 69–85. Springer.
- Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9556–9567.
- Zhang, P.; Wang, X. D. B.; Cao, Y.; Xu, C.; Ouyang, L.; Zhao, Z.; Ding, S.; Zhang, S.; Duan, H.; Yan, H.; et al. 2023. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*.
- Zhou, B.; Hu, Y.; Weng, X.; Jia, J.; Luo, J.; Liu, X.; Wu, J.; and Huang, L. 2024a. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*.
- Zhou, Z.; Lin, Y.; Jin, D.; and Li, Y. 2024b. Large language model for participatory urban planning. *arXiv preprint arXiv:2402.17161*.
- Zhu, S.; Yang, T.; and Chen, C. 2021. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3640–3649.