Article:
Huang, W., Wang, J. and Cong, G. (2024) Zero-shot urban function inference with street view images through prompting a pretrained vision-language model. International Journal of Geographical Information Science, 38 (7). pp. 1414-1442. ISSN 1365-8816

https://doi.org/10.1080/13658816.2024.2347322

Zero-shot urban function inference with street view images through prompting a pretrained vision-language model

Weiming Huang[a], Jing Wang[b] and Gao Cong[a]

[a]School of Computer Science and Engineering, Nanyang Technological University, Singapore; [b]Future Cities Laboratory, Singapore-ETH Centre, Singapore

ARTICLE HISTORY
Compiled May 31, 2024

ABSTRACT

Inferring urban functions using street view images (SVIs) has gained tremendous momentum. The recent prosperity of large-scale vision-language pretrained models sheds light on addressing some long-standing challenges in this regard, e.g., heavy reliance on labeled samples and computing resources. In this paper, we present a novel prompting framework for enabling the pretrained vision-language model CLIP to effectively infer fine-grained urban functions with SVIs in a zero-shot manner, i.e., without labeled samples and model training. The prompting framework UrbanCLIP comprises an urban taxonomy and several urban function prompt templates, in order to (1) bridge the abstract urban function categories and concrete urban object types that can be readily understood by CLIP, and (2) mitigate the interference in SVIs, e.g., street-side trees and vehicles. We conduct extensive experiments to verify the effectiveness of UrbanCLIP. The results indicate that the zero-shot UrbanCLIP largely surpasses several competitive supervised baselines, e.g., a fine-tuned ResNet, and its advantages become more prominent in cross-city transfer tests. In addition, UrbanCLIP's zero-shot performance is considerably better than the vanilla CLIP. Overall, UrbanCLIP is a simple yet effective framework for urban function inference, and showcases the potential of foundation models for geospatial applications.

KEYWORDS
Urban land use; prompt engineering; CLIP; foundation model; street view image

1. Introduction

Delineating the spatial distribution of urban functions (functional land use) in our cities has been a primary focus in the communities of geographic information science and urban studies. This task is fundamental to effective urban management and sustainable development (Zhang *et al.* 2018, Srivastava *et al.* 2020, Qiao and Yuan 2021, Huang *et al.* 2022). Specifically, our cities are organic and dynamic complexes, which are composed of numerous functional areas that afford various socioeconomic activities and needs of human habitation, e.g., *residential*, *commercial*, and *industrial* areas. Today, in view of the massive human efforts required in field or airborne survey, data mining techniques using varying urban sensing data sources, e.g., remote sensing data (Bai *et al.* 2023) , points of interest (Huang *et al.* 2023), human trajectories (Hu *et al.* 2021), and SVIs (Xu *et al.* 2022), are widely applied to urban function recognition.

Among various data sources, SVIs have shown significant potential in discerning

urban functions, in virtue of their increasing spatial coverage, human perspective of photographing, and abundant visual information (Biljecki and Ito 2021). The visual information carried by SVIs has been proved to be effective and has favorable discrimination power for this task, e.g., the façades of *residential* and *commercial* areas can be readily discerned in SVIs, which is challenging for aerial imagery (Srivastava *et al.* 2020). In addition, SVIs provide insights into urban function at granular levels, such as scene or single building scales (Kang *et al.* 2018, Zhang *et al.* 2022).

From a methodological perspective, many recent studies treat fine-grained urban function inference using SVIs as an image classification problem, and train (fine-tune) a deep vision model, e.g., a ResNet (He *et al.* 2016), for recognizing scene- or building-level urban function (Kang *et al.* 2018, Srivastava *et al.* 2020). This strategy has been useful, but it has an inherent limitation: data hungriness. Usually, a large number of labels (ground truth data) are required for sufficiently training or fine-tuning a deep vision model, in order to yield competent inference capability. This limitation makes the task often difficult in real-world practice, when ground truth data is not available. Even if labels could be collected with substantial endeavors, the transferability of the trained models is questionable (Zhang *et al.* 2022), as different cities (urban areas) often exhibit distinctive traits in their built environments, and target at inferring different urban function categories. These limitations undermine the underlying incentive of the task. Furthermore, we believe that the previously utilized visual models (e.g., ImageNet-pretrained ResNets) are not well suited in an urban context, because they are trained in image classification or segmentation tasks using pre-determined sets of object categories (e.g., plane, car, dog, etc.). Transferring such models to an urban context could lead to compromised capability.

The recent prosperity of pretrained vision-language foundation models unfolds a promising avenue to tackling the aforementioned limitations (Du *et al.* 2022). A key advantage of such models is that their training and inference are not confined to a fixed set of pre-determined object categories. Instead, they are trained through coupling visual and language models (encoders) and the objectives like finding the correct image-text pairs. Such properties unleash the potential of using a pretrained vision-language model to answer urban questions, e.g., inquire it for the embodied urban functions in an SVI. In this way, we could accomplish zero-shot learning, i.e., inferring urban functions with no labeled samples and no training of the model, which is a previously impossible mission. In addition, such vision-language models are often pretrained with a great diversity and number of image-text pairs, which makes them more competent in tackling the domain shift from general-purpose natural image tasks to an urban context. Today, a representative visual-language model is CLIP (Contrastive Language-Image pretraining) (Radford *et al.* 2021), which is trained with an enormous number (400 million) image-text pairs and a contrastive learning strategy. CLIP has demonstrated remarkable capabilities in several vision tasks, particularly in zero-shot image classification.

With the promising capability of CLIP, it is a natural idea to leverage it for zero-shot urban function inference. For example, we could simply ask what urban function does an SVI reflect through matching SVIs with textual descriptions, e.g., *residential*, *industrial*, etc. This process is denoted prompting for vision-language models (Zhou *et al.* 2022b). However, we find that simply prompting CLIP with the raw urban function category names does not work well. The first impediment is that urban function categories imply high-level, abstract, and sometimes polysemous semantics, whereas CLIP is more capable of handling concrete concepts, and often fails to understand abstract language phrases (Radford *et al.* 2021, Liao *et al.* 2023). The underlying rea-

son is the inherent ambiguity of visual representations of abstract terms. Each urban function encompasses a myriad of structures and visual cues, e.g., *residential* could be associated with vastly different visual forms like terrace house, urban village, and condominium. A recent investigation reveals that CLIP struggles with comprehending abstract concepts, regardless of its model size (Liao *et al.* 2023). The second obstacle stems from the overwhelming presence of common yet potentially distracting elements in real-world SVIs, due to their *street* view nature, e.g., vehicles, road surface, transportation facilities like bus stops, street-side landscape like trees, etc (Biljecki and Ito 2021). Such frequently appearing visual information is only weakly indicative for inferring urban functions, but may divert CLIP's attention, e.g., CLIP could erroneously believe that an SVI with several cars reflects a parking lot.

In this context, we develop the prompting framework UrbanCLIP to tailor CLIP for urban function inference. The key design principle of UrbanCLIP is *simplicity*, which requires no labeled samples and no model training, and can be readily used in practice. To this end, UrbanCLIP performs fine-grained (scene-level) urban function inference in a zero-shot manner. The zero-shot UrbanCLIP prompting is empowered by two key components, including an urban taxonomy that maps abstract urban function categories (e.g., *residential*) to concrete urban object types (UOTs; e.g., *condominium*), and several urban function prompt templates to mitigate the interference and noise in SVIs. We conduct extensive experiments in three different settings. First, we utilize UrbanCLIP to infer the most predominating urban function (primary function) in each SVI. Second, UrbanCLIP is applied to scenes carrying multiple functions, e.g., *residential* and *commercial*. These two settings are conducted in the main study area of Shenzhen, China. In the third setting, we test UrbanCLIP's transfer capacity in Singapore and London. In all three settings, we compare zero-shot UrbanCLIP with a series of competitive supervised baseline models, and the results demonstrate that UrbanCLIP has superior zero-shot capacity. For primary and multiple function inference, zero-shot UrbanCLIP largely outperforms most of the supervised baselines (e.g., a fine-tuned ResNet101) regardless of how many labeled samples are used to train them. In addition, UrbanCLIP more prominently prevails in the transfer experiment, i.e., it performs well in two other cities, while other baselines trained with the Shenzhen dataset exhibit compromised performances.

Following the introduction, we provide the background and review related works in Section 2. In Section 3, we elaborate on the details of the proposed framework UrbanCLIP. In Section 4, we demonstrate the results of fine-grained urban function inference using UrbanCLIP, compare the results with several baseline methods, and present an ablation study along with an error analysis. The paper ends with a discussion in Section 5 and conclusions in Section 6.

2. Background and related work

2.1. *Urban function inference with SVIs*

Utilizing SVIs for urban function inference has attracted tremendous attention in recent years. SVIs bring up the opportunity of mapping urban function at fine spatial scales. For example, Kang *et al.* (2018) fine-tuned several visual models like VGG and ResNet for classifying building types with SVIs. Srivastava *et al.* (2020) carried out building-level land use classification using SVIs, in which they utilized a CNN in a Siamese-like architecture to use SVIs from varying views (angles) for the task. Zhang

et al. (2022) transformed SVIs into textual descriptions using an image captioning model, and fed the image captions to a language model BERT (Devlin *et al.* 2019) for urban scene classification. Zhu *et al.* (2019) used geo-referenced images from social media platforms for land use classification, and they proposed a two-stream model (an object stream and a scene stream) for the task.

In addition, SVIs have been used for region-level urban function inference. For example, Xu *et al.* (2022) extracted visual features from SVIs using a semantic segmentation model, and fed SVI features into a graph convolutional network (GCN) (Kipf and Welling 2017) for urban function classification. Wang *et al.* (2020) proposed an unsupervised region embedding technique Urban2Vec using SVIs and POIs, in which they fine-tuned a ResNet using a spatial proximity-based contrastive learning method. Although not tested in their paper, Urban2Vec can be potentially used for urban function inference.

These studies have fostered prominent strides in SVI-based urban function inference. However, most of them encounter the difficulty of collecting a considerable quantity of labeled samples to sufficiently train a machine learning model. In reality, such ground truth data in fine scales is rarely available. To this end, the most common resort is the object tags from OpenStreetMap (OSM), whereas it has been revealed that OSM tags have several shortcomings, including misalignment between images and buildings, incorrect annotation of tags, and notable incompleteness (Qiao and Yuan 2021, Srivastava *et al.* 2020). This implies that the ground truth data from OSM tags could lead to a considerable number of noisy samples. Furthermore, such models are potentially lacking in the transferability across different cities (urban areas), and often require massive resources for training. In this paper, we present the UrbanCLIP framework to overcome the limitations.

2.2. *Vision-language pre-training and CLIP*

Vision-language pre-training pertains to the practices of utilizing large-scale image-text pairs to learn multi-modal foundation models that can facilitate various vision-language tasks, e.g., image-to-text retrieval, and visual question answering (Li *et al.* 2021). The state-of-the-art vision-language foundation models generally have two prominent strands (Wang *et al.* 2021). The first strand leverages dual uni-modal encoders to separately learn image and text representations, and uses cosine similarity or linear projection to model their interactions (Radford *et al.* 2021). Dual-encoder models are competent for image-text retrieval, and their inference is highly efficient. CLIP is a representative model in this strand. Another influential model is ALIGN (Jia *et al.* 2021), which scaled up the dual-encoder models with billion-level noisy image-text pairs. The second strand models the multi-modal interactions with additional encoders such as VL-BERT (Su *et al.* 2019), UNITER (Chen *et al.* 2020), and ViLT (Kim *et al.* 2021). These models have advantages in more complex vision-language tasks, e.g., visual question answering and visual grounding, while their inference is slower than dual-encoder models. For our task, scene-level urban function inference is essentially an image-text retrieval task (retrieve relevant text samples, namely urban function types, given SVIs as queries), and requires high inference efficiency, as a city usually has enormous SVIs for fine-grained inference. In this context, CLIP (Radford *et al.* 2021), arguably the most recognized dual-encoder vision-language pretrained model, is a natural choice.

CLIP is composed of an image encoder and a text encoder. The image encoder allows

for different choices, including various ResNets (He *et al.* 2016) or Vision Transformers (ViT) (Dosovitskiy *et al.* 2020). The text encoder is based on Transformer (Vaswani *et al.* 2017, Radford *et al.* 2019). CLIP's training follows a contrastive learning approach: in each training batch, the two encoders separately transform images and their paired textural descriptions into vector representations; then the two encoders are trained using the objective of maximizing the cosine similarities between the embeddings of the images and their associated textual descriptions, while minimizing the cosine similarities of the unmatched pairs. In order to sufficiently train the two encoders, Radford *et al.* (2021) constructed a new dataset of 400 million image-text pairs collected from the Web. They attempted to make the dataset cover as a broad set of visual concepts as possible, so as to learn semantically-rich and transferable visual encoders. With the pretrained image and text encoders, CLIP is able to carry out zero-shot image classification by finding the text description which has the largest cosine similarity to the query image in the embedding space.

2.3. *Prompt engineering*

Several machine learning domains are undergoing a sea change to the "*pre-train, prompt, and predict*" paradigm (Liu *et al.* 2021), where we could simply ask the large-scale pretrained models questions, i.e., prompting them, so as to obtain desired answers without further fine-tuning the pretrained models. Essentially, prompt engineering processes are reformulating downstream tasks to make them similar to those solved during pre-training (Liu *et al.* 2021). In natural language processing (NLP), prompting usually appears in two major forms: (1) cloze prompts, i.e., the "fill-in-the-blank" cloze tests, and (2) prefix prompts that induce language models to continue a string prefix, as commonly used in GPT-family models. Prompting appears in different forms for vision-language models. For CLIP, it is prompted to perform zero-shot image classification using text embeddings, i.e., given a search space of text embeddings and an image, finding the text description that is the most similar to the image in the embedding space. A simplistic way of generating such text embeddings is directly transforming category names (usually single words or few-word phrases) into embeddings using CLIP's pretrained text encoder. However, this naive manner usually leads to sub-optimal performance, and Radford *et al.* (2021) identified two major reasons for this problem. The first issue is polysemy, and CLIP's encoder is unable to differentiate between varying senses of a word with only the category name. The second issue is distribution shift, as CLIP is pretrained mostly with full sentences describing the images, and single-word descriptions are rare.

Radford *et al.* (2021) found that simply using the prompt template "`A photo of a {label}.`" could slightly lift CLIP's zero-shot performance on ImageNet. They finally utilized prompt ensembling on 80 prompt templates to cover different scenarios, e.g., "`A bright photo of a {label}.`" and "`A blurry photo of a {label}.`". With these 80 prompt templates ensembled, they observed nearly 5% accuracy improvement on ImageNet in a zero-shot setting, compared to merely using category names. In addition, they found that for specialized image classification tasks, it helps to specify the nature of the images. For example, the prompt template "`A satellite photo of a {label}.`" is useful for handling remote sensing imagery in EuroSAT.

Such studies provide valuable insights into our study, whereas their prompting methods do not align the requirement of this task, which goes beyond only addressing polysemy and distribution shift. First, prompting on the basis of function categories is

not ideal, as their semantics are generally ambiguous and polysemous. In this context, using function categories, even supplemented with carefully crafted prompt templates, would lead to substantial difficulties for CLIP. Second, we need to alleviate the irrelevant and sometimes overwhelming foreground information in SVIs, e.g., vehicles, street-side trees, etc. This necessitates prompt templates that can guide CLIP to emphasize the most relevant clues while sidelining the distracting ones. These two obstacles underscore the demand of prompt engineering of our task. The prompt engineering studies for vision-language models like CLIP are few and in an early stage. Nevertheless, several directions in prompt engineering in NLP (Liu *et al.* 2021) – a relatively ripe research field – shed light on tackling our challenges.

A relevant direction is prompt answer engineering, which aims to develop an answer space and map to the desired outputs, e.g., categories in a classification problem. For example, in sentimental analysis, one might use multiple bearing words like "excellent", "fabulous", "wonderful" to represent a single class "++" (the most positive class). This is enlightening for overcoming the problem that the urban function categories are abstract and difficult to comprehend for CLIP. We could develop an additional answer space with concrete visual concepts to establish a more direct linkage between visual clues in SVIs and corresponding descriptions, and the answers are eventually mapped to urban function categories. In NLP, prompt answer engineering largely relies on manual design, and there are some studies on automatically search ideal answers, e.g., using answer paraphrasing and prune-then-search, which are also partially built upon manual design (Liu *et al.* 2021). In this paper, given the relatively uncharted territory of automated prompt answer engineering for vision-language models, we choose to handcraft an urban taxonomy as the "prompt answers" for the UrbanCLIP framework. This design is primarily rooted in domain knowledge and existing categorizations from varying urban data sources.

Another pertinent area is prompt template engineering. Originally, the prompt templates for CLIP are largely handcrafted, drawing from domain expertise and tailored to the characteristics of the target datasets and categories they aimed to deduce. Recently, the idea of learning prompt templates instead of engineering them is gaining traction, in view of the labor-intensiveness and difficulty in guaranteeing optimal performance from handcrafting templates, e.g., Zhou *et al.* (2022a,b), Jia *et al.* (2022), Lu *et al.* (2022). Despite their advantages, they are misaligned with our goal. The first reason is that most of these methods require training samples (few-shot). Even if we were to adapt to a few-shot setting, they can hardly be utilized in conjunction with our urban taxonomy. As revealed in our experiments, the urban taxonomy is a major driver of performance enhancements. The cause of the issue is that the supervision signals cannot be readily propagated when we match SVIs to UOTs in the taxonomy while the labels are associated with urban function types. The second reason is that the automatically discovered prompts are in the form of embeddings that can hardly be converted to tangible real-world text. As a pioneering study in using CLIP for urban function inference, we believe that it is desirable to develop human-readable templates, so as to manifest how our domain knowledge in urban studies and the insights gleaned from the datasets can be harnessed to steer a pretrained foundation model.

Over the past year, a few studies have emerged to utilize foundation models by prompting them for different urban and geospatial applications. Yong *et al.* (2023) prompted CLIP for both zero-shot and few-shot building defect detection and classification using the textual definitions in construction jargon dictionaries as domain knowledge. Haas *et al.* (2023) fine-tuned CLIP to enhance its capacity in image geolocalization, in which they devised a prompt template for fune-tuning and several tem-

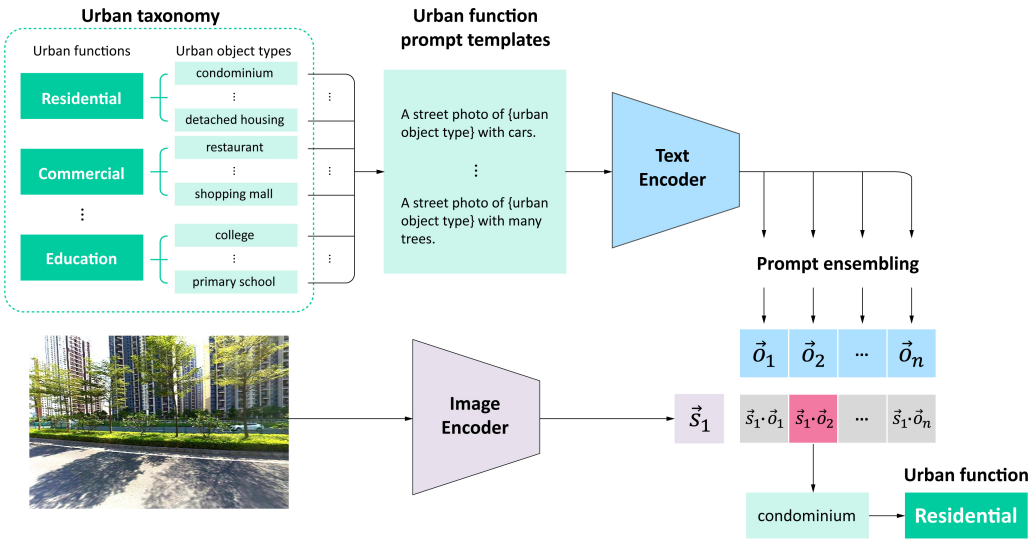

Figure 1. The architecture of UrbanCLIP zero-shot urban function inference.

plates for inference, e.g., "`A street view photo in {country}`". Hu *et al.* (2023) harnessed geospatial knowledge, e.g., location types like door number address and street name, to prompt GPT-family models for few-shot location extraction from social media messages. These prior studies provide valuable insights that utilizing domain knowledge is a key to effectively prompt foundation models in urban and geospatial applications. In addition, similar to the scope of this study, Wu *et al.* (2023) utilized CLIP for mixed land use inference through ensembling several prompt templates, which specifies the contexts of the target analysis and locations, e.g., "`{label} use in New York.`" and "`{label} purpose in New York.`", and they found that incorporating more granular location descriptions decreases the performance for this task. This study is inspiring, while it directly matches land use categories with SVIs, and does not mitigate the interference in SVIs, leaving room for improvements.

3. UrbanCLIP

We propose the prompting framework UrbanCLIP for SVI-based urban function inference at the scene level, i.e., for individual SVIs. With UrbanCLIP, we infer the primary urban function and multiple functions (if available) that each SVI reflects. Specifically, UrbanCLIP complements CLIP with (1) an urban taxonomy that maps ten fine-grained urban function categories to hundreds of concrete UOTs, which could be well understood by CLIP , and (2) several urban function prompt templates that are ensembled to mitigate the interfering information in SVIs, e.g., vehicles and street-side trees. With these two components, urban function inference can be performed in a zero-shot manner. The UrbanCLIP framework includes the pretrained CLIP as its foundation, and leverages the urban taxonomy and urban function prompt templates to prompt CLIP for effective urban function inference. The architecture of the zero-shot urban scene understanding using the UrbanCLIP prompting framework is illustrated in Figure 1.

In essence, the development of urban taxonomy and urban function prompt tem-

plates are respectively practices in prompt answer engineering and prompt template engineering. To be best of our knowledge, UrbanCLIP is a pioneering endeavor in prompt answer engineering for pretrained vision-language models, and previous studies on engineering an answer space in vision-language models are rare. In addition, the developed prompt templates are a novel way to enforce CLIP to pay less attention to the foreground interfering visual clues. Overall, we believe that the proposed UrbanCLIP framework entails novelty in both the communities of GIScience and prompt engineering for vision-language models.

3.1. *Urban taxonomy*

The first key component in the proposed UrbanCLIP framework is an urban taxonomy, which transforms abstract urban function categories to concrete UOTs, to facilitate CLIP to comprehend urban scenes.

The urban taxonomy contains ten fine-grained urban function categories $\mathcal{F} = \{f_1, \ldots, f_{10}\}$, i.e., (1) *residential*, (2) *commercial*, (3) *hotel*, (4) *industrial*, (5) *education*, (6) *health care*, (7) *civic, governmental, and cultural*, (8) *sports and recreation*, (9) *outdoors and natural*, and (10) *transportation*. This categorization is formulated by harmonizing the land use/urban function classifications prevalent in several major cities, including Singapore, Shenzhen, and New York City, to create a unified categorization. Among them, Singapore's system is rather detailed with over 30 function types, and the classifications in the other two cities are less detailed. The categorization is guided by two main considerations: (1) they mainly focus on built-up areas, where SVIs have sufficient coverage, and (2) they are used for understanding urban scenes, which is a spatially detailed scale, so they should be also semantically fine-grained to inform urban planning in a detailed manner, e.g., we use separate *education*, *health care*, and *civic, governmental, and cultural* categories instead of a generic *public service* category.

The urban taxonomy includes 354 UOTs ($\mathcal{O} = \{o_1, \ldots, o_{354}\}$) that are concrete visual concepts and can be readily understood by CLIP. The UOTs are merged from several well-known sources, including Amap POI categories [1], building classification system of New York City[2], example UOTs from the Urban Redevelopment Authority in Singapore [3], and the list of building types in Wikipedia [4]. The reason for integrating multiple information sources is that POIs categories are inclined to *commercial* venues, and fall short in representing some other functions like *residential* and *industrial*, which constitute most of the urban areas. In this context, the other two sources provide indispensable complement to the POI categories in forming the urban taxonomy, especially the building classification system in New York City offers a detailed categorization for residential buildings. Behind this process, the overall guiding principles are: (1) UOTs should have a generally comprehensive coverage for the visual elements in the built-up environment; (2) UOTs should be commonplace in urban environments, and extremely specialized or unique objects are generally avoided; (3) UOTs potentially with polysemy or multiple interpretations are normally not included. Specifically, the UOTs are selected in the following steps. First, we collate the aforementioned information sources and remove duplicated UOTs, whereas the UOTs are not necessarily semantically disjoint, so as to lift their semantic coverage, e.g., both

[1] https://lbs.amap.com/api/ios-sdk/guide/map-data/poi

[2] https://www1.nyc.gov/assets/finance/jump/hlpbldgcode.html

[3] https://www.ura.gov.sg/-/media/Corporate/Planning/Master-Plan/MP19writtenstatement.pdf?la=en

[4] https://en.wikipedia.org/wiki/List_of_building_types

shopping mall and *community shopping center* are included in the urban taxonomy. Second, domain experts are solicited to refine the urban taxonomy, mainly through removing some unnecessary UOTs, e.g., terms linked to rare objects (e.g., flex space), overly specific objects (e.g., office only 7-9 stories), and ambiguous descriptions (e.g., entertainment).

Finally, we establish *1-to-n* matching relations between the ten urban function categories and the 354 UOTs, namely each urban function category subsumes tens of UOTs. For instance, the urban function category *residential* subsumes *apartment*, *attached housing*, *bungalow* etc. This matching is performed based on the categorical hierarchy in their respective sources (e.g, in the building classification system in New York City, each building type is subsumed by a generic type), followed by domain expert engagement. We present the entire urban taxonomy in Appendix A.

We recognize that some UOTs (e.g., *flats with commercial uses at first storey*) could be linked to more than one function categories (e.g., *residential* and *commercial*), while we generally match such multi-function objects to their predominating functions (*residential* in this case). In addition, the urban taxonomy could be readily enriched (e.g., adding further UOTs), and re-mapped (e.g., match *filling station* to *commercial* instead of *transportation*) to adapt to different study areas and applications.

3.2. *Urban function prompt templates*

Another pivotal component of UrbanCLIP is a set of urban function prompt templates, which could alleviate the interference in SVIs for our task. As SVIs are mostly captured alongside streets, they commonly include visual information of road surfaces, vehicles, trees, etc. In fact, such visual information is indispensable for some other SVI-based analyses, e.g., the appearance of street-side landscapes like trees is key for analyzing urban greenery (Li *et al.* 2015). However, such visual clues are only weakly relevant to our task, and could possibly mislead CLIP in its inference. This is an especially notable problem, as such interference often appears as the salient foreground in SVIs. To tackle this challenge, we devise several prompts to help CLIP neglect the interference and concentrate on the useful information for our task, which often appears in the background.

In this context, we design a set of urban function prompt templates to enforce CLIP to focus on the most relevant information for our task. This implies that our incentive is divergent from Radford *et al.* (2021), where mitigating polysemy and distribution gap is the primary motivation. For example, we design a prompt template "`a street photo of {UOT} with many trees.`", and each of the 354 UOTs is used to replace the placeholder {`UOT`} in the template to form 354 sentences. In all these sentences, the information of "many trees" is presented with all UOTs, which waters down the importance of trees in zero-shot inference, thereby mitigating such interfering foreground objects. In addition, the information of "street photo" also accompanies every UOT, providing contextualization for SVI understanding. We devise six urban function prompts ($\mathcal{PT} = \{pt_1, \ldots, pt_6\}$) and demonstrate the prompt templates and their rationales in Table 1.

With the UOTs and urban function prompt templates, we could then generate UOT textual embeddings using CLIP's text encoder and prompt ensembling. To this end, each UOT replaces all the placeholder {`UOT`} in the prompt templates $\mathcal{PT}$ to form six sentences (language phrases). The six sentences of each UOT then go through the pretrained CLIP text encoder to generate six textual embeddings of the UOT. The six

Table 1. Urban function prompt templates

Prompt template	Rationale
PT1: {UOT}	PT1 enforces the model to focus on the target UOT itself, which fits the SVIs with little noisy and interfering information.
PT2: A street photo of {UOT} in city.	PT2 provides the context of SVIs by explicitly stating the "street" and "city" natures of SVIs.
PT3: A street photo of {UOT} with many trees.	PT3 enforces the zero-shot inference to pay less attention to the street-side and on-street trees, which are usually in the foreground of SVIs.
PT4: A street photo of {UOT} with cars.	PT4 enforces the zero-shot inference to pay less attention to the interfering information of vehicles on streets, which are usually in the foreground of SVIs.
PT5: A street photo of {UOT} on the road.	PT5 enforces the zero-shot inference to pay less attention to the visual information of streets/roads themselves.
PT6: A street photo of {UOT} with parking lot.	PT6 alleviates the zero-shot errors that an SVI with many cars reflects a parking lot (in reality they are moving though).

textual embeddings are finally ensembled through element-wise averaging to generate a single textual embedding for the UOT. Formally, the process can be defined as:

$$\vec{o}_i = \frac{1}{n_{pt}} \sum_j \phi_t \left(pt_j \left(o_i \right) \right) \tag{1}$$

where o_i is a particular UOT, e.g., *condominium*, pt_j is a certain urban function prompt template, ϕ_t is the pretrained text encoder of CLIP, n_{pt} is the number of urban function prompt templates, and $\vec{o}_i$ is the generated text embedding for o_i through prompt ensembling.

3.3. *Zero-shot urban function inference*

Till this point, the designed urban taxonomy and urban function prompt templates are transformed into hundreds of text embeddings. In the meantime, each urban scene (individual SVI) goes through the pretrained CLIP image encoder to generate an image (SVI) embedding, i.e., $\vec{s}_i = \phi_i \left(s_i \right)$, where ϕ_i is the pretrained image encoder in CLIP, and $\vec{s}_i$ is the embedding of an SVI s_i. Then we can readily carry out zero-shot urban function inference by calculating cosine similarity values between image and text embeddings. For each SVI, we can obtain a ranking list of UOTs (most similar to least similar), e.g., {*condominium*, *terrace house*, *restaurant*,...}. All the UOTs in the ranking list are subsequently replaced by their corresponding urban function types to form a ranking list of urban functions, e.g., {*residential*, *residential*, *commercial*,...}. We then de-duplicate the ranking list of urban functions to form a ranking set of urban functions concerning this SVI (only the first appearance of each urban function is kept), e.g., {*residential*, *commercial*, *education*,...}. For the inference of the primary function, the first element in this ranking set of urban functions, e.g., *residential*, is used as the prediction. In the scenario of multi-function inference, the *top-k* (e.g., *top-2*) elements are taken as the prediction, e.g., *residential* and *commercial*.

4. Experiments and results

4.1. *Study area and data*

We utilize the proposed prompting framework UrbanCLIP for fine-grained urban function inference in the main study area of Shenzhen, China. In Shenzhen, we obtain 226,881 SVIs from Baidu Map API[5] with a sampling interval of one point/$10,000m^2$, mainly in built-up areas. In order to understand the zero-shot capability of UrbanCLIP, we select a representative subset (i.e., the urban scene dataset) of the SVIs, and annotate the ground truth functions of each image. The annotation for each image can be a single function label, or multiple labels if entailed by that SVI. Specifically, the urban scene dataset contains 1,518 SVIs, among which 1,179 are annotated with a single urban function, while 339 are annotated with two different functions, e.g., an urban scene reflects both *residential* (primary) and *commercial* (secondary) functions. The categories of urban functions follow the ten fine-grained functions designed in the urban taxonomy. Figure 2 demonstrates the distribution of primary functions in the dataset, and a typical SVI for each function.

[5]https://lbsyun.baidu.com/

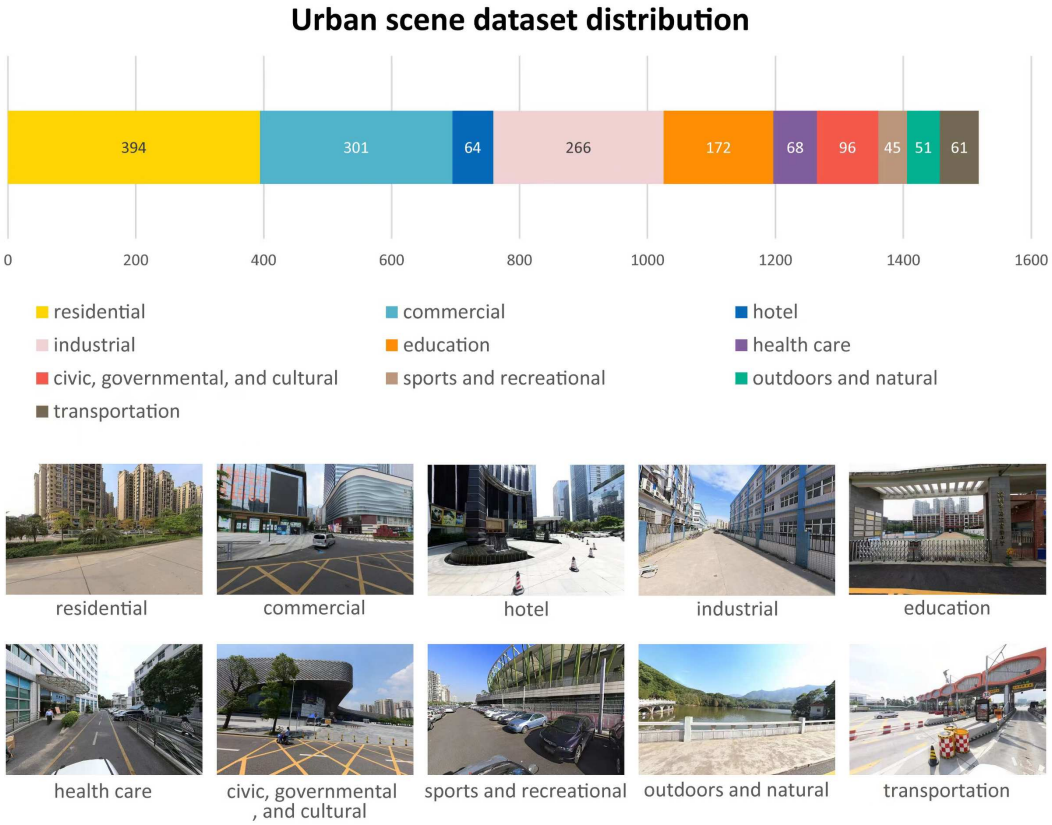

Figure 2. The distribution of primary function labels and example SVIs in the urban scene dataset.

For selecting the 1,518 SVIs, we focus on scenes reflective of the built environment, with smaller emphasis on less urbanized scenes and SVIs dominated by roads. We strive for a balanced representation of various urban functions in the dataset, to counter-balance the natural prevalence of residential areas with less commonly captured functions like hospitals. During the annotation process, we first analyze visible indicators such as architectural styles, signage, and specific urban elements to establish initial function labels. This is performed side-by-side with cross-referencing using detailed planning maps and online map services with POIs (from Baidu and Google) for validation.

To further understand the level of difficulty of the task, an urban planning expert is asked to gauge the urban functions in each SVI, meaning that each SVI is annotated with both ground truth label(s) and the inferred urban function(s) from the human expert. In addition, in order to investigate the transfer capacity of UrbanCLIP, we harvest 100 SVIs from Google Map, in Singapore and London, respectively, and manually annotate the ground truth labels for the 200 SVIs. In the datasets of Singapore and London, the labels are distributed evenly across the ten function categories, i.e., in each dataset, ten images are annotated with each urban function category. In fact, the previous studies of Kang *et al.* (2018), Zhao *et al.* (2021) released a dataset for building-level urban function recognition, while we choose to create and release our own dataset mainly based on three reasons: (1) their datasets are oriented to buildings, while other types of spaces (e.g., parks) are not included; (2) the semantic granularity of their function categories is coarse with four categories *residential*, *commercial*, *public*, and *industrial*; (3) the labels in their dataset are from OSM, and thus include

noise (Qiao and Yuan 2021).

4.2. *Experimental settings*

We evaluate the capacity of zero-shot UrbanCLIP in three different settings: (1) primary function classification: infer the primary function reflected in each SVI; (2) multiple function classification: we test UrbanCLIP's capacity in discovering multiple functions (primary and secondary functions) reflected within individual SVIs; and (3) cross-city transfer: we test the performance of UrbanCLIP and baseline models in two other cities Singapore and London, so as to compare their transfer capabilities.

4.3. *Baseline models*

We compare the zero-shot performance of UrbanCLIP with the following categories of baseline models:

(1) Wu *et al.* (2023): The prompting strategy for the same task from Wu *et al.* (2023) is used as a zero-shot baseline. We ensemble (average) the prompt templates proposed in this work for zero-shot urban function inference, in which the prompt templates are attached with city names as spatial contexts.
(2) Supervised CLIP: We report the performance of using CLIP for this task in a supervised manner, in three different ways: (a) CLIP-MLP, which extracts the SVI embeddings from CLIP's image encoder (ViT-L/14@336px) and feeds the SVI embeddings into a multilayer perceptron (MLP) for inference; (b) CLIP-RN101-MLP, which is the same as CLIP-MLP, except that it uses another image encoder of CLIP, i.e., a ResNet101; (c) CLIP-Finetune, which fine-tunes the entire CLIP's (ViT-L/14@336px) image encoder for the task.
(3) ViT, ResNet101, and Place365: We use the ViT pretrained on ImageNet-21k (specifically ViT-large-patch16-384) as the baselines in two different ways: (a) ViT-MLP, which extracts SVI embeddings from the pretrained ViT, and feeds the embeddings into an MLP for prediction; (b) ViT-Finetune, which fine-tunes the ViT pretrained on ImageNet21k for this task. We also utilize ResNet101 and Place365 (Zhou *et al.* 2017) (a ResNet50 trained using the scene classification dataset Place365) in the same ways, so as to form the baselines ResNet101-MLP, ResNet101-Finetune, Place365-MLP, and Place365-Finetune.
(4) Urban2Vec-Vision: This baseline is the vision model in the region representation learning method Urban2Vec (Wang *et al.* 2020), which fine-tunes an ImageNet pretrained ResNet101 by making spatially adjacent SVIs also close in the embedding space. The SVI embeddings are finally fed into an MLP for prediction.

4.4. *Implementation details*

The prompting framework UrbanCLIP is implemented in a zero-shot manner with the fully-fledged version of CLIP (ViT-L/14@336px). For the baselines that utilize an MLP for prediction, we use a unified MLP architecture. The MLP has a single hidden layer, which is 2048-dimensional and with a sigmoid activation function. For the baselines that fine-tune a vision model, all layers are fine-tuned. We do not adopt a few-shot linear probe protocol as used in Radford *et al.* (2021), as we find that this setting is generally not powerful enough, making it difficult to reveal the capacity of UrbanCLIP.

For the urban scene dataset in Shenzhen (1,518 SVIs), we use the entire dataset to test UrbanCLIP in the primary function classification setting, and use the SVIs with multiple labels for multi-function classification. For UrbanCLIP, no training data is used. For comparison, supervised baselines are tested across incremental proportions of labeled samples (10%, 20%, ..., to 80%), with 80% of these samples allocated for training/fine-tuning and 20% for validation. For instance, using 60% of labeled samples implies 48% of the dataset is for training, 12% for validation, with the remaining 40% as the test set. This procedure, repeated 100 times for each proportion, aims to determine the minimum labeled data needed to match UrbanCLIP's zero-shot performance. For fine-tuning baselines, we conduct data augmentation with horizontal flip. Different from MLP baselines, we exclude scenarios using less than 50% labeled samples and repeat the dataset split 20 times to accommodate the longer training time and increased needs of training samples.

For the cross-city transfer experiments, UrbanCLIP and the baseline method from Wu *et al.* (2023) still conduct zero-shot inference for the SVIs in Singapore and London. As to other baselines, their training and validation sets remain the same, while the test set is replaced by the SVIs in Singapore and London. This implies that the baseline models are all trained and validated with the dataset in Shenzhen, and tested in other cities.

4.5. *Results on primary function classification*

The primary urban function inference at the scene level is essentially an image classification task. Therefore, we utilize the evaluation metrics of weighted F1 score (in view of the unbalanced distribution of the primary functions in the Shenzhen dataset) and accuracy.

The performance of scene-level primary function inference is presented in Figure 3. We observe that The proposed UrbanCLIP considerably surpasses most of the baseline models, including the zero-shot method proposed in Wu *et al.* (2023). And the rich content of Figure 3 can be interpreted from several perspectives:

(1) CLIP-MLP forms a strong baseline that can slightly outperform zero-shot UrbanCLIP when using more than 50% (roughly 700) of the labeled samples. This is unsurprising, as this baseline utilizes the powerful image embeddings from CLIP, and training with many labeled samples is able to make fairly accurate predictions. But such a performance comes at the cost of more than 700 high-quality human labels, and we believe that this number would increase if employing OSM labels. In this regard, this finding does not undermine the effectiveness of UrbanCLIP, which is a zero-shot method. We observe that fine-tuned CLIP slightly under-performs CLIP-MLP. This finding aligns with the recent investigation in Dong *et al.* (2022), which finds that fine-tuning CLIP needs a large dataset and tedious hyper-parameter tuning.

(2) We observe that CLIP-based methods (i.e., CLIP-MLP and CLIP-RN101-MLP) largely excel others, including the large ViT-based models, and the fine-tuned vision models. This implies that the SVI embeddings from CLIP are more effective than from other models. The superiority of CLIP's image embeddings stems from not being confined by pre-determined categories, and the enormous number of image-text pairs. The more powerful image encoder ViT is also a key ingredient, as CLIP-MLP notably outperforms CLIP-RN101-MLP.

(3) Among other baselines, Urban2Vec-Vision and Place365-based methods are bet-

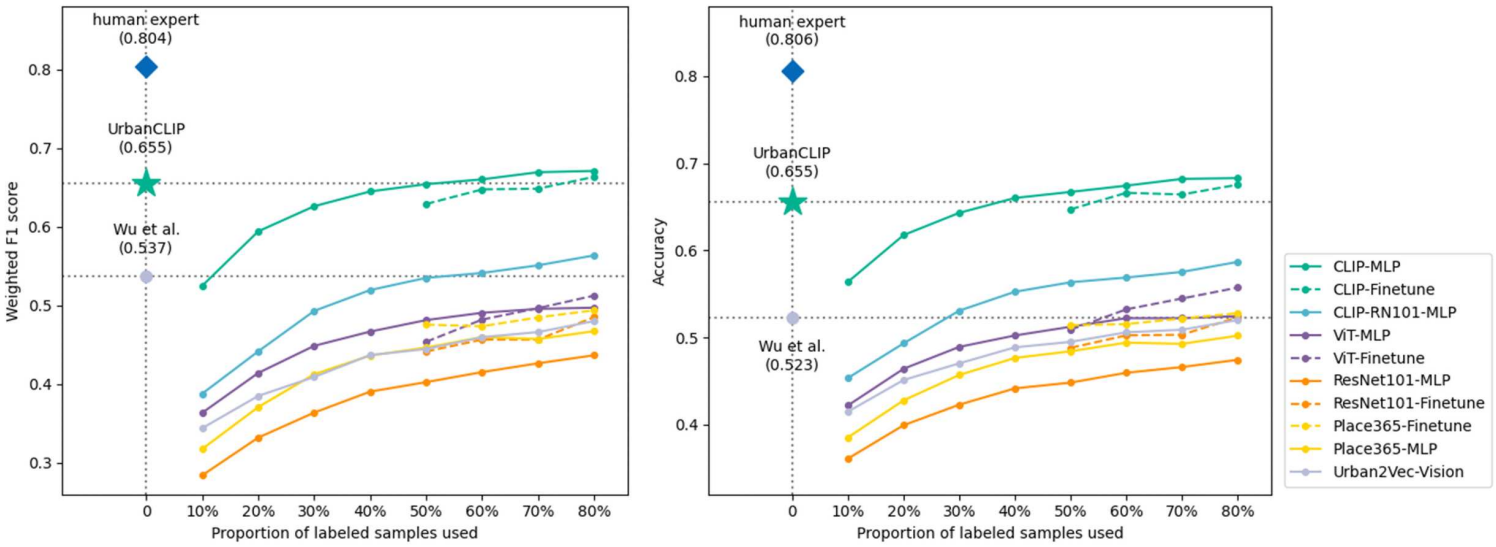

Figure 3. Results of primary function classification, including the performance of UrbanCLIP, and baseline models.

ter than the rest, which have been either trained with a spatial proximity-based objective or trained by a large scene classification dataset. Further fine-tuning Place365 using our dataset yields slightly better performance than Place365-MLP, while still cannot approach the performance of CLIP-based methods. ResNet101 comes after Place365, and it could also slightly benefit from fine-tuning.

Overall, UrbanCLIP is competent for zero-shot urban function inference at the scene level, which largely outperforms fine-tuning some deep vision models (e.g., ResNet101) with more than 1,000 high-quality labeled samples. In the meantime, we recognize that UrbanCLIP, despite its excelling performance, still cannot reach the competency of human expert, and there is still a large room for improvement.

In order to further examine the effectiveness of UrbanCLIP, we derive the confusion matrix from the zero-shot urban function inference and visualize it in Figure 4. We observe that UrbanCLIP performs well for the functions *residential* (F1: 0.82), *commercial* (F1: 0.72), and *outdoors and natural* (F1: 0.71). The competency in these functions is encouraging, as such functions are generally predominating in cities, and correctly recognizing them largely benefits fine-grained urban function mapping. We observe that *commercial* could be misrecognized as *hotel* and *industrial*, as the visual appearances of these functions are sometimes similar.

UrbanCLIP yields moderate performance for *transportation* (F1: 0.66), *sports and recreation* (F1: 0.62), *industrial* (F1: 0.60), and *education* (F1: 0.57). We observe that *industrial* buildings in Shenzhen could be misclassified as *education* facilities like schools; *education* and *residential* can also be confused, possibly because that education institutes like kindergartens are often situated in residential areas.

However, UrbanCLIP has comparatively low performance for *hotel* (F1: 0.46), *health care* (F1: 0.44), and *civic, governmental, and cultural* (F1: 0.34). The errors for *hotel* are mainly due to the confusion with *commercial*. *Health care* could be mis-interpreted as various functions, e.g., *residential*, *education*, and *industrial*. The main challenge for *civic, governmental, and cultural* is confusion with *industrial*, which is likely because

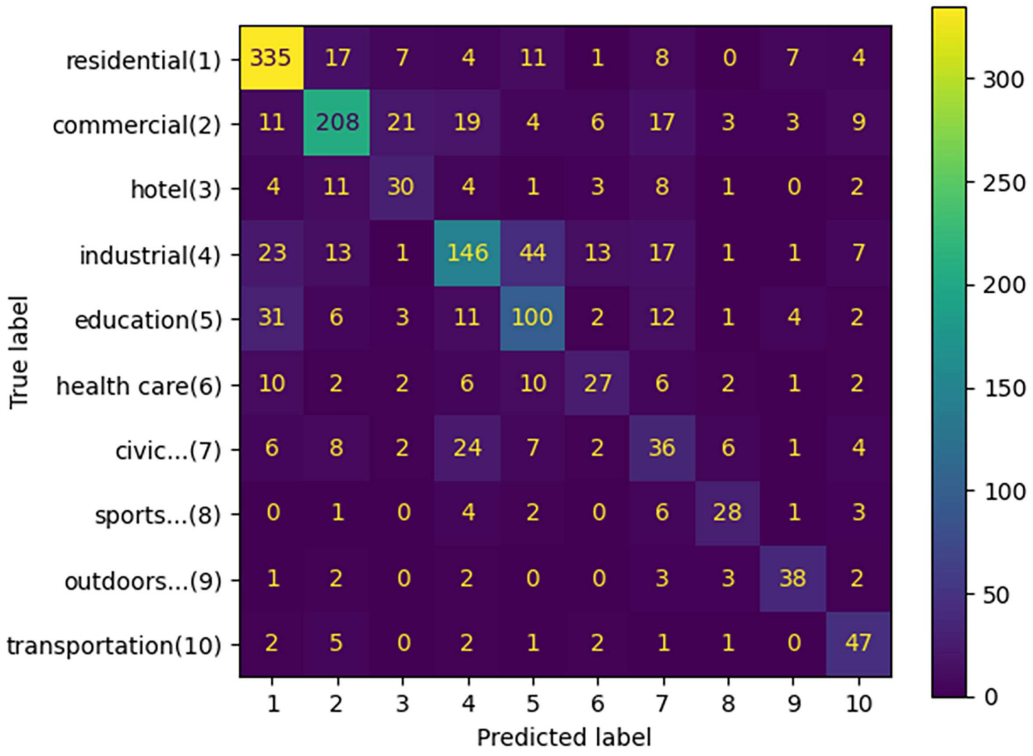

Figure 4. Confusion matrix of UrbanCLIP zero-shot primary function classification.

the building scales of the two functions can be similar.

4.6. *Results on multiple function classification*

Multiple function recognition for urban scenes can be boiled down to a multi-label image classification problem, so we utilize the two evaluation measures of weighted F1 score and Jaccard index, which is, for two label sets, the proportion of their interaction relative to their union (Murphy 1996). The results in this setting are demonstrated in Figure 5.

The results follow a similar pattern as the results for primary function inference (Figure 3). UrbanCLIP's zero-shot inference for multiple functions still surpasses the zero-shot baseline and most of the supervised baselines. We observe:

(1) UrbanCLIP's performance is comparable to CLIP-MLP with 10%-20% labeled samples (roughly 150-300 samples). In a way, UrbanCLIP's competency declines compared to its capacity in primary function inference. This is likely due to that multiple functions usually come with different objects in an SVI. For example, there are two buildings in an SVI, in which one represents *commercial*, and the other is *residential*. However, UrbanCLIP's zero-shot inference could mainly focus on the more salient commercial building, thereby possibly making the inference of *commercial* and *hotel* (both are targeted at the salient commercial building). Nevertheless, UrbanCLIP could still produce good multi-function inference without the need of labeled samples. In fact, high-quality multi-function

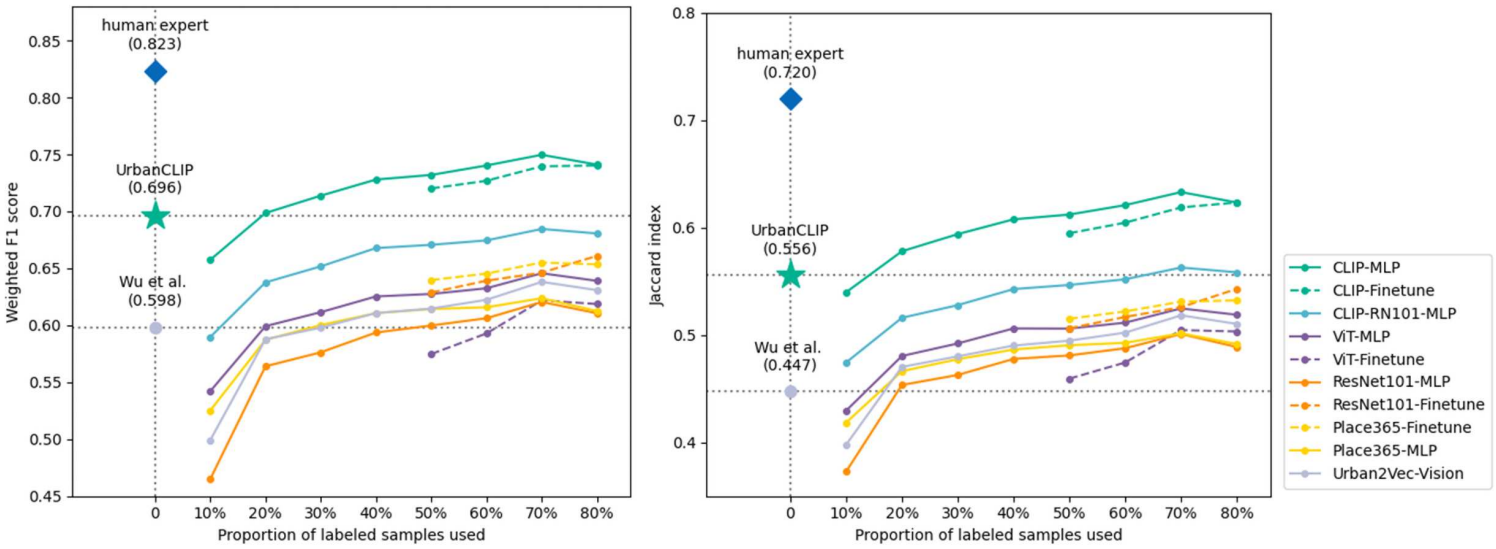

Figure 5. Results of multiple function classification, including the performance of UrbanCLIP, and baseline models.

ground truth labels at the fine-grained scene level can be barely obtained in scale, implying that it is often impractical to rely on a large number of labels in this setting. This observation strengthens the usefulness of the zero-shot UrbanCLIP.

(2) Human expert performs considerably better than UrbanCLIP in this setting. This could potentially be ascribed to that we use the machine learning models in an image classification manner, determining each SVI as a whole. However, human perception could easily discover different objects in each SVI, thus yielding wiser inferences. In this regard, one possible future work is to combine UrbanCLIP and object detection to improve multi-function inference.

4.7. *Results on cross-city model transfer*

The previous two experiments focus on a single city. Considering that collecting a certain amount of training labels for each city comes at great expenses, it is desirable to have a generalizable model that requires no further training to be applied in a new city. In this context, we test the cross-city transfer capacity of UrbanCLIP and the baseline models. We infer the primary urban functions in Singapore and London. The supervised baselines are trained and validated using the dataset in Shenzhen, and tested in Singapore and London, while UrbanCLIP still carries out zero-shot inference.

The results of cross-city model transfer are presented in Figure 6, from which we observe that UrbanCLIP significantly outperforms all baseline models. Specifically, we can draw the following observations:

(1) Zero-shot UrbanCLIP outperforms all baselines, regardless of how many labeled samples in Shenzhen are used. This situation differs from the primary function and multi-function inference experiments in the same city, where supervised CLIP baselines surpass UrbanCLIP if using a large or moderate number of labeled samples. The underlying driver of the difference is that although CLIP's image embeddings are powerful, the MLP or the entire image encoder (in the fine-tuning cases) is trained in Shenzhen, and thus biased towards the style of the

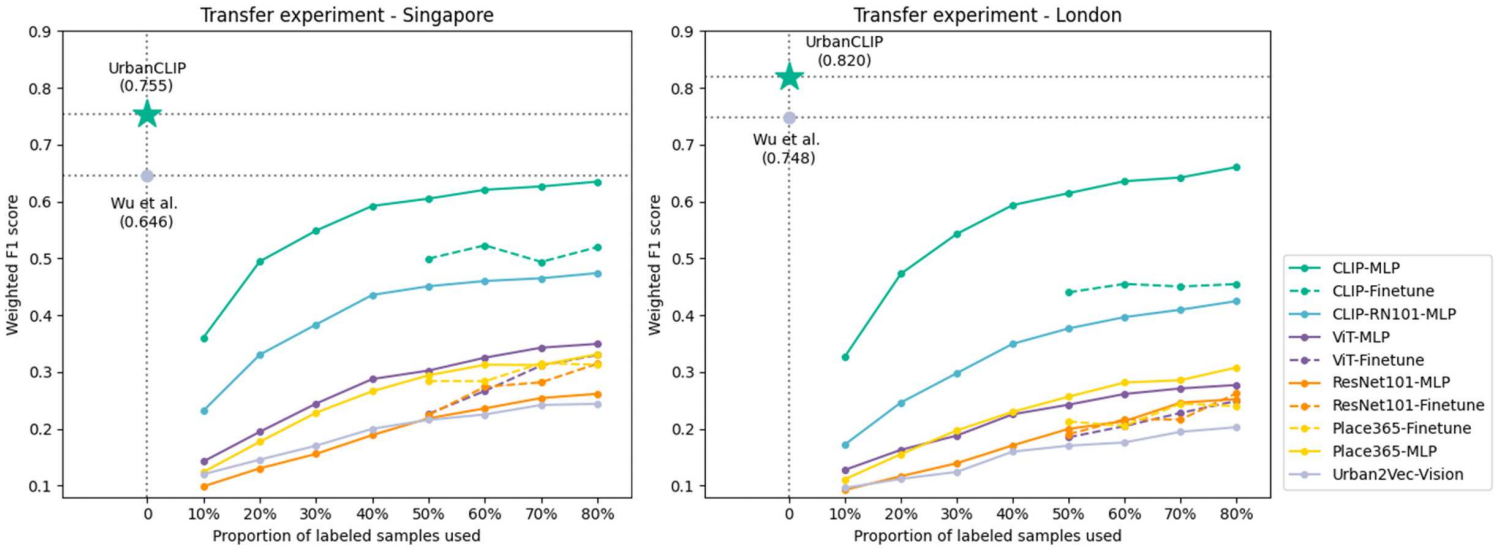

Figure 6. Results of cross-city transfer experiments, including the performance of UrbanCLIP, and baseline models. The baselines are trained with the urban scene dataset in Shenzhen, while tested in Singapore and London.

particular city. Singapore and London have considerably different visual styles and urban function distributions from Shenzhen. Using the trained models biased towards Shenzhen for inference in two other cities provides compromised effectiveness. Nevertheless, we recognize that there are still commonalities among the three cities, as increasing the number of training samples does benefit the performance in Singapore and London.

(2) Fine-tuning the deep vision models of CLIP, ViT, Place365 and ResNet101 does not benefit the performance in comparison to their MLP-based counterparts. In fact, fine-tuning, in most cases, results in performance decline. The declines are more notable for larger models like CLIP and ViT. Such evidence hints on that fitting a deep vision model with numerous parameters makes the fine-tuned models highly tailored to the training dataset, while it generally does not benefit, or could hurt, the inference in other study areas.

(3) UrbanCLIP continues to perform better than the prompting method in Wu *et al.* (2023), while the margins become smaller in this setting, especially in London. The reason behind this could be partially ascribed to the usefulness of spatial context in these prominent and English-background cities, for which we provide in-depth analysis in Section 4.8.

Overall, UrbanCLIP demonstrates notable advantages compared to all baseline models, in the cross-city transfer scenario. It is encouraging to see that the proposed UrbanCLIP is effective in various different cities in a zero-shot manner. In contrast, the fine-tuning strategy does not perform well when tested in vastly different study areas.

4.8. *Ablation study and UrbanCLIP variants*

The proposed prompting framework UrbanCLIP has two major components: the urban taxonomy and the urban function prompt templates. We thoroughly investigate the

effectiveness of the two components in an ablation study through replacing each component with its alternatives, and thus form various combinations of the prompts for zero-shot urban function inference. The results are presented in Table 2. The ablation study is carried out in all three settings, i.e., primary function inference (Primary), multi-function inference (Multiple), and cross-city transfer (Singapore and London). In each setting, we test both the scenarios of using the urban taxonomy (UOT) and using only the category names of urban functions (FN), e.g., *residential*. In terms of prompt templates, we test the urban function prompt templates (UrbanCLIP-PT) proposed in this study; the ensembling (through averaging) of the templates from Wu *et al.* (2023); the ensembling of the designed 80 templates from Radford *et al.* (2021) (CLIP 80); the simple template "`A photo of a {label}.`" (A photo of); and the scenario of only using categories (UOTs or function names) without any prompt template (No template). To assess the usefulness of spatial context, we append city names to the designed urban function prompt templates (except for PT1), e.g., for inference in Shenzhen, PT2 becomes "`A street photo of {UOT} in Shenzhen.`", and PT3 becomes "`A street photo of {UOT} with many trees in Shenzhen.`"; such templates are denoted UrbanCLIP-PT + SC. We also remove the spatial context (city name) from the templates designed in Wu *et al.* (2023) to form Wu et al. w/o SC. In addition, we test the Zero-shot Prompt Ensembling (ZPE) method (Allingham *et al.* 2023) for prompt template ensembling, to replace the averaging ensembling used in UrbanCLIP, and form UrbanCLIP-PT + ZPE. From the results, the following observations are drawn:

Table 2. Results (weighted F1 scores) of the ablation study. The best performance of each setting is bolded, while the second best is underscored.

Templates	Primary		Multiple		Singapore		London	
	UOT	FN	UOT	FN	UOT	FN	UOT	FN
UFPT	**0.655**	0.519	<u>0.696</u>	0.548	0.755	0.674	0.820	0.674
Wu et al.	0.546	0.537	0.648	0.598	<u>0.776</u>	0.646	**0.841**	0.748
CLIP 80	0.558	0.368	0.681	0.363	0.731	0.591	<u>0.840</u>	0.655
A photo of	0.533	0.371	0.652	0.383	0.684	0.629	0.801	0.627
No template	0.502	0.363	0.672	0.393	0.686	0.569	0.763	0.633
UFPT + SC	0.579	0.533	0.617	0.553	**0.797**	0.691	0.839	0.642
Wu et al. w/o SC	0.527	0.482	0.676	0.552	0.721	0.662	0.809	0.766
UFPT + ZPE	<u>0.632</u>	0.514	**0.709**	0.522	0.770	0.618	0.809	0.663

(1) The usefulness of the proposed urban taxonomy is prominent, and is the major source of performance gain in all settings. This is another evidence that CLIP is more capable of understanding concrete concepts than abstract ones (Liao *et al.* 2023), which is an important experience for prompting pretrained models for geospatial analyses where abstract and specialized concepts are prevalent.

(2) The proposed urban function prompt templates are useful in all settings, which implies that our incentive of using these templates for mitigating the interfering information in SVIs is valid. The templates from Wu *et al.* (2023) are also tested in conjunction with the urban taxonomy designed in our study. Their templates

bring slight benefits for the inferences in Shenzhen, while are more useful in Singapore and London, leading to better performances than using the prompt templates proposed in our study.

(3) The usefulness of spatial context varies in different cities. In the settings of primary and multi-function inference, adding the city name Shenzhen mostly undermines the performance. However, it is the opposite in two other cities. The incorporation of spatial context is also one of the underlying reasons of that the templates from Wu *et al.* (2023) perform better than ours in Singapore and London, when combined with the proposed urban taxonomy. This finding aligns with the recent investigation in Nwatu *et al.* (2023), which revealed that images from lower-income places and in non-Western styles can be more challenging for CLIP.

(4) Prompt template ensembling through simple averaging generally provides comparable performance to ZPE, which is encouraging to keep UrbanCLIP simple. We believe that the underlying reason for this finding is that urban function inference is a specialized application of CLIP, making the introduction of the overall signal from the pre-training dataset (LAION 400M in this case) less meaningful. In addition, the six urban function prompt templates in UrbanCLIP are carefully crafted leveraging domain knowledge, so each template generally holds value, which is different from the scenario of ensembling a large number of prompt templates in Allingham *et al.* (2023).

We also investigate the usefulness of each of the proposed urban function prompt templates. In primary function inference, all templates are proved beneficial, with their removal leading to a decline in performance. In other settings, the necessity of some templates varies. For example, using PT3 slightly diminishes the performance in the transfer tests in Singapore and London, as there is not much interference from road-side landscape in this setting, and the signal "many trees" becomes somewhat distracting. In multi-function inference, removing PT6 results in a slight increase in F1 score, as the signal of "parking lot" potentially overshadows its actual occurrence as the second function in certain areas.

Overall, the thorough ablation study verifies the design of UrbanCLIP. We recognize that the design of UrbanCLIP may not be optimal in certain cases. Nevertheless, it is generally effective across all settings, which is crucial for practical applications. Particularly, we speculate that incorporating spatial context could lift the performance in Western-style and higher-income cities, but it could possibly negatively impact performance in other cities. Therefore, we choose not to introduce spatial context in UrbanCLIP by default, because fine-grained mapping of urban functions in cities of emerging and developing economies (which may have fewer resources) could be more meaningful, and higher-income cities often have better curation of urban function (land use) data. Furthermore, the ablation study unveils promising pathways to adjust the proposed prompt templates, e.g., PT3 might not be necessary for SVIs with not much appearance of road-side landscape, and ZPE could potentially be utilized when multi-function inference is prioritized.

4.9. *Error analysis*

We carry out a thorough analysis of the inference errors of UrbanCLIP, and we select nine representative examples. The selected cases are demonstrated in Figure 7, in which (a)-(f) are representative errors in primary function inference, (g) is an error in multi-

Figure 7. Error analysis of misclassified urban scenes. The labels indicate the first two inferred UOTs. For (g), the first inferred UOT for each function is provided.

function inference, and (h)-(i) are two errors in Singapore and London respectively.

Figure 7(a) is a commercial complex in Shenzhen, which is transformed from a cruise. UrbanCLIP is misled by the appearance and assigns the largest cosine similarity to port in the urban taxonomy (*transportation*). Figure 7(b) is a factory, while UrbanCLIP believes that the most similar UOT to it is secondary school due to notable visual similarity, which belongs to the function of *education*. Figure 7(c) is a residential compound, and its wall is captured. The wall has a commercial advertisement for a restaurant, so interestingly UrbanCLIP is misled by the advertisement and recognizes it as a food center (*commercial*). Figure 7(d) is the building of a governmental agency in Shenzhen which looks like an industrial park, and thus UrbanCLIP believes this is most likely an *industrial* area. In fact, the name of the administration is attached to the building and visible in the SVI, while UrbanCLIP still does not recognize it, probably because that it has limited capacity in understanding Chinese characters. Figure 7(e) is a shopping mall in Shenzhen with salient patterns of waves on the building's facade. In this case, UrbanCLIP assigns the highest similarity to the UOT swimming complex (*sports and recreation*). Figure 7(f) is a pavilion within a

park (*outdoors and natural*), which has a classic Chinese roof. In this case, UrbanCLIP misunderstands it to be a religious facility (*civic, governmental, and cultural*). Figure 7(g) entails both *residential* and *commercial*. UrbanCLIP provides the correct answer for its primary function, whereas believes that the secondary function is *outdoors and natural* in view of the distant mountain in the SVI. Figure 7(h) in a science park (*industrial*) in Singapore, while UrbanCLIP recognizes it as office towers. This could be because the firms in this area belong to the research and development industry, rather than manufacturing, so the working spaces are more similar to office buildings. Figure 7(i) is a hospital (*health care*) in London, while UrbanCLIP believes that it is most likely a university (*education*). In fact, such classic building facades sometimes do not have explicit linkages to urban functions, as "form follows function" is a modern architectural design principle.

Overall, we summarize the reasons for several representative errors in UrbanCLIP's zero-shot urban function inference: (1) uncommon building appearances (cases a and e); (2) similar design/planning forms shared by different urban functions (cases b, d, f, and h); (3) being distracted by side information (cases c and g); (4) indirect linkages between appearances and urban function in classical architectures (case i).

4.10. *Fine-grained urban function mapping*

In this section, we map out the scene-level urban functions inferred from each SVI, thus generating a fine-grained functional land use map of Shenzhen. 226,881 SVIs points are used in total, and the mapping results in the main urban area are demonstrated in Figure 8.

From the overview mapping, we observe that the overall spatial distribution of urban functions is well captured, e.g., *residential* serves as the background of the city, which is often mixed with *commercial*. Upon such background, there appear some *commercial* and *industrial* functional clusters. In addition, the skeleton of the major roads is apparent, which is delineated by linear aggregations of *transportation* and *outdoors and natural*, as there are usually green buffers or natural areas besides major roads.

We zoom in and choose four representative areas to further demonstrate the mapping results. Case 1 is a residential neighborhood by the sea, and UrbanCLIP successfully identifies the major occupation of *residential*, which is mixed with *commercial* and *education*. Along the coastline, the function of *outdoors and natural* is also well captured. Case 2 shows the urban spine of Shenzhen, in which the north (upper) side is dominated by museums and exhibition halls (*civic, governmental and cultural*), while the south side is a mixture of luxury hotels and commercial venues. Case 3's center lies a stadium (*sports and recreation*). UrbanCLIP successfully outlined the sports center, which complements the land use classification found in Shenzhen's detailed plan [6]. Case 4 maps a university campus, which has a cluster of university residence halls (bottom left side), a cluster of education buildings mixed with open spaces (upper left side), and a cluster of science parks affiliated with the university (upper right side).

The visualizations provide encouraging evidence of UrbanCLIP, proving the framework is capable of mapping fine-grained urban functions without any labeled samples and any model training. This urban function map has a fine granularity both semantically and spatially, which can be particularly useful in various urban planning and management practices.

[6] http://pnr.sz.gov.cn/d-xgmap/

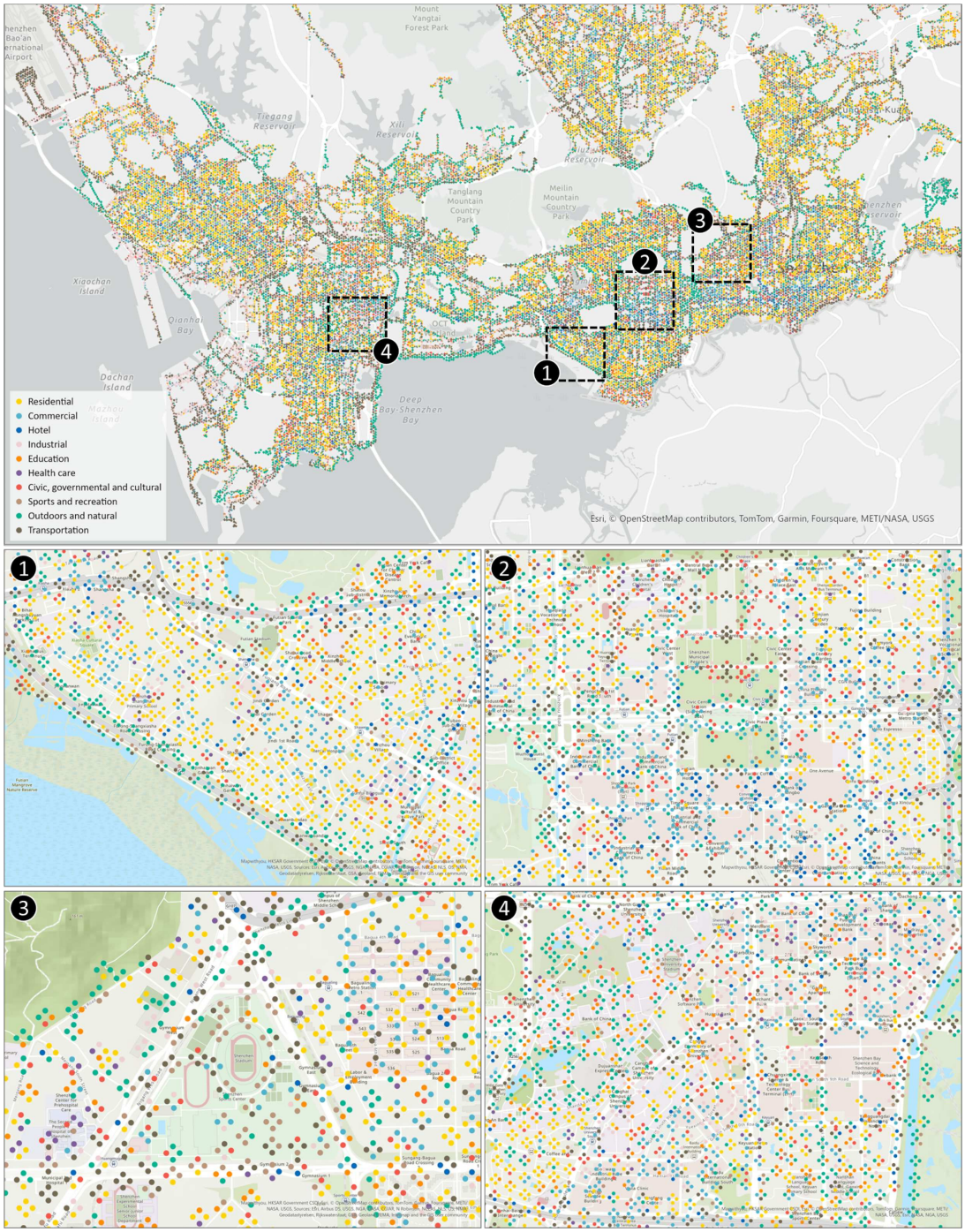

Figure 8. Fine-grained and zero-shot urban function mapping with UrbanCLIP.

5. Discussion

Today, large-scale pretrained models (foundation models) leveraging the unbridled powers of data and computing has become the frontier of machine learning research. Some representative foundation models include GPT-family models (e.g., ChatGPT[7]), DALL-E-2 (Ramesh *et al.* 2022), and CLIP. Mai *et al.* (2024) claimed that foundation models hold great potentials for geospatial applications that are yet to be unleashed. There have been some attempts to use foundation models for geospatial problems. For example, Xue *et al.* (2022) fine-tuned several language foundation models for human mobility prediction; Balsebre *et al.* (2023) fine-tuned the pretrained Bert for completing missing relations in geospatial knowledge graphs. Such studies yielded promising results, while they still fall under the paradigm dubbed "*pre-train, fine-tune*". In contrast, the proposed UrbanCLIP relies on an emerging new paradigm dubbed "*pre-train, prompt, and predict*" (Liu *et al.* 2021).

A key feature of the new paradigm is that it relies on minimal computing resources and human labors for data annotation. This does not mean that UrbanCLIP's superior performance comes for free, and we have to remember that the foundation models like CLIP are pretrained at enormous expenses. The foundation models are highly capable of general domain understanding, while they need to be guided to understand geospatial problems, which are often highly specialized. In our experiments, we demonstrate that further fine-tune CLIP does not guarantee performance lift, especially in a small dataset regime. In addition, fine-tuning entails tedious hyper-parameter tuning and is usually only feasible with expensive computing infrastructures. Instead, we unlock the potential of the pretrained CLIP in the target task with our instructions and guidance, i.e., prompting. In this regard, we design the urban taxonomy and urban function prompt templates to guide CLIP, enabling it to infer urban functions at low cost.

A crucial message that can be distilled from this study is the indispensable role of geospatial knowledge in prompting (guiding) the foundation models to understand geospatial questions. This observation aligns with the recent study in geo-knowledge-guided prompting for location description extraction from social media data, in which they found that guiding GPT with a few examples in different categories results in notable performance enhancement (Hu *et al.* 2023). In this study, we incorporate several types of geospatial knowledge in the UrbanCLIP framework, including (1) knowledge embedded in the categorization of geospatial data (e.g., POIs) supplemented with expertise in urban studies to develop the urban taxonomy, so as to help CLIP understand the abstract urban function types; (2) knowledge in geospatial ontology design (e.g., Kuhn 2001) that informs the process of developing the hierarchical urban taxonomy; (3) knowledge in the nature of SVIs to discover the potential interfering visual clues, which is used to guide the design of urban function prompt templates.

We believe that UrbanCLIP is one of the pioneering works in prompting engineering for geospatial applications, which has vastly unexplored potentials. In this paper, we tailor CLIP to the task of urban function inference, while it is likely that it could also be tailored to other investigations like analyzing urban greenery, road condition, and urban noise pollution. We speculate that we may need to develop distinct ways to prompt CLIP to understand different applications. For example, one could develop a taxonomy for urban noise analysis, including the concepts like "noisy area", "quiet area", and "traffic-intensive area". In this context, we envision that UrbanCLIP could be extended

[7] https://openai.com/blog/chatgpt/

to support (zero-shot) inference for various urban applications, with a knowledge base (Huang and Harrie 2020) encapsulating various taxonomies and prompt templates, which can be organized in ontologies to foster wide adoption and reusability.

Despite the promising results, UrbanCLIP has several limitations. First, UrbanCLIP's multi-function inference has substantial room for improvement, as it sometimes pays most attention to the most predominating object in each SVI, but not multiple objects. In this regard, UrbanCLIP could be further developed to incorporate an object detection pipeline to help multi-function inference. Second, our framework is still heavily reliant on domain knowledge-informed manual design, which is laborious. The performance can also be impacted by different choices of language phrases (Zhou *et al.* 2022a). Further developments can be carried out in developing city-specific urban taxonomies out of a comprehensive overall taxonomy, possibly through analyzing the interpretation of image embeddings in a specific city (Bhalla *et al.* 2024). Another possible future work is to learn prompt templates to be combined with the urban taxonomy, instead of hand-crafting them (Zhou *et al.* 2022b). In addition, UrbanCLIP inherents the limitations of its foundation CLIP, which is biased towards wealthy and Western-style cities. In this regard, a vision-language model pretrained with global-scale city images, if viable, would be desirable (Klemmer *et al.* 2023).

6. Conclusions

In this paper, we propose the prompting framework of UrbanCLIP to enable the pretrained vision-language model CLIP for zero-shot urban function inference with SVIs, i.e., it requires no labeled training samples and no model training. UrbanCLIP comprises an urban taxonomy and several urban function prompt templates, in order to prompt CLIP to understand the specialized target application. Through extensive experiments, we find that the zero-shot UrbanCLIP outperforms several competitive supervised models with many labeled samples, e.g., it largely outperforms a fine-tuned ResNet, and the advantages become more prominent in cross-city transfer scenarios. In addition, UrbanCLIP's zero-shot performance is considerably better than using the vanilla CLIP without our prompting framework. In summary, UrbanCLIP is a simple but effective framework for urban function inference with SVIs, and it could substantially diminish the demand for labeled samples and computing resources.

Acknowledgment

We appreciate Yi Li at Nanyang Technological University for his technical inputs, and James Allingham at the University of Cambridge for his help in implementing the prompt ensembling method ZPE.

Funding

This work was partially funded by the Knut and Alice Wallenberg Foundation (KAW 2019.0550), the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No. AISG2-TC-2021-001), a Singapore MOE AcRF Tier-2 grant (MOE-T2EP20221-0015), and a Singapore MOE AcRF Tier-1 project (RT6/23). The work was also partially funded by the Future Cities Lab Global programme. Fu-

ture Cities Lab Global is supported and funded by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme and ETH Zurich, with additional contributions from the National University of Singapore, Nanyang Technological University, and the Singapore University of Technology and Design.

Notes on contributors

Weiming Huang received his PhD in Geographical Information Science at Lund University, Sweden in 2020. He is a Wallenberg-NTU Postdoctoral Fellow at Nanyang Technological University, Singapore. His research interests mainly include spatial data mining and geospatial knowledge graphs.

Jing Wang is an urban planner and a researcher at Future Cities Laboratory, Singapore-ETH Centre. Her research interests lie in knowledge discovery from spatial data to inform sustainble planning and design.

Gao Cong is currently a Professor in the School of Computer Science and Engineering at Nanyang Technological University (NTU). He received his PhD degree from the National University of Singapore in 2004. His current research interests include spatial data management, machine learning for databases, spatial-temporal data mining, and recommendation systems.

Data and codes availability statement

The data and codes that support the findings of this study are available at https://github.com/RightBank/UrbanCLIP.

References

Allingham, J.U., *et al.*, 2023. A simple zero-shot prompt weighting technique to improve prompt ensembling in text-image models. *In*: *International Conference on Machine Learning*. PMLR, 547–568.

Bai, L., *et al.*, 2023. Geographic mapping with unsupervised multi-modal representation learning from vhr images and pois. *ISPRS Journal of Photogrammetry and Remote Sensing*, 201, 193–208.

Balsebre, P., *et al.*, 2023. Mining geospatial relationships from text. *Proceedings of the ACM on Management of Data*, 1 (1), 1–26.

Bhalla, U., *et al.*, 2024. Interpreting clip with sparse linear concept embeddings (splice). *arXiv preprint arXiv:2402.10376*.

Biljecki, F. and Ito, K., 2021. Street view imagery in urban analytics and gis: A review. *Landscape and Urban Planning*, 215, 104217.

Chen, Y.C., *et al.*, 2020. Uniter: Universal image-text representation learning. *In*: *European conference on computer vision*. Springer, 104–120.

Devlin, J., *et al.*, 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *In*: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, June, Minneapolis, Minnesota. Association for Computational Linguistics, 4171–4186.

Dong, X., *et al.*, 2022. Clip itself is a strong fine-tuner: Achieving 85.7% and 88.0% top-1 accuracy with vit-b and vit-l on imagenet. *arXiv preprint arXiv:2212.06138*.

Dosovitskiy, A., *et al.*, 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *In*: *International Conference on Learning Representations*.

Du, Y., *et al.*, 2022. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*.

Haas, L., Alberti, S., and Skreta, M., 2023. Learning generalized zero-shot learners for open-domain image geolocalization. *arXiv preprint arXiv:2302.00275*.

He, K., *et al.*, 2016. Deep residual learning for image recognition. *In*: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

Hu, S., *et al.*, 2021. Urban function classification at road segment level using taxi trajectory data: A graph convolutional neural network approach. *Computers, Environment and Urban Systems*, 87, 101619.

Hu, Y., *et al.*, 2023. Geo-knowledge-guided gpt models improve the extraction of location descriptions from disaster-related social media messages. *International Journal of Geographical Information Science*, 37 (11), 2289–2318.

Huang, W., *et al.*, 2022. Estimating urban functional distributions with semantics preserved poi embedding. *International Journal of Geographical Information Science*, 36 (10), 1905–1930.

Huang, W. and Harrie, L., 2020. Towards knowledge-based geovisualisation using semantic web technologies: A knowledge representation approach coupling ontologies and rules. *International Journal of Digital Earth*, 13 (9), 976–997.

Huang, W., *et al.*, 2023. Learning urban region representations with pois and hierarchical graph infomax. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196, 134–145.

Jia, C., *et al.*, 2021. Scaling up visual and vision-language representation learning with noisy text supervision. *In*: *International conference on machine learning*. PMLR, 4904–4916.

Jia, M., *et al.*, 2022. Visual prompt tuning. *In*: *European Conference on Computer Vision*. Springer, 709–727.

Kang, J., *et al.*, 2018. Building instance classification using street view images. *ISPRS journal of photogrammetry and remote sensing*, 145, 44–59.

Kim, W., Son, B., and Kim, I., 2021. Vilt: Vision-and-language transformer without convolution or region supervision. *In*: *International Conference on Machine Learning*. PMLR, 5583–5594.

Kipf, T.N. and Welling, M., 2017. Semi-Supervised Classification with Graph Convolutional Networks. *In*: *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17.

Klemmer, K., *et al.*, 2023. Satclip: Global, general-purpose location embeddings with satellite imagery. *arXiv preprint arXiv:2311.17179*.

Kuhn, W., 2001. Ontologies in support of activities in geographical space. *International Journal of Geographical Information Science*, 15 (7), 613–631.

Li, J., *et al.*, 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34, 9694–9705.

Li, X., *et al.*, 2015. Assessing street-level urban greenery using google street view and a modified green view index. *Urban Forestry & Urban Greening*, 14 (3), 675–685.

Liao, J., Chen, X., and Du, L., 2023. Concept understanding in large language models: An empirical study. *In*: *ICLR 2023 Tiny Papers*.

Liu, P., *et al.*, 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Lu, Y., *et al.*, 2022. Prompt distribution learning. *In*: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5206–5215.

Mai, G., *et al.*, 2024. On the opportunities and challenges of foundation models for geoai (vision paper). *ACM Transactions on Spatial Algorithms and Systems*.

Murphy, A.H., 1996. The finley affair: A signal event in the history of forecast verification. *Weather and forecasting*, 11 (1), 3–20.

Nwatu, J., Ignat, O., and Mihalcea, R., 2023. Bridging the digital divide: Performance variation across socio-economic factors in vision-language models. *arXiv preprint arXiv:2311.05746*.

Qiao, Z. and Yuan, X., 2021. Urban land-use analysis using proximate sensing imagery: a survey. *International Journal of Geographical Information Science*, 35 (11), 2129–2148.
Radford, A., *et al.*, 2021. Learning transferable visual models from natural language supervision. *In*: *International Conference on Machine Learning*. PMLR, 8748–8763.
Radford, A., *et al.*, 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1 (8), 9.
Ramesh, A., *et al.*, 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.
Srivastava, S., *et al.*, 2020. Fine-grained landuse characterization using ground-based pictures: a deep learning solution based on globally available data. *International Journal of Geographical Information Science*, 34 (6), 1117–1136.
Su, W., *et al.*, 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.
Vaswani, A., *et al.*, 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
Wang, W., *et al.*, 2021. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*.
Wang, Z., Li, H., and Rajagopal, R., 2020. Urban2vec: Incorporating street view imagery and pois for multi-modal urban neighborhood embedding. *In*: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 34, 1013–1020.
Wu, M., *et al.*, 2023. Mixed land use measurement and mapping with street view images and spatial context-aware prompts via zero-shot multimodal learning. *International Journal of Applied Earth Observation and Geoinformation*, 125, 103591.
Xu, Y., *et al.*, 2022. Application of a graph convolutional network with visual and semantic features to classify urban scenes. *International Journal of Geographical Information Science*, 1–26.
Xue, H., Voutharoja, B.P., and Salim, F.D., 2022. Leveraging language foundation models for human mobility forecasting. *In*: *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*. 1–9.
Yong, G., *et al.*, 2023. Prompt engineering for zero-shot and few-shot defect detection and classification using a visual-language pretrained model. *Computer-Aided Civil and Infrastructure Engineering*, 38 (11), 1536–1554.
Zhang, X., Du, S., and Wang, Q., 2018. Integrating bottom-up classification and top-down feedback for improving urban land-cover and functional-zone mapping. *Remote Sensing of Environment*, 212, 231–248.
Zhang, Y., Zhang, F., and Chen, N., 2022. Migratable urban street scene sensing method based on vision language pre-trained model. *International Journal of Applied Earth Observation and Geoinformation*, 113, 102989.
Zhao, K., *et al.*, 2021. Bounding boxes are all we need: street view image classification via context encoding of detected buildings. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–17.
Zhou, B., *et al.*, 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40 (6), 1452–1464.
Zhou, K., *et al.*, 2022a. Conditional prompt learning for vision-language models. *In*: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16816–16825.
Zhou, K., *et al.*, 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130 (9), 2337–2348.
Zhu, Y., Deng, X., and Newsam, S., 2019. Fine-grained land use classification at the city scale using ground-level images. *IEEE Transactions on Multimedia*, 21 (7), 1825–1838.

Appendix A. Urban taxonomy

Urban function	Urban object type
Residential	apartment, attached housing, bungalow, central-passage house, chattel house, condominium, cottage, courtyard house, detached housing, dormitory, duplex house, elevator apartment, elevator apartment luxury type, elevator apartment with stores, elevator cooperative, flats, flats with commercial uses at first storey, housebarn, loft apartment, loft with stores, log house, mansion, mobile homes/ trailer parks, multi-family housing, permanent living quarter, primarily family residence with store/ office, primarily residence - mixed use, quadplex house, residential area, residential shop, residential with shop first floor, retirement housing, semi-detached house, serviced apartments, shanty town, shophouse, single-family housing, single/ multiple dwelling with stores/ offices, strata-landed housing, student hostel, suburban residence, summer cottage, tenement, terrace house, townhouse, triplex house, urban village, walk-up apartment, walk-up cooperative
Commercial	bakery, bank, banking facilities, bath & massage center, beauty & hairdressing store, big box retail, big box store, business office, business park, car repair, car sales/ rental, car sales/ rental lots without showroom, car sales/ rental with showroom, car wash/ lubritorium facility, car washes, cinema, clothing store, coffee house, commercial buildings, commercial street, community shopping center, comprehensive market, convenience store, daily life shop, dessert house, entertainment center, fast food restaurant, finance & insurance institution, food & beverages, food & beverages shop, food centre, foreign trade mission, franchise store, home building materials market, home electronics hypermarket, icecream shop, laundry, leisure food restaurant, lifestyle center, logistics service, lottery store, market, miscellaneous store building, mother & baby store, moving service, multi-story department store, multi-story retail building, neighborhood shopping center, office building with commercial uses at first storey, office towers, offices, one story retail building, personal care shop, photo studio, plant & pet market, pop-up retail, power center, predominant retail with other uses, regional shopping center, repair store, restaurant, retail, retail outlet, shared device, shopping, shopping center with/ without parking, shopping mall, shopping plaza, sports store, stand-alone food establishment, stationary/ office supply store, store buildings, strip/ convenience shopping center, super-regional shopping center, supermarket, tea house, trading house, travel agency
Hotel	backpackers hostel, boutique hotels, boutique rooms, economical chain hotel, extended stay hotels, five-star hotel, four-star hotel, full service hotels, hostel, hotel, luxury hotel, miscellaneous hotel, motel, resort hotels, three-star hotel, two-star hotel, youth hostel
Industrial	clean industry, contractors warehouse, distribution warehouse, distribution/ fulfillment centers, factory, general industry, heavy manufacturing, industrial area, industrial enterprises, industrial estate, industrial machinery, industrial park, innovation park, light industry, light manufacturing, manufacturing, manufacturing quarter, metal frame warehouse, miscellaneous warehouse, science & technology park, self-storage warehouses, warehouses
Education	city university, college, driving school, elementary school, faculty building, foreign system school, high school, institute of technical education, junior college, kindergarten, middle school, miscellaneous educational facility, nursery school, other college & university, parochial school, polytechnic, primary school, private school, public elementary, junior/ senior high school, religious school/ institute, research institution, school, science, culture & education service, secondary school, special education school, training institution, training school, university, university academic building, university cafeteria, university canteen, university library
Healthcare	adult care facility, clinic, dental clinic, disease prevention institution, dispensary, emergency center, health care facility, health center, hospital, hospitals & health facilities, infirmary, inpatient building, medical center, medical service, mental institution, miscellaneous hospital, pharmacy, polyclinic, sanitarium, special hospital, veterinary clinic, veterinary hospital

Civic, governmental and cultural	archives hall, art gallery, arts organization, assembly, church, civic institutions, community center, community hall, community institutions, concert hall, consulate, convent, convention & exhibition center, court house, cultural institutions, customs building, embassy, exhibition hall, fire station, foreign organization, government/ city departments, governmental & social groups, governmental administration building, governmental association premises, governmental organization, governmental organization & social group, jail, media organization, military & naval installation, miscellaneous indoor public assembly, miscellaneous religious facility, mosque, museum, observatory, parliament house, performing arts centre, planetarium, police station, post office, prison, public library, public security organization, religious facilities, science & technology museum, temple, theatres
Sports and recreation	amusement park, baseball field, bowling alleys, campsite, ferris wheel, golf course, golf driving range, gym, marina, yacht club, miscellaneous outdoor recreational facility, Olympic stadium, outdoor recreational facilities, outward bound school, playground, recreation club, sports & recreation, sports & recreation places, sports complex/ indoor stadium, sports stadium, stadium, swimming complex, tennis court, theme park, water sports centre
Outdoors and natural	bay, beach, botanic gardens, campground, canal, castle, cave, cemetery, coast, farm, fountain, hiking trail, historic & protected site, hot spring, island, lake, memorial archway, memorial site, monument, national park, natural park, nature preserve, park & square, park, plaza, picnic area, pond, regional park, reservoir, river, rock climbing spot, scenery spot, scenic lookout, swamp area, tourist attraction, volcano, waterfront, windmill, zoological gardens
Transportation	airport airfield terminal, airport related, bus depot/ terminal, car tunnel, coach station, filling station, gas station only with/ without small kiosk, gas station with retail store, gas station with service/auto repair, harbor, licensed parking lot, light rail station, marina, MRT/LRT marshalling yard/depot, other energy station, parking lot, petrol station/ kiosk, pier dock bulkhead, port, railway station, ropeway station, subway station, toll gate, trailer park, transport depot, transportation