Mixed Land Use Detection via Vision-Language Multi-modal Learning

Meiliu Wu[1], Qunying Huang[1], and Song Gao[1]

[1]Affiliation not available

December 31, 2022

Mixed Land Use Detection via Vision-Language Multi-modal Learning

Meiliu Wu[1], Qunying Huang[1], Song Gao[1]

[1]Department of Geography, University of Wisconsin-Madison, Madison, WI.

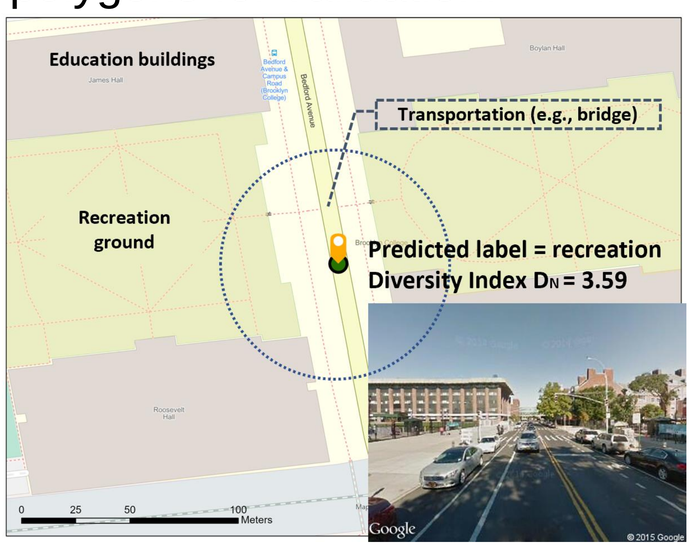
Abstract: While land use classification and mapping based on visual interpretation of aerial images have been extensively studied over decades, such overhead imagery can hardly determine land use(s) accurately in complicated urban areas (e.g., a building with different functionalities). Meanwhile, images taken at the ground level (e.g., street view images) are more fine-grained and informative for mixed land use detection. Considering land use categories are often used to describe urban images, mixed land use detection can be regarded as the Natural Language for Visual Reasoning (NLVR) problem. As such, this study develops a vision-language multimodal learning model with street view images for mixed land use detection, which is based on the contrastive language-image pre-training (CLIP) model and further improved and tailored by two procedures: 1) prompt tuning on CLIP, which not only learns the visual features from street view images, but also integrates land use labels to generate textual features and fuses them with the visual ones; and 2) calculating the Diversity Index (DI) from the fusions of visual and textual features, and using the DI value to estimate the mixed level for each image. Our experiments demonstrate that simply leveraging the street view image itself with tailored prompt engineering is effective for mixed land use detection, reaching the degree of matching from 71% to 84% between the predicted labels and the OpenStreetMap ones. Moreover, a land-use map with mixture information represented as probabilities of different land-use types is produced, paving the way for fine-grained land-use mapping in urban areas with heterogeneous functionalities.

1. Introduction

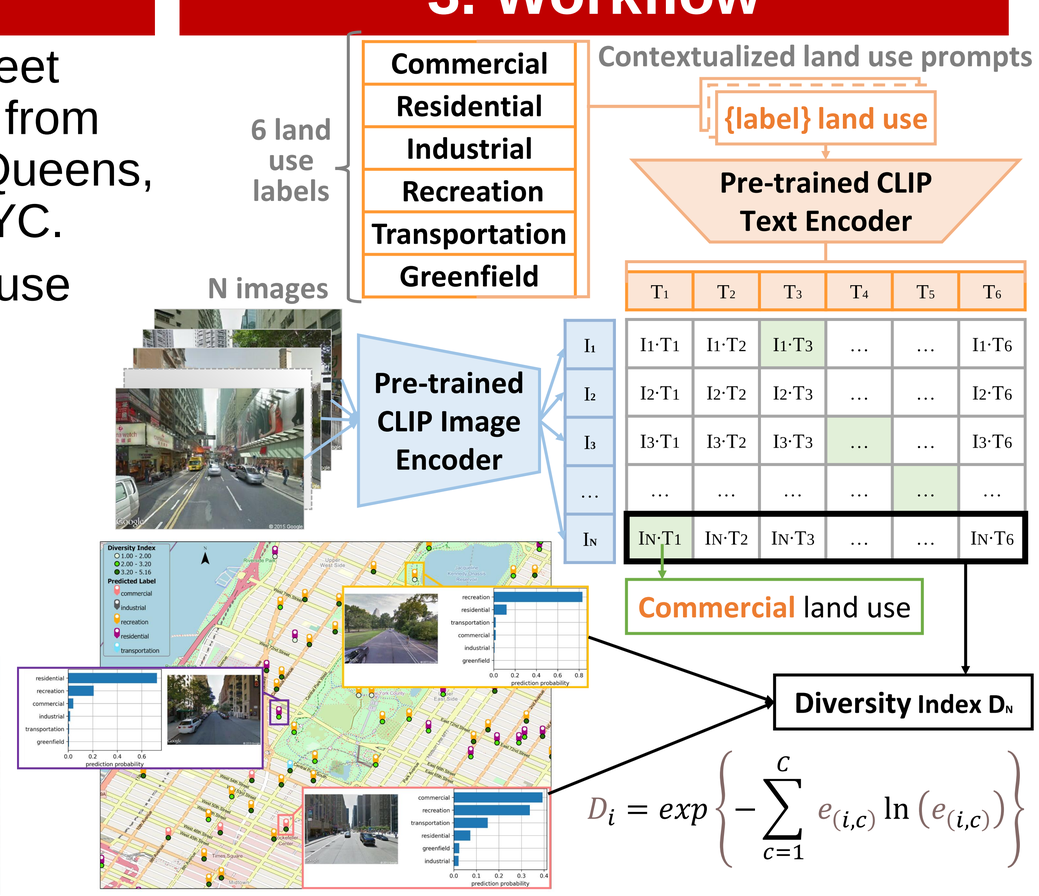
- Traditional measurements and mapping of mixed land uses degree rely on laborious field surveys or visual interpretation of satellite images.
- One common limitation of existing methods is that only a single label is recorded for each ground area.
- This study uses street view images that capture more detailed, representative, and heterogeneous visual characteristics of land uses.
- Vision-language multimodal learning is leveraged, given land-use scenarios are to describe the human use of land.

2. Datasets

- 3,398 geotagged Google Street View images uniformly sampled from four main boroughs (Brooklyn, Queens, Manhattan, and the Bronx) in NYC.
- >308k OpenStreetMap land-use polygons for validation.



3. Workflow



4. Experimental Results & Analysis

#1 Prompt engineering

Contextualized prompts	Degree of matching
Prompt #1: "{label} place"	18.98%
Prompt #2: "{label} area"	35.29%
Prompt #3: "{label} land use"	39.25%
Prompt #4: "for {label}"	53.66%
Prompt #5: "{label} use"	61.43%
Prompt #6: "{label} purpose"	63.34%
no prompt (i.e., "{label}")	64.02%
Prompt ensembling (softmax weighted)	**64.56%**

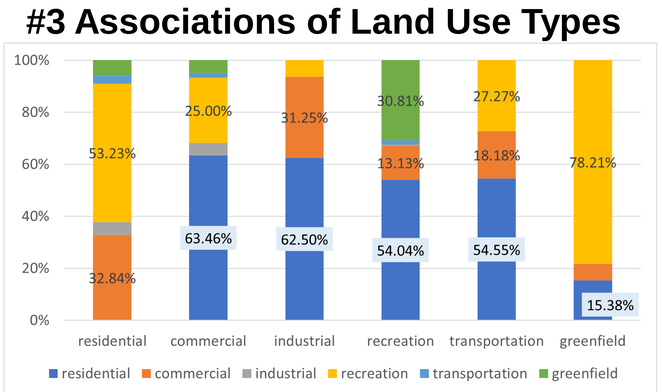
#2 Mixed Land Use Detection

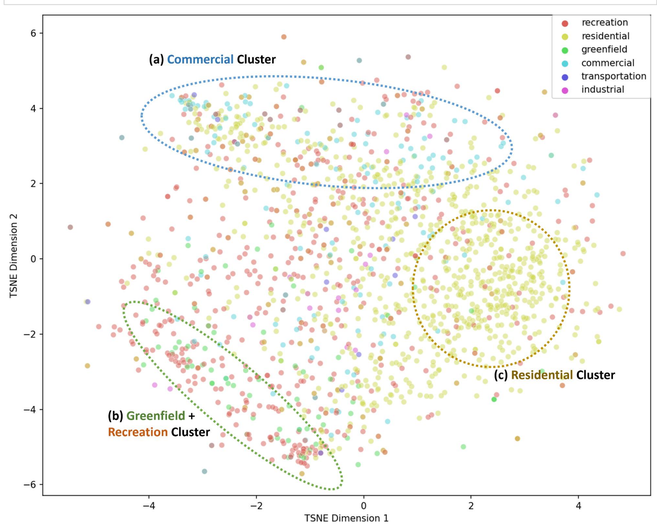
Group	Num.	Degree of matching	Avg. DI
Non-mixed	1,058	62.67%	1.94
Mixed (2-labels)	229	71.18%	2.34
Mixed (3-labels)	25	84.00%	2.55
Overall	1,312	64.56%	2.02

Land Use Type	Num.	Degree of matching	Avg. DI*
Residential	811	94.08%	1.75
Greenfield	114	21.05%	2.22
Recreation	461	38.18%	2.30
Transportation	13	38.46%	2.61
Industrial	20	75.00%	2.79
Commercial	172	40.12%	2.88

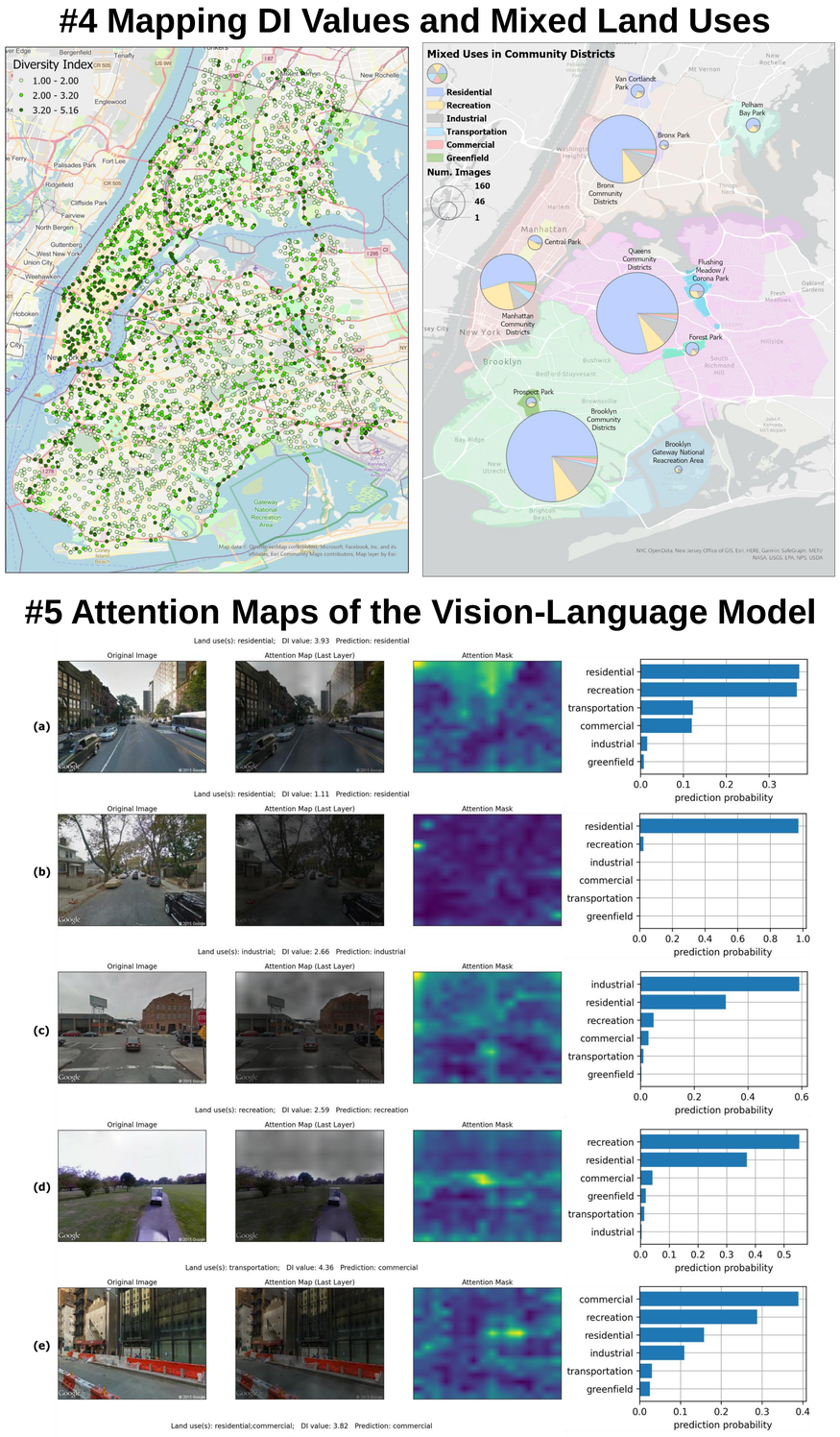
*Sorted by the average DI values.

#3 Associations of Land Use Types





#4 Mapping DI Values and Mixed Land Uses



#5 Attention Maps of the Vision-Language Model

5. Conclusions & Future Work

- Pioneeringly demonstrating the effectiveness of vision-language multimodal learning for mixed land use detection with geo-tagged street view images, capturing multiple functionalities of any ground feature;
- All datasets globally available, and methodology applicable to other cities;
- Providing insights of prompt tuning to contextualize land use labels;
- Mapping mixed land uses at a point level, providing flexibility for different administrative aggregation (e.g., census tracts and zoning districts) if needed.
- Future work: (1) using {land use text, street view image} pairwise datasets to fine-tune the model; and (2) measuring the area proportion of each land use type in each location's buffer zone as another DI evaluation metric.

- Selected References:

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark et al., "Learning transferable visual models from natural language supervision," in International Conference on Machine Learning. PMLR, 2021, pp. 8748–8763.

A. Dubey, N. Naik, D. Parikh, R. Raskar, and C. A. Hidalgo, "Deep learning the city: Quantifying urban perception at a global scale," in European conference on computer vision. Springer, 2016, pp. 196–212.

A. Chao, C.-H. Chiu, and L. Jost, "Phylogenetic diversity measures and their decomposition: a framework based on hill numbers," Biodiversity Conservation and Phylogenetic Systematics, vol. 14, 2016.

- Acknowledgements: This study is funded by National Science Foundation (1940091) and National Institute of Food and Agriculture (WIS04084).