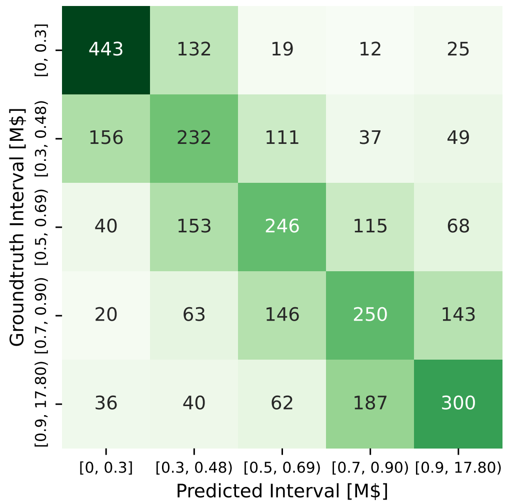


Figure 9. Confusion matrix of property price classification with LGBM [12] *across* scenes.

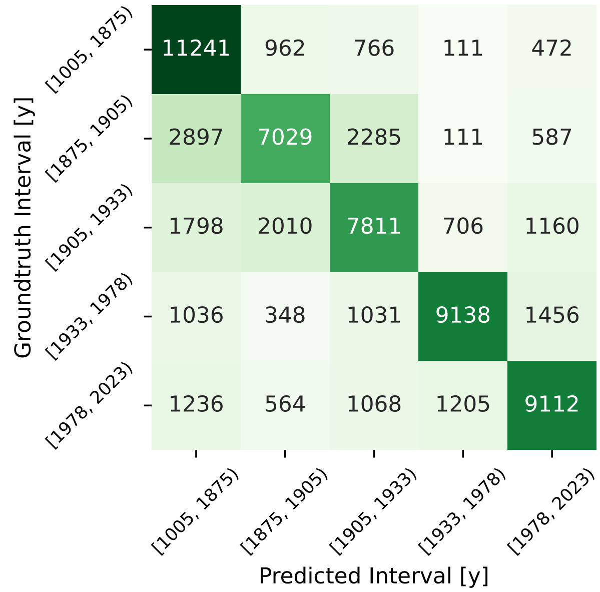


Figure 10. Confusion matrix of building age classification with LGBM [12] *across* scenes.

A. Additional Results

A.1. Evaluation Across and Within Scenes

In the main paper, we presented results for building age prediction *within* cities and property price estimation *across* scenes. The complementary results for building age *across* cities and property price *within* scenes are presented in Tables 1 and 6, featuring additional metrics. Furthermore, confusion matrices are visualized in Figures 9 and 10.

A.2. Evaluation with more Training Data

We find that the results *across* scenes can be significantly boosted when training with more than 30% of the dataset. Figures 11 and 12 visualize this effect.

A.3. Ablation: 3D Point Cloud vs. Flat Grid

Although only evaluating on a 2D grid we find that the usage of a 3D point cloud is beneficial for feature fusing. In table 7 we demonstrate that performance degrades significantly if projecting to a flat 2D point grid instead. We believe that this is caused by the imprecise attribution of points to masks.

B. Implementation Details

B.1. Dataset Creation

We sample positions based on a 2D grid, adding random offsets on all axes. The angle to the z-xis is sampled between 0 and 90 degrees to avoid sky-facing perspectives. The other angles are sampled uniformly at random. RGB-D images with depth closer than 50m and images with infinite depth in more than 20% of the pixels are discarded. See Table 9 for details on the scenes.

B.2. Projection to Point Cloud

The point cloud is first downsampled to 1M points (0.5M if only the coarsest level was processed) to reduce memory consumption. Following OpenMask3D [28], point visibility is determined based on depth.

However, we filter the masks before projection. As most segments only cover a handful of pixels, we retain only those that cover at least 0.25% of the image. This leads to the removal of roughly 60% of all segments and speeds up the overall processing time by 40%.

B.3. Prompting the Point Embeddings

As mentioned in the main paper, we prompt the model with *positive* and *negative* queries. We find that the choice of negatives can have a strong impact on performance. For building segmentation, the full set of negatives was: `'tree'`, `'road'`, `'park'`, `'river'`, `'car'`, `'sea / lake / canal'`, `'parking lot'`, `'urban scene'`, and `'city'`.

B.4. Estimation

We use scikit-learn [18] to build unweighted KNN regressors and classifiers ($k = 5$). Each point and feature level provides a data point. As for LightGBM [12], we use the official package with default settings. We find that classifiers on building age, crime rate, noise levels, and popula-

F1 Score	Overall	Amsterdam	The Hague	Eindhoven	Groningen	Maastricht	Rotterdam	Utrecht
lgbm	**0.67**	**0.54**	**0.47**	**0.81**	**0.75**	**0.60**	**0.76**	**0.59**
linear	0.61	0.52	0.38	0.76	0.66	0.56	0.55	0.53
knn	0.61	0.51	0.43	0.78	0.70	0.54	0.70	0.49
dummy	0.20	0.23	0.21	0.28	0.24	0.21	0.23	0.21
Spearman Correlation								
lgbm	**0.73**	**0.32**	**0.56**	**0.40**	**0.84**	**0.65**	**0.76**	**0.68**
linear	0.67	0.29	0.46	0.32	0.70	0.61	0.57	0.60
knn	0.67	0.25	0.46	0.33	0.77	0.56	0.67	0.52
dummy	0.00	-0.01	0.01	-0.02	0.03	-0.00	-0.01	0.01
MAE [y]								
lgbm	**50.85**	**122.23**	**57.99**	**12.64**	**18.26**	**63.50**	**15.65**	**60.62**
linear	62.84	137.46	88.79	13.09	25.09	82.48	22.31	68.57
knn	55.62	125.30	62.59	14.46	24.12	67.76	18.67	72.12
dummy	102.95	166.55	93.28	77.03	88.49	106.14	75.75	109.80
MAPE [%]								
lgbm	**3.03**	**8.28**	**3.11**	**0.64**	**0.94**	**3.43**	**0.81**	**3.72**
linear	3.63	8.94	4.71	0.66	1.29	4.40	1.15	4.10
knn	3.30	8.53	3.36	0.73	1.23	3.67	0.96	4.31
dummy	5.85	11.10	5.03	3.94	4.51	5.82	3.93	6.33

Table 5. OpenCity3D few-shot results for construction year prediction trained *across* various cities in the Netherlands.

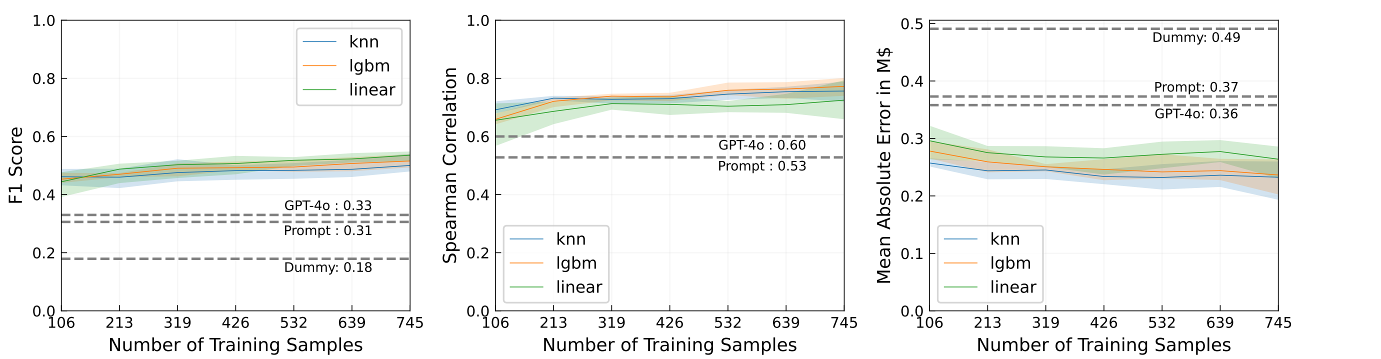


Figure 11. Property price estimation results against dataset size for experiment *across* scenes. Zero-shot MAE baselines were obtained from scores by matching quantiles.

tion density benefit strongly from reducing noise by averaging the embeddings of the relevant area before training and inference.

B.5. Projection of Scores to Ground Truth Scale

We experiment with methods to convert the scores into estimates matching the scale of the ground truth distribution. To that end, we compute the q quantiles of the predicted and the ground truth distribution. Then we assign a prediction in the i-th quantile of the score distribution the mean of the values in the i-th quantile of the true distribution. We implement this strategy with $q = 5$

B.6. GPT-4o Integration

We use GPT-4o to produce one score per prompt and image. The obtained score is then fused into the point cloud analogously to the embeddings. Due to cost and time constraints, we only process full images (coarsest level) and no individual masks. Table 8 shows the used prompts for the GPT experiments (GPT4o). For the property price and

	Mean	Detroit	Miami	San Juan	Boston	San Fran.	Seattle	Los Angeles
F1 Score								
lgbm	**0.34**	0.33	**0.25**	**0.38**	**0.34**	0.34	**0.33**	0.40
linear	0.28	0.30	0.19	0.33	0.29	0.24	0.22	0.38
knn	0.32	**0.34**	0.19	0.36	0.31	**0.35**	0.26	**0.45**
dummy	0.20	0.20	0.20	0.19	0.22	0.21	0.17	0.18
Spearman Correlation								
lgbm	0.49	0.55	0.24	**0.45**	**0.49**	0.57	0.39	0.75
linear	**0.51**	0.55	**0.30**	0.38	0.44	**0.68**	0.43	**0.79**
knn	**0.51**	**0.59**	0.29	0.39	0.41	0.63	**0.46**	0.77
dummy	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MAE [M$]								
lgbm	0.34	0.19	1.03	0.39	**0.14**	0.17	0.32	0.14
linear	0.37	0.21	1.10	0.45	0.16	0.16	0.35	0.13
knn	**0.32**	**0.17**	**0.97**	**0.37**	**0.14**	**0.14**	**0.30**	**0.11**
dummy	0.52	0.28	1.29	0.55	0.20	0.39	0.51	0.39
RMSE [M$]								
lgbm	0.58	0.28	2.20	0.56	**0.17**	0.24	0.42	0.19
linear	0.60	0.31	2.23	0.64	0.21	0.22	0.44	0.17
knn	**0.55**	**0.26**	**2.15**	**0.54**	**0.17**	**0.19**	**0.39**	**0.15**
dummy	0.80	0.38	2.60	0.74	0.25	0.48	0.64	0.50

Table 6. OpenCity3D few-shot results for property price prediction trained *within* various cities in the US. This experiment was conducted using 50% of the samples as training data. The small training set size (down 30 samples) can otherwise lead to overfitting.

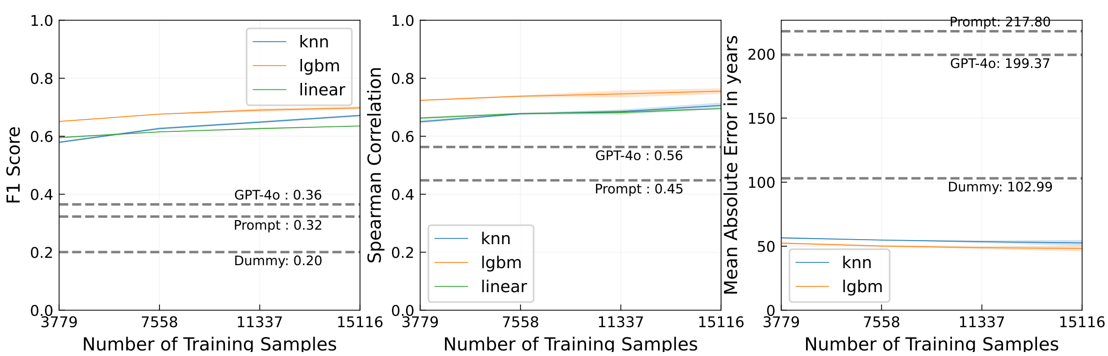


Figure 12. Building age estimation results against dataset size for experiment *across* scenes. Note how quantile matching fails to produce meaningful zero-shot baselines, producing MAE significantly worse than chance.

building age experiments, the rating has been grounded by providing reference values for ratings 3, 6, and 9. These reference values are obtained by binning the ground truth data into 10 bins. Despite this grounding, the resulting scores only match the ground truth distribution to a limited extent. We therefore evaluate them analogously to the similarity scores. The induced prompting cost scales with the number and quality of images as well as the length of the response. Our experiments with 7k to 10k images per scene cost 10-20$ per query. At the time of creation (September 2024), the inference time was roughly at 4-8h per scene.

B.7. Evaluation

Unless stated otherwise, the 3D point cloud is projected to 2D and then interpolated linearly to a regular grid. Correlation is computed on the points (not the districts/buildings). The validation set of the KNN estimators is uniformly randomly downsampled to 20k points per scene to reduce in-

Geometry Type	ROC-AUC [4]	F1 Score
3D Point Cloud		
+ prompt	**0.946**	**0.813**
+ KNN	0.828	0.625
Flat Geometry		
+ prompt	0.904	0.724
+ KNN	0.789	0.591

Table 7. Comparison of building segmentation performance in Groningen with a 3D point cloud vs. using a flat point grid.

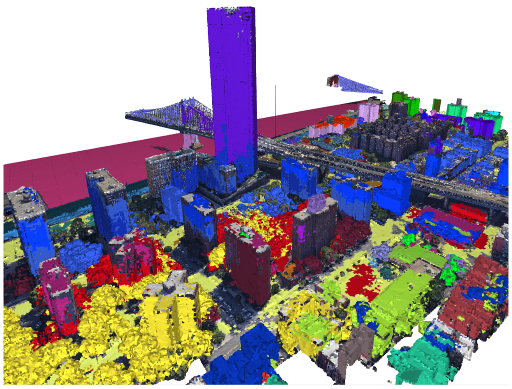


Figure 13. An example segmentation of a city area using Segment3D

ference time. Preliminary experiments showed that this has no significant effect on the results.

C. Experiment: OpenMask3D for Urban Point Clouds

One of the key characteristics of OpenMask3D [28] is that it segments the input point cloud and then stores one feature per 3D segment. This greatly boosts storage and memory efficiency, making it well-suited for city-scale input.

Unfortunately, Mask3D [25], the 3D segmentation model used by OpenMask3D, failed to generate meaningful segments for our 3D city scenes. Neither OpenMask3D's Scannet200 [24] and STPLS3D [9] checkpoints, nor the more recent Segment3D [11] - a model claimed to have superior generalization performances compared to Mask3D - remedied the situation (see Fig. 13).

In particular, we find that the models display high sensitivity to the density and scale of the point clouds.

D. Additional Visualizations

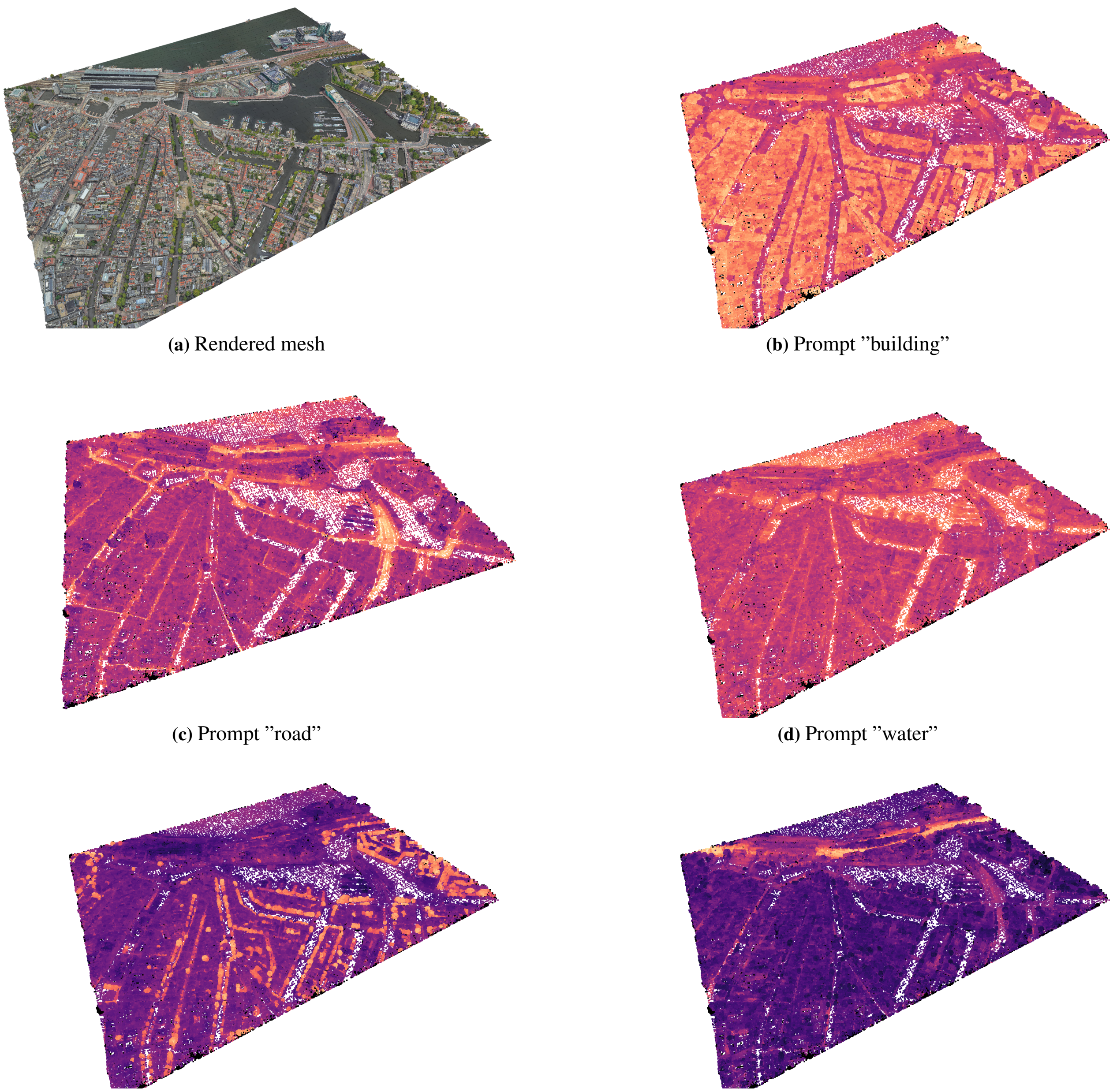
We provide qualitative results for open-set segmentation in Fig. 14. Figures 15 and 16 visualize the complete results for property price prediction, whereas figures 17 and 18 display the ones for building age prediction.

Experiment	**Prompt**
Noise Levels, Population Density and Dangerous Neighborhoods	Estimate the noise level, population density and how dangerous the neighborhood might be of the area shown in this image from 0 to 10. return the result without explanation
Property Prices	Estimate the average property value of the area in the US from a scale from 0 to 10: 3 meaning around 250k$ 6 meaning around 600k$ 9 meaning around 1.5m$ return the result without explanation
Building Age	Estimate the average building age of the area from a scale from 0 to 10: 3 meaning around 1739 6 meaning around 1883 9 meaning around 1987 return the result without explanation

Table 8. GPT4-o experiments and their corresponding prompts.

Scene	Area (km^2)	Latitude Bounds	Longitude Bounds	Sampling Year	Rendered Images
Buenos Aires (Argentina)	5.20	[-58.3801, -58.3593]	[-34.6041, -34.5803]	2021 - 2023	14261
Rotterdam (Netherlands)	1.68	[51.9088, 51.9194]	[4.4542, 4.4741]	2019 - 2023	5704
Amsterdam (Netherlands)	1.99	[52.3698, 52.3809]	[4.8937, 4.9174]	2021 - 2023	6597
The Hague (Netherlands)	1.70	[52.0782, 52.0887]	[4.3073, 4.3285]	2020 - 2023	6520
Utrecht (Netherlands)	1.78	[52.0818, 52.0929]	[5.0987, 5.1197]	2017 - 2019	6527
Eindhoven (Netherlands)	1.35	[5.42727, 5.44250]	[51.43233, 51.44241]	2015 - 2023	8946
Groningen (Netherlands)	1.10	[6.57495, 6.59036]	[53.21107, 53.21964]	2024	7310
Maastricht (Netherlands)	2.20	[5.68648, 5.70744]	[50.8425, 50.8525]	2011 - 2023	12390
San Juan (Puerto Rico)	3.45	[-66.0883, -66.0707]	[18.4475, 18.4642]	2016	9369
Detroit (USA)	4.12	[-83.0038, -82.9789]	[42.3467, 42.3648]	2019 - 2023	9649
Miami Beach (USA)	3.18	[-80.1444, -80.1272]	[25.7664, 25.7831]	2018 - 2022	9377
Seattle (USA)	2.10	[-122.39508, -122.36096]	[47.49694, 47.51248]	2018 - 2023	12834
Boston (USA)	3.83	[-70.99674, -70.96593]	[42.36831, 42.39076]	2018 - 2021	14800
San Francisco (USA)	1.98	[-122.16672, -122.15059]	[37.67978, 37.69241]	2022 - 2023	9822
Los Angeles	2.67	[-117.71718, -117.69846]	[33.61083, 33.62591]	2017 - 2024	7610

Table 9. Scene information. Sampling year indicates the time underlying footage for the reconstruction was taken according to Google Earth [1].



(a) Rendered mesh

(b) Prompt ”building”

(c) Prompt ”road”

(d) Prompt ”water”

(e) Prompt ”tree”

(f) Prompt ”train tracks”

Figure 14. Qualitative results for open-set segmentation in Amsterdam. We can see that buildings 14b, trees 14e and train tracks 14f are recognized with high precision, but the model has difficulties for water 14d and roads 14c

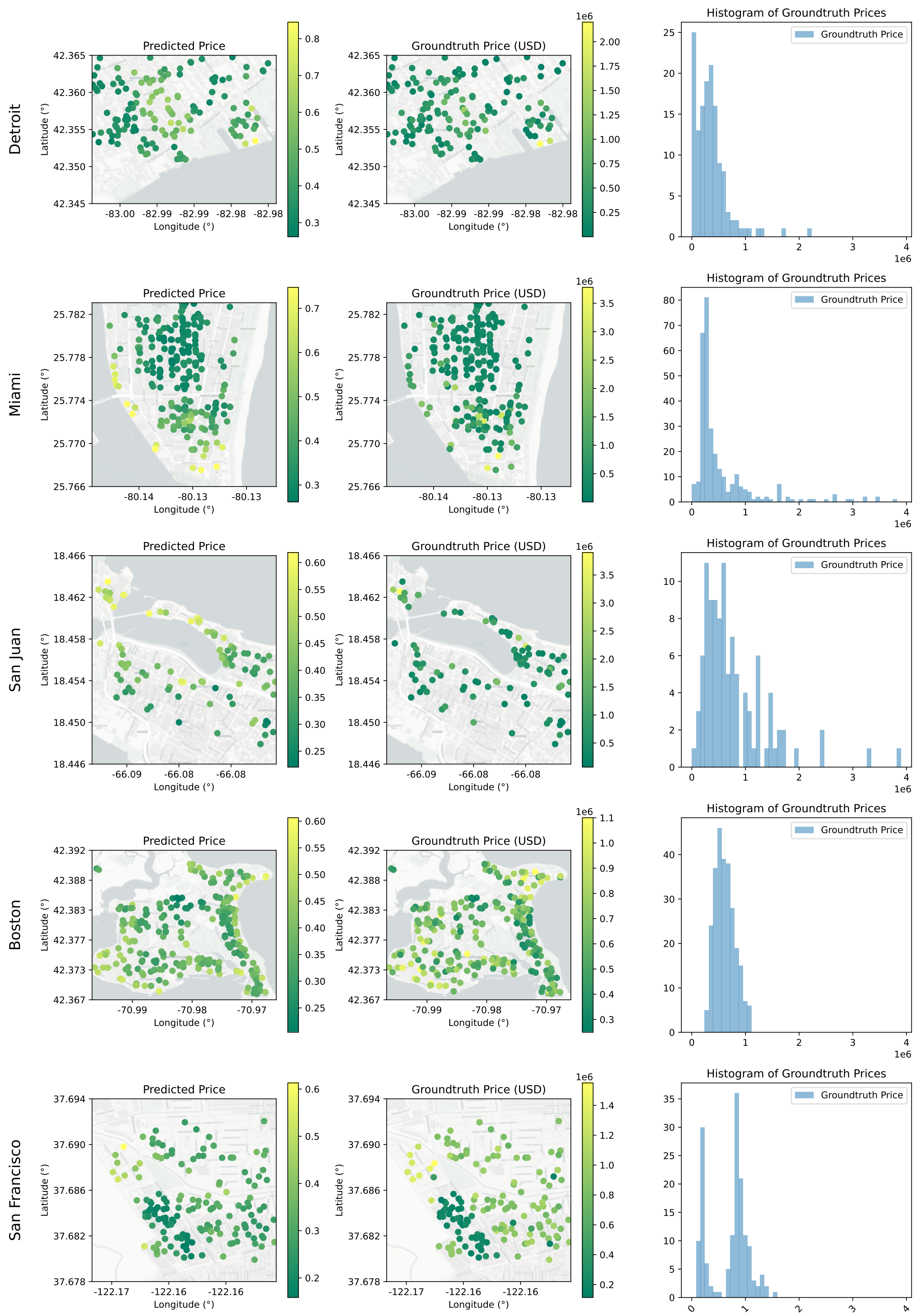


Figure 15. Visualization of zero-shot property price predictions (left) vs ground truth (right) by OpenCity. Basemaps are from CartoDB [8].

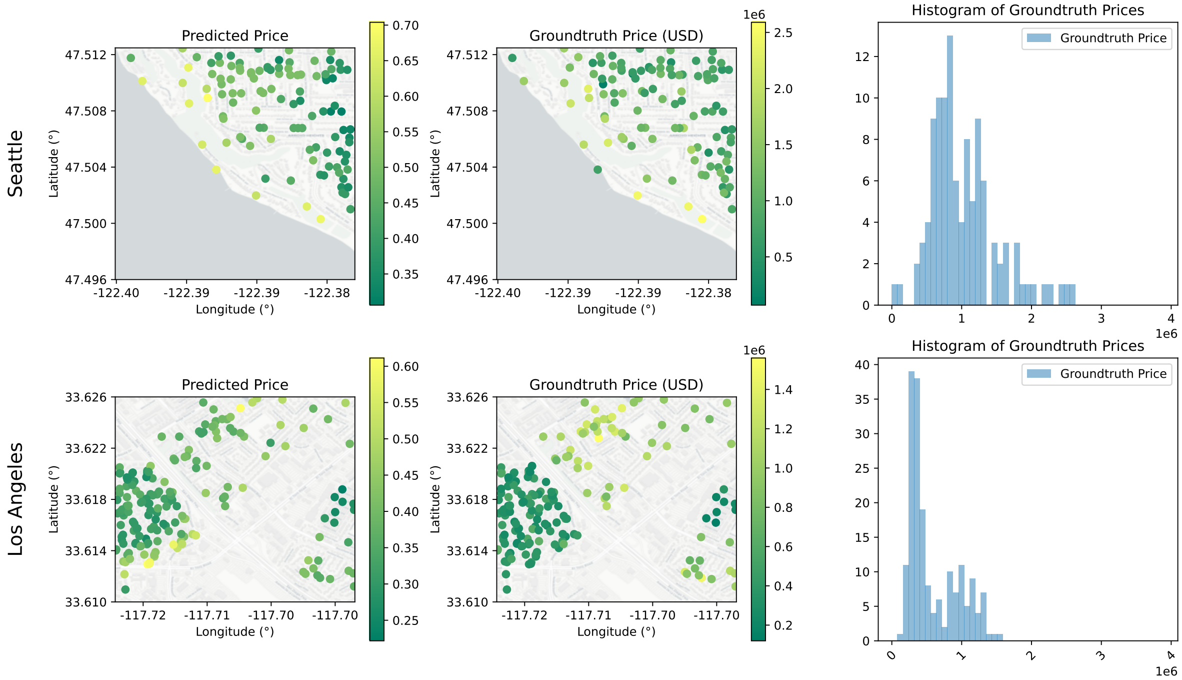


Figure 16. Visualization of zero-shot property price predictions (left) vs ground truth (right) by OpenCity. Basemaps are from CartoDB [8].

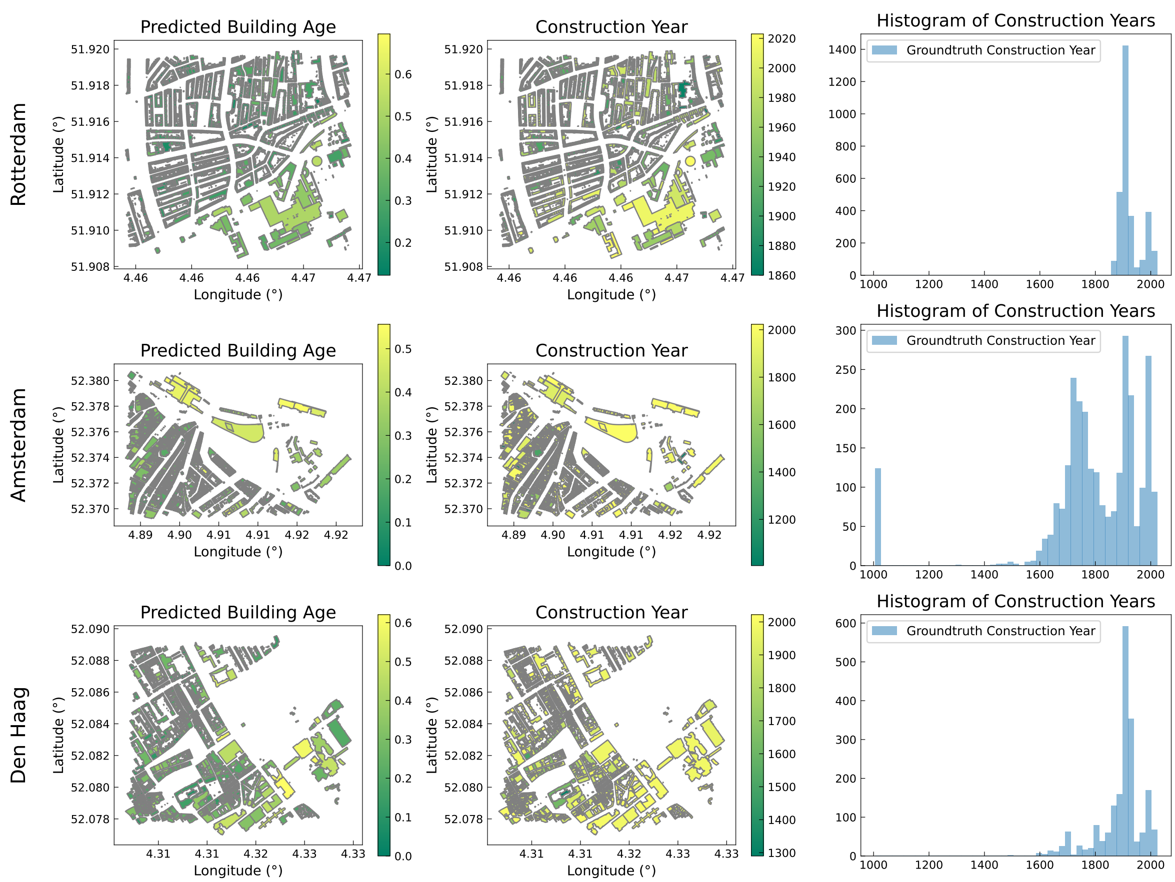


Figure 17. Visualization of zero-shot building age predictions (left) vs ground truth (right) by OpenCity. Basemaps are from CartoDB [8].

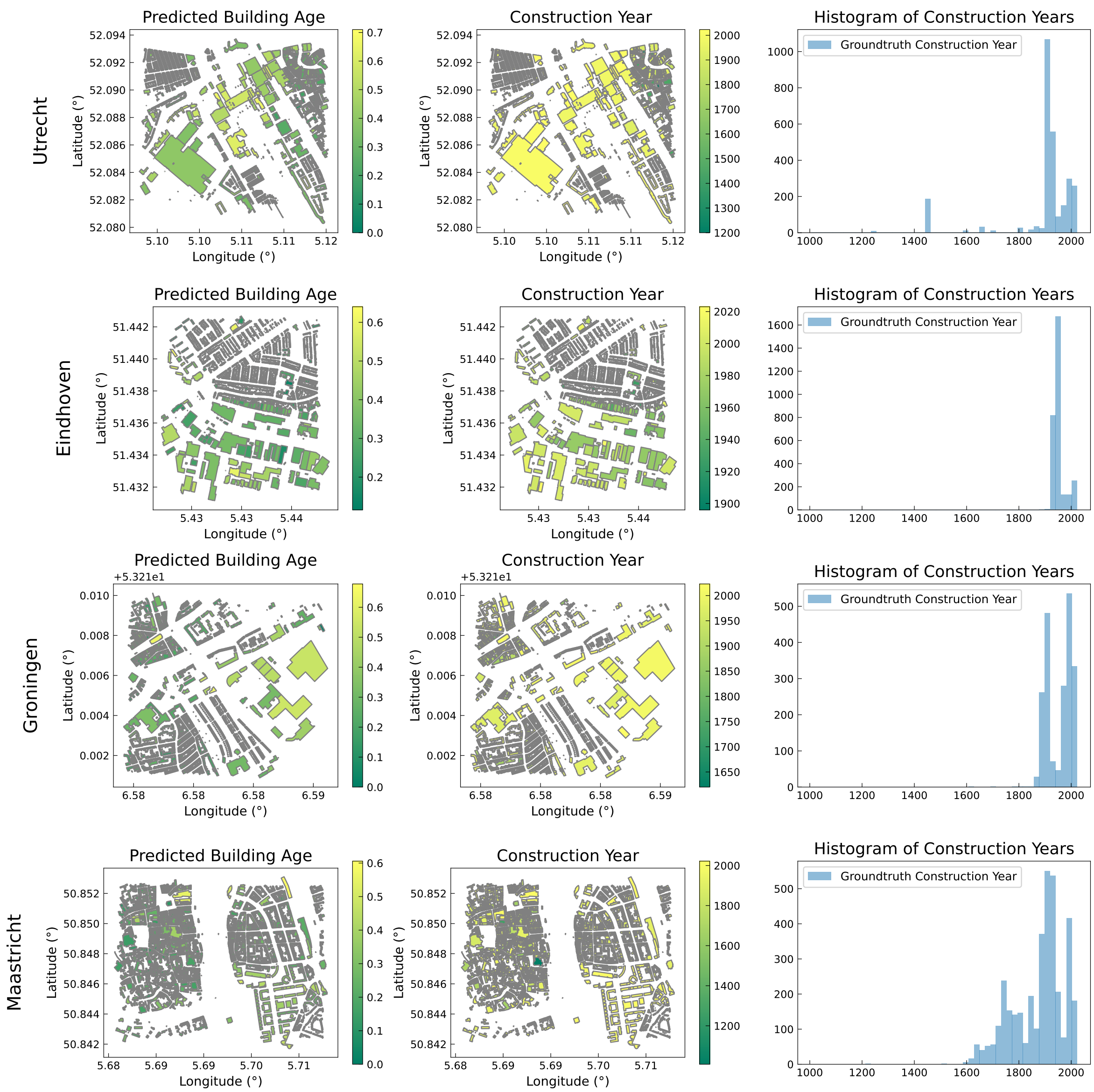


Figure 18. Visualization of zero-shot building age predictions (left) vs ground truth (right) by OpenCity. Basemaps are from CartoDB [8].

References

[1] Google 3d tiles. https://www.google.com/3dtiles/. Accessed: 2024-06-15. 2, 4, 6, 15

[2] Zillow group, inc. https://www.zillow.com/homes/. Accessed: 2024-07-13. 4, 7

[3] Bits and Bricks. Buenos aires population density. https://bitsandbricks.github.io/data/CABA_rc.geojson, 2024. Accessed: 2024-06-15. 6

[4] Andrew P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997. 4, 5, 14

[5] Buenos Aires City Government. Buenos aires government open data portal, 2021. Accessed: 2024-05-18. 6

[6] Buenos Aires City Government. Buenos aires government open data portal, 2021. Accessed: 2024-05-18. 6, 7

[7] Angela Burden. Imputing data for the zestimate. https://www.zillow.com/tech/imputing-data-for-the-zestimate/. Accessed: 2024-06-15. 6

[8] CARTO. Carto basemap styles. Accessed: 2024-07-16. 7, 17, 18, 19

[9] Meida Chen, Qingyong Hu, Thomas Hugues, Andrew Feng, Yu Hou, Kyle McCullough, and Lucio Soibelman. Stpls3d: A large-scale synthetic and real aerial photogrammetry 3d point cloud dataset, 2022. 2, 14

[10] Francis Engelmann, Fabian Manhardt, Michael Niemeyer, Keisuke Tateno, Marc Pollefeys, and Federico Tombari. OpenNerf: Open Set 3D Neural Scene Segmentation with Pixel-Wise Features and Rendered Novel Views. In *ICLR*, 2024. 2

[11] Rui Huang, Songyou Peng, Ayca Takmaz, Federico Tombari, Marc Pollefeys, Shiji Song, Gao Huang, and Francis Engelmann. Segment3d: Learning fine-grained class-agnostic 3d segmentation without manual labels. *ECCV*, 2023. 14

[12] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154, 2017. 3, 4, 11

[13] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. 1

[14] Justin* Kerr, Chung Min* Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *International Conference on Computer Vision (ICCV)*, 2023. 1, 2, 3

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2

[16] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 2

[17] OpenAI. Open ai. hello gpt-4o. https://openai.com/index/hello-gpt-4o. Accessed: 2024-09-9. 4

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 11

[19] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *CVPR*, 2023. 1, 2

[20] Ravi Peters, Balázs Dukai, Stelios Vitalis, Jordi van Liempt, and Jantien Stoter. Automated 3d reconstruction of lod2 and lod1 models for all 10 million buildings of the netherlands, 2022. 4, 7

[21] Angéline Pouget, Lucas Beyer, Emanuele Bugliarello, Xiao Wang, Andreas Peter Steiner, Xiaohua Zhai, and Ibrahim Alabdulmohsin. No filter: Cultural and socioeconomic diversityin contrastive vision-language models. *arXiv preprint arXiv:2405.13777*, 2024. 8

[22] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. *arXiv preprint arXiv:2312.16084*, 2023. 1, 2, 3, 7

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2

[24] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 14

[25] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8216–8223. IEEE, 2023. 2, 14

[26] Ashish Seth, Mayur Hemani, and Chirag Agarwal. Dear: Debiasing vision-language models with additive residuals, 2023. 8

[27] Corinne Stucker, Bingxin Ke, Yuanwen Yue, Shengyu Huang, Iro Armeni, and Konrad Schindler. ImpliCity: City Modeling from Satellite Images with Deep Implicit Occupancy Fields. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2022. 2

[28] Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1, 2, 11, 14

[29] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P. Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis, 2022. 2

[30] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. 1, 2, 3