3D Building Reconstruction from Monocular Remote Sensing Images with Multi-level Supervisions

Supplementary Material

In this supplementary material, we first provide additional details of our proposed MLS-BRN model (Sec. 1). Then we provide additional details of our newly released datasets, as well as the sample diversity used in our study (Sec. 2). Last, we provide additional experimental results of building footprint segmentation, offset angle prediction, and 3D building reconstruction (Sec. 3).

1. Additional details of methods

1.1. Additional training details

In our proposed model, different levels of samples are supervised with different training strategies. Consequently, the ground truth of different levels of samples is utilized differently (Fig. 1). The PBC module employs the building footprint and height ground truth of $\mathcal{X}^H$ to compute the pseudo building bboxes, while the building footprint and height ground truth of $\mathcal{X}^{OH}$ are not used by PBC since their building bbox ground truth is already known. However, PBC uses the off-nadir angle and offset angle ground truth of $\mathcal{X}^{OH}$ for supervising the training of the two angle heads. Furthermore, PBC cannot calculate the pseudo building bboxes for $\mathcal{X}^N$ since they have no building height ground truth. Instead, the pseudo building bboxes of $\mathcal{X}^N$ are calculated by enlarging the building footprint ground truth by a certain percentage.

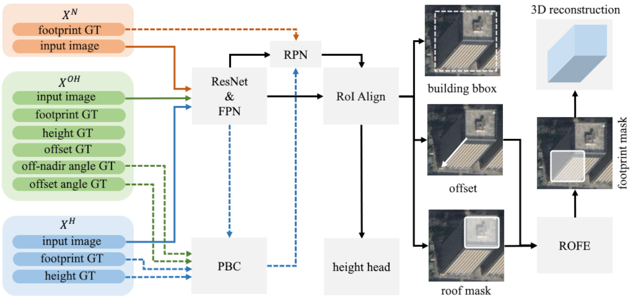

Figure 1. The utilization details of the ground truth of samples with different supervision levels. The green dotted lines indicate the supervision of the off-nadir angle head and offset angle head in PBC using the ground truth provided by $\mathcal{X}^{OH}$. The blue dotted lines denote the calculation of the pseudo building bbox of $\mathcal{X}^H$. The orange dotted line denotes the calculation of the pseudo building bbox of $\mathcal{X}^N$.

1.2. Additional implementation details

In our proposed model, the feature map sent to PBC for calculating the pseudo building bbox is the largest layer from FPN (*i.e.* the layer with the size of 256 × 256). The off-nadir angle head of PBC is composed of 4 Conv layers and 3 FC layers, while the off-nadir angle head of PBC is composed of 8 Conv layers and 6 FC layers.

1.3. Additional details of 3D model reconstruction

We apply the method outlined in [12] to regularize the predicted building footprint mask obtained from our MLS-BRN. Subsequently, we use the Douglas–Peucker algorithm [3] to simplify the regularized polygons by reducing extraneous vertices. Furthermore, the raster polygons are converted to vector data format for visualization. Lastly, the vectorized polygons are combined with the predicted building height to complete the 3D building reconstruction.

2. Additional details of datasets

2.1. Details of existing building datasets

Tab. 1 lists some popular building footprint extraction and 3D reconstruction datasets (with offset or height annotations). The public building footprint extraction datasets far exceed the 3D reconstruction datasets in terms of both the number of images and the number of building instances. Our MLS-BRN demonstrates the great potential of leveraging large-scale footprint segmentation datasets to improve 3D building reconstruction performance and reduce the need for 3D annotations.

Dataset	#Images	#Instances	Off-Nadir	Foot.	Offset	Height
Microsoft [7]	-	1,240M	×	✓	×	×
Open Bld. [9]	-	1,800M	×	✓	×	×
CrowdAI [8]	340K	2,915K	×	✓	×	×
WHU [5]	8.2K	120K	×	✓	×	×
SpaceNet [2]	24.6K	303K	×	✓	×	×
MVOI [11]	60K	127K	✓	✓	×	×
OmniCity [6]	75K	2,573K	✓	✓	×	✓
DFC19 [1]	3.2K	500K	✓	✓	✓	✓
ATL-SN4 [1]	8K	1,100K	✓	✓	✓	✓
BONAI [10]	3.3K	269K	✓	✓	✓	✓
ISPRS 3D [4]	0.033K	-	✓	✓	✓	✓

Table 1. A summary of popular building footprint segmentation and 3D reconstruction datasets. Foot. is the abbreviation for footprint.

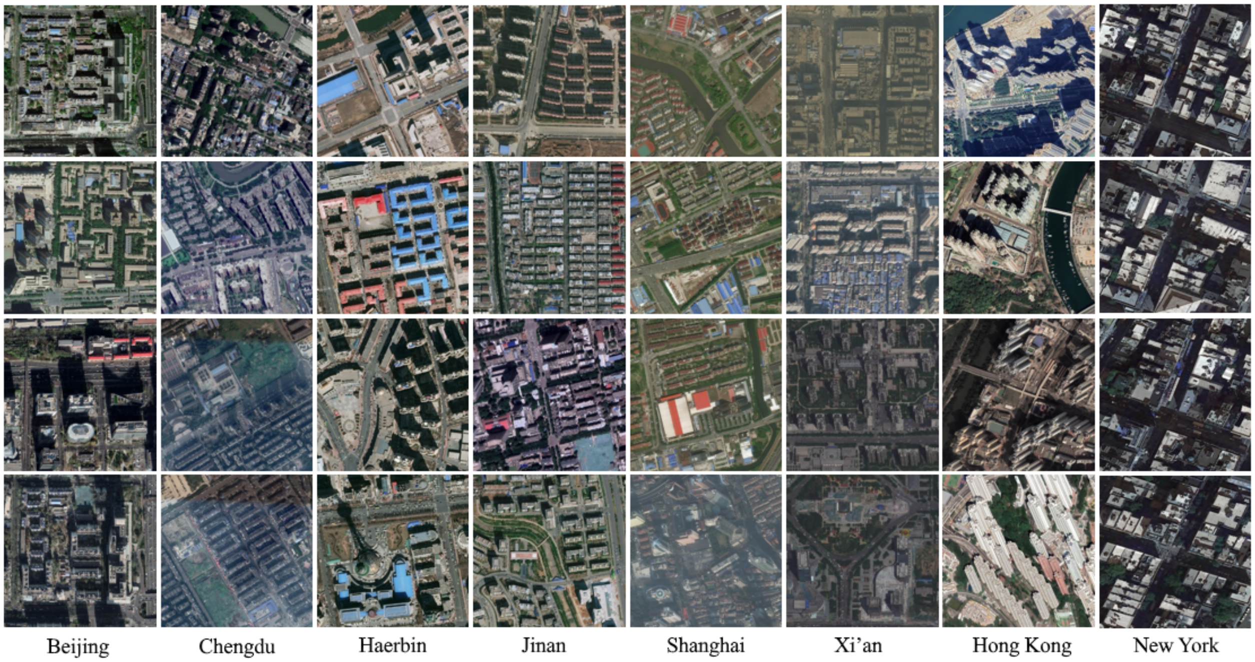

Figure 2. Remote sensing images of 8 cities. The remote sensing images of Beijing, Chengdu, Harbin, Jinan, Shanghai and Xi’an are chosen from the BONAI dataset. The images of New York are chosen from the OmniCity-view3 dataset. The images of Hong Kong are chosen from the HK dataset.

2.2. Details of samples of each city

In Fig. 2, we provide some examples of the remote sensing image samples used in our datasets, which demonstrate a high diversity of each city in terms of the off-nadir angle, offset angle, as well as the building density, areas, height, etc.

2.3. Additional details of newly released dataset

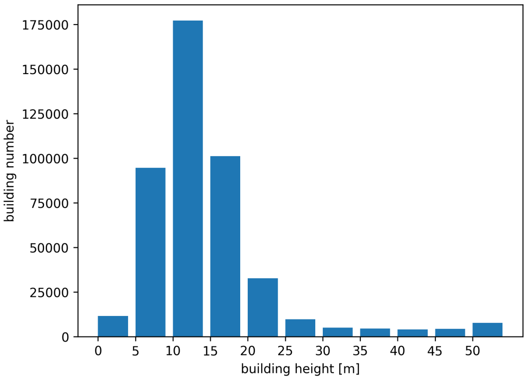

Figure 3. The building height distribution of OmniCity-view3.

In this study, we provide additional offset annotations for the view3 subset of OmniCity (denoted by OmniCity-view3) since this subset contains images with the largest off-nadir angles. Specifically, we annotate roof-to-footprint offsets for 17,092 and 4,929 images from trainval and test sets, respectively. Fig. 3 demonstrates the building height distribution of OmniCity-view3 dataset. We also release a new dataset collected from Hong Kong (denoted by HK dataset), containing 500 remote sensing images for the trainval set and 119 images for the test set, all of which are annotated with building footprint, roof-to-footprint offset, and building height. The remote sensing images are cropped to 1024 × 1024 and contain 24,851 annotated buildings in total. Fig. 4 demonstrates the building height distribution of HK dataset.

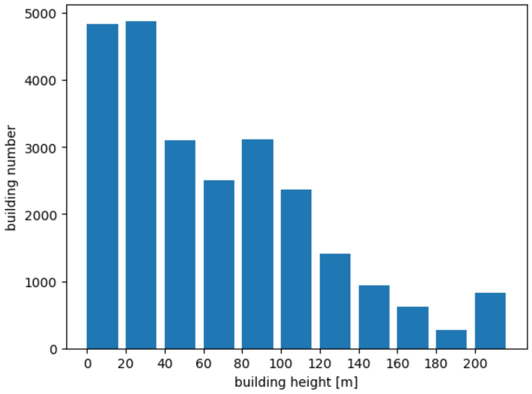

Figure 4. The building height distribution of HK.

3. Additional experimental results

3.1. Ablation study on multi-level sample division

Tab. 2 displays the footprint segmentation and offset prediction performance of our method trained on datasets with different proportions of $\mathcal{X}^{OH}$ and $\mathcal{X}^{H}$ samples. The performance of LOFT-FOA [10] trained only on the $\mathcal{X}^{OH}$ samples are also listed for better demonstrating the performance gains from introducing different percentages of $\mathcal{X}^{H}$ samples. The results show that the building footprint segmentation performance difference between LOFT-FOA [10] and our method is getting smaller with the increase in the proportion of $\mathcal{X}^{OH}$ samples. In the main paper, we opt for the ratio of 30%:70% since the building footprint performance of our method, trained on $BN_{30/70}$, surpasses that of LOFT-FOA [10] trained on BN_{100}.

dataset	Model	F1	Precision	Recall	EPE
BN_{10}	LOFT-FOA	53.91	53.28	54.55	7.42
$BN_{10/90}$	Ours	63.18	65.05	61.42	6.14
BN_{20}	LOFT-FOA	59.65	59.05	60.27	5.79
$BN_{20/80}$	Ours	64.47	67.71	61.52	5.49
BN_{30}	LOFT-FOA	61.35	61.84	61.65	5.70
$BN_{30/70}$	Ours	65.50	66.94	64.11	5.39
BN_{40}	LOFT-FOA	63.17	62.79	63.56	5.26
$BN_{40/60}$	Ours	65.78	66.16	65.40	5.22
BN_{100}	LOFT-FOA	64.31	63.37	65.29	4.94
BN_{100}	Ours	66.36	65.90	66.83	4.76

Table 2. The experimental results of datasets with different proportions of $\mathcal{X}^{OH}$ and $\mathcal{X}^{H}$ samples. As described in the main paper, $BN_{x/y}$ means x% of BONAI trainval samples are of $\mathcal{X}^{OH}$ type and y% are of $\mathcal{X}^{H}$ type. The results of LOFT-FOA and our method trained on BN_{100} are also listed for better comparison with our methods trained on datasets composed of both $\mathcal{X}^{OH}$ and $\mathcal{X}^{H}$ samples.

3.2. Additional results on footprint segmentation

Fig. 5 and Fig. 6 demonstrate the additional building footprint segmentation results of four different cities (*i.e.* Shanghai, Xi'an, New York, and Hong Kong) from different models trained on solely $\mathcal{X}^{OH}$ samples. Fig. 7 display the building footprint segmentation results of two different cities (*i.e.* New York and Hong Kong) from LOFT-FOA [10] and our method trained on datasets containing $\mathcal{X}^{OH}$ and $\mathcal{X}^{H}$ samples.

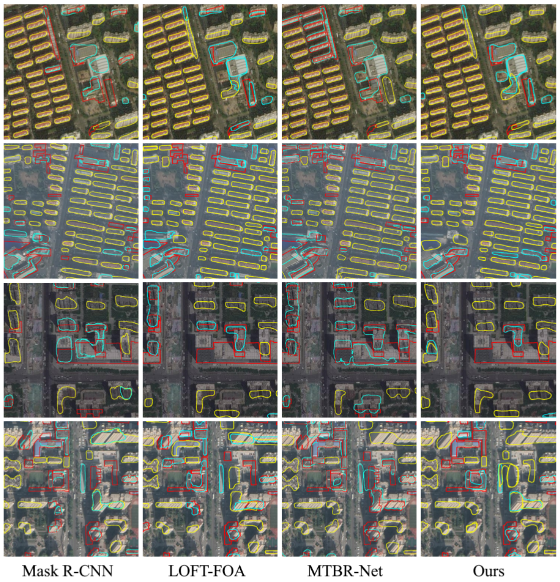

Figure 5. The footprint segmentation results of Shanghai and Xi'an from models trained on BN_{100}. The first two rows display the results of Shanghai, and the last two rows display the results of Xi'an.

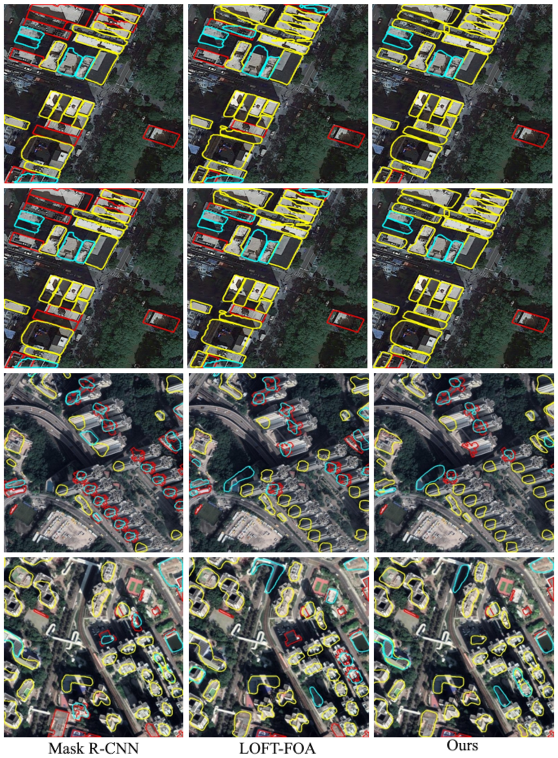

Figure 6. The footprint segmentation results of different models trained on OC_{100} and BH_{100}, respectively. The first two rows display the results of New York (OmniCity-view3), and the last two rows display the results of Hong Kong (HK dataset).

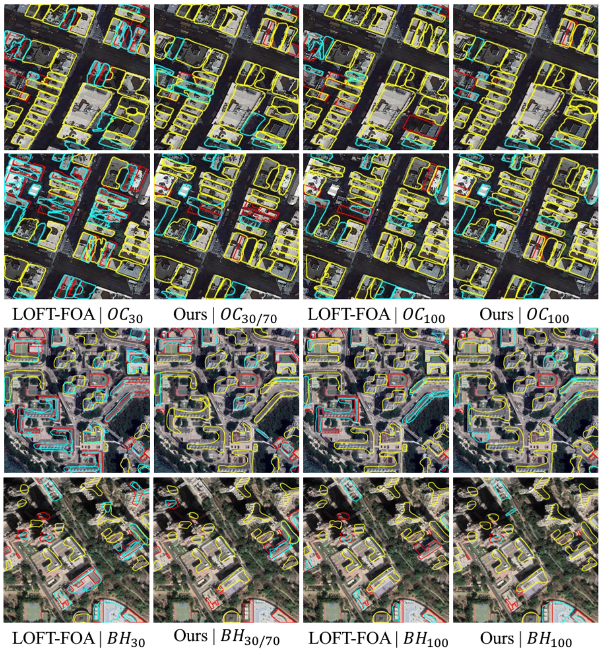

Figure 7. The footprint segmentation results of New York (the first two rows) and Hong Kong (the last two rows) from LOFT-FOA and our method trained on OC_x and BH_x, respectively. Note that LOFT-FOA$|OC_{30}$ means the results of LOFT-FOA trained on OC_{30}.

3.3. Additional offset angle prediction results

Fig. 8 demonstrates the offset angle prediction results of our method. To aid comprehension, a vector is used to represent the offset angle, with the vector direction pointing from the footprint to the roof. For example, a vector pointing horizontally to the right denotes a 0 degree angle, whereas a vector pointing downwards vertically denotes a 90 degree angle.

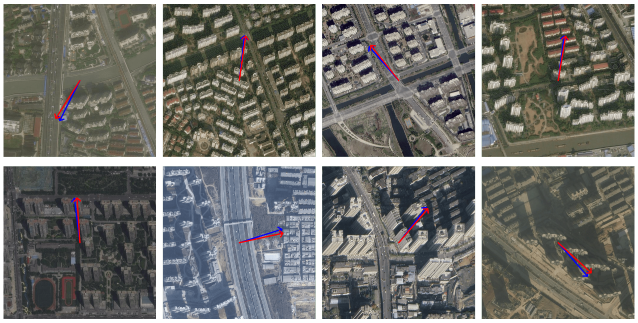

Figure 8. The offset angle prediction results of Shanghai (the first row) and Xi'an (the second row). The red line with the arrow denotes the offset angle ground truth, while the blue line with the arrow denotes the predicted offset angle.

3.4. Failure case analysis

Fig. 9 displays some typical failure cases obtained from our method. The most common failure cases include: (1) the mixing up of the building roof and facade (the first column); (2) inaccurate segmentation of a complex building roof (the second column); and (3) the misinterpretation of multiple roofs as one roof, or vice versa (the third column).

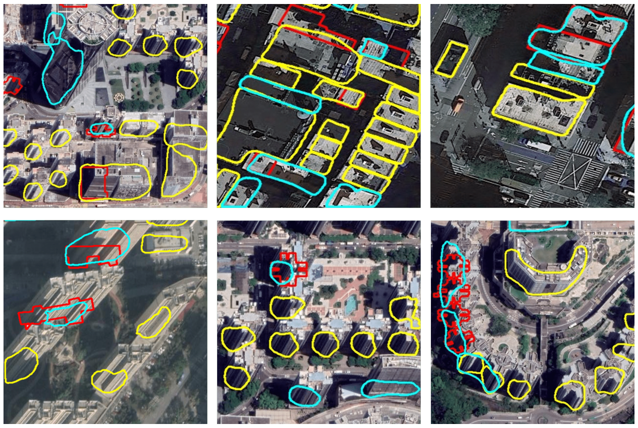

Figure 9. Some typical failures of footprint segmentation results. The yellow, cyan, and red polygons denote the TP, FP, and FN.

3.5. Additional 3D building reconstruction results

Fig. 10 shows additional 3D reconstruction results of four different cities from our method, alongside their corresponding ground truth. Moreover, in order to demonstrate the generalization performance of our method in new regions, Fig. 11 shows the 3D reconstruction results of two new cities, i.e., Shenzhen and Guangzhou. The results indicate that our model has a good generalization performance in terms of 3D building reconstruction tasks.

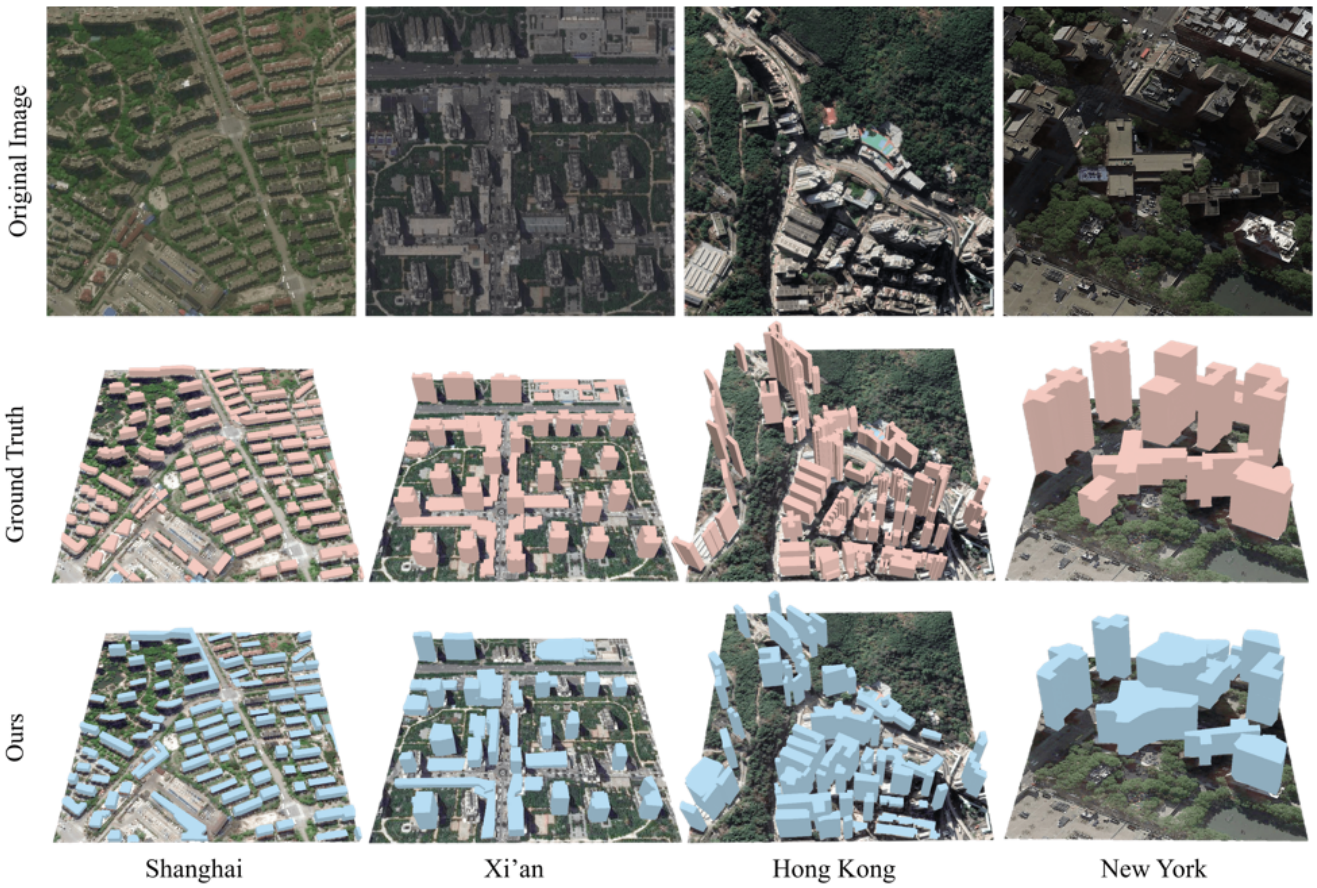

Figure 10. The 3D reconstruction results of Shanghai, Xi’an, Hong Kong, and New York.

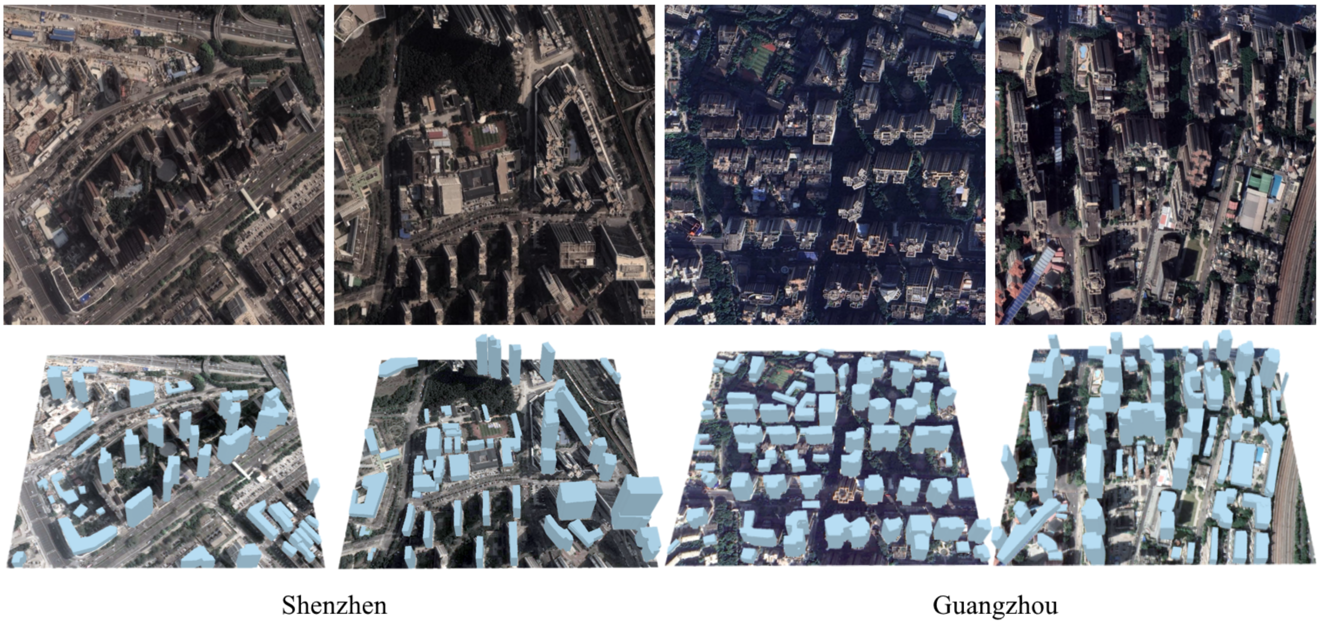

Figure 11. The 3D reconstruction results of Shenzhen and Guangzhou.

References

[1] Gordon Christie, Rodrigo Rene Rai Munoz Abujder, Kevin Foster, Shea Hagstrom, Gregory D Hager, and Myron Z Brown. Learning geocentric object pose in oblique monocular images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14512–14520, 2020. 1

[2] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018. 1

[3] David H Douglas and Thomas K Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: the international journal for geographic information and geovisualization*, 10 (2):112–122, 1973. 1

[4] ISPRS. ISPRS 3D Semantic Labeling Contest. `https://www.isprs.org/education/benchmarks/UrbanSemLab/3d-semantic-labeling.aspx`, 2022. 1

[5] Shunping Ji, Yanyun Shen, Meng Lu, and Yongjun Zhang. Building instance change detection from large-scale aerial images using convolutional neural networks and simulated samples. *Remote Sensing*, 11(11):1343, 2019. 1

[6] Weijia Li, Yawen Lai, Linning Xu, Yuanbo Xiangli, Jinhua Yu, Conghui He, Gui-Song Xia, and Dahua Lin. Omnicity: Omnipotent city understanding with multi-level and multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17397–17407, 2023. 1

[7] Microsoft. Microsoft Global Building Footprints. `https://github.com/microsoft/GlobalMLBuildingFootprints`, 2023. 1

[8] Sharada Prasanna Mohanty. Crowdai dataset: the mapping challenge. https://www.aicrowd.com/challenges/. 2018. 1

[9] Wojciech Sirko, Sergii Kashubin, Marvin Ritter, Abigail Annkah, Yasser Salah Eddine Bouchareb, Yann Dauphin, Daniel Keysers, Maxim Neumann, Moustapha Cisse, and John Quinn. Continental-scale building detection from high resolution satellite imagery. *arXiv preprint arXiv:2107.12283*, 2021. 1

[10] Jinwang Wang, Lingxuan Meng, Weijia Li, Wen Yang, Lei Yu, and Gui-Song Xia. Learning to extract building footprints from off-nadir aerial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1294–1301, 2022. 1, 3

[11] Nicholas Weir, David Lindenbaum, Alexei Bastidas, Adam Van Etten, Sean McPherson, Jacob Shermeyer, Varun Kumar, and Hanlin Tang. Spacenet mvoi: a multi-view overhead imagery dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 992–1001, 2019. 1

[12] Stefano Zorzi, Ksenia Bittner, and Friedrich Fraundorfer. Machine-learned regularization and polygonization of building segmentation masks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3098–3105. IEEE, 2021. 1