The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-1/W5-2025
13th International Conference on Mobile Mapping Technology (MMT 2025)
"Mobile Mapping for Autonomous Systems and Spatial Intelligence", 20–22 June 2025, Xiamen, China

A training-free method for estimating the relative height of buildings

Zhe Chen[1], Chengjie Li[1], Fuxun Liang[1,2], Peiling Tong[1], Chen Long[1], Zhen Dong[1], Bisheng Yang[1]

[1] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China
[2] School of Urban Design, Wuhan University, Wuhan, China

Keywords: 3D urban morphology, Building height, Monocular depth estimation, Graph optimization.

Abstract

Urbanization's vertical shift underscores the need for accurate building height estimation to support sustainable planning. Existing methods, limited by low-resolution data and poor generalization, cannot resolve individual buildings. We propose a training-free approach leveraging the foundation model Depth Anything V2 for relative depth estimation from high-resolution remote sensing (RS) imagery. To address GPU memory constraints, RS images are cropped into overlapping patches, and their depth predictions are unified using a height-weighted graph optimization with Levenberg–Marquardt refinement, prioritizing building-related errors. A viewpoint bias filter, modeled as terrain variation, converts relative depth to height by subtracting a morphologically derived DEM. Experiments on Google satellite imagery with 0.5 m resolution over 20 km² in Wuhan, validated against airborne LiDAR, show an R^2 of 0.73 for building heights, significantly outperforming the state-of-the-art MLSBRN whose R^2 is only 0.25 and which underestimates tall buildings. Without annotation or training, our scalable method accurately estimates individual building heights, generalizes across complex urban morphologies, and provides a robust solution for 3D urban studies.

1. Introduction

Urbanization is increasingly characterized by vertical (3D) development due to limited land resources, shifting from historical horizontal (2D) expansion (Wang et al., 2023). Buildings, as the fundamental components of 3D urban morphology, have become key indicators of urban development and the evolution of spatial structures. For instance, numerous urban factors, such as urban energy consumption and carbon emissions (Chen et al., 2024), exhibit significant correlations with buildings. Limited by data availability, most existing studies on urban expansion have focused on exploring the horizontal expansion of urban areas (Gong et al., 2020). In contrast, relatively few studies have addressed vertical urban expansion, i.e., accurately estimating and characterizing building heights and their dynamic changes.

In recent years, researchers have developed various methods to extract building heights at different scales based on diverse remote sensing data sources, including LiDAR, monocular optical imagery, multi-view optical imagery, SAR, InSAR, and street-view imagery. Generally, the level of detail in building height extraction is proportional to the costs of data acquisition and processing.

At the national scale, the high costs of airborne LiDAR, multi-view stereo imagery, and high-precision SAR limit large-scale, continuous building height estimation. Leveraging freely available, globally distributed data such as Sentinel-1 SAR, Sentinel-2 multispectral, and the ALOS Global Digital Surface Model (AW3D30 DSM), researchers have achieved large-scale height estimation (Li et al., 2020a, Li et al., 2020b, Frantz et al., 2021). Li et al. demonstrated the effectiveness of Sentinel-1 data for height estimation, developing a method based on the novel VVH index to estimate building heights in seven major U.S. cities at 500m resolution (Li et al., 2020b). However, due to multiple echoes from complex spatial layouts and high reflectivity of certain metallic materials, Sentinel-1's backscatter coefficients are unreliable in some areas (Li et al., 2016). Moreover, the estimated heights represent the average of buildings and non-buildings, failing to fully reflect actual building heights. To address unreliable radar signals, Li et al. applied a random forest model with hybrid data to map continental-scale 3D building structures (area, height, and volume) (Li et al., 2020a). However, this method shows estimation biases in areas with small building footprints, and its coarse 1000m resolution cannot capture fine structural differences. Subsequent studies improved height inversion resolution (Frantz et al., 2021, Wu et al., 2023). Frantz et al. combined Sentinel-1 and Sentinel-2 time-series data, using a machine learning regression model to generate Germany's building height map at 10m resolution (Frantz et al., 2021). Wu et al. used a random forest model with all-weather earth observation data (radar, optical, and nighttime light) to estimate China's 2020 building heights at 10m resolution (Wu et al., 2023). Morphological methods have also been applied to process global open DSMs (e.g., AW3D 30). Huang et al. estimated building heights from AW3D30 DSM, introducing a slope correction algorithm to mitigate terrain effects, producing China's height map at 30m resolution and analyzing spatial patterns (Huang et al., 2022). He et al. combined GAIA, WSF 2015, and AW3D 30 data to extract 3D building morphology (He et al., 2023). With advances in deep learning, Yadav et al. proposed the T-SwinUNet network, utilizing Sentinel-1 SAR and Sentinel-2 MSI time-series data to achieve 10m-resolution height mapping with strong generalization (Yadav et al., 2025). These methods enable grid-level height estimation from national to global scales, but low-resolution data limits their ability to accurately reflect individual building heights.

At the urban scale, point clouds from airborne LiDAR directly provide high-precision 3D height information (Bisheng et al., 2017, Park and Guldmann, 2019), but high acquisition costs restrict application scope and update frequency. High-resolution SAR imagery enables height estimation by relating extracted features (e.g., double-bounce lines, layovers, shadows) to building heights (Sun et al., 2022). Additionally,

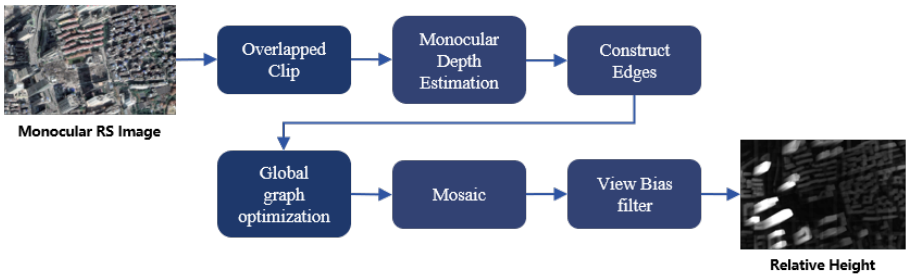


Figure 1. The proposed pipeline.

high-resolution optical imagery estimates heights by analyzing shadow lengths and sun-sensor geometry, though accuracy is limited when buildings are occluded or mixed with other objects, posing challenges in complex urban areas (Liasis and Stavrou, 2016). Multi-view optical imagery mitigates this by providing rich spatial information. For example, Liu et al. generated DSM from ZY-3 imagery using semi-global matching (SGM) and computed normalized DSM (nDSM) via morphological top-hat methods (Liu et al., 2017). Similarly, Wang et al. used SGM and adaptive closing algorithms to estimate building heights in parts of Beijing from GF-7 imagery (Wang et al., 2021). These studies highlight that DSM from traditional stereo matching often suffer from insufficient matching features, complex scenes, and occlusions, leading to height underestimation (Wang et al., 2021). Zhang et al. proposed a constrained stereo matching algorithm to estimate building heights in Yingde and Xi'an from GF-7 imagery, improving high-rise height estimation (Zhang et al., 2022). Cao and Huang used ZY-3 multi-view imagery to estimate heights across 42 Chinese cities, employing a multi-task deep network to regress relationships between ZY-3 imagery and reference heights from Amap (Cao and Huang, 2021). Chen et al. used GF-7 imagery, integrating photogrammetry and deep learning in the BHEPD model to reduce errors and improve reliability across building types (Chen et al., 2023a). These methods are effective in urban scenarios but are limited by high data acquisition costs, hindering broader application. Advances in deep neural networks enable learning geometric features from monocular imagery. Studies have explored deep learning for height estimation from high-resolution orthorectified imagery (Karatsiolis et al., 2021, Li et al., 2021). For instance, Karatsiolis et al. proposed IMG2nDSM to infer object heights from single RGB images, tested on Manchester and DFC 2018 datasets (Karatsiolis et al., 2021). HTC-DC Net combines classification and regression tasks with a ViT encoder to integrate local and global contexts, achieving state-of-the-art performance (Chen et al., 2023b). Additionally, single-building reconstruction from publicly available nadir monocular imagery relaxes data constraints, but data biases in deep learning methods lead to poor generalization for mid- and high-rise buildings, limiting accurate urban 3D morphology depiction (Li et al., 2021, Li et al., 2024). Overall, height estimation from monocular imagery is innovative but remains a complex and challenging problem.

To address these challenges, we propose a novel framework tailored for high-resolution Remote Sensing (RS) imagery that leverages a state-of-the-art monocular relative depth estimation model. Specifically, to manage the computational constraints posed by the large size of RS images, which often exceed available computer memory, we divide the imagery into smaller, overlapping patches that can be processed independently. These patches are processed by a robust foundation model for relative depth estimation to generate depth, inspired by recent advancements in computer vision (Fu et al., 2024, Yang et al., 2024), which excels in precise boundary delineation and generalizes effectively across diverse scenes. Then, we introduce a Height-Weighted Graph Optimization (HWGO) method. This approach unifies the depth predictions from individual patches, accounting for their varying scales and offsets, by modeling the relationships between overlapping regions as a graph and optimizing for consistency across the entire image. Finally, the optimized depth maps are fused to produce a coherent, high-fidelity building height map, enabling accurate and scalable height estimation while preserving fine-grained details. To bridge the gap between relative depth (representing distances to the viewpoint) and absolute building height (defined relative to an orthographic baseline), we design a View Bias Filter (VBF) to achieve the conversion from depth to height. This methodology effectively balances the trade-offs among computational cost, processing scale, and prediction granularity, offering a practical solution for large-scale urban mapping applications.

2. Methods

The overall framework of the method in this paper is shown in Fig. 1, mainly consisting of three parts: Relative Depth Estimation, Height-Weighted Graph Optimization, and Relative Height Extraction with View Bias Filter. Each part is described in detail below.

2.1 Relative Depth Estimation

To estimate the relative depth of large-scale remote sensing (RS) imagery, we employ the Depth Anything V2 (DAv2) (Yang et al., 2024). This advanced depth estimation framework leverages a two-stage training strategy to achieve robust and generalizable performance. Initially, a teacher model is trained using synthetic data, which provides controlled depth annotations to establish a foundational understanding of depth cues. Subsequently, the model is refined using large-scale real-world imagery annotated with pseudo-labels, enhancing its ability to generalize across diverse scenes. This approach ensures high-quality depth predictions for a wide range of visual inputs. However, DAv2 is primarily designed for natural images with relatively small dimensions, typically 518×518 pixels, which are processed efficiently on standard GPU hardware. In contrast, RS imagery often encompasses significantly larger spatial extents, with resolutions reaching thousands of pixels in both width and height. Directly applying the model to such large images is infeasible due to GPU memory constraints. To address this limitation, we adopt a patch-based processing strategy, dividing the input RS imagery into manageable segments while preserving spatial continuity.

To enable depth estimation on large-scale RS imagery, we partition the input image into a set of overlapping patches, denoted as ($X = x_i \mid i = 1, \ldots, n$), where each patch x_i has a fixed size of 1024×1024 pixels. To ensure seamless transitions and mitigate boundary artifacts during subsequent depth fusion, we incorporate a substantial overlap of 75% between adjacent patches. This overlap facilitates the alignment of depth predictions across patches, as overlapping regions provide shared spatial context for subsequent optimization. For each patch (x_i), we apply the Depth Anything V2 model to compute a corresponding depth map (d_i), resulting in a set of depth predictions ($D = d_i \mid i = 1, \ldots, n$). Each depth map (d_i) represents the relative depth of pixels within the patch, defined as the distance from the pixel to the observation viewpoint under a perspective projection model. However, due to variations in scene content across patches (e.g., differences in terrain, vegetation, or man-made structures), the depth predictions exhibit scale and offset inconsistencies. These inconsistencies arise because the Depth Anything V2 model estimates relative depths independently for each patch without a unified global reference.

2.2 Height-Weighted Graph Optimization for Depth Alignment

To address the scale and offset disparities among the depth patches, we propose a Height-Weighted Graph Optimization (HWGO) method to unify the depth predictions into a globally consistent depth map. The optimization process aligns all depth patches to a common reference frame, ensuring that the fused depth map is coherent across the entire RS image.

Selection of a Global Baseline. We begin by randomly selecting one depth patch from the set D, denoted as d_j, to serve as the global baseline. This baseline patch defines the reference scale and offset for the entire depth map. For all other patches $\{d_i \mid i \neq j\}$, we assume a linear transformation to align their depth values to the global baseline. Specifically, the aligned depth for patch i, denoted d_i^{global}, is computed as:

$$d_i^{\text{global}} = s_i \cdot d_i + b_i, \tag{1}$$

where s_i is the scale factor and b_i is the offset for patch i. The goal of the optimization is to determine the optimal values of $\{s_i \mid i \neq j\}$ and $\{b_i \mid i \neq j\}$ that minimize discrepancies in the overlapping regions of the depth patches.

Graph-Based Optimization Framework. We formulate the alignment problem as a graph optimization task, where the nodes represent the scale and offset parameters $\{s_i, b_i \mid i \neq j\}$. The edges correspond to the overlapping regions between pairs of patches, and the error function quantifies the discrepancy between aligned depth values in these regions. For each pair of overlapping patches (i, k), let O_{ik} denote the set of pixels in their overlapping region. The error for a pixel $p \in O_{ik}$ is defined as the squared difference between the aligned depth values:

$$e_{ik}(p) = \left((s_i \cdot d_i(p) + b_i) - (s_k \cdot d_k(p) + b_k)\right)^2. \tag{2}$$

To prioritize regions with significant height variations, which are critical for applications such as building height extraction, we introduce a height-weighted error function. Specifically, we define a threshold $\delta = 10$m, based on empirical analysis, to emphasize errors in regions where the relative depth difference exceeds this threshold. The weight for each pixel (p) is computed as:

$$w(p) = \begin{cases} w_{\text{high}} & \text{if } |d_i(p) - d_k(p)| > \delta \\ w_{\text{low}} & \text{otherwise} \end{cases}, \tag{3}$$

where $w_{\text{high}} > w_{\text{low}}$ ensures that regions with larger depth disparities, typically corresponding to elevated structures, contribute more significantly to the optimization. The total error function for the graph is the weighted sum of squared errors across all overlapping regions:

$$E = \sum_{(i,k)} \sum_{p \in O_{ik}} w(p) \cdot e_{ik}(p), \tag{4}$$

Optimization Using Levenberg-Marquardt. To minimize the error function E, we employ the Levenberg-Marquardt (LM) algorithm, a robust non-linear least-squares optimization method that balances the efficiency of gradient descent with the precision of the Gauss-Newton method. The LM algorithm iteratively adjusts the parameters $\{s_i, b_i \mid i \neq j\}$ to find the optimal solution that minimizes E. Upon convergence, the optimized parameters are used to transform each depth patch d_i into the global reference frame, yielding d_i^{global}.

Depth Fusion. After optimization, the aligned depth patches $\{d_i^{\text{global}} \mid i = 1, \ldots, n\}$ are fused to produce a single, globally consistent relative depth map d. The fusion process averages the depth values in overlapping regions, weighted by their proximity to the patch center to reduce boundary artifacts. The resulting depth map d represents the relative distance from each pixel to the observation viewpoint, adhering to the perspective projection relationship.

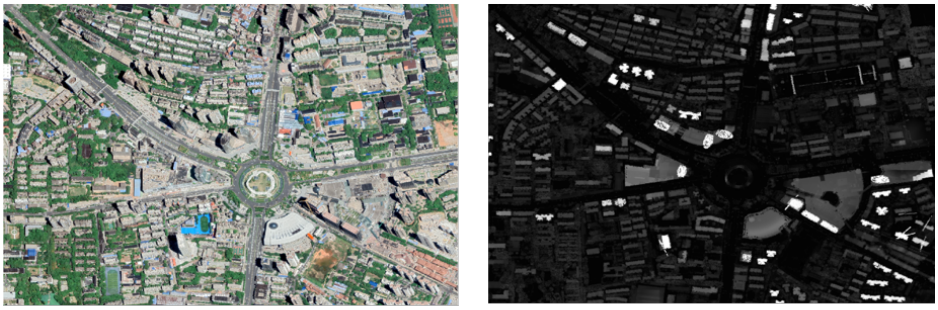
2.3 Relative Height Extraction with View Bias Filter

The relative depth map d provides distances under a perspective projection, which includes the influence of the observation viewpoint. In contrast, the relative height h represents the vertical distance from each pixel to a reference plane, following an orthographic projection model. To extract h we must account for the continuous and gradually varying effect of the viewpoint, which we model as "terrain variation," and extract the viewpoint bias for each pixel using the following pseudocode.

Algorithm 1 View Bias Filter

Require: Relative depth map d, sliding window size s, ground ratio r, image height H, image width W
Ensure: Relative Height h
  for $i = 1, 2, \ldots, H$ **do**
    for $j = 1, 2, \ldots, W$ **do**
      $window = D(i - s : i + s, j - s : j + s)$
      $window = \text{Sort}(\text{Flatten}(window))[0 : r * s^2]$
      $bias(i, j) = \text{Mean}(window)$
      $h(i, j) = d(i, j) - bias(i, j)$
    end for
  end for

View Bias Modeling. We treat the view bias as a smooth, low-frequency component of the depth map, representing the underlying topography of the scene. To isolate this component, we design a morphological method to extract the view bias, which approximates the terrain surface. The morphological approach applies a series of dilation and erosion operations to smooth the depth map while preserving large-scale topographic features.



(a) GES image (a) nDSM generated by ALS

Figure 2. The study area.

Height Computation. The view biash s is subtracted from the relative depth map d to obtain the relative height map h. This operation removes the terrain variation, leaving the vertical distances of objects (e.g., buildings, trees) relative to the reference plane. The resulting height map h is suitable for applications such as urban modeling, infrastructure monitoring, and environmental analysis.

3. Experiments and discussion

3.1 Data Collection and Metrics

Data Collection. To validate the effectiveness of our proposed method, we conducted experiments in a 2.58 km² area within the Optics Valley of Wuhan, Hubei Province, as shown in Fig. 2a. We downloaded the satellite imagery with a 0.5-meter resolution for this area from the Google Earth platform (https://earth.google.com/), which offers several advantages, including fast updates, high resolution, and wide coverage, making it highly suitable for dynamic, high-precision, large-scale applications. Additionally, we used airborne LiDAR data collected in 2020 as the validation ground truth. The point cloud was processed using ENVI LiDAR software to generate nDSM data (Fig. 2b).

Metics. Relative Root Mean Square Error (RRMSE) is a widely used metric for evaluating the accuracy of predictive models, particularly for quantifying relative errors between predicted and actual values. RRMSE is calculated as the ratio of the Root Mean Square Error (RMSE) to the mean of the actual values, providing a normalized measure of error that facilitates comparison across datasets with different scales. Its formula is defined as:

$$RRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\bar{y}}, \tag{5}$$

where y_i denotes the actual value, $\hat{y}_i$ the predicted value, n the number of samples, and $\bar{y}$ the mean of the actual values. Additionally, We employ R^2 to evaluate the goodness-of-fit of a predictive model. It quantifies the proportion of the variance in the dependent variable that is explained by the model, providing insight into its predictive capability. R^2 is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}i)^2}{\sum i = 1^n (y_i - \bar{y})^2}, \tag{6}$$

3.2 Implementation Details

When using Google satellite imagery, the data is typically off-nadir, resulting in a displacement of the estimated roof height relative to the building footprint position. To restore accurate relationships between relative building heights, we utilized MLSBRN (Li et al., 2024) to extract building instance roofs and their offsets relative to the footprints. The height of each individual building instance is determined by the 90th percentile of the relative height within the predicted roof instance mask, and this mask is then adjusted by the offset to accurately locate the building footprint. Furthermore, few studies can achieve height estimation at the individual building level. MLS-BRN can also predict the height of each building mask and represents the current state-of-the-art level, making it a suitable comparison method (Li et al., 2024).

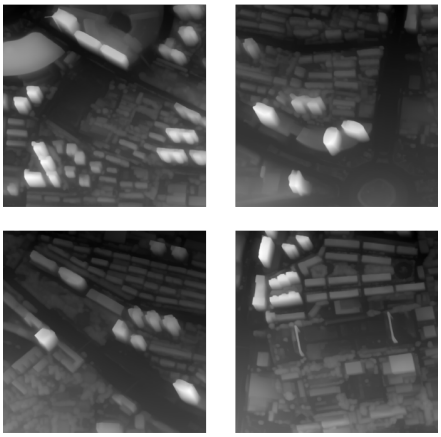


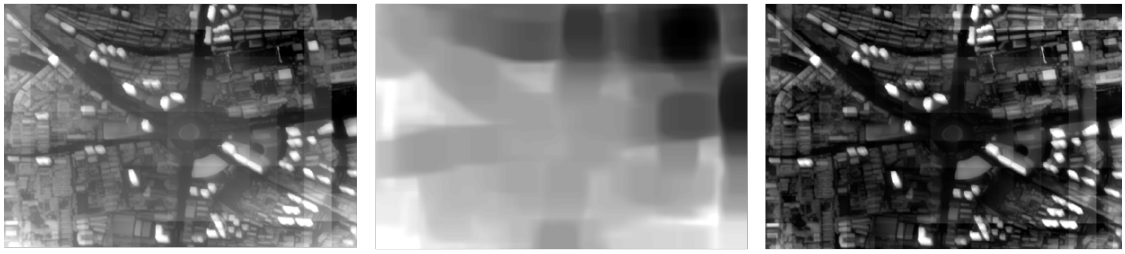
Figure 3. The examples predicted by DPv2.

Table 1. Quantitative Comparison for the proposed method and MLS-BRN.

Metric	R^2	RRMSE
MLS-BRN	0.29	0.53
Ours	0.77	0.25

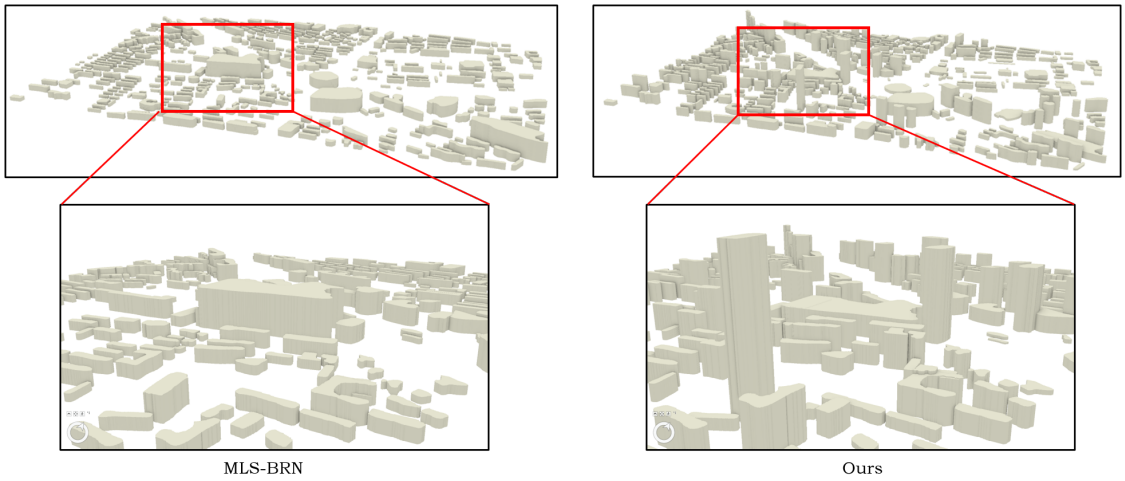
3.3 Results

The depth prediction examples for each patch using our proposed method are shown in Fig. 3. Leveraging the strong generalization capability of DAv2, it demonstrates superior depth prediction performance on remote sensing imagery, capable of obtaining pixel-level relative heights at the same resolution.



(a) Relative heigh (w/o VBF) (b) View bias (c) Relative heigh (w/ VBF)

Figure 4. The visualization result of the View Bias Filter.



MLS-BRN Ours

Figure 5. The visualization result of 3D building model using different methods.

Due to the high resolution, building boundaries are clearly defined, enabling height estimation at the individual building level. After depth alignment through HWGO, all image patches can be fused into a continuous whole, as shown in 3a. However, a height reference bias due to the distance from the estimated viewpoint is clearly observable. Our proposed View Bias Filter treats this viewpoint-induced bias as terrain undulation (Fig. 3b) and effectively eliminates its impact (Fig. 3c).

The quantitative comparison results between our proposed method and MLS-BRN are detailed in Table 1, highlighting significant performance differences. Our method achieves an R^2 of 0.77 and an RRMSE of 0.25, demonstrating high accuracy and reliability in estimating individual building heights across diverse urban settings. In contrast, MLS-BRN performs poorly, with an R^2 of 0.29 and an RRMSE of 0.53, reflecting its limited capability in this task.

Few studies currently excel at precise height estimation for individual buildings, especially in complex urban environments with varied architectural styles. MLS-BRN's reliance on supervised training is a key drawback, as its training dataset lacks sufficient high-rise building samples. This scarcity causes MLS-BRN to consistently underestimate the heights of taller structures while also struggling to capture the intricate 3D morphology of urban landscapes, resulting in poor generalization across different scenarios, as shown in Fig. 5. On the other hand, our method requires no data annotation, making it more practical and scalable. By utilizing a high-capacity large-scale relative depth estimation model, our approach effectively represents buildings of varying heights—from low-rise to high-rise—with exceptional accuracy and robust generalization, making it a promising tool for applications like urban planning and 3D city modeling.

4. Conclusion

Due to limited land resources, urbanization is rapidly shifting towards vertical development, highlighting the urgent need for accurate and scalable building height estimation to support sustainable urban planning and 3D urban morphology analysis. Traditional methods, constrained by low-resolution data and limited generalization across diverse architectural styles, often fail to provide precise height estimates at the individual building level. To address these challenges, we have developed a novel training-free method for relative height estimation from high-resolution remote sensing (RS) imagery. By segmenting large RS images into overlapping patches to overcome computational memory limitations, our method processes each patch independently using a robust foundational model that excels in boundary delineation and scene generalization. By integrating a Height-Weighted Graph Optimization (HWGO) approach, we ensure consistent depth alignment across patches, while an innovative View Bias Filter (VBF) effectively converts relative depth into accurate relative height, eliminating distortions caused by viewing angles. This framework eliminates the need for large annotated datasets, offering a practical and efficient solution. Our method achieves state-of-the-art performance with an R^2 of 0.77 and an RRMSE of 0.25. Relying on widely available RS imagery, combined with its strong generalization capability and high-resolution output, our approach presents a promising tool for urban planning, 3D city modeling, and environmental monitoring. However, the current method only re-

covers relative building heights. Future work could focus on enhancing the method's adaptability to varying imaging conditions and integrating additional data sources to obtain absolute building heights.

4.1 Acknowledgements

This work was supported by the National Key Research and Development Program of China under Grant 2023YFF0725200.

References

Bisheng, Y., Fuxun, L., Ronggang, H., 2017. Progress, challenges and perspectives of 3D LiDAR point cloud processing. *Acta Geodaetica et Cartographica Sinica*, 46(10), 1509.

Cao, Y., Huang, X., 2021. A deep learning method for building height estimation using high-resolution multi-view imagery over urban areas: A case study of 42 Chinese cities. *Remote Sensing of Environment*, 264, 112590.

Chen, P., Huang, H., Liu, J., Wang, J., Liu, C., Zhang, N., Su, M., Zhang, D., 2023a. Leveraging Chinese GaoFen-7 imagery for high-resolution building height estimation in multiple cities. *Remote Sensing of Environment*, 298, 113802.

Chen, S., Shi, Y., Xiong, Z., Zhu, X. X., 2023b. Htc-dc net: Monocular height estimation from single remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–18.

Chen, Z., Yang, B., Zhu, R., Dong, Z., 2024. City-scale solar PV potential estimation on 3D buildings using multi-source RS data: A case study in Wuhan, China. *Applied Energy*, 359, 122720.

Frantz, D., Schug, F., Okujeni, A., Navacchi, C., Wagner, W., van der Linden, S., Hostert, P., 2021. National-scale mapping of building height using Sentinel-1 and Sentinel-2 time series. *Remote Sensing of Environment*, 252, 112128.

Fu, X., Yin, W., Hu, M., Wang, K., Ma, Y., Tan, P., Shen, S., Lin, D., Long, X., 2024. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. *European Conference on Computer Vision*, Springer, 241–258.

Gong, P., Li, X., Wang, J., Bai, Y., Chen, B., Hu, T., Liu, X., Xu, B., Yang, J., Zhang, W. et al., 2020. Annual maps of global artificial impervious area (GAIA) between 1985 and 2018. *Remote Sensing of Environment*, 236, 111510.

He, T., Wang, K., Xiao, W., Xu, S., Li, M., Yang, R., Yue, W., 2023. Global 30 meters spatiotemporal 3D urban expansion dataset from 1990 to 2010. *Scientific data*, 10(1), 321.

Huang, H., Chen, P., Xu, X., Liu, C., Wang, J., Liu, C., Clinton, N., Gong, P., 2022. Estimating building height in China from ALOS AW3D30. *ISPRS Journal of Photogrammetry and Remote Sensing*, 185, 146–157.

Karatsiolis, S., Kamilaris, A., Cole, I., 2021. Img2ndsm: Height estimation from single airborne rgb images with deep learning. *Remote Sensing*, 13(12), 2417.

Li, H., Li, Q., Wu, G., Chen, J., Liang, S., 2016. The impacts of building orientation on polarimetric orientation angle estimation and model-based decomposition for multilook polarimetric SAR data in Urban areas. *IEEE Transactions on Geoscience and Remote Sensing*, 54(9), 5520–5532.

Li, M., Koks, E., Taubenböck, H., van Vliet, J., 2020a. Continental-scale mapping and analysis of 3D building structure. *Remote Sensing of Environment*, 245, 111859.

Li, W., Meng, L., Wang, J., He, C., Xia, G.-S., Lin, D., 2021. 3d building reconstruction from monocular remote sensing images. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12548–12557.

Li, W., Yang, H., Hu, Z., Zheng, J., Xia, G.-S., He, C., 2024. 3d building reconstruction from monocular remote sensing images with multi-level supervisions. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27728–27737.

Li, X., Zhou, Y., Gong, P., Seto, K. C., Clinton, N., 2020b. Developing a method to estimate building height from Sentinel-1 data. *Remote Sensing of Environment*, 240, 111705.

Liasis, G., Stavrou, S., 2016. Satellite images analysis for shadow detection and building height estimation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 119, 437–450.

Liu, C., Huang, X., Wen, D., Chen, H., Gong, J., 2017. Assessing the quality of building height extraction from ZiYuan-3 multi-view imagery. *Remote Sensing Letters*, 8(9), 907–916.

Park, Y., Guldmann, J.-M., 2019. Creating 3D city models with building footprints and LIDAR point cloud classification: A machine learning approach. *Computers, environment and urban systems*, 75, 76–89.

Sun, Y., Mou, L., Wang, Y., Montazeri, S., Zhu, X. X., 2022. Large-scale building height retrieval from single SAR imagery based on bounding box regression networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184, 79–95.

Wang, J., Hu, X., Meng, Q., Zhang, L., Wang, C., Liu, X., Zhao, M., 2021. Developing a method to extract building 3D information from GF-7 data. *Remote Sensing*, 13(22), 4532.

Wang, Y., Li, X., Yin, P., Yu, G., Cao, W., Liu, J., Pei, L., Hu, T., Zhou, Y., Liu, X. et al., 2023. Characterizing annual dynamics of urban form at the horizontal and vertical dimensions using long-term Landsat time series data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 203, 199–210.

Wu, W.-B., Ma, J., Banzhaf, E., Meadows, M. E., Yu, Z.-W., Guo, F.-X., Sengupta, D., Cai, X.-X., Zhao, B., 2023. A first Chinese building height estimate at 10 m resolution (CNBH-10 m) using multi-source earth observations and machine learning. *Remote Sensing of Environment*, 291, 113578.

Yadav, R., Nascetti, A., Ban, Y., 2025. How high are we? Large-scale building height estimation at 10 m using Sentinel-1 SAR and Sentinel-2 MSI time series. *Remote Sensing of Environment*, 318, 114556.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H., 2024. Depth anything v2. *Advances in Neural Information Processing Systems*, 37, 21875–21911.

Zhang, C., Cui, Y., Zhu, Z., Jiang, S., Jiang, W., 2022. Building height extraction from GF-7 satellite images based on roof contour constrained stereo matching. *Remote sensing*, 14(7), 1566.