BldgWeaver: An appearance-contingent generation solution with a three-dimensional automated creation of building digital cousins model using pre-trained transformer architecture

Lingfeng Liao[1], Chenbo Zhao[1], Yoshihide Sekimoto[1], Yoshiki Ogawa[1]

[1] Center for Spatial Information Science, The University of Tokyo - (lfliao, zhaocb, sekimoto, ogawa)@csis.u-tokyo.ac.jp

Keywords: building digital cousins, urban digital twins, 3D reconstruction, generative AI.

Abstract

This paper introduces BldgWeaver, a novel adaptive generative model for creating 3D building digital cousin (BDC) models using pre-trained Transformer architecture. Unlike traditional approaches that require complete 3D reconstruction with extensive visual data, BldgWeaver approximates building geometries using artificial intelligence-generated content to address data deficiencies in urban digital twin development. The proposed method employs a token-based approach to convert triangle mesh coordinates into discrete tokens for auto-regressive prediction, incorporating parallel conditional controls and an optimized footprint-masked training strategy. Experiments conducted on the PLATEAU dataset demonstrate our model's capability to generate Level of Detail 2 (LoD2) building models with diverse roof structures, achieving an average 49% improvement in geometric proximity compared to basic LoD1 representations. The proposed model effectively addresses challenges in wide-range urban mapping by reducing data dependencies while maintaining satisfactory architectural fidelity.

1. Introduction

Buildings are the most important human-made constructions in global urban environments. While creating and managing wide-range digital building models in three-dimensional (3D) cases has accelerated the development of urban digital twins, recent decades have witnessed the rapid growth of developments in modeling 3D urban buildings by incorporating many advanced technologies, such as photogrammetric reconstruction or detailed laser and mobile scanning (Deng et al., 2021, Lei et al., 2023, Abdelrahman et al., 2025). To generate smooth and applicable 3D urban models in universal triangle meshes, existing photogrammetric methods generally employ multiview matching based on merging aerial images captured from aircraft and drones for the overall architecture and street-view vehicle images for detailed information (Shao et al., 2016, Yu et al., 2021). The collected images are matched and used for feature extraction via mature approaches, such as structure from motion (SfM) (Schonberger and Frahm, 2016), from which dense 3D points are computed using multi-view stereo (MVS) (Schönberger et al., 2016) as references for mesh triangularization. Novel methods based on deep learning without multiview references have been increasingly developed in recent years (Zhou et al., 2019, Nikoohemat et al., 2020, Feng and Atanasov, 2021, Li et al., 2024). However, the following significant data constraints are constantly mentioned as limitations of these approaches: (1) the poor quality of source data or occasional occlusions of vegetation usually result in difficulties in reconstructing high-level building details, such as the orientations of the roof surface, especially in SfM-related methods that rely on spectral feature extraction and matching, while it is not viable to collect highly detailed visual data for each individual building; and (2) deep-learning-based methodologies usually suffer from deficiencies in the training data, but model performance relies largely on the quality of such data. These limitations indicate the urgent demand for an alternative approach to address the data-related challenges in the aforementioned methods.

Therefore, research has focused on automated creation of digital cousins (ACDC) (Dai et al., 2024) to approximate visual object appearance using artificial intelligence-generated content (AIGC) and generative large model (GLM) methodologies, instead of fully recovering details. By leveraging a series of approaches that require only parameters and hints, ACDC can address existing challenges, while providing technical foundations for improving the extensive mapping of buildings, regardless of data availability. Building Digital Cousins (BDC) is derived from existing theorization that utilizes a simulated geometric approximation of real-world building appearances. principally addressing horizontal (footprint), vertical (height), and cubic (roof variation) geometries. While few approaches aim specifically at BDC creations, related methods for 3D asset generation provide strong references for BDC development. Recent 3D asset generation approaches can be categorized into two routes: (1) diffusion probabilistic models (DPM), which employ sequential denoising procedures to synthesize the desired outputs, such as point clouds or triangle meshes (Alliegro et al., 2023, Wei et al., 2023); and (2) auto-regressive GLMs that apply a transformer architecture and encode 3D structures into discrete tokens resembling natural languages (Siddiqui et al., 2023, Gao et al., 2025). While both approaches are effective in creating general 3D assets, they face challenges in terms of modeling the geometries of BDCs for the following reasons: (1) a building mesh model contains fewer individual triangles than other 3D models, which maintain minor spatial details; (2) building models should be created under specific standards, such as within a certain range (footprint) and with vertical facade surfaces, but this may result in a worse outcome when universally trained GLMs are used; and (3) DPMs, which were originally developed for image synthesis under an integer pixel space, may not be able to satisfy the precision requirements of Euclidean coordinate space.

Consequently, this study aimed to derive a precise generative model, BldgWeaver, from transformer-based GLMs for mapping BDC at various levels of details (LoDs). Section 2 presents

recent related work in the field of 3D urban digital model development and discusses the merits of BDC to replace full-level reconstruction labor. Section 3 introduces our approach for creating BDC models by incorporating the next-token prediction approach for triangular mesh generation, implementing parallel conditional controls, and an optimized training strategy. Section 4 illustrates the results of our generation of BDC instances and quantitative comparisons between previous 3D urban building modeling methodologies and our methodology. Finally, Section 5 concludes the study and discusses its limitations.

2. Background and Related Works

2.1 Automatic modeling of urban buildings using visual priors

Visual data are essential for the reconstruction of urban structures with intricate details. Conventional reconstruction approaches primarily utilize multiview matching approaches that integrate SfM or MVS techniques by leveraging aerial or street-view imagery as data sources (Zhou et al., 2019, Pepe et al., 2022, Li et al., 2023). Point clouds serve as another critical reference for high-quality 3D building reconstruction by complementing multiview image sources (Wang et al., 2020, Huang et al., 2022, Ogawa et al., 2024b). In parallel with point-based representations, neural radiance field (NeRF) (Mildenhall et al., 2021) represents 3D scenes containing building structures as continuous volumetric functions (f_v) that map 3D coordinates to color and density values. Similarly, 3D Gaussian splatting (Kerbl et al., 2023) employs Gaussian mixture models (GMMs) (Reynolds et al., 2009) to characterize 3D scenes as collections of probabilistic distributions with anisotropic covariance matrices. Both approaches offer the advantage of subsequent triangulation into mesh representations for further applications.

Despite their impressive results, these reference methods face significant challenges in terms of their computational complexity and data requirements, which constrain the quality of the reconstructed output. For example, NeRF models typically require hours to days of training on high-end graphics processing units (GPUs) and dense multiview imagery with precise camera poses. Similarly, although 3D Gaussian splatting offers faster optimization than NeRF, it requires substantial input data and computational resources for high-fidelity reconstruction. Consequently, to address these difficulties, this study advances the BDC concept, which approximates building geometries without full-level recovery of details. This approach satisfies the key requirements for urban investigation applications, while significantly reducing data dependencies and computational demands.

2.2 Non-visual 3D modeling using generative AI approaches

Generative AI methods have demonstrated remarkable capabilities for creating diverse targets. Prominent approaches include the generative adversarial network (GAN) (Goodfellow et al., 2020), variational autoencoder (VAE) (Yan et al., 2016), denoising diffusion model (DDM) (Rombach et al., 2022, Peebles and Xie, 2023), and autoregressive (AR) model based on the transformer architecture. GANs employ a competitive framework in which a generator and discriminator are adversarially trained to approximate the target distributions of 3D shapes. VAEs operate by encoding inputs into probability distributions in the latent space and sampling these distributions to generate novel instances through a decoder network. More recently, DDMs have advanced probabilistic modeling by learning to iteratively denoise a Gaussian distribution into a desired plain target through a Markov chain of denoising steps. Hybrid approaches that integrate VAEs with diffusion models leverage encoder-decoder architectures to enable denoising procedures in the latent space, thereby enhancing stability and computational efficiency. This has proven to be effective for generating multi-view images as a reference for 3D reconstruction (Anciukevičius et al., 2023).

Beyond these generative approaches, AR models have been developed to transfer scalability from natural language processing to vision tasks. Unlike the DDM or GAN methods, which map a probabilistic distribution for fixed-sized targets, a typical transformer AR model predicts the distribution of discrete tokens. This enables the generation of sequences with flexible lengths, thereby providing a significant advantage for 3D models with uncertain numbers of points, voxels, or triangles. Recent research has extensively explored the token-based generation of mesh models, such as PolyGen (Nash et al., 2020), MeshGPT(Siddiqui et al., 2023), MeshXL(Chen et al., 2024), and MARS (Gao et al., 2025). These approaches apply advanced strategies to convert mesh triangles into discrete token sequences, in order to generate high-quality BDC representations for real-world urban constructions. For the principal purpose of this study, which sought to efficiently map wide-range building targets, we adopted the MeshXL codebase, which converts triangle-wise coordinates into sequential tokens and trains a transformer model to learn the token distribution.

2.3 Aligning building representations in terms of LoD standards

Urban 3D building models are predominantly categorized according to different LoDs, which quantify the geometric and semantic richness of digital building representations (Biljecki et al., 2016, Ogawa et al., 2024a). The CityGML universal standard defines building LoDs across five levels, ranging from 0 to 4. Figure 1 illustrates the characteristic appearances and variations in the building models from LoD 0 to LoD 3. LoD 4, which primarily concerns interior spatial configurations, was excluded from the scope of this study because our focus was on exterior architectural representation.

For urban investigation and exploration, an appropriate LoD classification is essential for effective building model utilization. LoD 0 typically represents buildings as 2D footprints, LoD 1 as simple extruded white boxes, LoD 2 as structures with differentiated roof shapes and thematic surfaces, and LoD 3 as architecturally detailed models including openings (windows, doors) and smaller exterior features. However, previous generative approaches have largely addressed building representation without explicitly considering LoD standards, thereby creating a disconnect between generative outcomes and established urban data management frameworks. This oversight limits the applicability of the generated models to standardized urban analysis workflows that rely on consistent LoD classification.

In addition, the availability of high-quality training data presents significant challenges, particularly for LoD 3 models, which remain scarce because of prohibitive acquisition and modeling costs. The detailed façade elements and architectural features required at LoD 3 demand substantially more modeling effort compared to lower LoD representations. When facing training data deficiencies, we resorted to maintaining a wide-range LoD 2 data foundation of our generated BDCs. We established

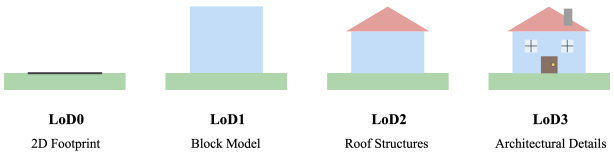

Figure 1. Illustration of building LoDs from level 0 to 3, from which we selected level 2 as our target for model generation.

a scalable means for generating models across multiple LoDs, while suggesting the potential to systematically append opening information (windows, doors, and other façade elements) to LoD 2 structures to implement a procedural framework for LoD 3 creation. This strategy balances practical data-availability constraints with the need for standardized models that conform to established urban data specifications.

3. Methods

3.1 Overview of the approach

BldgWeaver was essentially developed based on a prevailing transformer-based autoregressive next-token prediction scheme. To create building cases with a generally interpretable distribution in the coordinate space, all coordinates of the involved building instances were normalized to a fixed range. Figure 2 illustrates the overall framework. Because the transformer architecture can understand and predict discrete tokens, both the training and generation procedures were initiated with a customized token discretization procedure to convert continuous coordinates into the discrete token space. We referred to the advanced token discretization strategy proposed in (Chen et al., 2024). This approach allowed for converting the continuous vertex coordinates into discrete tokens distributed between [0, n_{token}], in which n_{token} refers to the range of available tokens configured in advance. The token mapping procedure can be abstracted using the following equation (Equation 1):

$$t_i = round(\frac{c_i - c_{i_{min}}}{c_{i_{max}} - c_{i_{min}}} * n_{token}), i \in \{x, y, z\} \quad (1)$$

Where c_i and t_i denote the original and discretized coordinate values on x, y, z axes, respectively. $c_{i_{min}}$ and $c_{i_{max}}$ denote the minimum and maximum coordinate values from a individual building on a single axis, and $round$ retains the integral parts to discretize the normalized values. Therefore, the mesh-based 3D building model is converted into a token sequence $\{t_x^v, t_y^v, t_z^v\}_{v\in\{1,2,3\}}^{tri}$ (tri indicates the triangle set of an individual building), in which the vertices are aligned in an ascending order of the vertical dimension (y in this paper), as illustrated in Figure 3. During the training procedure, the Transformer decoder reads and predicts all discretized tokens based on an auto-regressive masking strategy, which has been universally applied in the previous LLM training approaches. Meanwhile, a constant masking strategy adaptive to the building complexity is additionally involved to ignore the prediction of footprint tokens, and a parallel conditioner is employed to control the buildings' detailed appearances. Subsequently, the generation procedure imports multiple parallel building footprints as the initial conditions for the trained decoder to predict remaining portions of building models. Last, a reversed tokenizing module is applied to convert discrete tokens into continuous coordinates for output.

3.2 Next-token prediction for generating building meshes

As discussed in Section 2, transformer-based AR models have demonstrated their ability to map token distributions in both language and visual processing tasks. Principally, an AR model solves an autoregressive problem by sequentially predicting the probabilistic distribution of subsequent tokens, using previously generated tokens as conditions. While the various components of a building, such as the façade, roof, and attachments, can be partially or conditionally constrained by the lower components, especially the footprint shape, it is feasible to utilize autoregressive theory to convert the building mapping problem into a universal sequence processing task. For the model architecture, we applied the open pretrained transformer (OPT) as our basic model, with a pretrained model checkpoint available for more efficient fine-tune training. We employed the checkpoint with 350M parameters to balance efficiency and quality, while training configuration was further introduced in Section 4. By incorporating the token discretization strategy introduced in Section 3.1, we could achieve a robust mapping of building geometries using a plain autoregressive approach similar to MeshXL.

Furthermore, we manually appended an <sos> token at the beginning and an <eos> token at the end, as start and end identifications of the sequence. As shown in Figure 2, the standard version of cross-entropy loss was computed between a predicted token sequence and its corresponding ground truth, which is universally applied in general LLM training, as Equation 2:

$$\mathcal{L}_{ce} = -\frac{1}{|seq|}\sum_{i=1}^{|seq|}\sum_{j=1}^{n_{token}} \mathbf{1}[y_i, j] log P(p_{ij}) \quad (2)$$

where $|seq|$ denotes the discretized token length, y_i indicates the predicted token, p_{ij} denotes the probability of token j predicted at the position i, and $\mathbf{1}[y_i, j]$ is a binary function (1 if y_i equals to j else 0). The loss computation automatically omits <sos> and <eos> tokens since no predictions are required.

3.3 Parallel controls for building appearance regularization

In addition to this fundamental framework that can achieve sequence mapping of building meshes, regularization of the buildings' appearance requires an external embedder to allow integrating additional parameters such as roof types, in order to achieve a more precise approximation of building architectures. Therefore, along with the default decoder-only transformer architecture, we inserted an additional cross-attention layer between the self-attention and feedforward layers in each decoder layer, enabling cross-computation of internal token distributions and externally embedded feature distribution to regularize the appearance of the predicted result in the latent space. Figure 4 illustrates the details of the implementation of the orange blocks for cross-attention insertion.

Based on the original CE loss computation, the transformer decoder with parallel cross-attention control was further used to optimize the following cross-attention loss target using Equation 3 as follows:

$$\mathcal{L}_{ce}^{cross} = -\sum_{i=1}^{|seq|}\sum_{k=1}^{|text|} log P(p_{ik}) \quad (3)$$

where p_{ik} denotes the probabilistic distribution derived from the cross-attention mechanism, $\sum_{i=1}^{|seq|}\sum_{k=1}^{|text|}$ indicates a token-

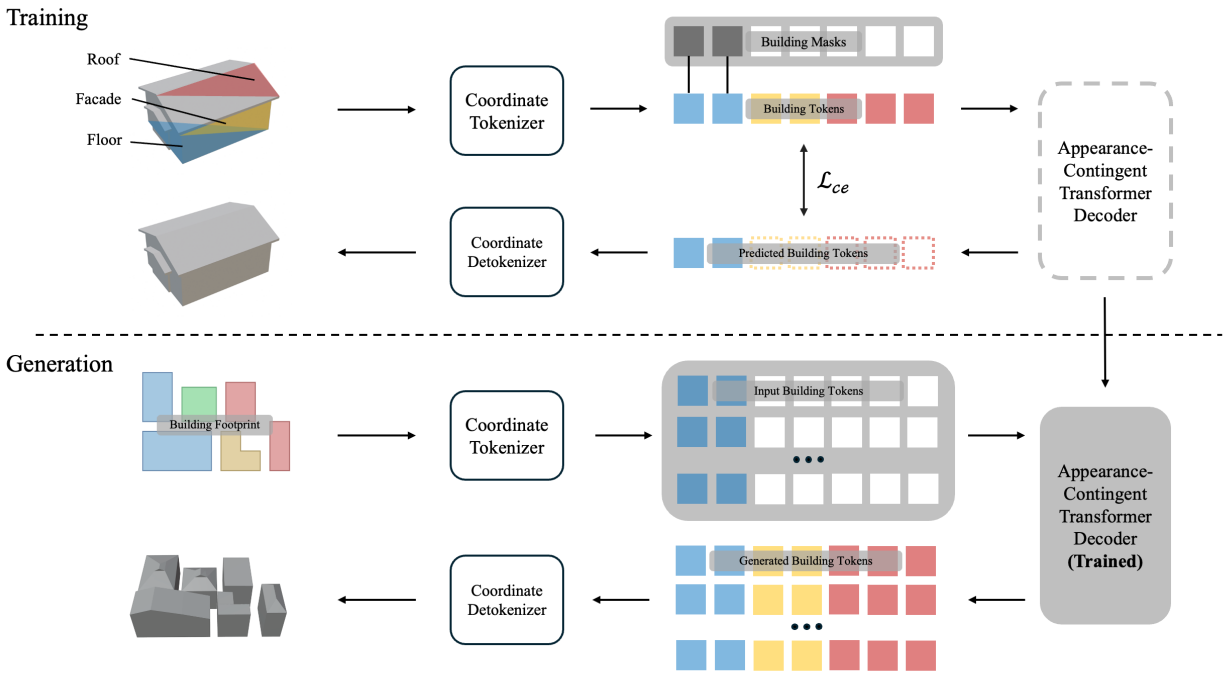

Figure 2. Overall training and generation images of BldgWeaver.

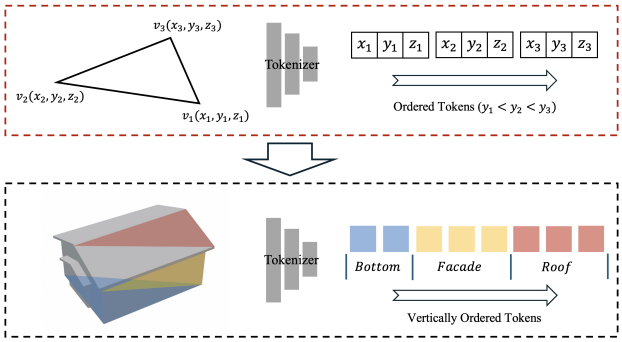

Figure 3. Illustration of the token discretization strategy.

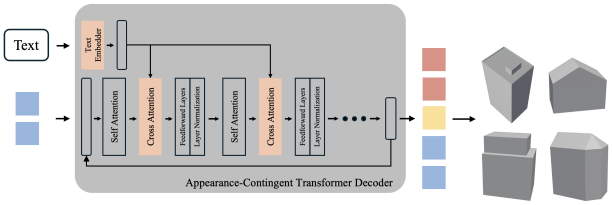

Figure 4. Illustration of the decoder-only Transformer architecture with parallel cross-attention controls.

wise probability evaluation between the embedded textual features and the latent token distribution.

3.4 Adaptive optimization of footprint-masking training

Alongside the next-token prediction with an additional appearance-contingent mechanism, it is critical to initialize an appropriate preliminary condition for each autoregressive generation procedure to ensure the quality of the generated results. While previous approaches have generally adopted additional modules for condition encoding, this study proposes an appearance-contingent masking strategy to add adaptive conditions to various building shapes during the training procedure. As each triangle can be represented by nine discretized tokens, the model automatically generates a parallel mask to constantly reveal the ground-truth footprint token distribution. Consequently, the loss is not computed within this token section during the entire training procedure involving the building data. Figure 5 illustrates the optimization strategy.

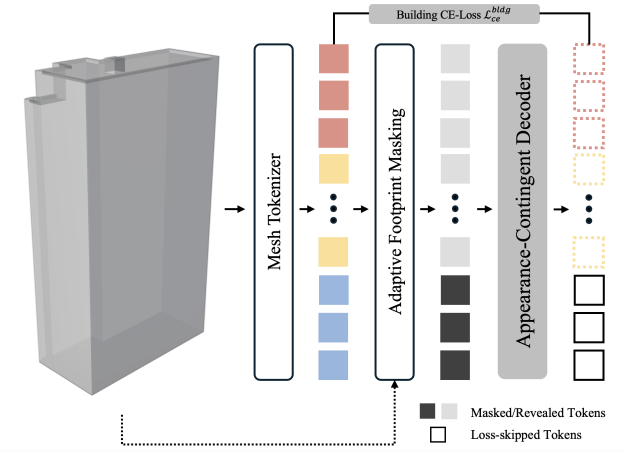

Figure 5. Illustration of the optimized footprint-masked training.

The computation of the optimized cross-entropy loss L_{ce}^{bldg} for the building-oriented training can be mathematically expressed,

as shown in Equation 4:

$$\mathcal{L}_{ce(f)} = -\frac{1}{|seq|}\sum_{i=1}^{|seq|} logP(p_{i,y_i})\mathbf{1}_{\{t\}_{footprint}}(i) \quad (4)$$
$$\mathcal{L}_{ce}^{bldg} = \mathcal{L}_{ce(f)} + \lambda\mathcal{L}_{ce}^{cross}$$

where $\mathbf{1}_{\{t\}_{footprint}}(i)$ represents another binary function. which equals to 1 when i is in the set $\{t\}_{footprint}$ that includes all the corresponding footprint tokens. When the remaining part is identical to Equation 2, this optimization simply implements a token-wise filter to omit the unnecessary computation upon the original cross-entropy loss. λ is a contribution parameter for a weighted control of the cross-attention term and is gradually tuned during the training procedure. We set the tuning section as [0.1, 0.5] following the common approach.

4. Experimental Results

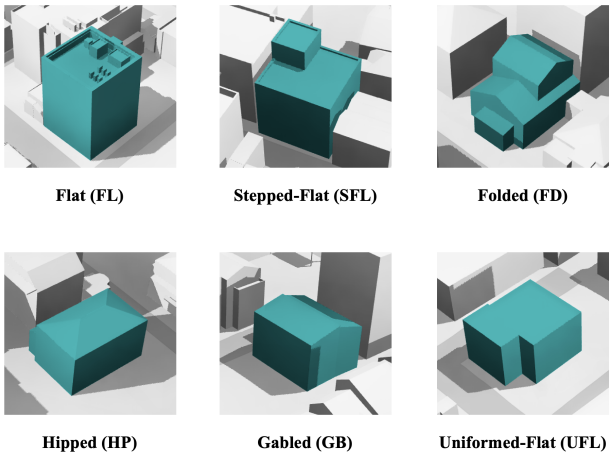

Figure 6. Illustration of the selected six roof types generally visible in Japanese urban built-up areas.

4.1 Data and Implementation

To train the BldgWeaver for wide-range BDC generation, we obtained a building dataset by sampling approximately 30,000 building mesh models in LoD 2 from the PLATEAU dataset developed by the Ministry of Land, Infrastructure, Transport, and Tourism of Japan (MLIT) (MLIT, 2022). We conducted fine-tuning of the MeshXL-350M model checkpoint using this dataset, which was pretrained on a combination of four general open datasets with more than 20 million 3D object instances. For data augmentation, we assigned random translations in six directions and horizontal rotations to each building instance. Furthermore, we normalized all data into [-0.95, 0.95] and translated the floor face to the bottom, where the height coordinates were constant at -0.95, to ensure identical generation patterns. Building roofs were categorized into six general types in relation to Japanese urban scenes: flat (FL), stepped-flat (SFL), folded (FD), hipped (HP), gabled (GB), and uniform-flat (UFL), to allow for learning the corresponding optimal roof types for various building footprint shapes. Figure 6 illustrates the prototypes of the six selected roof types. The pre-defined roof types were converted into pure texts and integrated via cross attention layers as shown in Figure 4.

In order to precisely estimate the proximity of our BDC instances, we used root mean square error (RMSE, σ_h) assessment to evaluate the global errors in relation to mesh geometries for quantitative evaluation, which was calculated using Equation 5 as follows:

$$\sigma_h = \mu_n \sum_{i\in m, k\in n} \|z_i - z_{ik}\| \quad (5)$$

where m and n represent the sampled point cloud from the two sets, z_i and z_{ik} indicate the pairwise height value between two point clouds, and μ_n indicates the building-wise average RMSE.

Both the model training and BDC generation experiments were conducted on a single NVIDIA RTX 4090 GPU owing to the selection of a smaller model size. Similar to the MeshXL pattern, the model was trained using bfloat16. We used the AdamW optimizer with a learning rate decaying from 1e-4 to 1e-6 and a weight decay of 0.1. In the prediction of the BDC instances, we generated 3D meshes using top-k and top-p sampling strategies with k = 50 and p = 0.95.

	Area (a)	Area (c)	Area (d)	Area (e)	AVG.
	RMSE (m)				
LoD1	0.37	0.49	0.47	0.34	0.41
LoD2	0.22	0.25	0.19	0.17	**0.21**

Table 1. Quantitative evaluation of BldgWeaver in four selected testing areas. (a), (c), (d), and (e) correspond to the numbering in Figure 8.

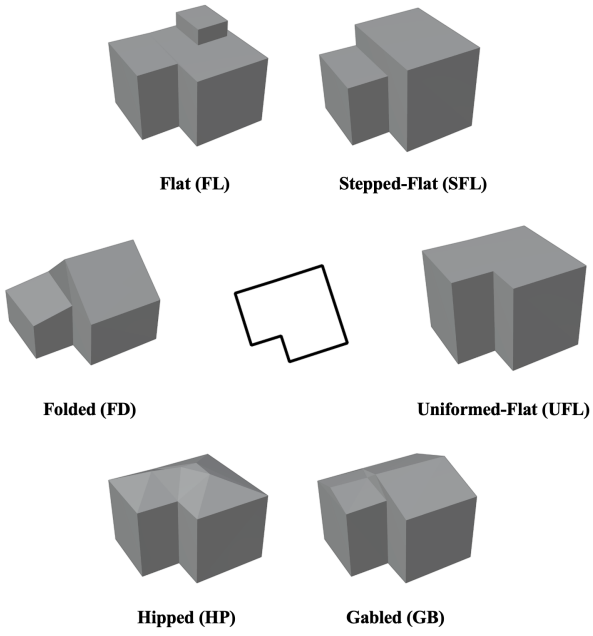

Figure 7. Results of single-footprint BDC generation under the configuration of six roof types. The polygon positioned at the center indicates the footprint involved.

4.2 Results from PLATEAU dataset

We conducted a visual evaluation of the selected footprint and several test areas around Meguro Ward, Tokyo, and Kashiwa City, Chiba Prefecture, Japan, to demonstrate the capability of BldgWeaver in mapping various building geometries regardless of the complexity of urban environments.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W6-2025-145-2025 | © Author(s) 2025. CC BY 4.0 License.

Figure 8. Visualization of the generated BDCs in six testing areas. (a)-(c) are from the Meguro downtown areas and (d)-(f) are from the country areas of Kashiwa City.

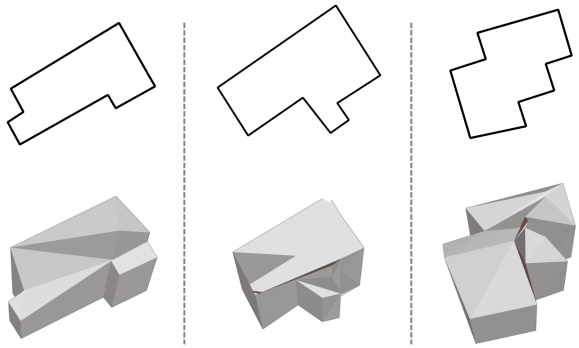

Figure 9. Exceptional errors in generating BDCs due to irregular or over-complex footprint shapes. The upper footprints correspond to the lower generated results by column.

Figure 7 illustrates the single-footprint evaluation of the generated BDC instances configured with the six preset roof types. The proposed model successfully mapped the corresponding geometric distributions for each roof type, such as the minor protrusions in the FL case and slope-like structures in the HP and GB cases.

Based on the reliable results demonstrated, we conducted quantitative and qualitative experiments on six selected testing areas: three from downtown districts and three from outlying country areas. Figure 8 illustrates the visual BDC results. This visualization demonstrated the capability of the proposed model to create a building geometric representation in satisfactory proximity to real-world building architectures. In addition, Table 1 lists the quantitative evaluation results from the following four representative areas: (a), (c), (d), and (e) for a sampled evaluation, compared with the geometries in LoD 1 in white boxes, while areas (b) and (f) showed similar performance patterns.

The results in Table 1 indicated an average improvement of 49% in geometric proximity, demonstrating the robustness of our method. It was illustrated that simpler roof types (FL, GB, HP) show more consistent generation quality, while complex types (FD, SFL) occasionally produce geometric inconsistencies, particularly for irregular footprints. However, some exceptional error cases were encountered when attempting to map the mesh units with over-complex footprint shapes containing irregular shapes or various edges, such as the zigzagging case in the middle of area (c) and the uncommon horizontal protrusion on the left of area (e), as shown in Figure 9, which reveals a magnified view of these unexpected errors. Although errors were observed in less than 5% of the overall cases. Nevertheless, this still indicated that there are limitations to our proposed method, which need to be addressed in future work.

5. Conclusion

To resolve issue related to data deficiencies and generalization problems in creating urban digital twins of buildings, this study proposed a novel and reliable generative model, BldgWeaver, for mapping wide-range urban BDC instances using GLM theory, attempting to replace 3D reconstruction in scenes with data deficiencies to approximate building geometries instead of complete reconstructions. The proposed model employs an advanced tokenizing strategy to discretize the continuous mesh coordinates into tokens that are understandable to our applied transformer AR model. In addition, we contribute to the field of urban 3D mapping by integrating a parallel controlling module to embed appearance parameters and reveal footprint priors during the training procedure, which enables the appearance-contingent generation of respective BDC representations for individual buildings with greater robustness. Our proposed method

achieved satisfactory performance with an average improvement in geometric proximity of 49%, as shown by hybrid experiments conducted in six independent testing areas. However, the proposed method still has some limitations in terms of generalization, such as the failure to map reasonable upper-mesh geometries when encountering over-complex footprints. Nevertheless, we were able to integrate the horizontal geometric features of the footprint shapes with existing priors to further improve the generalized understanding of 3D architectures for GLMs.

Acknowledgements

This study was supported by grants from the "Advanced AI Talent Development to Lead the Next-Generation Intelligent Society (BOOST NAIS)" of the University of Tokyo by the Broadening Opportunities for Outstanding young researchers and doctoral students in STrategic areas (BOOST) of the Japan Science and Technology Agency (JST) and the Project BRIDGE from the Japan Ministry of Land, Infrastructure, Transport and Tourism (MLIT).

References

Abdelrahman, M., Macatulad, E., Lei, B., Quintana, M., Miller, C., Biljecki, F., 2025. What is a Digital Twin anyway? Deriving the definition for the built environment from over 15,000 scientific publications. *Building and Environment*, 274, 112748.

Alliegro, A., Siddiqui, Y., Tommasi, T., Nießner, M., 2023. Polydiff: Generating 3d polygonal meshes with diffusion models. *arXiv preprint arXiv:2312.11417*.

Anciukevičius, T., Xu, Z., Fisher, M., Henderson, P., Bilen, H., Mitra, N. J., Guerrero, P., 2023. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12608–12618.

Biljecki, F., Ledoux, H., Stoter, J., 2016. Generation of multi-lod 3d city models in citygml with the procedural modelling engine random3dcity. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*.

Chen, S., Chen, X., Pang, A., Zeng, X., Cheng, W., Fu, Y., Yin, F., Wang, Y., Wang, Z., Zhang, C., Yu, J., Yu, G., Fu, B., Chen, T., 2024. Meshxl: Neural coordinate field for generative 3d foundation models.

Dai, T., Wong, J., Jiang, Y., Wang, C., Gokmen, C., Zhang, R., Wu, J., Fei-Fei, L., 2024. Automated creation of digital cousins for robust policy learning.

Deng, T., Zhang, K., Shen, Z.-J. M., 2021. A systematic review of a digital twin city: A new pattern of urban governance toward smart cities. *Journal of management science and engineering*, 6(2), 125–134.

Feng, Q., Atanasov, N., 2021. Mesh reconstruction from aerial images for outdoor terrain mapping using joint 2d-3d learning. *2021 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 5208–5214.

Gao, J., Liu, W., Sun, W., Wang, S., Song, X., Shang, T., Chen, S., Li, H., Yang, X., Yan, Y., Ji, P., 2025. Mars: Mesh autoregressive model for 3d shape detailization.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2020. Generative adversarial networks. *Communications of the ACM*, 63(11), 139–144.

Huang, J., Stoter, J., Peters, R., Nan, L., 2022. City3D: Large-scale building reconstruction from airborne LiDAR point clouds. *Remote Sensing*, 14(9), 2254.

Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4), 139–1.

Lei, B., Janssen, P., Stoter, J., Biljecki, F., 2023. Challenges of urban digital twins: A systematic review and a Delphi expert survey. *Automation in Construction*, 147, 104716.

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, Z., Wu, B., Li, Y., Chen, Z., 2023. Fusion of aerial, MMS and backpack images and point clouds for optimized 3D mapping in urban areas. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202, 463–478.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., Ng, R., 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1), 99–106.

MLIT, 2022. Plateau. Accessed on Mar. 15th, 2025.

Nash, C., Ganin, Y., Eslami, S. M. A., Battaglia, P. W., 2020. Polygen: An autoregressive generative model of 3d meshes.

Nikoohemat, S., Diakité, A. A., Zlatanova, S., Vosselman, G., 2020. Indoor 3D reconstruction from point clouds for optimal routing in complex buildings to support disaster management. *Automation in construction*, 113, 103109.

Ogawa, Y., Nakamura, R., Sato, G., Maeda, H., Sekimoto, Y., 2024a. End-to-End Framework for the Automatic Matching of Omnidirectional Street Images and Building Data and the Creation of 3D Building Models. *Remote Sensing*, 16(11), 1858.

Ogawa, Y., Sato, G., Sekimoto, Y., 2024b. Geometric-based approach for linking various building measurement data to a 3D city model. *Plos one*, 19(1), e0296445.

Peebles, W., Xie, S., 2023. Scalable diffusion models with transformers. *Proceedings of the IEEE/CVF international conference on computer vision*, 4195–4205.

Pepe, M., Fregonese, L., Crocetto, N., 2022. Use of SfM-MVS approach to nadir and oblique images generated throught aerial cameras to build 2.5 D map and 3D models in urban areas. *Geocarto International*, 37(1), 120–141.

Reynolds, D. A. et al., 2009. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 3.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Schonberger, J. L., Frahm, J.-M., 2016. Structure-from-motion revisited. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.

Schönberger, J. L., Zheng, E., Frahm, J.-M., Pollefeys, M., 2016. Pixelwise view selection for unstructured multi-view stereo. *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, Springer, 501–518.

Shao, Z., Yang, N., Xiao, X., Zhang, L., Peng, Z., 2016. A multi-view dense point cloud generation algorithm based on low-altitude remote sensing images. *Remote Sensing*, 8(5), 381.

Siddiqui, Y., Alliegro, A., Artemov, A., Tommasi, T., Sirigatti, D., Rosov, V., Dai, A., Nießner, M., 2023. Meshgpt: Generating triangle meshes with decoder-only transformers.

Wang, C., Wen, C., Dai, Y., Yu, S., Liu, M., 2020. Urban 3D modeling using mobile laser scanning: A review. *Virtual Reality & Intelligent Hardware*, 2(3), 175–212.

Wei, Y., Vosselman, G., Yang, M. Y., 2023. Buildiff: 3d building shape generation using single-image conditional point cloud diffusion models. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2910–2919.

Yan, X., Yang, J., Sohn, K., Lee, H., 2016. Attribute2image: Conditional image generation from visual attributes. *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, Springer, 776–791.

Yu, D., Ji, S., Liu, J., Wei, S., 2021. Automatic 3D building reconstruction from multi-view aerial images with deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 171, 155–170.

Zhou, Y., Wang, L., Love, P. E., Ding, L., Zhou, C., 2019. Three-dimensional (3D) reconstruction of structures and landscapes: a new point-and-line fusion method. *Advanced Engineering Informatics*, 42, 100961.

https://doi.org/10.5194/isprs-annals-X-4-W6-2025-145-2025 | © Author(s) 2025. CC BY 4.0 License.