Closing the Loop: Perceptually-Guided Iterative Generation of LoD2.0 3D Building Meshes for Building Digital Cousin development

Lingfeng Liao[1], Chenbo Zhao[1], Yoshihide Sekimoto[1], Yoshiki Ogawa[1, 2]

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

Keywords: building reconstructions, urban digital cousins, generative AI, building mesh quality index

Abstract

This study proposes an advanced generative framework for creating digital building representations, building upon our previously developed autoregressive Building Digital Cousin (BDC) generator. Previous approaches to building reconstruction have relied heavily on extensive visual data references, which often lead to significant data deficiencies. Our earlier method partially addressed these limitations by incorporating generative modeling based on textual and image targets. However, this approach still suffered from instability during multi-round generation and from limitations associated with manually configured conditioning. This proposal advances the configuration by introducing a simple but novel perceptual evaluation index, Building Mesh Quality Index (BMQI), for building instances. Additionally, an expanded vocabulary for vertex discretization and a latent image-based conditioning mechanism are introduced to refine the previous pipeline. By setting the generation target to LoD2.0 without minor-level details, experiments on the PLATEAU dataset demonstrate the capability of our model to generate a wide range of building models that conform to the image description while maintaining outstanding perceptual quality in more than 99% of cases. An average improvement of 13% in geometric proximity and a BMQI close to 0.99 further demonstrate the effectiveness of our method in addressing previous deficiencies related to data dependency and appearance conformity in urban building reconstruction.

1. Introduction

The demand for three-dimensional (3D) urban models has grown rapidly in recent years, driven by increasing applications in urban planning and infrastructure management, such as urban digital twin (UDT) development. However, conventional approaches rely on photogrammetric methods and airborne laser scanning (ALS), combined with labor-intensive manual validation by domain experts (Lei et al., 2023, Abdelrahman et al., 2025), rendering large-scale model production prohibitively costly for thousands of buildings. A representative photogrammetric solution for reconstructing building mesh models with high-quality details involves the combination of structure from motion (SfM) (Schonberger and Frahm, 2016) and multi-view stereo (MVS) (Schönberger et al., 2016). SfM efficiently extracts image-based features and precise camera positions, from which dense point clouds are predicted, reconstructed, and positioned for subsequent triangulation. Although more recent approaches have partially adopted single-view reference methods to reduce computational cost (Feng and Atanasov, 2021, Li et al., 2024), these methods still exhibit a strong dependence on source data, such as the pixel resolution of aerial images and the spatial resolution of ALS point clouds. Low-quality input data often produce models with insufficient appearance detail or unexpected geometric errors. However, collecting sufficiently large visual datasets for training deep learning models remains challenging, indicating the need for alternative approaches with reduced dependence on visual data.

Instead of faithfully reconstructing real-world details in digital twin representations, robotics research has recently explored generative methods based on digital cousins (DCs), which provide robots with visual priors by generating alternative 3D assets with similar appearances rather than exact reconstructions (Dai et al., 2024). Considering the application requirements of 3D urban assets, simplified geometric representations are sufficient in many scenarios, such as approximate heat-flow simulations, in which minor structural details have limited impact on simulation outcomes. Meanwhile, the subdivision theory of building levels of detail (LoD) provides an empirical framework for simplifying building geometries (Biljecki et al., 2016). In this context, LoD2.0, which excludes minor attachments and secondary structures, can naturally serve as a reference representation for creating DCs. Figure 1 illustrates the characteristics of LoD2.0 and LoD2.1/2.2 roof representations. In addition, DC generation requires only simple single-view visual inputs with less stringent precision requirements. Available aerial imagery sources therefore provide adequate visual information for generating DC-based building representations. Building upon these theoretical foundations, we extend our previous proposal, BldgWeaver (Liao et al., 2025), by developing an updated Building Digital Cousins (BDC) framework that adapts DC generation through a type-based conditioning mechanism for controlling building appearance. According to the original BDC definition, the framework is expected to support

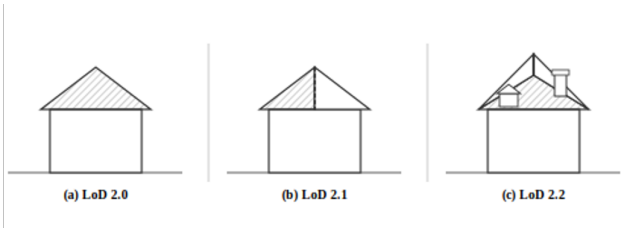

Figure 1. Visual comparison of LoD2.0 and LoD2.1/2.2, from which we selected level 2.0 as our primary target for model generation.

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-XII-4-W1-2026-227-2026

efficient updates within short time intervals. The initial DC proposal employed a combined pipeline that utilized large models (LMs) for both textual and visual scenes to retrieve spatial information from a single RGB image. Although this approach creates robust spatial references for robotic training environments, it requires substantial computational resources. Besides, with the rapid progress in the field of automatically detecting building contexts in the aerial images (Chen et al., 2023, Yuan et al., 2025), advanced conditions provided by the building-wise visual priors become viable. Furthermore, buildings exhibit unique structural characteristics that limit the applicability of simple geometric evaluation metrics. Conventional metrics such as Chamfer Distance (CD) assess point-level accuracy but fail to capture perceptual irregularities specific to architectural geometry, including watertightness violations, non-vertical facade faces, and irregular protrusions, which are structurally unacceptable under the LoD2.0 standard. These limitations highlight the need for a complementary evaluation metric capable of estimating perceptual mesh quality. To address this need, we propose the Building Mesh Quality Index (BMQI), a LoD2-centric assessment scheme that jointly evaluates roof adjacency and geometric abundance, facade verticality, and overall enclosure conformity relative to the building footprint.

This paper advances BldgWeaver by refining the tokenization strategy and incorporating building-specific image conditions, thereby replacing naïve type-based definitions in the building generator. In addition, an empirical evaluation index is proposed alongside geometric metrics, and a recurrent generation mechanism is introduced to iteratively optimize the appearance of generated buildings. The remainder of this paper is organized as follows: Section 2 reviews previous work on digital twin creation, maintenance, and assessment; Section 3 presents the theoretical definition of the proposed building evaluation index and describes improvements to the BDC generator; Section 4 presents experimental results demonstrating the effectiveness of the proposed index and generative framework for producing and evaluating a wide range of BDCs; finally, Section 5 concludes the paper and discusses limitations and future research directions.

2. Background and Related Works

2.1 Reducing References in Creating Buiding Digital Twin Models

Traditional reconstruction methods for urban buildings rely heavily on prior information provided by visual inputs, such as multi-view images captured from aerial or road cameras and laser-scanned point clouds (Pepe et al., 2022, Li et al., 2023, Ogawa et al., 2024). Neural radiance fields (NeRF) (Mildenhall et al., 2021) and their explicit successor, 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023), introduced a novel paradigm for urban model creation by encoding scene geometry and appearance into continuous volumetric functions or representing urban scenes as collections of anisotropic Gaussians (Marí et al., 2022, Lee et al., 2025). Both the SfM–MVS and rendering-based paradigms provide triangulation pathways for subsequent applications involving mesh-based 3D urban models. (Liu et al., 2024, Kulhanek et al., 2025) further applied the 3DGS framework to urban-scale scene reconstruction, enabling multiple LoDs for rendering.

Nevertheless, both paradigms fundamentally rely on dense multi-view imagery and often produce rendering-oriented representations or LoD categorizations that do not align with standardized building LoD definitions. In contrast, generative methods reduce this dependence on dense visual inputs by learning structural priors directly from training datasets. (Wei et al., 2023) implemented building-specific 3D generation using point cloud data within a conditioned diffusion model guided by single aerial images. (Rasoulzadeh et al., 2025) proposed a two-stage pipeline combining an autoregressive (AR) transformer architecture for generating coarse shapes with hierarchical diffusion-based upsampling to recover architectural details. At the city scale, (Xie et al., 2024) introduced a compositional neural field generator that synthesizes building instances and background elements to create unbounded 3D cities from semantic maps, whereas (Xie et al., 2025) implemented a Gaussian splatting representation that achieved a 60-fold speedup. However, these generative approaches primarily focus on photo-realistic appearance synthesis and remain based on implicit or voxel-based representations, producing outputs that are not structured as polygonal meshes. Moreover, these methods often require substantial per-model optimization or multi-stage inference. Diffusion-based pipelines require iterative denoising across hundreds of steps, while AR models scale poorly with increasing scene complexity. Consequently, these approaches are not well suited for efficient updates of digital twin representations. To balance computational efficiency and geometric complexity, our previous work introduced the BDC theory and adopted MeshXL, an advanced AR mesh-generation method, to accelerate the generation process. In this framework, LoD2.0, which captures the fundamental architectural characteristics of building instances, is incorporated to standardize the control of BDC generation. In addition, an optimized tokenization strategy is employed to balance dimensional categorization and generation efficiency.

2.2 Assessing Comprehensive Quality of Building Models

Regardless of the target data format, geometric fidelity to ground-truth references has long served as the dominant criterion for evaluating 3D object reconstruction quality (Yu et al., 2021, Huang et al., 2022, Li et al., 2022). Point-wise distance metrics therefore form the de facto evaluation standard across the field. (Yu et al., 2021) used Mean Absolute Error (MAE) to compare reconstruction performance across five baseline methods, while (Huang et al., 2022) employed Root Mean Square Error (RMSE) to evaluate vertex-wise precision relative to ground-truth building models across multiple public datasets. Consequently, these approaches achieved improvements over previous methods in terms of both quantitative metrics and qualitative geometric proximity.

However, this fidelity-centric paradigm conflates geometric completeness—how much of the surface is recovered—with geometric correctness—whether the reconstructed model satisfies structural constraints such as watertight enclosures and vertical facades. These represent fundamentally different evaluation criteria, yet the former is often used as a proxy for both. With the emergence of generative reconstruction approaches, several studies have adopted perceptual metrics from image synthesis, such as Fréchet Inception Distance (FID) and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018), to evaluate 3D generation performance (Shu et al., 2019, Cheng et al., 2025, Luan et al., 2025). (Luan et al., 2025) introduced a textured geometric evaluation method that extends LPIPS for evaluating textured 3D meshes without relying on rendered images. (Shu et al., 2019) proposed Fréchet Point

Cloud Distance (FPD) as a 3D adaptation of FID for assessing the distributional realism of generated objects, and (Caytuiro and Sipiran, 2025) further summarized and extended this formulation for broader generative evaluation scenarios.

Although these universal evaluation metrics have supported significant methodological progress in 3D reconstruction research, they often overlook architectural properties specific to building structures, such as watertightness and facade verticality, while focusing primarily on minimizing pairwise spatial distance between reconstructed objects and reference models. Critically, neither fidelity-centric metrics nor perceptual evaluation frameworks align their results with standardized geometric representations such as the City Geography Markup Language (CityGML), which provides a comprehensive definition of building LoDs (Biljecki et al., 2016). This limitation reduces the applicability of generated models in downstream urban workflows that require consistent LoD classification. This gap motivates the development of our proposed building-specific index for evaluating the perceptual quality of building constructions. By targeting LoD2.0 with sufficiently differentiated roof geometries and computable conformity criteria, the proposed index provides an architecture-aware and computationally tractable measure of perceptual quality that aligns directly with standardized urban data frameworks.

3. Methods

The proposed framework builds upon our previous work, BldgWeaver, which adopts a next-token prediction paradigm following MeshXL (Chen et al., 2024), a state-of-the-art method for generating low-poly triangular mesh assets by mapping sequences of discretized mesh vertex coordinates. The previous framework receives building footprints and peripheral parameters as conditions for mapping building geometries, while maintaining consistent performance across multiple generation rounds remained challenging, which limited its effectiveness in practical applications. Moreover, generation control relied solely on textual embeddings, which restricted the model's ability to describe building roof geometries composed of multiple structural types. To address these limitations, we introduce the BMQI, a perceptual evaluation metric designed to assess the appearance quality of diverse building mesh assets. By integrating this metric into the generation pipeline, the framework enables recurrent refinement of generated building mesh models. Figure 2 illustrates the proposed guided iterative generative framework, including the involved data and the feedback-loop generation procedure.

3.1 Building Mesh Quality Index (BMQI)

BMQI is essentially defined as an additional metric to the conventional geometric evaluation, aiming to assess the perceptual quality of BDC models. As the principal target of our BDC generative framework is derived from the general LoD2.0 standard **in CityGML**, we designed the LoD2-centric assessment scheme for BMQI, which specifically focuses on roof geometries besides the fundamental building structure such as vertical facade faces. We consider the following division for comprehensively evaluating building stereotypes:

3.1.1 Ratio of Roof Adjacency and Abundance (r_{AA}): Inspired by conventional watertightness estimation methods, this factor extends the concept to roof components of building models. Given a set of roof triangle meshes, three aspects are considered for comprehensive quality assessment: the adjacency relationship of triangles in all directions, indicating the watertightness of roof components; the local smoothness of roof vertices, which reflects roof regularity by detecting irregular edges relative to the alignment of footprint shapes; and the abundance of roof geometries, which prevents cubic buildings from being incorrectly represented as LoD2.0 models by controlling roof complexity. When defining V, E, and F as the vertex, edge, and face sets of an individual building mesh, respectively, the following equations describe the three evaluation factors:

$$\begin{aligned} \mathcal{S}_{adj} &= \frac{1}{|E|}\sum_{e\in E}\mathbf{1}[adj(e)=2] \\ \mathcal{S}_{smo} &= 1-\frac{|\{f\in F_{roof}\mid \alpha_f<\varepsilon\cdot\bar{\alpha}_{roof}\}|}{|F_{roof}|} \\ \mathcal{S}_{geom} &= -\alpha\sum_i p_i\log p_i+\beta(1-\frac{|\{n:n||\hat{z}\}|}{|F|}) \end{aligned} \tag{1}$$

The three variables correspond to the three evaluation aspects. The first term calculates the proportion of edges that connect exactly two adjacent mesh faces. The second term measures the smoothness of roof components by identifying degenerated roof faces with near-zero areas, where α indicates the face area and $\varepsilon = 10^{-6}$ represents the degeneracy threshold. The third geometric factor evaluates roof complexity by analyzing the distribution of face normals and the proportion of flat faces. Here, p represents directional bin divisions, where i indexes the bins and p_i denotes the proportion of faces whose normal vectors fall within the i-th bin. $\hat{z}$ represents a vertical unit vector used as a reference for identifying horizontal faces. The ratio r_{AA} is computed as the geometric mean of these three factors.

3.1.2 Ratio of Façade Verticality (r_{FV}): Standardized LoD2.0 building models should maintain fully vertical facade faces, consistent with LoD1 models. Therefore, the second component focuses on facade mesh faces, for which the following variable is computed as the evaluation factor:

$$r_{FV}=\frac{|\{n:n\perp\hat{z}\}|}{|F_{facade}|} \tag{2}$$

By referring to the proportion defined in $\mathcal{S}_{geom}$ while changing the criterion from parallel to perpendicular orientation relative to $\hat{z}$, the ratio of vertical facade faces is calculated. This ensures a structurally consistent perpendicular facade configuration for the overall building models.

3.1.3 Ratio of Overall Enclosing (r_{OE}): According to the official definition, exterior attachments and major protrusions are not permitted in LoD2.0 building models. Therefore, all components are expected to remain within the range of their footprint shapes, from which we derive the following component of the BMQI:

$$r_{OE}=\frac{T_{in}}{T_{bldg}} \tag{3}$$

where T_{in} denotes the number of mesh components whose horizontal projections lie entirely within the footprint boundary, and T_{bldg} denotes the total number of mesh components in the same building model.

A visual representation of the BMQI composition is illustrated in Figure 3. As a supplementary evaluation index to geometric metrics, BMQI helps identify building mesh outputs that achieve satisfactory geometric accuracy but exhibit collapsed or unrealistic appearances. In our experiments, it is applied to-

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-XII-4-W1-2026-227-2026 | © Author(s) 2026. CC BY 4.0 License.

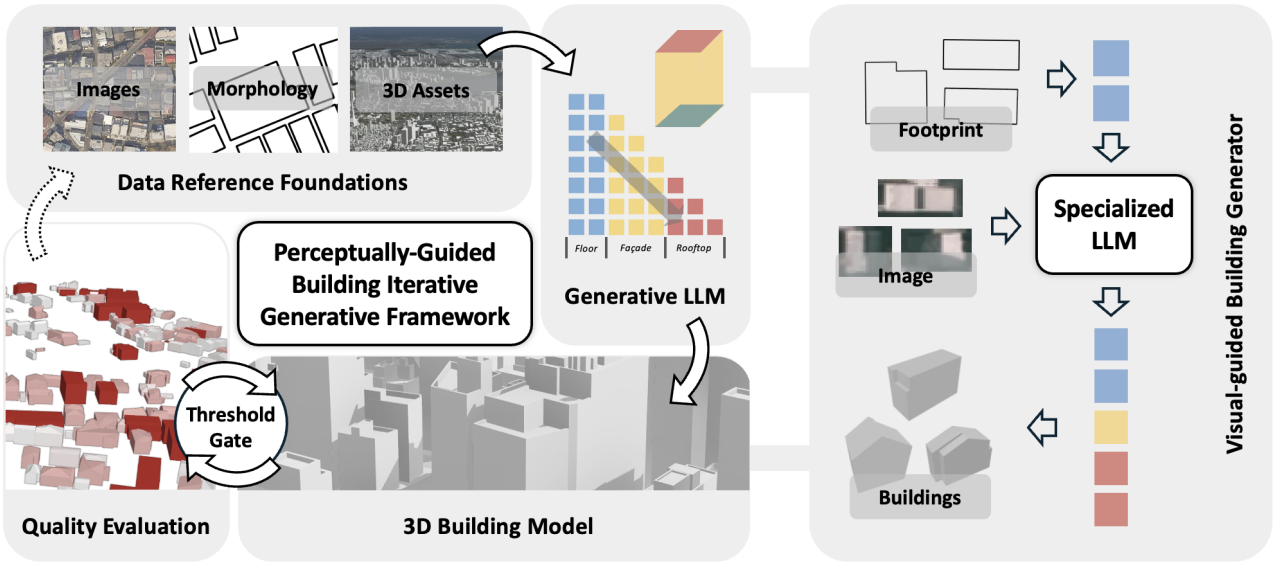

Figure 2. Overall illustration of our proposed framework for iterative generation of building models.

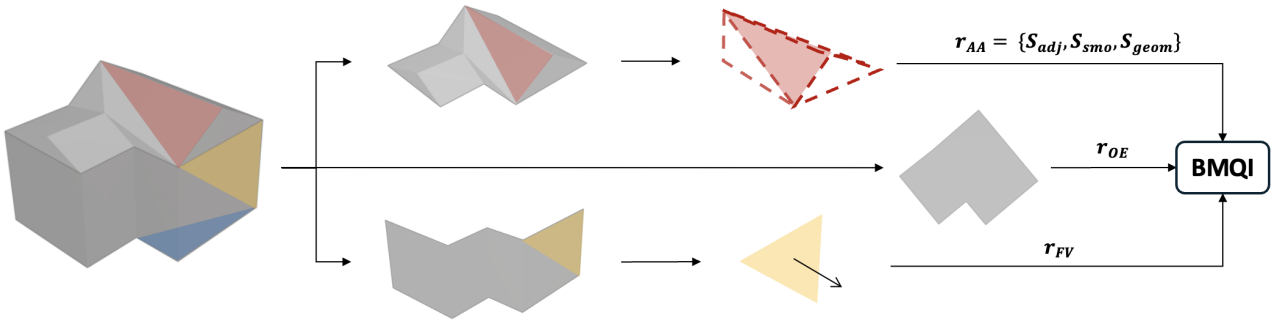

Figure 3. Illustration of BMQI definition, including the roof adjacency and abundance factor, the facade verticality factor, and the overall enclosing factor.

gether with the CD to evaluate geometric accuracy while improving the overall perceptual quality of generated building mesh assets.

3.2 Transformer-based Building Mesh Generator

The decoder-only transformer architecture has demonstrated strong performance across computer vision tasks, including 3D mesh generation (Siddiqui et al., 2023, Chen et al., 2024). Our earlier contribution, BldgWeaver, generated building models using a predefined set of roof types; however, this approach suffered from limited descriptive richness when characterizing diverse building structures and precisely describing roof orientations. Consequently, we extend the framework by introducing more powerful guidance for the generation process and refining the coordinate value discretization strategy to improve both global and fine-grained geometric accuracy.

3.2.1 Independent Coordinate Discretization: Existing approaches that adopt voxel-based discretization or follow MeshXL typically unify mesh vertex coordinate values across all axes into a single universal span. This design results in identical discrete tokens across dimensions, which are difficult for the model to distinguish. For example, when a decoder-only transformer predicts the next token, previously generated tokens may contain identical coordinate values originating from different axes, making it difficult for the model to interpret the dimensional relationships correctly. To address this limitation, we implement an axis-independent vocabulary that allows the model to distinguish tokens based on their respective spatial dimensions. As shown in Figure 4, coordinate values are discretized into three independent spans corresponding to the x, y, and z axes, which together form a continuous coordinate space. Let v_i denote the indexed value of this offset list: $\{x : 0, y : 1, z : 2\}$. The discretization strategy is defined as follows:

$$t_i = round(\frac{c_i - c_{i_{min}}}{c_{i_{max}} - c_{i_{min}}} * n_{token}) + v_i * c_{i_{max}} \quad (4)$$

Here, c_i and t_i denote the original and discretized coordinate values, respectively. $c_{i_{min}}$ and $c_{i_{max}}$ represent the minimum and maximum coordinate values of a building along a given axis, and $round$ extracts the integer component of the normalized value for discretization. This formulation introduces an implicit positional separation across coordinate dimensions, allowing the model to learn dimension-specific patterns in the coordinate value distribution during generation.

3.2.2 Visual-based Conditional Guidance: Naïve textual guidance relies on manually defined categorical labels during generation, which limits the ability to represent complex roof

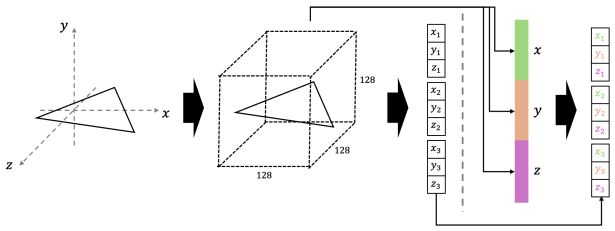

Figure 4. Illustration of our axis-distributed tokenization strategy. x, y, and z coordinate values are distributed into three independent discrete spaces.

geometries composed of multiple structural types. To overcome this limitation, we incorporate latent conditional guidance by embedding visual information extracted from aerial images, thereby enabling type-free guidance during the generation process. Specifically, we employ the image encoder from CLIP, a model pre-trained on large-scale image datasets for extracting visual feature embeddings. Because building footprint geometries are available, aerial images are cropped into building-level patches that serve as conditional inputs. These cropped images are encoded by the CLIP model and integrated with the mesh generation process through cross-attention layers within the decoder-only transformer architecture. The image-guided module minimizes the following loss function:

$$\mathcal{L}_{ce}^{cross} = -\sum_{i=1}^{|seq|}\sum_{k=1}^{|E_{CLIP}|} logP(p_{ik}) \tag{5}$$

where p_{ik} denotes the probability distribution produced by the cross-attention mechanism, and $\sum_{i=1}^{|seq|}\sum_{k=1}^{|E_{CLIP}|}$ represents token-wise probability evaluation between CLIP-derived visual features and the latent token distribution.

3.2.3 Iterative Generative Optimization via BMQI: Due to the AR nature of transformer-based models, multiple rounds of generation often introduce uncertainty. Moreover, the decoder-only architecture reconstructs coordinate sequences by minimizing token-level prediction errors but does not explicitly guarantee that the resulting geometric structure is globally coherent. Therefore, in addition to the aforementioned improvements, we incorporate a parallel BMQI evaluation after each generation round to ensure structurally reasonable building meshes. This evaluation functions as an implicit filtering mechanism that accepts or rejects generated results based on perceptual quality. Specifically, BMQI evaluates mesh connectivity, edge regularity, and facade verticality, thereby ensuring that generated buildings satisfy structural properties commonly observed in human-designed 3D building models.

By integrating these three optimization strategies, the proposed framework improves the performance of transformer-based mesh generation for building modeling tasks. Building upon the original BldgWeaver formulation, the optimized model minimizes the following training objective $\mathcal{L}^{optim}bldg$:

$$\mathcal{L}_{ce} = -\frac{1}{|seq|}\sum_{i=1}^{|seq|} logP(p_{i,y_i})\mathbf{1}_{\{t\}_{footprint}}(i)$$
$$\mathcal{L}_{bldg}^{optim} = \mathcal{L}_{ce} + \lambda\mathcal{L}_{ce}^{cross} \tag{6}$$

The first equation follows the original BldgWeaver formulation, while $\mathcal{L}_{bldg}^{optim}$ represents the linear combination of the token prediction loss and the cross-modal guidance loss. The tuning parameter λ is gradually adjusted within the range [0.1, 0.5], following common training practice.

4. Experimental Results

4.1 Data Preparation and Training Implementation

We constructed a dataset containing approximately 50,000 building instances derived from the PLATEAU dataset developed by the Ministry of Land, Infrastructure, Transport, and Tourism of Japan (MLIT) (MLIT, 2022). To ensure consistent training conditions, PLATEAU aerial images were explicitly retrieved as visual inputs for building generation and were organized according to the default Japanese 8-digit tile indexing system. Six urban regions with available aerial imagery were selected to extract building instances, ensuring that the sampled data were geographically distributed across Japan. All sampled building meshes were normalized to the range [-0.95, 0.95] along all axes, while the bottom surface was aligned to a height of -0.95 before discretizing coordinate values into tokens. In addition, random horizontal rotations around the z axis were applied during training as a data augmentation strategy. To associate building meshes with image descriptions, aerial images were cropped using the footprint contours of the corresponding buildings, producing paired mesh–image training samples.

The building mesh generator was trained on a single NVIDIA RTX 4090 GPU with 24GB of memory, enabled by the use of a relatively compact backbone model. Training was performed using bfloat16 precision and the AdamW optimizer, with a cosine learning-rate decay schedule ranging from 1e-4 to 1e-5.

4.2 Practicality Proof of Recurrent Generation

We conducted four experiments using independent building instances randomly selected from the test dataset to demonstrate the consistency of generated meshes and their conformity to image-based conditioning.

The generator was configured using a top-p sampling strategy with $p = 0.95$. The maximum number of recurrent generation rounds was set to 10 as a loose constraint. Figure 5 illustrates the results, indicating that the peripheral constraint, BMQI, can regularize generation behavior without requiring modifications to the model architecture. For all selected cases, a building model aligned with the appearance suggested by the cropped aerial image tile was successfully generated and compared with the PLATEAU ground-truth references. The first and third samples illustrate the model's capability to refine unexpected holes, while the remaining two visually demonstrate that BMQI helps regularize roof orientations. These results demonstrate the effectiveness of BMQI in evaluating the perceptual quality of generated building assets.

4.3 Wide-range Experimental Results of PLATEAU datasets

Further experiments were conducted in three 400m^2 testing areas and one 1 km^2 testing area selected from Kyoto City to evaluate the generalizability of the proposed image-conditioned building mesh generator through both quantitative and qualitative analysis. The three smaller areas were used for quantitative

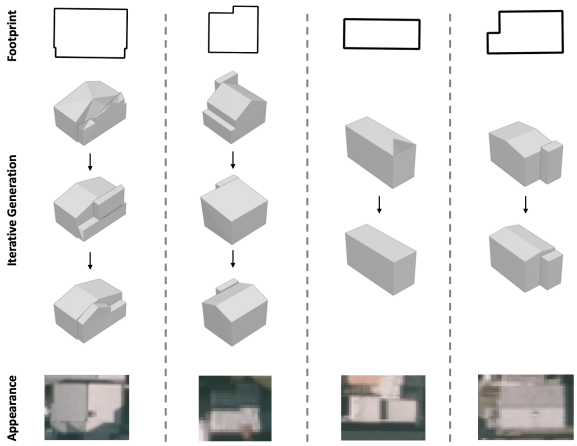

Figure 5. Illustration of four independent experiments of BMQI-threshold recurrent generation, demonstrating the capability of BMQI to regularize generation behavior in terms of model compactness and geometric alignment.

evaluation, in which the RMSE and BMQI were employed as evaluation metrics. Figure 6 visualizes the BMQI-based quantitative results, and Table 1 reports the corresponding quantitative scores for the three 400m^2 testing areas. An average improvement in RMSE of over 13% and BMQI values close to 0.99 indicate that the proposed model can generate building meshes with high geometric proximity to real-world instances while maintaining strong conformity with architectural standards.

The generated results for the 1 km^2 testing area are further visualized in Figure 7. The results demonstrate that the optimized pipeline can generate a wide range of building appearances conditioned on cropped aerial images without relying on additional type-based constraints, including high-rise buildings and residential structures with diverse roof geometries. Despite this demonstrated generalizability, the proposed framework exhibits limitations when encountering buildings with extremely large footprints or highly complex geometries. In particular, footprints with excessive concavities or irregular protrusions introduce ambiguities in token prediction, which may cause the AR decoder to generate geometrically inconsistent mesh components.

5. Conclusion

This study developed a novel evaluation metric for estimating the perceptual quality of building instances and advanced our previous AR building mesh generator in terms of both perceptual quality and generation capability. The previous BDC framework was extended by incorporating LoD2.0 constraints to improve generalization across larger urban areas. We further introduced an improved tokenization strategy by separating the vocabularies of the x-, y-, and z-axes to enable natural positional embeddings for vertex generation. In addition, visual conditioning derived from cropped aerial imagery through a CLIP encoder was integrated into the generation pipeline, and recurrent mesh generation was implemented to enable iterative evaluation and refinement of building appearance. By restricting the generation target to LoD2.0 without minor structures or attachments, the quantitative and qualitative experiments verified the effectiveness of the proposed framework for urban 3D building modeling. The results demonstrate an improvement

		Area (a)	Area (b)	Area (c)
MeshXL	RMSE(m)	0.35	0.41	0.36
	BMQI	0.79	0.69	0.65
BldgWeaver	RMSE(m)	0.25	0.27	0.25
	BMQI	0.86	0.86	0.92
Ours	RMSE(m)	0.22	0.25	0.23
	BMQI	0.99	0.96	0.98

Table 1. Quantitative evaluation of our framework in three selected testing areas using RMSE and BMQI in parallel.

of 13% in geometric proximity compared with baseline methods and consistently high BMQI scores across diverse roof typologies in wide-ranging urban areas. Despite the observed limitations when handling high-rise buildings with highly complex footprint geometries, the proposed framework opens several promising research directions. Extending the training distribution to accommodate larger and more geometrically complex footprints remains an important objective, alongside exploring adaptive tokenization strategies that generalize more effectively across building scales. Furthermore, integrating procedural façade elements, such as windows and doors, into the current LoD2.0 generation pipeline presents a viable pathway toward automated LoD3 model generation, thereby further narrowing the gap between generative urban modeling and standardized digital twin deployment.

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. https://doi.org/10.1016/j.buildenv.2025.112748.

Biljecki, F., Ledoux, H., Stoter, J., 2016. An improved LOD specification for 3D building models. *Computers, environment and urban systems*, 59, 25–37. https://doi.org/10.1016/j.compenvurbsys.2016.04.005.

Caytuiro, N., Sipiran, I., 2025. 3d shape generation: A survey. *arXiv preprint arXiv:2506.22678*. https://doi.org/10.48550/arXiv.2506.22678.

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. https://doi.org/10.48550/arXiv.2405.20853.

Chen, S., Ogawa, Y., Zhao, C., Sekimoto, Y., 2023. Large-scale individual building extraction from open-source satellite imagery via super-resolution-based instance segmenta-

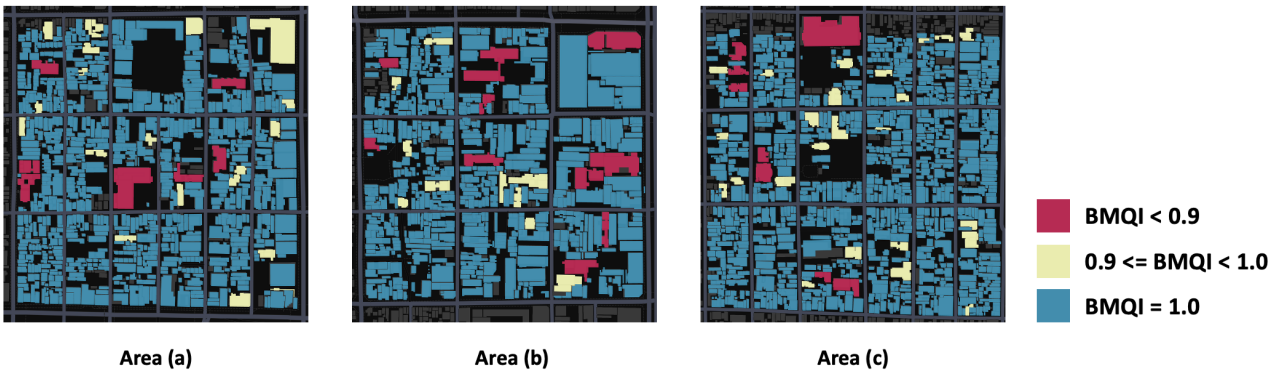

Figure 6. Quantitative results of area-based experiments under cropped-image conditions based on BMQI scoring.

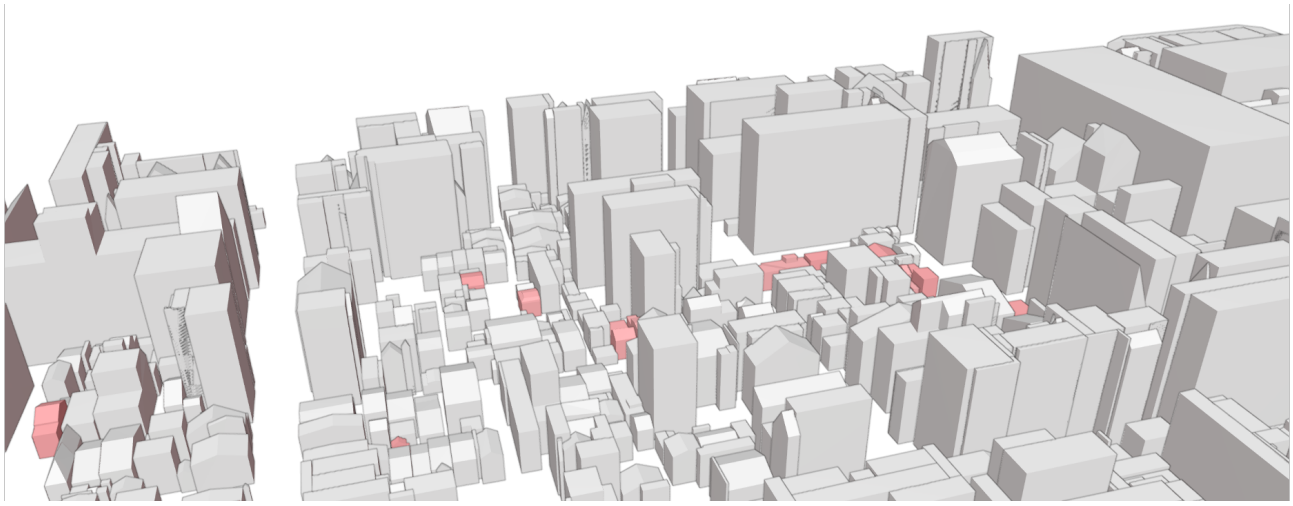

Figure 7. Visualization of the generated buildings in the 1km^2 area of Kyoto City. Instances with low BMQI scores are highlighted with red color.

tion approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, 195, 129–152. https://doi.org/10.1016/j.isprsjprs.2022.11.006.

Cheng, B., Luo, L., He, Z., Zhu, C., Tao, X., 2025. Perceptual point cloud quality assessment for immersive metaverse experience. *Digital Communications and Networks*, 11(3), 806–817. https://doi.org/10.1016/j.dcan.2024.07.001.

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. https://doi.org/10.48550/arXiv.2410.07408.

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. https://doi.org/10.1109/ICRA48506.2021.9561337.

Huang, J., Stoter, J., Peters, R., Nan, L., 2022. City3D: Large-scale building reconstruction from airborne LiDAR point clouds. *Remote Sensing*, 14(9), 2254. https://doi.org/10.3390/rs14092254.

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. https://doi.org/10.1145/3592433.

Kulhanek, J., Rakotosaona, M.-J., Manhardt, F., Tsalicoglou, C., Niemeyer, M., Sattler, T., Peng, S., Tombari, F., 2025. LODGE: Level-of-detail large-scale Gaussian splatting with efficient rendering. *arXiv preprint arXiv:2505.23158*. https://doi.org/10.48550/arXiv.2505.23158.

Lee, J.-Y., Liu, Y.-R., Tsai, S.-R., Chang, W.-C., Wu, C.-H., Chan, J., Zhao, Z., Lin, C. H., Liu, Y.-L., 2025. Skyfall-gs: Synthesizing immersive 3d urban scenes from satellite imagery. *arXiv preprint arXiv:2510.15869*. https://doi.org/10.48550/arXiv.2510.15869.

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. https://doi.org/10.1016/j.autcon.2022.104716.

Li, L., Song, N., Sun, F., Liu, X., Wang, R., Yao, J., Cao, S., 2022. Point2Roof: End-to-end 3D building roof modeling from airborne LiDAR point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 193, 17–28. https://doi.org/10.1016/j.isprsjprs.2022.08.027.

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. https://doi.org/10.1109/CVPR52733.2024.02619.

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. https://doi.org/10.1016/j.isprsjprs.2023.07.010.

Liao, L., Zhao, C., Sekimoto, Y., Ogawa, Y., 2025. BldgWeaver: An appearance-contingent generation solution with a three-dimensional automated creation of building digital cousins model using pre-trained transformer architecture. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-4/W6-2025, 145–152. https://doi.org/10.5194/isprs-annals-X-4-W6-2025-145-2025.

Liu, Y., Guan, H., Luo, C., Fan, L., Wang, N., Peng, J., Zhang, Z., 2024. Citygaussian: Real-time high-quality large-scale scene rendering with gaussians. https://doi.org/10.48550/arXiv.2404.01133.

Luan, T., Feng, X., Zhu, Z., Nuney, P., Liu, S., Gong, X., Doermann, D., Qiao, C., Yuan, J., 2025. Textured Geometry Evaluation: Perceptual 3D Textured Shape Metric via 3D Latent-Geometry Network. *arXiv preprint arXiv:2512.01380*. https://doi.org/10.48550/arXiv.2512.01380.

Marí, R., Facciolo, G., Ehret, T., 2022. Sat-nerf: Learning multi-view satellite photogrammetry with transient objects and shadow modeling using rpc cameras. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1311–1321. https://doi.org/10.1109/CVPRW56347.2022.00137.

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. https://doi.org/10.1145/3503250.

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

Ogawa, Y., Sato, G., Sekimoto, Y., 2024. Geometric-based approach for linking various building measurement data to a 3D city model. *Plos one*, 19(1), e0296445. https://doi.org/10.1371/journal.pone.0296445.

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. https://doi.org/10.1080/10106049.2019.1700558.

Rasoulzadeh, S., Stigsen, M. B., Kovacic, I., Schinegger, K., Rutzinger, S., Wimmer, M., 2025. ArchComplete: Autoregressive 3D architectural design generation with hierarchical diffusion-based upsampling. *Computers & Graphics*, 104477. https://doi.org/10.1016/j.cag.2025.104477.

Schonberger, J. L., Frahm, J.-M., 2016. Structure-from-motion revisited. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113. https://doi.org/10.1109/CVPR.2016.445.

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. https://doi.org/10.1007/978-3-319-46487-9_31.

Shu, D. W., Park, S. W., Kwon, J., 2019. 3d point cloud generative adversarial network based on tree structured graph convolutions. *Proceedings of the IEEE/CVF international conference on computer vision*, 3859–3868. https://doi.org/10.1109/ICCV.2019.00396.

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. https://doi.org/10.48550/arXiv.2311.15475.

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. https://doi.org/10.1109/ICCVW60793.2023.00313.

Xie, H., Chen, Z., Hong, F., Liu, Z., 2024. Citydreamer: Compositional generative model of unbounded 3d cities. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9666–9675. https://doi.org/10.1109/CVPR52733.2024.00923.

Xie, H., Chen, Z., Hong, F., Liu, Z., 2025. Generative gaussian splatting for unbounded 3d city generation. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 6111–6120. https://doi.org/10.1109/CVPR52734.2025.00573.

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. https://doi.org/10.1016/j.isprsjprs.2020.11.011.

Yuan, Y., Shi, X., Gao, J., 2025. Building extraction from remote sensing images with deep learning: A survey on vision techniques. *Computer Vision and Image Understanding*, 251, 104253. https://doi.org/10.1016/j.cviu.2024.104253.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595. https://doi.org/10.1109/CVPR.2018.00068.