Automatic Generation of Post-Redevelopment 3D City Models Using Satellite Imagery

Integrating image generative AI, segmentation, and building height estimation models

Yuki Mugita[1], Tomohiro Fukuda[2], Nobuyoshi Yabuki[3]
[1,2,3]The University of Osaka[3]Tokyo City University
[1]mugita@it.see.eng.osaka-u.ac.jp [2]fukuda.tomohiro.see.eng@osaka-u.ac.jp
[3]yabukin@tcu.ac.jp

Digital city models play a pivotal role in urban planning and landscape design. In addition to reproducing existing cities, there is a growing demand for generating city models of redeveloped areas. Previous studies have explored the generation of 3D city models by integrating open-source building footprints data and elevation data, as well as estimating digital surface models (DSM) from monocular satellite imagery. However, these methods rely on existing urban data and still require manual modeling to generate urban models of redeveloped areas where such data is unavailable. This study proposes a novel method to quickly and automatically generate both post-redevelopment virtual 2D satellite imagery and 3D city models using only satellite imagery. The proposed approach integrates three deep learning models, including an image generative model, a semantic segmentation model, and a height estimation model, along with image processing techniques to enable efficient and cost-effective urban planning simulations. For quantitative evaluation, the proposed method was assessed in terms of both 3D model generation time and accuracy. The results demonstrated that a 3D city model could be generated in an average of approximately 10.9 seconds. Additionally, the accuracy of building footprints extraction, evaluated by intersection over union (IoU), reached an average of approximately 78%. However, the building height estimation accuracy, evaluated by structural similarity (SSIM), remained below 0.5, indicating the need for further improvements through fine-tuning of the model. This study presents a method for generating urban models after redevelopment using only satellite imagery. In the future, this method will contribute to the visualization of landscapes and flexible urban simulations for disasters and other situations in urban planning.

Keywords: *digital city model, image generative AI, satellite imagery, semantic segmentation, building height estimation, urban planning*

INTRODUCTION

The visualization and reproduction of future urban space after redevelopment is very important in urban planning for various purposes such as landscape design, consensus building, and urban simulation (Lovett et al. 2015). In the past, simple methods such as sketching and photomontage (Lever 1973) were proposed to

visualize urban space after redevelopment. More recently, 2D image generation based on text input using image generative AI (Mugita, Fukuda, and Yabuki. 2023) has been proposed. In addition, the demand for digital 3D urban modeling is increasing, as simulation of the physical environment and spatial understanding after redevelopment are also necessary in actual urban planning.

Girindran et al. (2020) proposed a method for 3D modeling of existing cities by integrating open-source 2D building footprints and known elevation data. Mao et al. (2023) proposed a method for estimating digital surface models (DSM) from monocular satellite imagery. However, both methods were applied to existing cities for which geographic and satellite data exist. For post-redevelopment cities, there is a mixture of existing areas with existing buildings and non-existing areas that have been newly demolished or reconstructed. Therefore, the creation of post-redevelopment city models without available data still presents challenges, as it requires conventional manual modeling methods, which are inefficient.

This study aims to propose a novel method for quickly and automatically generating a virtual 2D satellite image and a virtual 3D city model at level of detail (LOD1) of post-redevelopment, using only satellite imagery of the existing city. The proposed method integrates deep learning models such as an image generative model, a segmentation model, and a height estimation model with image processing techniques. This research enables the creation of post-redevelopment urban models in a short period of time using only satellite images, without the need for specific software or detailed geographic data. In the future, this approach will enable more efficient and versatile urban simulations, encompassing not only rapid and cost-effective visualizations of post-redevelopment urban landscapes, but also physical simulations such as sunlight and flood analyses.

LITERATURE REVIEW

The research that has attempted to generate 3D city models for real cities has long been conducted in the field of city modeling. Girindran et al. (2022) proposed a method for generating 3D city models by combining open-source 2D building footprints data with known satellite-based elevation datasets, and Qian et al. (2023) created a 3D model by estimating the 3D density field of the area from satellite imagery. Yu et al. (2020) used deep learning to extract buildings automatically from multi-view aerial photographs and integrated known DSMs to create a 3D model. Mao et al. (2023) proposed Building3D as a new approach for 3D reconstruction of buildings from single view remote sensing images using deep learning.

Thus, although there have been studies using deep learning for some of the building footprints extraction and height estimation (DSM estimation), there have been no studies that combine the two, automate the entire pipeline or focus on the speed of model generation.

On the other hand, one method for generating non-existent 3D city models is the procedural modeling method. This method can randomly generate buildings and urban landscapes according to user-defined rules and algorithms (Rogla, Patow, and Pelechano. 2021) and can semi-automatically generate city models.

However, for cities after redevelopment, some existing buildings in real cities are demolished or reconstructed, so the reality and unreality are mixed. Therefore, methods that simply convert existing cities into 3D models or generate non-existent cities struggle to accurately reproduce post-redevelopment city models. As a result, conventional manual modeling approaches, which rely on specific software and are often inefficient, remain necessary. However, these methods are time-consuming and labor-intensive. To address this challenge, this study proposes a method for generating post-redevelopment 3D urban models automatically

and in a short time using only satellite imagery, without dependence on the specific software.

PROPOSED METHOD

This study proposes a method that integrates three deep learning models and two image processing techniques to quickly and automatically generate virtual 2D satellite images and virtual 3D city models that simulate post-urban redevelopment conditions, as shown in Figure 1.

The method takes three types of input images: an RGB satellite image, an IRRG (Infrared-Red-Green) satellite image, and a binary mask image specifying the redevelopment area.

First, in Step (1), a virtual satellite image representing the post-redevelopment state is generated using an image generative model, which takes the RGB satellite image and the redevelopment mask as input.

Next, Step (2) through (5) are used to generate a 3D city model of the redeveloped area. In step (2), building footprints are extracted from the RGB satellite image using semantic segmentation. Since the entire city is targeted, the size of the input image is very large, and segmentation as it would take a long time and would be inaccurate. Therefore, the image is divided into multiple images, each of which is segmented, and the results are later reassembled to form a complete footprint map of the entire city.

In Step (3), a building height map is generated from the IRRG image using a height estimation model. This model predicts building heights as grayscale values with 256 levels, representing relative heights within a single image rather than absolute real-world measurements. Similar to Step (2), the IRRG image is divided, processed, and then reconstructed into a single height map.

In Step (4), as shown in Figure 2(a), microfootprints are removed as noise, and building contours from Step (2) are refined using linear edge approximation. These are then merged with the redevelopment mask to generate final building footprints reflecting post-redevelopment conditions.

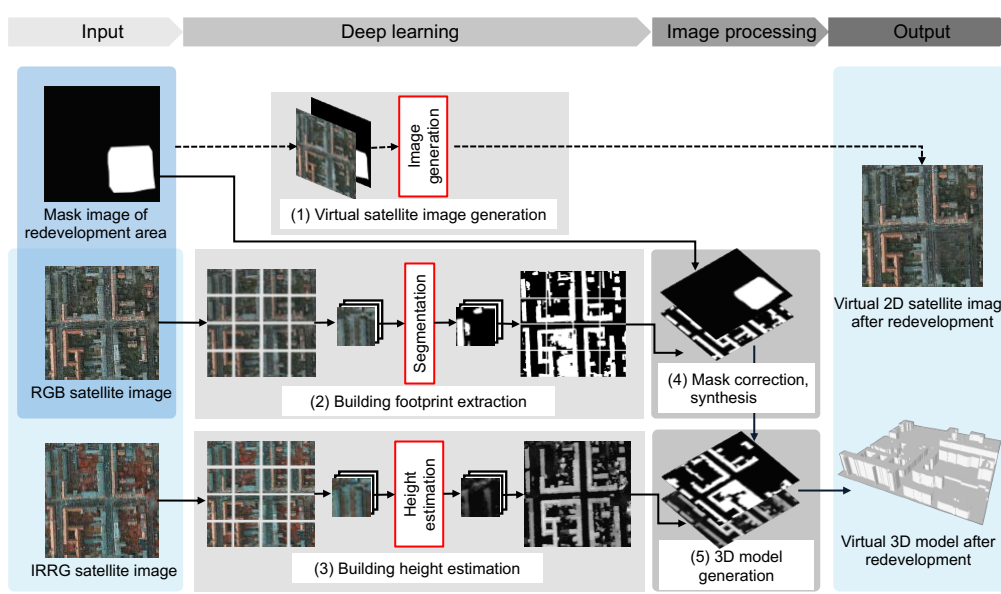


Figure 1
Overview of the proposed method

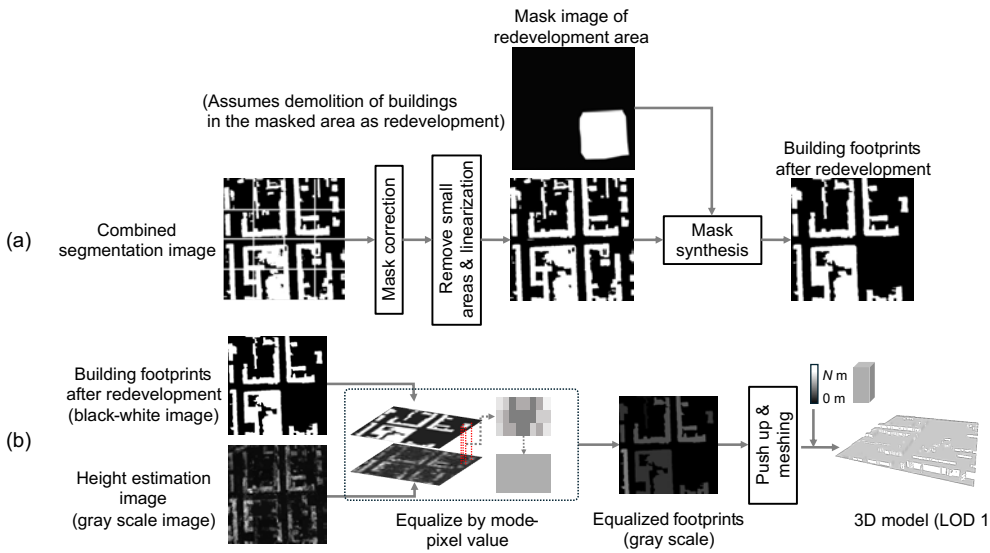


Figure 2
Some details of the proposed method
(a) Step (4)
(b) Step (5)

In step (5), the height information obtained in Step (3) is added to the redeveloped building footprints in step (4) as shown in Figure 2 (b). Since this information is relative height information as mentioned above, a 3D box model of LOD1 is generated by setting the real height according to 256 levels of grayscale in advance, pushing up each footprint in the vertical direction accordingly, and then meshing it.

PROTOTYPE SYSTEM

The system was developed in Python, integrates multiple deep learning models and open-source libraries, and was implemented in the environment shown in Table 1. The image generative model in Step (1) was large mask inpainting model (LaMa) (Suvorov et al. 2022), and the segmentation model in Step (2) was YOLOv11 (Redmon 2016), selected for its high-speed processing capability.

OS	Windows 11 Education 23H2
CPU	AMD Ryzen 7 7700X 8-Core Processor
GPU	NVIDIA RTX A4000 (16.0 GB)
RAM	32.0 GB

Table 1
Development environment (Desktop PC)

Segmentation model	YOLOv11n-seg
Source of custom training dataset	Vienna in Inria Aerial Image Labeling Dataset
Number of training images	Train: 15,000 images Validation: 5,000 images
Number of epochs	100
Image size	256 x 256 pixels

Table 2
Training conditions

Aerial photographs and corresponding mask images of the Vienna area from the Inria Aerial Image Labeling Dataset (Maggiori et al. 2017) were preprocessed by splitting and resizing them to 256×256 pixels, followed by data augmentation. This preprocessing resulted in a custom dataset comprising 20,000 images. The dataset was subsequently used for fine-tuning training under the training conditions in Table 2.

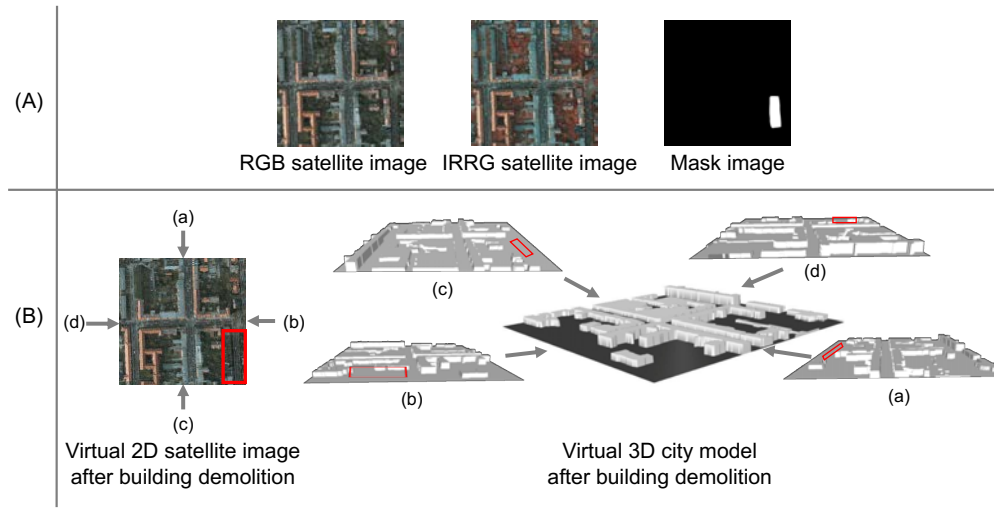


Figure 3
An example of the result
(A) Input images
(B) Results

To assess the detection accuracy of the fine-tuned model, COCO evaluation metrics (Lin et al., 2014) were used. The results demonstrate a high level of accuracy, with an mAP^{50} score of 0.8454, indicating that the model achieves robust performance at intersection over union (IoU) thresholds above 50%, confirming its suitability for prototype implementation.

Additionally, for the height estimation model in Step (3), this system utilized Unpaired Simulation of Digital Surface Models (U-IMG2DSM) (Paoletti et al., 2020), which can be developed in a local environment. For 3D modeling in Step (5), the Open3D library in Python was employed to enable efficient 3D mesh generation.

Figure 3 illustrates an example output from the prototype system. In this scenario, partial demolition of existing buildings is assumed as part of an urban redevelopment plan. Given the input shown in Figure 3(a), the system automatically generated a virtual 2D satellite image and a virtual 3D city model representing the post-redevelopment condition, as shown in Figure 3(b).

In the generated satellite image, buildings specified for removal in the mask were replaced seamlessly with vegetation such as grass and trees, with no visible artifacts. The virtual 3D model also accurately retained the structural details of areas outside the redevelopment zone, confirming the qualitative effectiveness of the proposed method.

VERIFICATION

This study conducted verification of the proposed method from the following three perspectives.

1. Generation time of the 2D image and 3D model
2. Accuracy of building footprints extraction
3. Accuracy of building height estimation

Verification 1 evaluated the temporal efficiency of the proposed method, while verification 2 and verification 3 evaluated the accuracy of the generated model.

For all verifications, the RGB and IRRG satellite image pairs from the Potsdam city dataset provided by the ISPRS (International Society for

Photogrammetry and Remote Sensing) (ISPRS, n.d.) were used as the validation dataset. This dataset was chosen because it contains not only satellite imagery but also building footprint annotations and grayscale DSM data, enabling direct comparison for validation.

Verification 1: generation time

This verification measured the processing time required to generate a post-redevelopment virtual 2D satellite image and 3D city model, from image input to final output, within the system environment described in Table 1.

As shown in Figure 4, 20 pairs were randomly selected from the validation dataset and a common mask image simulating a post-redevelopment area was applied. Processing times for the entire pipeline and for each individual step were measured using Python.

(a) (b) (c)

Figure 4
An example of input image for verification
(a) RGB satellite image
(b) IRRG satellite image
(c) Mask image

Step	(1)	(2)	(3)	(4)	(5)	Total	3D time
Average time (second)	6.86	4.47	3.56	0.20	2.67	17.76	10.90

Table 3
Results of generation time (3D time is the generation time of the 3D model only)

Average times across the 20 cases are presented in Table 3, with each step corresponding to the workflow stages shown in Figure 1. The results show that the entire 3D model generation process takes approximately 10.9 seconds on average.

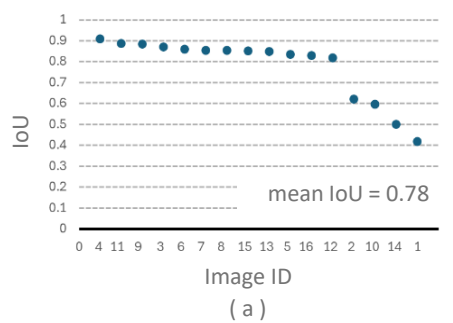
Verification 2: accuracy of building footprints extraction

This verification used a fine-tuned YOLOv11 model trained on a custom dataset to assess the accuracy of building segmentation, or footprints extraction. As the validation dataset, the RGB satellite image shown in Figure 3 (A) was used. Each image was divided into 16 tiles, and segmentation was performed on each tile, following the same procedure as the proposed method. The segmentation results were evaluated using IoU by comparing them with the ground truth data in the validation dataset.

The evaluation results are shown in Figure 5 (a). The mean IoU across the 16 tiles was 0.78, indicating that the fine-tuned segmentation model trained on the custom dataset achieved approximately 78% accuracy in building footprints extraction. Furthermore, excluding the four images with low IoU values from the 16 evaluated images, all remaining images achieved an IoU exceeding 0.8. These results indicate that the trained model is both stable and reliable in extracting building footprints.

Verification 3: accuracy of building height estimation

This verification assessed the accuracy of building height estimation using a height estimation model applied to IRRG satellite imagery. Height estimation was performed on the IRRG satellite image in Figure 3 and was quantitatively compared with the grayscale ground truth image from the validation dataset. Since the model provides only relative height values within a single image, a direct pixel-by-pixel comparison with the ground truth is not feasible. Instead, accuracy was evaluated by comparing the structural composition of the estimated height map with the ground truth using structural similarity (SSIM) (Wang et al. 2004), which considers brightness, contrast, and structure. A higher SSIM value (closer to 1) indicates higher accuracy.



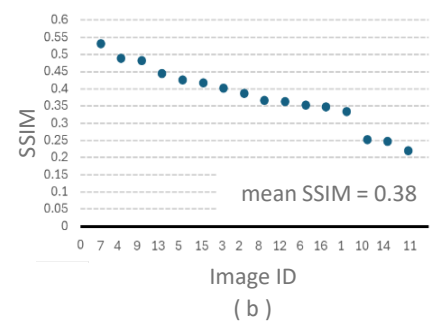


Figure 5
Results of the verification
(a) Accuracy of building footprints extraction
(b) Accuracy of building height estimation

The results are shown in Figure 5 (b). The highest SSIM achieved was approximately 0.53, with an average of 0.38, indicating that the accuracy of the height estimation remains limited.

DISCUSSION

This section discusses the temporal efficiency of the 3D model from Verification 1 and the accuracy of the 3D model from Verifications 2 and 3,based on both quantitative results and qualitative observations.

Temporal efficiency

The 3D model was generated in a short time of approximately 10.9 seconds. This processing time is significantly shorter than that required when using the 3D modeling software. The 20 input image pairs used in this verification differed in terms of the number of buildings, their shapes, and their layouts within the images. This finding suggests that stable and rapid automatic generation is possible regardless of the type of input image. In contrast to conventional 3D modeling, which typically requires significant time for operation due to the use of specific software such as game engines, the proposed method achieves 3D modeling through a simple data processing technique. This involves vertically extruding the building footprints with height information on a pixel-by-pixel basis. As shown in Table 3 (5), this process was completed in an average of approximately 2.67 seconds.

However, the problem was observed where the pixel-level roughness of the building footprints became pronounced on the sides of the 3D model after extrusion, as shown in Figure 6. This issue can likely be alleviated by applying smoothing processing and mesh optimezation to the generated 3D model.

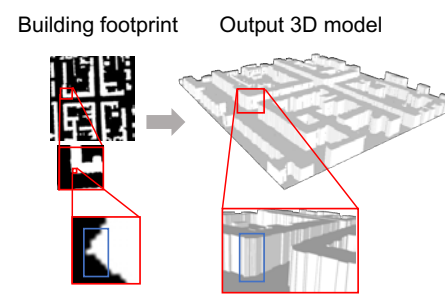


Figure 6
An example of side roughness

3D model accuracy

The accuracy of the generated 3D models is discussed in terms of both building footprints extraction and height estimation.

Building footprints extraction. The mean IoU result from Verification 2 indecated that building footprints could be extracted with approximately 78% accuracy. As shown in Figure 7, a qualitative comparison between the final building footprints and the ground truth confirms that the general shapes of buildings are appropriately captured.

However, extraction errors were observed in certain areas, such as the region outlined by the black frame in Figure 7. Specifically, as illustrated in Figure 8, buildings with orange-colored roofs were accurately detected, while frequent false positives occurred for buildings whose roof colors closely resembled those of the surrounding roads.

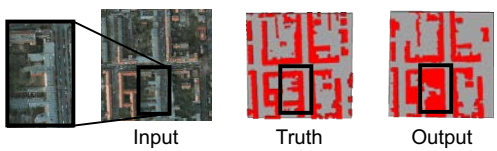


Figure 7
Comparison of the building footprints results with ground truth

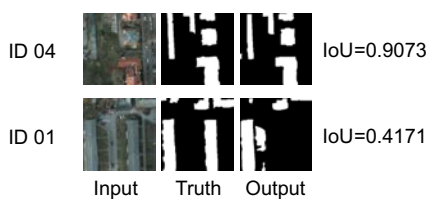


Figure 8
Comparison of high and low IoU

In this study, 20,000 images were already used for fine-tuning. Therefore, rather than simply increasing the number of training images, further improvement in accuracy can be expected by enhancing the diversity of the dataset. While the current training focuses on building images from Potsdam, incorporating data from other urban areas with varying architectural characteristics and scenarios is likely to yield better generalization.

Building height estimation. In Verification 3, the mean SSIM was approximately 0.38, with the highest value reaching only 0.53, indicating that the accuracy of height estimation requires improvement. Figure 9 presents a comparison between the images with the highest and lowest SSIM scores. In the case of Image ID07, areas with significant building heights were roughly estimated; however, the edges appear blurred. Since SSIM is sensitive to structural features such as edges, the resulting score was relatively low.

Nevertheless, qualitative assessment indicates that the overall distribution of building heights did not deviate significantly from the ground truth. On the other hand, Image ID11, which had the lowest SSIM, exhibited inaccurate height estimations. One possible reason for this issue is the similarity in color between roads and rooftops, as observed in the segmentation process too.

Ground truth 3D models were created using accurate height and footprint images from the validation dataset, and the relative heights between buildings were compared.

The relative heights between different buildings were compared. As shown in Figure 10, many discrepancies were found in the height order of the buildings. The height estimation model used in the prototype system, U-IMG2DSM, is characterized by an unsupervised learning approach. As a result, it is expected that further accuracy can be achieved through fine-tuning based solely on satellite images, without requiring ground truth height data.

In this study, emphasis was placed on high-speed processing, and therefore U-IMG2DSM, among multiple height estimation models, was selected for its capability of fast processing. Future improvements in accuracy may be realized by adopting models with more precise estimation capabilities.

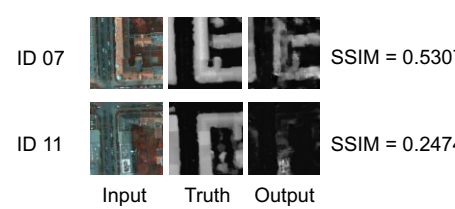


Figure 9
Comparison of high and low SSIM

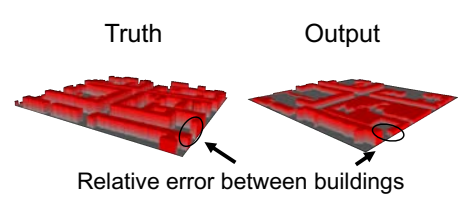


Figure 10
Comparison of relative heights between buildings

CONCLUSIONS

This study aimed to automatically and rapidly generate virtual 2D satellite images and 3D city models of post-redevelopment areas from satellite imagery. The following findings were derived from this research:

- A method was proposed that integrates multiple deep learning models, including image generation, semantic segmentation, and building height estimation, along with image processing techniques.
- Regarding temporal efficiency, the virtual 2D satellite image was generated in approximately 6.86 seconds, and the virtual 3D city model in approximately 10.9 seconds, demonstrating the capability for rapid generation.
- In terms of model accuracy, the building footprints extraction achieved approximately 78% accuracy, while the height estimation accuracy reached only about 38%, highlighting the need for further refinement of the estimation model.

This study demonstrates the feasibility of generating post-redevelopment urban models solely from satellite imagery, thereby contributing to the development of flexible and scalable urban simulation frameworks. Enhancing the accuracy of the generated models further opens the possibility of extending the approach to higher levels of LOD representations. These advancements promote cost-effective simulations from multiple perspectives, including urban landscapes and disaster scenarios, in urban planning by utilizing rapidly generated post-redevelopment models.

ACKNOWLEDEMENT

This work was partially supported by JSPS KAKENHI Grant Number 23K11724.

During the preparation of this work, the authors used ChatGPT and DeepL Write to improve the readability and language of the text. After using this tool/service, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

REFERENCES

Girindran, R., Boyd, D. S., Rosser, J., Vijayan, D., Long, G. and Robinson, D. (2020) 'On the reliable generation of 3D city models from open data', *Urban Science*, 4(4), 47. Available at: https://doi.org/10.3390/urbansci4040047 (Accessed: 30 January 2025).

ISPRS (n.d.) *ISPRS Urban Semantic Labeling Benchmark* [online]. Available at: https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/default.aspx (Accessed: 30 January 2025)

Lovett, A., Appleton, K., Warren-Kretzschmar, B. and Von Haaren, C. (2015) 'Using 3D visualization methods in landscape planning: An evaluation of options and practical issues', *Landscape and Urban Planning*, 142, 85–94. Available at: https://doi.org/10.1016/j.landurbplan.2015.02.021 (Accessed: 30 January 2025).

Lever, P. D. (1973) 'A photomontage system for site planning', *Computer-Aided Design*, 5, 103–104. Available at: https://doi.org/10.1016/0010-4485(73)90005-5 (Accessed: 30 January 2025).

Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C. L. (2014) 'Microsoft COCO: Common Objects in context', in *Lecture notes in computer science,* pp. 740–755. Available at : https://doi.org/10.1007/978-3-319-10602-1_48 (Accessed: 30 January 2025).

Maggiori, E., Tarabalka, Y., Charpiat, G. and Alliez, P. (2017) 'Can semantic labeling methods generalize to any city? The Inria aerial image labeling benchmark', *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. Available at:

https://doi.org/10.1109/IGARSS.2017.8127684 (Accessed: 30 January 2025).
Mao, Y., Chen, K., Zhao, L., Chen, W., Tang, D., Liu, W., Wang, Z., Diao, W., et al. (2023) 'Elevation estimation-driven building 3-D reconstruction from single-view remote sensing imagery', *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–18. Available at: https://doi.org/10.1109/tgrs.2023.3266477 (Accessed: 30 January 2025).
Mugita, Y., Fukuda, T. and Yabuki, N. (2023) 'Future landscape visualization by generating images using a diffusion model and instance segmentation', *Proceedings of the International Conference on Education and Research in Computer Aided Architectural Design in Europe, 2023*, 2, pp. 549–558. Available at: https://doi.org/10.52842/conf.ecaade.2023.2.549 (Accessed: 30 January 2025).
Paoletti, M. E., Haut, J. M., Ghamisi, P., Yokoya, N., Plaza, J. and Plaza, A. (2020) 'U-IMG2DSM: Unpaired simulation of digital surface models with generative adversarial networks', *IEEE Geoscience and Remote Sensing Letters*, 18(7), 1288–1292. Available at: https://doi.org/10.1109/lgrs.2020.2997295 (Accessed: 30 January 2025).
Qian, M., Xiong, J., Xia, G. and Xue, N. (2023) 'SAT2Density: Faithful density learning from Satellite-Ground image pairs'. Available at: https://doi.org/10.1109/iccv51070.2023.00341 (Accessed: 30 January 2025).
Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2016) 'You only look once: unifed, real-time object detection', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.779–788. Availabe at: https://doi.org/10.1109/CVPR.2016.91 (Accessed: 30 January 2025).
Rogla, O., Patow, G. A. and Pelechano, N. (2021) 'Procedural crowd generation for semantically augmented virtual cities', *Computers &Amp; Graphics*, 99, 83–99. Available at: https://doi.org/10.1016/j.cag.2021.06.014 (Accessed: 30 January 2025).
Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., et al. (2022) 'Resolution-robust Large Mask Inpainting with Fourier Convolutions', *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV),* 3172–3182. Available at: https://doi.org/10.1109/wacv51458.2022.00323 (Accessed: 30 January 2025).
Wang, Z., Bovik, A., Sheikh, H. and Simoncelli, E. (2004) 'Image Quality Assessment: From Error Visibility to Structural Similarity', *IEEE Transactions on Image Processing*, 13(4), 600–612. Available at: https://doi.org/10.1109/tip.2003.819861 (Accessed: 30 January 2025).
Yu, D., Ji, S., Liu, J. and Wei, S. (2020) ,Automatic 3D building reconstruction from multi-view aerial images with deep learning', *ISPRS Journal of Photogrammetry and Remote Sensing*, 171, 155–170. Available at: https://doi.org/10.1016/j.isprsjprs.2020.11.011 (Accessed: 30 January 2025).