The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLIX-B2-2026
XXV ISPRS Congress 2026 “From Imagery to Understanding”, Commission II, 4–11 July 2026, Toronto, Canada

Comparison of Different Object Detection Methods for Automatic Facade Enrichment of Existing Building Models from Aerial Images

Johannes Otepka[1], Günter Sükar[2], Martin Kerschner[2], Gerald Forkert[2], Norbert Pfeifer[1]

[1] TU Wien, Department of Geodesy and Geoinformation, Vienna, Austria - (johannes.otepka, norbert.pfeifer)@tuwien.ac.at
[2] UVM Systems GmbH, Wien, Austria - (martin.kerschner, guenter.suekar, gerald.forkert)@uvmsystems.com

Keywords: Object Detection, Window, Facade Modeling, Aerial Images, Neural Network, Segment Anything

Abstract

This study investigates the enrichment of existing building models using deep learning-based window detection from oblique aerial imagery acquired by a high-end multi-camera sensor system. While many cities maintain LOD2 building models at Level of Detail 2, higher levels of detail require the integration of facade elements such as windows. Three detection strategies are evaluated using 3D reference building models to assess accuracy and completeness. The test site is located in Vienna and consists of multiple large residential buildings with varying facade characteristics. The evaluated methods include zero-shot object detection with Grounding DINO combined with Segment Anything Model 2, applied to both oblique images and facade orthophotos, as well as a SAM2-UNeXT network requiring minimal training. Results indicate that zero-shot detection on orthophotos achieves the best performance, with a precision of 0.95 and an F1 score of 0.85. In contrast, the SAM2-UNeXT approach shows lower precision and F1 scores but slightly higher recall. The investigation shows that detection performance is influenced by facade viewing angles. Steeper viewing angles generally improve detection quality but increase susceptibility to occlusions, particularly in dense urban environments. The article concludes with a detailed outlook on future work, including the extension of the approach to more complex three-dimensional building structures.

1. Introduction

Building and city models are standard data sets that are used for visualization, planning, modeling, administration, and a variety of other tasks. In contrast to mesh or digital surface models (DSM), which can be generated by highly automated workflows, the generation of building models typically requires a significant amount of manual labor, especially if high quality models are required. An advantage of building models compared to ordinary mesh models is the larger range of applications for which they can be applied and the possibility of easy incremental updates. Therefore many cities (e.g. Basel, Cologne, Darmstadt, Hamburg, Helsinki, Linz, Luxembourg, Prague, Vienna, Zurich) take the effort in generating and maintaining up-to-date high-quality building model databases (Wysocki et al., 2024). Such models are typically realized within the CityGML standard (Kolbe et al., 2021) which not only allow describing buildings in different level of details (LoD) but also roads, rivers, bridges, terrain, vegetation and city furniture.

The scientific development of algorithms for automatic building model generation peaked around the year 2000 (Haala and Kada, 2010). At that time, the primary focus was on generating LoD2 models. In LoD2, a building is represented by wall surfaces, roof surfaces, and a ground plane (Hensel et al., 2019), with roof geometries typically simplified. To further enrich building models with facade elements such as windows and doors, a high level of detail (LoD3) is required, as these explicitly represent surface openings and architectural structures. The relevance of such detailed building models is expected to increase, as they enable more accurate analyses of solar potential, facade greening, green roof suitability, and overall urban energy balance. Those key indicators will help to fight or rather adept to regional climate change especially in cities (Adilkhanova et al., 2024).

In general automatic modeling processes work well in rural areas and areas with free-standing or regular arranged houses and use point clouds and/or aerial images as input. However, in densely populated inner-city areas that have evolved over centuries, automatic modeling algorithms still lack accuracy and completeness, particularly when aiming to achieve levels of detail beyond LoD2. Information within facades and roof surfaces, and additional structures on the outer building hull are typically not contained in models of inner cities, because 1) manual acquisition is very laborious, 2) occlusions may hinder a full coverage, and 3) automation for detailed, multi-element enrichment (LoD3) has not been successful up to now. However, current investigations confirm the relevance of this research topic. In (Pantoja-Rosero et al., 2022) the addition of facade openings for free standing houses from ground based images is shown. (Huang, 2025) suggested a semi-automatic approach to generate building models with windows and doors in the facades from multi-view-stereo meshes. Using terrestrial laser scanning point clouds (Hanke et al., 2025) demonstrate a method to add openings to existing building models. (Hensel et al., 2019) use deep learning and mixed integer linear programming (MILP) to detect and model facade openings in textured building models, to upgrade LoD2 models to LoD3.

Within the last decade analyzing facades has developed into a separate research topic called ‘facade parsing’ based on facade ortho photos without actually 3D modeling the results. Research in that area typically make strong use of deep learning which is why, several annotated facade datasets are publicly available, such as ECP, eTRIMS, Graz50, ArtDeco, and the CMP dataset. Another aspect of facade parsing is regularization strategies since facades typically show a lot of symmetries and repetitive structures. (Liu et al., 2020) suggest a special symmetric loss function that penalizes window predictions that are non-rectangles.

https://doi.org/10.5194/isprs-archives-XLIX-B2-2026-517-2026 | © Author(s) 2026. CC BY 4.0 License.

Whereas most research on LoD3 modeling make use of high resolution data from Laser Scanner (LS) or images captured by terrestrial, mobile, and UAS (Unmanned Aircraft System) platforms, we focus on professional aerial image sensors that capture nadir and oblique images for several reasons. Although terrestrial and mobile mapping systems provide high spatial resolution data, their capability to capture roof structures and inner courtyards is limited. Furthermore, data acquisition using UAS in densely populated urban areas is often constrained by legal and regulatory restrictions. In contrast, oblique aerial imagery offers comprehensive spatial coverage but is associated with a lower ground sampling distance due to higher acquisition altitudes, as well as unfavorable viewing geometries for facades in narrow street canyons. Vegetation further introduces occlusions, particularly under leaf-on conditions; while this limitation affects all sensor platforms, it is generally more pronounced for terrestrial systems.

Most importantly, aerial imagery is acquired systematically in many countries, with nationwide coverage updated at regular intervals; in most European countries, this occurs every 2–3 years. Furthermore, urban areas are typically captured using oblique imaging systems. Consequently, this work focuses on enriching building models using oblique aerial imagery, as suitable datasets are generally available without the need for dedicated acquisition campaigns.

To reach greater versatility, we suggest to use existing AI methods, which require little or no training, to extended traditional modeling algorithms for enriching building models with a high degree of automation. In this paper we investigate three different strategies to detect windows in oblique aerial images. For evaluating accuracy and completeness of the three methods we use a manually created 3D reference models that are projected into image space. It is important to note that such an approach requires thoroughly carried out image orientation and camera calibration otherwise the rendered 3D building model will not match the corresponding image content especially for shallow viewing angles. The developed framework provides bidirectional transition between image and object space, which will used in further work for detecting and modeling non-planar complex 3D objects like dormers, balconies, solar panels, chimneys or other roof-type stockpiling. Nevertheless, this study is restricted to windows (i.e. planar objects), and within the utilized pipeline, only the projection from object space to image space is required.

2. Data

The investigation uses data from the city of Vienna, Austria which was captured with *Vexcel UltraCam Osprey 4.1* sensor flying at an altitude of ca. 1700 m in summer 2023. The nadir sensor has a resolution of 20,544 x 14,016 pixels with a 80 mm lens resulting in average ground sampling distance of approx. 7 cm. The four oblique sensors have a resolution of 14,144 x 10,560 pixels using a focal length of 120 mm. The bundle block adjustment (BBA) was performed with the *Trimble Inpho MATCH-AT* software. To illustrate the size and quality of oblique images from this sensor, Fig. 1 presents a full-resolution oblique image of the test site, along with a magnified view of a building.

The actual test site consists of an area of approx. 2 hectares with multiple large scale 5-story building locate ca. 4 km to the west of the city center (cf. Fig. 2). The 3D reference model consists of five independent building parts which were created using the *CityGRID Shaper* Software (cf. Fig. 3) from *UVM Systems*.

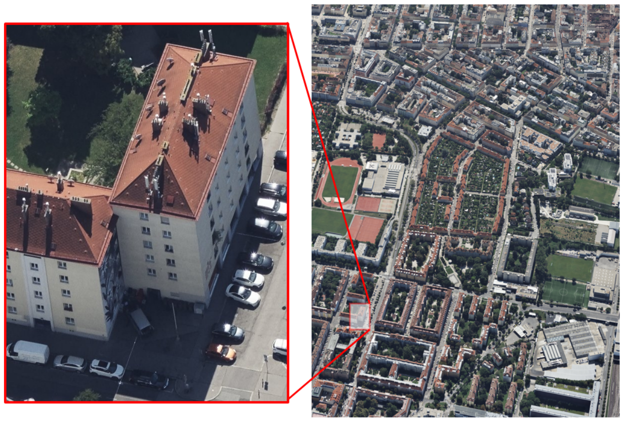
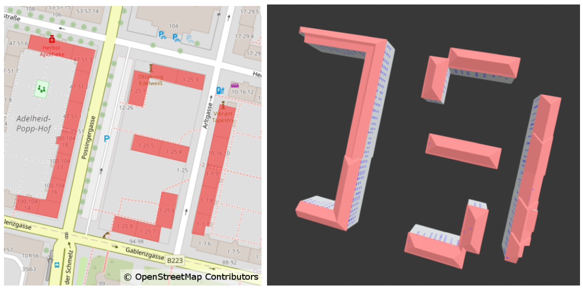
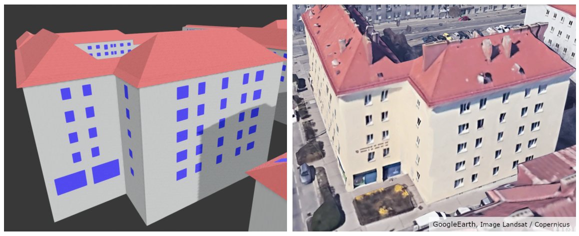


Figure 1. Full size oblique image (right side) with magnified section of one test site building



(a) (b)

Figure 2. Open Street Map of test site (a) and top view of building models (b)



(a) (b)

Figure 3. 3D view of building model (a) and textured triangulated *GoogleEarth* model (b)

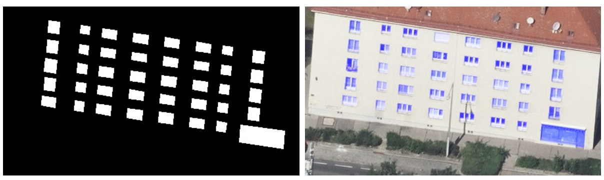
3. Methods

Detecting windows using object detection algorithms in image space is not sufficient for the enrichment of building models. As window boundaries generally do not align with the image grid, an additional segmentation step is required to accurately delineate their geometry. Once segmented, window masks—potentially derived from multiple images—can be vectorized and projected into object space, typically followed by a regularization step to ensure geometric consistency.

However, the subsequent modeling process is beyond the scope of this paper. Therefore, this study focuses on window detection and the evaluation of detection quality in image space (cf. Fig.

4) using different existing methods . As mentioned previously, three different deep learning-based strategies are compared to address this task.

1. Grounding DINO (Liu et al., 2023) is employed for zero-shot object detection, while Segment Anything Model 2 (SAM2) (Ravi et al., 2024) is used for subsequent segmentation. In this two-step approach, the bounding boxes predicted by the object detection model serve as input prompts for SAM2, enabling the precise delineation of window boundaries in the original oblique aerial images.
2. The same pipeline (Grounding DINO combined with SAM2) is also applied to facade orthophotos. In this case, windows are typically aligned with the image axes, such that the bounding boxes predicted by the object detector already provide a reasonable approximation of their geometry. This raises the question of whether an additional segmentation step is necessary for predominantly rectangular window shapes.
   In practice, however, the segmentation masks produced by SAM2 are generally more precise than the bounding boxes generated by Grounding DINO, particularly along object boundaries. Furthermore, SAM2 is capable of capturing non-rectangular or partially occluded openings more accurately. Applying the same processing strategy to both oblique images and facade orthophotos enables a direct and consistent comparison of detection performance.
   It is expected that Grounding DINO performs better on orthophotos, as training data for facade-related tasks are typically derived from orthorectified imagery, as discussed in the introduction.
3. SAM2-UNeXT (Xiong et al., 2025) is employed as a single-step segmentation approach. In this architecture, SAM2 serves as the encoder within a U-shaped segmentation network, while its representational capacity is further enhanced through the integration of an auxiliary DINOv2 encoder. This design enables the network to effectively combine general-purpose visual features with task-specific representations.
   The network requires comparatively little training data to achieve robust detection performance, making it well-suited for applications with limited annotated datasets.



(a) (b)

Figure 4. Ground truth mask (a) and overlay (b) of example image

4. Evaluation and results

Zero-shot object detection approaches, as used in Methods 1 and 2, offer the key advantage of not requiring task-specific training data. In these approaches, Grounding DINO is first applied to generate bounding boxes, which are subsequently used as prompts for SAM2 to obtain precise segmentation masks of the detected windows.

Method	F1	IoU	Precision	Recall
1) Zero-Shot	0.79	0.65	0.83	0.75
2) Zero-Shot Ortho	**0.85**	**0.75**	**0.95**	0.78
3) SAM2-UNeXT	0.72	0.57	0.69	**0.78**

Table 1. Average results of 5-fold validation

In contrast, SAM2-UNeXT follows a single-step segmentation strategy, directly predicting object masks without an intermediate detection stage. However, this approach requires training on annotated data to achieve reliable performance for the target object class.

For training the SAM2-UNeXT model, a total of 120 annotated ground truth images were available. These samples were extracted as image patches from the oblique aerial imagery and selected based on their viewing angles relative to the corresponding 3D building models. Due to the limited dataset size, only 12 images (10%) were reserved for evaluation. In general, a 5-fold cross-validation scheme was applied, with each test set further split into training and testing subsets. Following the recommendations of (Xiong et al., 2025), the model was trained for 20 epochs. Furthermore, a AdamW optimizer with an initial learning rate of 0.0001 and cosine learning rate decay to stabilize training was applied.

To assess the performance of the different methods multiple statistic measures using a 5-fold validation approach were determined (cf. Table 1). Considering the raw results, Method 2 (zero-shot object detection applied to orthophotos) achieves the best performance with a precision of 0.95 and an F1 score of 0.85. As discussed earlier, this outcome is expected, since such networks are typically trained on common datasets that predominantly contain facade orthophotos or terrestrial images. The unusual viewing direction of buildings in oblique image (cf. Fig. 1) reduce the performance of such networks. Preliminary analyses indicate that the shallower the viewing angle, the lower the window detection rate (cf. Fig. 5). A more thorough investigation is required, as it would necessitate a larger dataset categorized by incident viewing angle.

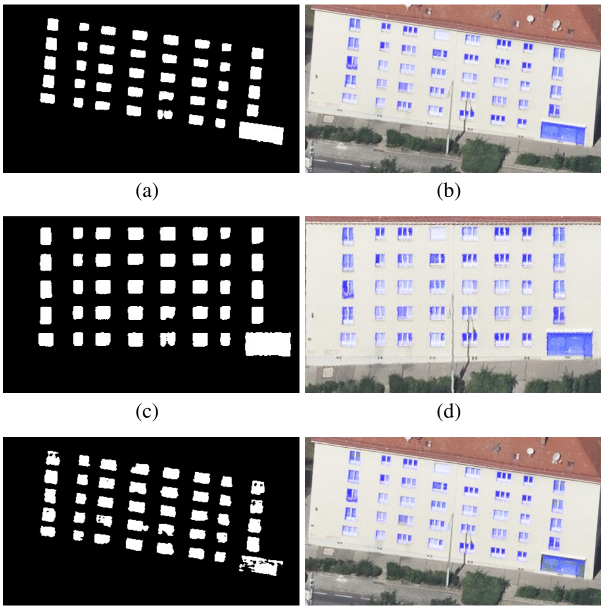
Our trained SAM2-UNeXT network has a significant lower precision (0.69) and F1 score (0.72) than method 2 but a minimal better recall (0.78). This effect can be seen in Fig. 6c. The bottom right large window is clearly detected, but not all pixels of the window area are correctly classified. The pixel-based statistical measures in Table 1 are certainly relevant; however, as noted earlier, window segmentation in image space is only the first step toward enriching 3D building models. Since regularization algorithms are required to produce uniform facade openings, higher recall may be more important than higher precision. A brief inspection of the SAM2-UNeXT results indicates that all windows were at least partially detected. Consequently, the final improved 3D building model could still achieve comparable accuracy. In general, the authors expect that a self-trained SAM2-UNeXT network could achieve better performance at shallower viewing angles, provided that suitable training datasets are available (cf. Fig. 5).

5. Discussion and Conclusion

In the following the research for improving results and the path from detection to modeling is discussed.

(a) (b) (c)

Figure 5. Reduced detection performance for shallow incident angles (*not* included in evaluation statistic); original image (a), zero-shot object detect (method 1, b) and SAM2-UNext (c)



(a) (b) (c) (d) (e) (f)

Figure 6. Mask and overlay results of zero-shot object detection (method 1, a & b), based on facade ortho images (method 2, c & d), and SAM2-UNeXT (method 3, e & f)

5.1 Viewing angle vs. occlusions for facade modeling

When using oblique aerial imagery for facade modeling, there is always a trade-off between a beneficial viewing angle and the risk of occlusion. As shown in Fig. 7 facades that are farther away from the camera position exhibit a less shallow (i.e., steeper) viewing angle. This will positively impact the object detection. However, increasing the distance may also introduce challenges such as reduced spatial resolution and greater susceptibility to occlusion from foreground objects.

In Fig. 7, the lowest row of windows is occluded by a foreground building. In such situations, it is essential to rely on existing building models that support visibility analysis, enabling the correct assignment of facade elements to their respective buildings. For this reason, previous research has employed textured LoD2 models (Hensel et al., 2019) or mobile sensing systems (Wysocki et al., 2023) to avoid or significantly reduce the issue of buildings occluding one another.

Future work should include a thorough investigation of the correlation between viewing angle and the detection performance of the proposed deep learning networks. By fine-tuning the models with specific near-nadir training data, it may be possible to improve detection performance for steeper viewing angles. This, in turn, could expand the range of viewing directions suitable for reliable object detection. For completeness, it should be noted that steeper viewing directions also affect the geometric accuracy of the enriched building models. Offsets in window detection, as well as inaccuracies in the building model itself, can lead to increasing displacement errors in the reconstructed facades as the viewing angle becomes steeper.

Figure 7. Better viewing angles often increase the risk of occlusion

5.2 From object detection to 3D modeling

Integrating the detected window segments into 3D building models represents the next step of this research. Combining observations from multiple images is expected to improve both completeness and geometric accuracy through multi-view redundancy and consistency constraints. The central idea is to move beyond reliance on predefined 3D building models and instead utilize 3D point clouds derived from LiDAR data or dense image matching to project and refine the detected segments in object space.

While this approach may be unnecessarily complex for planar features such as windows, it offers clear advantages for complex 3D structures, including dormers, balconies, solar panels, and chimneys. In particular, aerial imagery provides an ideal basis for modeling roof-based objects, especially when leveraging text-based deep learning models such as Grounding DINO.

Nowadays, a wide variety of annotated image datasets and 3D point clouds are available for training deep learning networks and developing building modeling algorithms. Typically, one must choose between working in image space or object space; however, identifying suitable hybrid datasets for addressing the aforementioned tasks remains challenging.

For example, the Technical University of Munich provides a comprehensive multi-sensor dataset of campus buildings, acquired using various platforms and designed to support a wide range of research applications (Wysocki et al., 2026). This dataset includes both, aerial imagery and 3D point clouds from a UAV-based measurement campaign (Anders et al., 2024). However, when high-quality image orientation and camera calibration are required to ensure consistency with the 3D point clouds, bundle block adjustment and orientation must still be performed by the user.

Therefore, the authors suggest that lens-distortion-free images, together with accurate image orientation parameters, should be provided to facilitate seamless integration between image and object space within the dataset.

5.3 Conclusion

The analysis of window detection in professional photogrammetric oblique images showed that current methods achieve accuracy between 70% and 80%. The best IoU of 0.75 was achieved for a combination of zero shot window detection using DINO and subsequent segmentation using SAM2 in orthorectified single facade images. While the use of specifically training the detection on oblique images suggests to be promising, we identified a lack of suitable training data. This also relates to the comparatively small number of buildings evaluated in this investigation.

These quality figures describe the performance in image space, but finally the quality will have to be evaluated in object space to provide results relevant for the exploitation of building models. It will be necessary to include other facade elements and go beyond the ‘hole in the facade’ definition of a window.

Acknowledgments

The authors thank the City of Vienna, Municipal Department 41 (MA41), for providing the image data for this manuscript.

References

Adilkhanova, I., Santamouris, M., Yun, G. Y., 2024. Green roofs save energy in cities and fight regional climate change. *Nature Cities*, 1(3), 238–249.

Anders, K., Wang, J., Wysocki, O., Huang, X., Liu, S., 2024. Uav laser scanning and photogrammetry of tum downtown campus (1.0). `https://doi.org/10.5281/zenodo.14548134`.

Haala, N., Kada, M., 2010. An update on automatic 3D building reconstruction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(6), 570-580. https://doi.org/10.1016/j.isprsjprs.2010.09.006. ISPRS Centenary Celebration Issue.

Hanke, F., Bieringer, A., Wysocki, O., Jutzi, B., 2025. CM2LoD3: Reconstructing LoD3 Building Models Using Semantic Conflict Maps. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-4/W6-2025, 81–88. https://isprs-annals.copernicus.org/articles/X-4-W6-2025/81/2025/.

Hensel, S., Goebbels, S., Kada, M., 2019. Facade Reconstruction for Textured LoD2 CityGML Models based on Deep Learning and Mixed Integer Linear Programming. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-2/W5, 37–44. https://isprs-annals.copernicus.org/articles/IV-2-W5/37/2019/.

Huang, J., 2025. Structure-aware 3D Building Reconstruction. PhD thesis, TU Delft.

Kolbe, T. H., Kutzner, T., Smyth, C. S., Nagel, C., Roensdorf, C., Heazel, C., 2021. OGC city geography markup language (CityGML) part 1: Conceptual model standard. https://docs.ogc.org/is/20-010/20-010.html.

Liu, H., Xu, Y., Zhang, J., Zhu, J., Li, Y., Hoi, S. C. H., 2020. DeepFacade: A Deep Learning Approach to Facade Parsing With Symmetric Loss. *IEEE Transactions on Multimedia*, 22(12), 3153-3165.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L., 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection.

Pantoja-Rosero, B., Achanta, R., Kozinski, M., Fua, P., Perez-Cruz, F., Beyer, K., 2022. Generating LOD3 building models from structure-from-motion and semantic segmentation. *Automation in Construction*, 141, 104430.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., Feichtenhofer, C., 2024. SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714*. https://arxiv.org/abs/2408.00714.

Wysocki, O., Schwab, B., Beil, C., Holst, C., Kolbe, T. H., 2024. Reviewing Open Data Semantic 3D City Models to Develop Novel 3D Reconstruction Methods. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-4-2024, 493–500. https://isprs-archives.copernicus.org/articles/XLVIII-4-2024/493/2024/.

Wysocki, O., Schwab, B., Biswanath, M. K., Greza, M., Zhang, Q., Zhu, J., Froech, T., Heeramaglore, M., Hijazi, I., Kanna, K., Pechinger, M., Chen, Z., Sun, Y., Segura, A. R., Xu, Z., AbdelGafar, O., Mehranfar, M., Yeshwanth, C., Liu, Y.-C., Yazdi, H., Wang, J., Auer, S., Anders, K., Bogenberger, K., Borrmann, A., Dai, A., Hoegner, L., Holst, C., Kolbe, T. H., Ludwig, F., Nießner, M., Petzold, F., Zhu, X. X., Jutzi, B., 2026. TUM2TWIN: Introducing the large-scale multimodal urban digital twin benchmark dataset. *ISPRS Journal of Photogrammetry and Remote Sensing*, 232, 810-830. https://doi.org/10.1016/j.isprsjprs.2025.12.013.

Wysocki, O., Xia, Y., Wysocki, M., Grilli, E., Hoegner, L., Cremers, D., Stilla, U., 2023. Scan2lod3: Reconstructing semantic 3d building models at lod3 using ray casting and bayesian networks.

Xiong, X., Wu, Z., Zhang, L., Lu, L., Li, M., Li, G., 2025. SAM2-UNeXT: An Improved High-Resolution Baseline for Adapting Foundation Models to Downstream Segmentation Tasks. *arXiv preprint arXiv:2508.03566*.