ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume X-M-2-2025
30th CIPA Symposium "Heritage Conservation from Bits:
From Digital Documentation to Data-driven Heritage Conservation", 25–29 August 2025, Seoul, Republic of Korea

Image-based Deep Learning Approaches for Point Cloud Classification for Heritage BIM Modelling

Aleksander Gil[1], Yusuf Arayici[1]

[1] Department of Architecture and Built Environment, Northumbria University, Newcastle-upon-Tyne, NE1 8ST
- (aleksander.gil, yusuf.arayici)@northumbria.ac.uk

Keywords: Heritage Building Information Modelling, Point Cloud Classification, Deep Learning, Image-Based Segmentation, Semantic Enrichment, Cultural Heritage

Abstract

This paper investigates the use of image-based deep learning methods to automate the segmentation and semantic classification of point clouds for Heritage Building Information Modelling (HBIM). In response to the limitations of classical machine learning approaches such as Random Forests, DBSCAN, and K-Nearest Neighbours, this paper proposes a hybrid pipeline combining 360° panoramic imagery with state-of-the-art computer vision models. The proposed solution leverages Meta's Segment Anything Model (SAM) for image-based segmentation and YOLO World for open-vocabulary object classification, with segmentation masks reprojected into 3D space to annotate point clouds data of heritage buildings.

Experiments were conducted on a high-resolution dataset from the Queen's House, Royal Museums Greenwich. Results show that SAM generalises well to equirectangular projections, particularly when applied to synthetic panoramas rendered from point clouds. YOLO World enhanced semantic labelling but showed reduced specificity in heritage contexts. The proposed hybrid pipeline produced spatially consistent and semantically enriched 3D segments, demonstrating potential for reducing manual labour in HBIM workflows.

Despite challenges related to projection ambiguity, occlusion, and semantic granularity, the research presented in the paper validates a novel paradigm for 3D heritage interpretation that fuses visual intelligence with geometric precision. With the results from the experiment presented in the paper, a future recommendation incorporating multi-view inputs, depth filtering, and ontology mapping is also provided to scale the approach toward practical HBIM adoption.

1. Introduction

Digital workflows are increasingly central to cultural heritage documentation and conservation, yet significant bottlenecks persist. International frameworks such as UNESCO's Sustainable Development Goals and the UN Urban Agenda 2030 emphasise the need for innovative technologies to support heritage safeguarding (Cotella, 2023). In response, 3D laser scanning and photogrammetry are now widely employed to record historic buildings, generating dense point clouds. However, these datasets remain scattered, disorganised and lack semantic richness (Zhao *et al.*, 2023), necessitating substantial manual input for meaningful interpretation.

This paper investigates a novel approach to automated segmentation and classification of point clouds for Heritage Building Information Modelling (HBIM), leveraging recent advances in image-based deep learning. Prior methods using classical machine learning such as Random Forests, RANSAC, DBSCAN, and K-Nearest Neighbours exhibit limitations in scalability, semantic fidelity, and generalisability. These constraints highlight the need for more effective techniques to reduce manual effort and enhance interpretative value. Automated semantic segmentation thus remains a critical challenge and bottleneck for HBIM, with implications for damage assessment, restoration planning, and beyond.

This study proposes using 360° panoramic images, derived from LiDAR scanner as intermediaries between 2D image segmentation and 3D point cloud analysis. It evaluates the performance of two foundation models: Meta's Segment Anything Model (SAM) for segmentation, and YOLO World for semantic classification. SAM, trained on over a billion masks, offers robust zero-shot generalisation, while YOLO World enables object detection using natural language labels. These models are applied to heritage datasets without the need for domain-specific training, enabling direct segmentation of architectural elements. Outputs are then reprojected onto the 3D point cloud to produce semantically enriched spatial models.

The methodology is demonstrated using datasets from the Queen's House at Royal Museums Greenwich via a four-stage pipeline: (i) segmentation with SAM, (ii) 3D back-projection, (iii) testing on real and synthetic panoramas, and (iv) semantic labelling using YOLO World. This proof-of-concept demonstrates the potential of 2D foundation models to enhance HBIM workflows, offering scalable solutions for point cloud interpretation, semantic enrichment, and improved interoperability across heritage systems.

The remainder of this paper is structured as follows: Section 2 reviews recent developments in image-based segmentation and AI-assisted heritage modelling. Section 3 synthesises key insights from previous classical machine learning experiments. Section 4 outlines the dataset and experimental methodology. Sections 5 and 6 present and discuss the findings respectively, and Section 7 concludes with reflections on limitations and future research directions.

2. Literature Review

Advances in deep learning have significantly enhanced image segmentation. Early methods based on handcrafted features and

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-M-2-2025-97-2025 | © Author(s) 2025. CC BY 4.0 License.

clustering algorithms struggled with the visual complexity typical of heritage contexts. The advent of Fully Convolutional Networks enabled end-to-end pixel-wise classification (Long, Shelhamer and Darrell, 2015), later refined by architectures such as U-Net (Ronneberger, Fischer and Brox, 2015) and DeepLab (Chen, Wei and Wang, 2018), which improved edge definition and detail retention. Subsequent innovations, including Mask R-CNN (He *et al.*, 2018) and vision transformers (Chen, Hsieh and Gong, 2022), further improved contextual understanding. Most notably, foundation models like Meta's Segment Anything Model (SAM), trained on over a billion masks now enable robust zero-shot generalisation across diverse object categories (Kirillov *et al.*, 2023). This is particularly pertinent in heritage domains where stylistic variability and limited annotated datasets make conventional supervised learning approaches infeasible (Remondino *et al.*, 2022).

Object detection has followed a parallel trajectory. The YOLO model family, particularly in later iterations (Wang, Bochkovskiy and Liao, 2022), has achieved a strong balance between inference speed and detection accuracy. YOLO World (Cheng *et al.*, 2024) marks a notable evolution, enabling open-vocabulary detection through integration with large-scale language models. This capacity to identify previously unseen object types is highly relevant in historic environments, where architectural features are often bespoke, undocumented, or stylistically hybrid (Croce, Caroti, Piemonte, *et al.*, 2021). YOLO-based approaches have been used to detect façade components (Zhao *et al.*, 2025) and diagnose structural pathologies (Chen, He and Wang, 2025) Likewise, SAM has recently been tested on indoor panoramic heritage imagery, successfully identifying architectural elements, albeit with performance degradation at image peripheries due to panoramic distortion (Zhong *et al.*, 2025).

Despite these advances, integration between 2D vision models and 3D point cloud workflows in heritage contexts remains limited. Cultural heritage documentation typically relies on laser scanning or photogrammetry, yielding dense but unstructured point clouds that underpin HBIM. Converting these into semantically structured BIM elements is a persistent challenge. Classical techniques have been applied to heritage datasets (Czerniawski *et al.*, 2018; Croce, Caroti, De Luca, *et al.*, 2021), but semantic richness is limited without extensive labelled training data. More recent deep networks, such as PointNet and SparseCNN (Huang *et al.*, 2022; Haznedar *et al.*, 2023), show improved performance but remain computationally demanding, domain-specific, and reliant on large, annotated datasets.

Recent reviews corroborate these limitations. Puerto *et al.*, (2024) identify a lack of generalisable automated segmentation methods suited to HBIM, with point-based networks struggling in ornamentally complex environments. Graph-based approaches applied to heritage datasets (Pierdicca *et al.*, 2020) show promise but have yet to produce BIM-ready outputs. PointNet applications, such as those by Haznedar *et al.*, (2023) on Gaziantep's heritage, required heavy augmentation for modest classification accuracy. Similarly, while Croce *et al.*, (2023) successfully classified façade imagery with 2D CNNs, these results have yet to be integrated into spatial models or HBIM workflows. A comprehensive review by Şentürk and Şimşek, (2024) confirms that no current AI system offers fully automated parametric object generation from 3D point clouds within heritage applications.

3. Key learnings from the Previous Machine Learning Experiments for Point Cloud Segmentation

Prior to adopting deep learning, a series of classical machine learning experiments were conducted to segment and classify point clouds from a representative heritage dataset. The evaluation of RANSAC, DBSCAN, HDBSCAN, K-Nearest Neighbours (KNN), and Random Forest (RF) revealed significant limitations in automation, generalisation, and semantic clarity, ultimately guiding the decision to explore image-based AI methods.

RANSAC, applied for planar segmentation, successfully extracted major flat surfaces such as walls and floors. However, it could not generalise to non-planar features like stairs, vaults, or decorative elements, and was prone to ambiguity in multi-planar environments. Its outputs were purely geometric, lacking semantic interpretation.

DBSCAN and HDBSCAN enabled clustering of discrete objects based on point density and scale. While effective in identifying standalone elements such as benches or sculptures, these algorithms required sensitive parameter tuning and often over-segmented continuous architectural features. HDBSCAN mitigated some tuning burdens but introduced its own subjective thresholds for defining cluster persistence.

K-Nearest Neighbours (KNN) was tested as a supervised classifier using manually labelled exemplars. Although simple to implement, its accuracy deteriorated in complex scenes due to proximity-based confusion (e.g., misclassifying adjacent paintings and walls) and its inability to scale to high-density point clouds.

Random Forests (RF) yielded the most promising supervised results. When trained on manually annotated segments, RF achieved high classification accuracy within the same room. However, it failed to generalise across spaces and required extensive feature engineering and labelled training data, limiting its automation potential.

Collectively, these trials underscored the need for a generalisable and semantically enriched approach. In particular, the lack of automated semantic interpretation and the labour-intensive training workflows prompted a shift toward deep learning models that leverage pre-trained visual understanding. This informed the adoption of panoramic image segmentation using SAM and YOLO World as explored in subsequent sections.

Consequently, a hybrid approach, where panoramic images mediate between 2D vision and 3D segmentation, offers a compelling middle ground. By leveraging robust pretrained models on panoramas, then projecting the results into 3D, it may be possible to semantically enrich point clouds with minimal manual labelling. This integration bridges the generality of modern computer vision with the precision demanded by HBIM workflows.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-M-2-2025-97-2025 | © Author(s) 2025. CC BY 4.0 License.

4. Materials and Methods

The research presented in the paper adopts an experimental approach to measure the effectiveness of image based Deep Learning classification systems.

4.1 Case Study Context: Royal Museums Greenwich

The Queen’s House in Greenwich, designed by Inigo Jones and completed in 1635, is recognised as the first fully classical building in England (Delman, 2021). Commissioned originally for Anne of Denmark and later completed under Queen Henrietta Maria, it exemplifies the application of Palladian architectural principles, symmetry, classical orders, and mathematical proportion in British architecture. The building now forms part of the Royal Museums Greenwich (RMG) estate and sits within the Maritime Greenwich UNESCO World Heritage Site, alongside the Royal Observatory and the National Maritime Museum (Smith, 2003).

Figure 1. Quens House, the case study building for point cloud segmentation.

Architecturally, the Queen’s House offers a unique mix of complex geometries, curved staircases, and decorative cornices, presenting an ideal testbed for the segmentation of heritage datasets. Its historical importance and preservation need underscore the relevance of advanced documentation methods such as HBIM. The dataset used in this study comprises a high-resolution terrestrial laser scan of the Queen’s House, totalling approximately 1.2 billion points. The scan includes interior and partial exterior coverage and is complemented by registered 360° panoramic images. These served as the visual input for segmentation and classification tasks, linking 2D and 3D representations.

4.2 Technical Specification

All experiments were conducted on a high-performance computing workstation specifically configured for processing large-scale 3D datasets and high-resolution image segmentation. The machine featured an AMD Ryzen Threadripper PRO 5965WX processor, operating at 3.8 GHz across 24 cores, enabling extensive parallelisation for computationally intensive tasks such as point cloud manipulation and segmentation. The system was equipped with 256 GB of ECC DDR4 RAM to support the simultaneous handling of multiple datasets and memory-heavy image processing operations. Deep learning inference, including the execution of Meta’s Segment Anything Model (SAM) and YOLO World, was accelerated using an NVIDIA RTX A6000 graphics processing unit, which provided 48 GB of dedicated VRAM. The use of a 4 TB NVMe solid-state drive ensured rapid input/output operations for large-scale laser scans and 360° imagery.

The core implementation environment was based on Python 3.10, with essential libraries including OpenCV for image manipulation, NumPy for array processing, and Matplotlib for visualisation. Segmentation was performed using the official ViT-h checkpoint of the SAM model, accessed through Meta’s GitHub repository, while semantic object detection was conducted using YOLO World, integrated via an open-vocabulary model based on CLIP embeddings. Visual inspection of results and validation of back-projected segments were performed using CloudCompare, which facilitated the overlay of segmentation outputs onto the original point clouds. In addition, synthetic panoramic images were rendered from the 3D scans using custom scripts, with camera calibration data derived from the original LiDAR metadata. This ensured a geometrically accurate correspondence between 2D image masks and their associated 3D coordinates during the reprojection stage.

4.3 Implementation Process Flow

The core of this research is a four-stage experimental workflow that investigates the feasibility of using image-based deep learning models for the segmentation and semantic classification of 3D point clouds in a heritage context. This workflow was designed to integrate 2D vision models with 3D geometric datasets, specifically those representing historically significant buildings. Each stage builds incrementally upon the preceding one to address key challenges in transitioning from raw point clouds to semantically enriched HBIM-ready outputs. The overall implementation plan is illustrated in Figure 2.

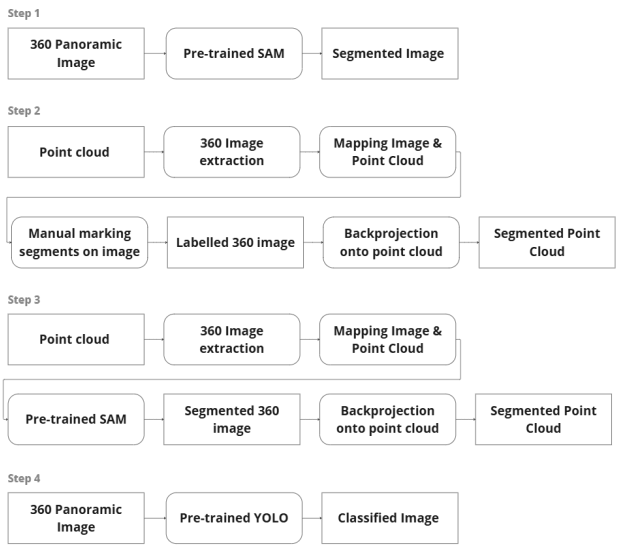

Figure 2. Experiment Implementation Workflow Plan

The workflow (pipeline) begins with applying a general-purpose segmentation model (SAM) to a panoramic image of the scene (Step 1). Next, the resulting 2D mask is *back-projected* into the 3D point cloud to test geometric alignment (Step 2). Step 3 evaluates the combined use of SAM segmentation and back-projection on both real and synthetic panoramic images to compare results. Finally, Step 4 introduces a semantic labelling step using the YOLO World detection framework to assign meaningful classes to the segments. Together, these steps form a pipeline that takes raw point cloud

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-M-2-2025-97-2025 | © Author(s) 2025. CC BY 4.0 License.

data and produces a segmented, labelled point cloud through the intermediate image processing. The detailed implementation of each step is described below.

4.3.1 Step 1: Feasibility of Meta SAM on Panoramic Images: The first stage explored the viability of using the Segment Anything Model (SAM), developed by Meta AI, to segment architectural features within panoramic images derived from laser scan data. SAM is a foundation model trained on over a billion masks and designed for general-purpose segmentation. The model was deployed using the ViT-h checkpoint within an open-source Python environment. The panoramic images used as input were formatted in equirectangular projection (4000×3000 pixels) and originated from both real scanner photography and synthetically rendered scenes based on point cloud data.

SAM's architecture comprises three modules: a Vision Transformer-based encoder that extracts image features, a prompt encoder that interprets spatial cues such as points and boxes, and a mask decoder that generates binary segmentation outputs. For this research, SAM was run in fully automatic mode, generating a grid of prompt points across the image to produce multiple candidate masks without manual intervention. The objective was to assess whether SAM, trained primarily on perspective photography, could identify structural and artefactual elements such as walls, doors, benches, and paintings in the highly distorted geometry typical of 360° imagery. The resulting masks were exported as both coloured overlays and individual binary mask files for further processing.

4.3.2 Step 2: Back-Propagation of Segmentation Results into 3D Space: Following segmentation, the second stage addressed the projection of the 2D segmentation results back into 3D space. This was achieved by exploiting the known geometric relationship between the panoramic images and the original point cloud from which they were derived. Each pixel in the 2D image has a corresponding 3D point in the LiDAR scan, allowing for a pixel-wise transfer of mask labels from the image to the point cloud.

The reprojection algorithm was implemented in Python using OpenCV and NumPy. For each binary mask, the image coordinate system was transformed back into spherical coordinates and then mapped to the appropriate 3D point. This mapping relied on the original scan metadata, including scanner position, rotation, and internal orientation parameters. Points corresponding to pixels within a given segmentation mask were assigned an identifier and coloured accordingly for visual validation.

This stage tested the geometric accuracy and viability of mask-to-point reprojection. Key challenges addressed included occlusion handling (i.e., ensuring points behind surfaces were not erroneously labelled), distortion compensation (due to equirectangular projection), and the granularity of segmentation boundaries. Visual validation was performed using CloudCompare to inspect whether segmented objects in the image corresponded to coherent point clusters in space.

4.3.3 Step 3: Integrated Segmentation and Reprojection: In the third step, the SAM-back-projection pipeline was applied to two types of panoramic inputs: real photographs captured by the laser scanner's onboard camera, and synthetic panoramas rendered from the 3D point cloud. The synthetic images were generated using a custom script that emulated the scanner's viewpoint and field of view, with uniform lighting, full surface coverage, and no optical artefacts.

The rationale for this comparison stemmed from the hypothesis that synthetic images, being free from exposure inconsistencies, glare, occlusions, and reflective surfaces, would yield more accurate segmentation results than real images. Both image types were processed through the same SAM model, and their respective segmentation masks were back-projected to generate labelled point clouds.

The results were analysed qualitatively by comparing the completeness, coherence, and boundary accuracy of the segmented objects. Special attention was given to difficult regions such as glossy floors, areas with shadow, and high-curvature elements like staircases and cornices. Quantitative analysis, where applicable, included point coverage per segment, mask fragmentation, and detection consistency across images. This step was instrumental in identifying whether synthetic imagery can act as a reliable intermediary for segmentation workflows, especially in heritage environments where real photographic conditions are often suboptimal.

4.3.4 Step 4: Semantic Classification with YOLO World: The final step introduced the semantic enrichment by applying YOLO World, a recent open-vocabulary object detection model that combines YOLO's real-time detection capabilities with CLIP-based semantic generalisation. Unlike traditional object detectors limited to fixed class labels, YOLO World can assign descriptors to previously unseen object types based on visual-linguistic embedding alignment. This is particularly valuable in heritage contexts where domain-specific elements such as statues, historical furniture, or bespoke architectural ornaments, may not appear in standard training datasets.

The same panoramic images used in SAM segmentation were input into the YOLO World model, which returned bounding boxes with associated labels and confidence scores. These outputs were then spatially matched with SAM-generated masks using intersection-over-union (IoU) heuristics. When a YOLO detection overlapped significantly with a SAM segment, the segment was assigned the corresponding class label. This process effectively merged pixel-wise segmentation with object-level semantic understanding, allowing point clusters to inherit meaningful labels such as "door", "painting", or "bench".

The resulting semantically segmented point cloud could then be used as a basis for further HBIM modelling or ontological mapping. Visual inspection of the output was again performed in CloudCompare, while misclassified or unlabelled segments were logged for discussion in the subsequent analysis section. This final stage demonstrated the potential of combining general-purpose image models to produce semantically structured 3D data in a scalable, largely automated manner.

5. Findings

The findings of this research are presented in alignment with the four experimental steps outlined in the implementation process flow (section 4.3). These results assess segmentation

completeness, reprojection fidelity, and semantic classification performance, with comparisons drawn between synthetic and real panoramic inputs.

5.1 Step 1 Findings

Meta’s Segment Anything Model (SAM) was successfully applied to both real and synthetic panoramic images without fine-tuning. On real scanner-derived imagery, SAM detected and delineated primary architectural elements such as walls, floors, ceilings, and large fixtures (Figure 3). It also produced masks for several freestanding objects including benches and framed paintings. However, segmentation quality was variable. Thin, reflective, or shadowed features such as lights, stair rails, or picture frames, were frequently omitted or fragmented. Curved ceiling mouldings, in particular, were only partially detected.

Figure 3. SAM Results on LiDAR Scanner Image.

SAM performed markedly better on synthetic panoramic inputs rendered from point cloud data (Figure 5). These images featured uniform lighting and controlled exposure, reducing visual noise. Segmentation masks derived from synthetic images were spatially complete and topologically consistent. The SAM model generated few fragmented regions, and large elements like walls and floors appeared as single, cohesive masks. Notably, few “unclassified” segments were generated in the synthetic case, confirming the hypothesis that image quality and visual uniformity directly affect segmentation reliability.

5.2 Step 2 Findings

Back-projection of 2D segmentation masks into the 3D point cloud validated the geometric accuracy of the reprojection logic. Using known camera positions and the panorama-to-point correspondence map, each SAM mask was successfully assigned to subsets of 3D points. As shown in Figure 4, large surfaces such as a wall-mounted painting or a gallery bench were clearly visible within the 3D cloud once back-projected, retaining the spatial coherence of the original mask.

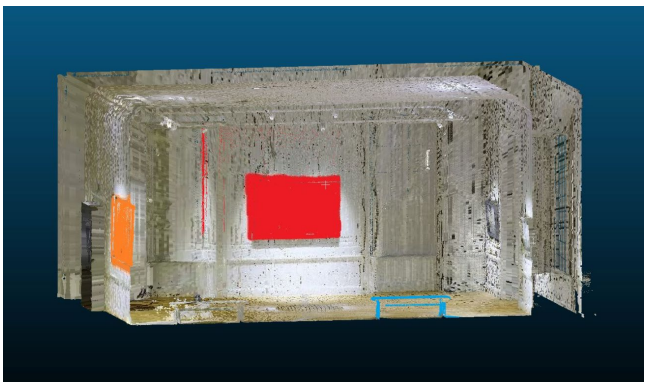

Figure 4. Back Propagation Results.

Nevertheless, some errors were introduced during projection. Due to the lack of depth-awareness in 2D segmentation, SAM masks occasionally "bled" onto background surfaces aligned along the same line of sight. For instance, lights suspended from the ceiling projected onto the ceiling surface itself. This issue was particularly evident in areas with overlapping geometries. Furthermore, a recurring artefact, the so-called “scanning circle”, was identified as a false segment. This refers to the circular void directly below the scanner, which appeared dark in the panorama and was often segmented as an object, despite corresponding to an absence of data.

5.3 Step 3 Findings

The comparative analysis of segmentation results between real and synthetic panoramas highlighted significant improvements in the latter. As illustrated in Figure 5, the synthetic input resulted in more unified and complete mask coverage than sole real input. For example, the gallery floor was segmented as a single contiguous region in the synthetic case, whereas it was fragmented across three segments in the real image. Similarly, SAM generated over 25 distinct mask regions on the real panorama, including many small unlabelled artefacts, compared to only 16 on the synthetic equivalent—indicating a 36% reduction in fragmentation.

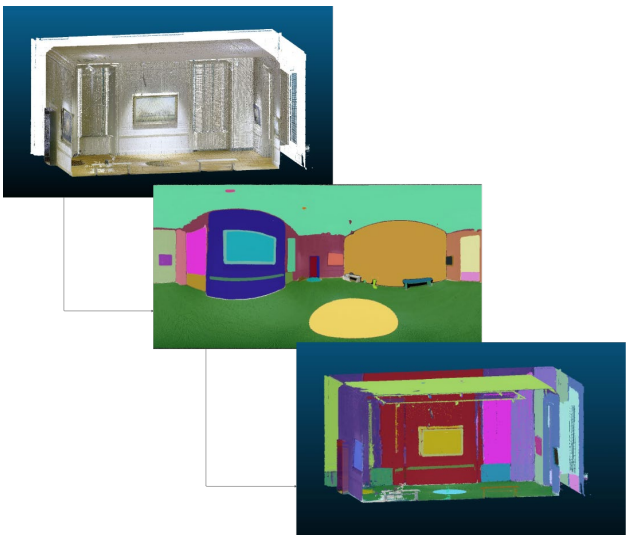

Figure 5. Segment Anything Meta Point Cloud Results.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-M-2-2025-97-2025 | © Author(s) 2025. CC BY 4.0 License.

Unclassified pixel coverage was also reduced by approximately 20% in the synthetic case. Shadows, specular reflections on floors and framed artwork, and lighting inconsistencies in the real panorama were identified as causes for mask fragmentation and omission. These effects were absent in the synthetic image, which suggests that point-cloud-derived panoramic imagery may serve as a robust input for point cloud segmentation in heritage.

5.4 Step 4 Findings

Semantic classification using YOLO World provided a critical interpretive layer. On both image types, YOLO accurately identified large structural elements such as walls, ceilings, floors, doors, windows, and seating as shown in Figures 6 and 7. Additionally, the model recognised several smaller objects such as paintings and sculptures, often labelling them as “picture”, “art”, or “sculpture” based on learned visual-linguistic embeddings.

Detection reliability improved in the synthetic image case. All four ceiling-mounted lights, missed in the real panorama, were correctly classified in the synthetic image. This is likely due to enhanced contrast and spatial separation in the synthetically rendered image. YOLO World also demonstrated increased classification confidence: the average confidence score for correctly labelled objects was 0.58 in the synthetic case, compared to 0.46 in the real case.

Processing times for YOLO detection were also observed. For 4000×3000-pixel inputs, semantic detection on real panoramas averaged 90 seconds per image, whereas synthetic images processed in under 50 seconds. This performance improvement is attributed to reduced visual complexity and texture uniformity in the synthetic panoramas.

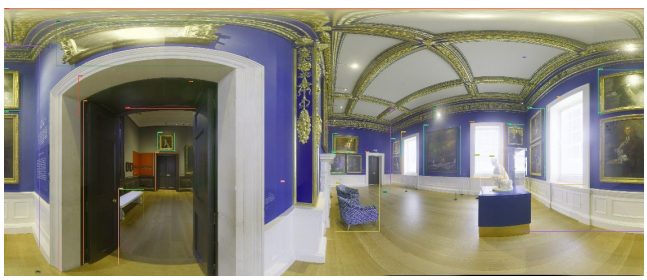

Figure 6. YOLO Results on LiDAR laser scanner panoramic image.

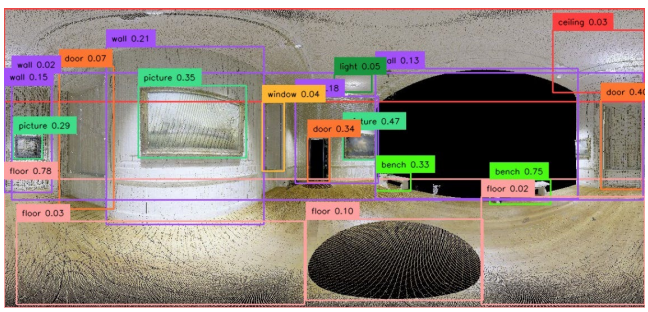

Figure 7. YOLO Result on Synthetically created panoramic image.

A confidence threshold of 30–40% yielded the optimal balance between false positives and false negatives. Lower thresholds introduced spurious labels (e.g. a statue misclassified as “person”), while thresholds above 50% missed smaller valid objects. Once bounding boxes were obtained, they were aligned with SAM mask regions using intersection-over-union (IoU) metrics. Each SAM mask was assigned the YOLO label with the highest spatial overlap, allowing class descriptors to propagate into the 3D point cloud.

6. Discussion

This research has established a proof-of-concept for combining image-based deep learning methods with 3D point cloud data to automate segmentation and semantic labelling processes within Heritage Building Information Modelling (HBIM) workflows. The image based deep learning implementation process flow, comprising SAM segmentation, 2D-to-3D reprojection, comparative testing of real and synthetic panoramas, and semantic integration via YOLO World, has surfaced a range of practical insights. This section critically analyses those findings, acknowledging both the contributions and constraints of the proposed pipeline approach, and positions the study within a wider framework for future HBIM automation.

A key contribution of the study is the demonstration that foundation vision models trained on generic 2D imagery can be repurposed, with minimal adaptation, to segment architectural features in panoramic representations of heritage interiors. The Segment Anything Model (SAM) exhibited strong generalisation, effectively delineating large surfaces and furnishing boundaries, even within highly distorted equirectangular projections. Notably, SAM achieved this without fine-tuning, indicating that the geometric abstraction of architectural elements is well-aligned with the image features these models are trained to recognise.

However, segmentation quality was heavily influenced by the nature of the input imagery. SAM performed significantly better on synthetic panoramas than on real panoramic photographs. Synthetic views, rendered from the point cloud under controlled lighting and perspective, reduced issues such as glare, exposure imbalance, and occlusion shadows. This finding supports the proposition that synthetic image generation can act as a valuable pre-processing step, increasing the fidelity and coherence of AI-driven segmentation, especially in complex heritage environments. The cleaner output also facilitated accurate back-projection into 3D, resulting in large, complete, and spatially consistent segment clusters.

The 2D-to-3D back-projection stage confirmed that a direct image-to-point mapping is computationally efficient and geometrically valid. Nonetheless, it also exposed important limitations. Chief among these was depth ambiguity: because panoramic images lack inherent depth awareness, projected segment masks sometimes “bled” onto surfaces behind the target object, especially in scenes with nested or overlapping geometries. One symptomatic error was the mislabelling of ceiling planes behind light fixtures. Another artefact, the “scanning circle” (a dark void beneath the scanner), was consistently interpreted by SAM as a valid object, indicating that models can misread data acquisition artefacts as physical elements. These issues suggest that naïve reprojection from image masks must be complemented by depth filters or visibility constraints to improve reliability.

The introduction of YOLO World in the final stage provided crucial semantic enrichment. The model’s open-vocabulary detection capabilities allowed it to assign meaningful class

labels to detected objects, bridging the semantic gap between geometric segmentation and HBIM ontology. YOLO World operated reliably on synthetic panoramas, showing improved performance in identifying subtle or high-level features such as sculptures, lights, or picture frames. Detection confidence was optimised within a 30–40% threshold, and spatial mapping of YOLO outputs to SAM segments proved effective in generating labelled 3D content. However, label granularity remained a challenge: while "bench" and "door" were consistently identified, specialised heritage terms, such as "pediment", "cornice", or "architrave", were either misclassified or absent entirely. This exposes a limitation of relying solely on generic language-image models in a domain that demands high-resolution architectural taxonomy.

Processing times for SAM (~60 seconds/panorama) and YOLO World (~50–90 seconds/panorama) were reasonable for prototype purposes but may prove prohibitive for large-scale deployment across hundreds of spaces. The results therefore highlight the need for workflow optimisation, potentially through model compression, parallelised computation, or more efficient segmentation proxies.

The most valuable outcome of this research is the validation of a novel architectural segmentation paradigm: one that shifts from traditional geometry-centric heuristics towards visual-semantic hybridisation. In this paradigm, foundational models provide object-awareness via learned priors, and point clouds supply geometric context. This approach introduces a scalable, replicable method for interpreting heritage environments with reduced human input, an increasingly critical need in conservation surveying, archival digitisation, and pre-refurbishment planning.

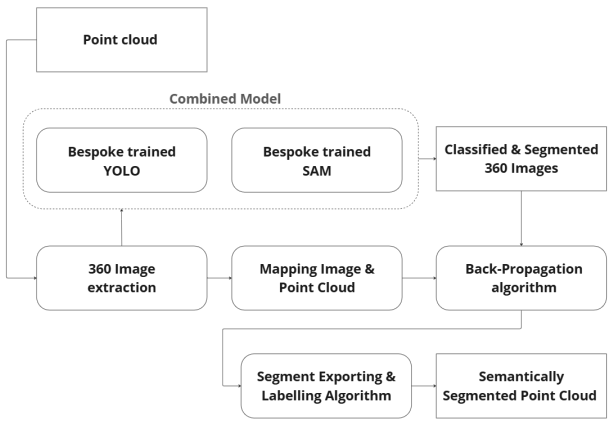

Figure 8. Proposed future investigation pipeline for image-based segmentation.

Looking forward, Figure 8 encapsulates a proposed pipeline to address current limitations and scale the approach. Central to this future pipeline is the multi-view segmentation strategy, wherein multiple panoramas are either captured or synthetically rendered per space. This would mitigate occlusion artefacts and enhance spatial redundancy. These images would be processed using efficient variants of SAM and YOLO World, potentially fine-tuned for architectural contexts. Segment outputs would be merged through depth-informed fusion techniques to ensure consistency across views.

The framework also integrates depth-aware back-projection algorithms, capable of occlusion filtering using Z-buffer logic or range maps, to restrict label propagation to visible surfaces. Such improvements would correct for "label bleeding" and enable the clean layering of semantic information onto occlusion-heavy scenes. Downstream, ontology mapping modules would interpret YOLO-derived labels into structured vocabularies aligned with IFC or Uniclass systems. These enriched outputs could then be automatically exported as object clusters or parametric placeholders for BIM authoring platforms.

This research confirms that the application of image based deep learning point cloud segmentation models to heritage datasets is not only feasible but also strategically advantageous. The integration of image-based segmentation with point cloud reprojection and open-vocabulary classification creates a promising pathway toward semi-automated HBIM generation. While refinement is needed, particularly around label precision, occlusion handling, and computational scalability, the methodological direction is sound. With the continued development, the proposed solution in the paper offers a scalable approach to semantic data enrichment in digital heritage practice, bridging the gap between visual perception and spatial modelling.

7. Conclusion

This study presents a technically grounded proof of concept for integrating image-based deep learning models into the segmentation and semantic labelling of 3D point cloud data within a heritage context. By combining the Segment Anything Model (SAM) for region extraction with YOLO World for object classification, the research proposes a hybrid pipeline (workflow) that bridges visual understanding and spatial precision, contributing to emerging approaches for automating Heritage Building Information Modelling (HBIM).

The proposed hybrid pipeline comprised four steps: i) segmentation of panoramic images using SAM, ii) reprojection of image masks into the point cloud, iii) comparative testing of synthetic versus real panoramas, and iv) semantic enrichment using YOLO World. The results confirmed that both models, despite being trained on conventional 2D datasets, could generalise effectively to panoramic representations of heritage interiors. In particular, synthetic images yielded more coherent and complete segmentations than real-world photographs, while semantic alignment via YOLO World facilitated the labelling of point cloud clusters with object-level descriptors such as "door", "bench", and "painting".

Nevertheless, the study also revealed critical limitations. The lack of depth-awareness in SAM led to projection artefacts, including mislabelling of occluded surfaces and the introduction of non-existent features such as the scanning circle. YOLO World, though semantically rich, struggled with fine-grained architectural classifications and occasionally produced inconsistent results in areas of image distortion. Both models were originally trained on large-scale, rectilinear datasets and were not optimised for the radial distortions or occlusion patterns inherent in 360° indoor imagery.

Addressing these constraints will require several technical extensions. Depth-informed back-projection, multi-view consistency enforcement, and the adoption of panoramic-aware detection models, such as Omnidirectional YOLO, may enhance

spatial accuracy and label robustness. Furthermore, mapping YOLO outputs to structured HBIM ontologies such as IFC or Uniclass could improve semantic relevance and facilitate integration into BIM environments. Incorporating domain-specific rules or knowledge graphs may also support the hierarchical labelling of architectural elements (e.g., distinguishing an “interior wall” from an “original load-bearing wall”), aligning outputs with heritage interpretation goals.

Looking forward, the future workflow outlined in Figure 8 offers a scalable pathway for automating HBIM. It envisions a multi-view, depth-aware, and semantically structured system in which panoramic inputs are processed through lightweight and tuned vision models, generating interoperable 3D models ready for downstream applications in conservation, refurbishment, and digital archiving.

In conclusion, this research demonstrates that fusing state-of-the-art image segmentation with spatially accurate point cloud data represents a promising strategy for automating the generation of semantically enriched HBIM datasets. While challenges remain in model adaptation, semantic depth, and computational scaling, the foundation laid in the paper provides a practical and forward-compatible approach to advancing digital heritage workflows.

References

Chen, L., Wei, G. and Wang, Z. (2018). ‘PointAGCN: Adaptive Spectral Graph CNN for Point Cloud Feature Learning’, in *2018 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC). 2018 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, Jinan, China: IEEE, pp. 401–406. Available at: https://doi.org/10.1109/SPAC46244.2018.8965522.

Chen, X., He, J. and Wang, S. (2025). ‘Deep learning-driven pathology detection and analysis in historic masonry buildings of Suzhou’, *npj Heritage Science*, 13(1). Available at: https://doi.org/10.1038/s40494-025-01783-y.

Chen, X., Hsieh, C.-J. and Gong, B. (2022). ‘When Vision Transformers Outperform ResNets without Pre-training or Strong Data Augmentations’. arXiv. Available at: https://doi.org/10.48550/arXiv.2106.01548.

Cheng, T. *et al.* (2024). ‘YOLO-World: Real-Time Open-Vocabulary Object Detection’. arXiv. Available at: https://doi.org/10.48550/arXiv.2401.17270.

Cotella, V.A. (2023). ‘From 3D point clouds to HBIM: Application of Artificial Intelligence in Cultural Heritage’, *Automation in Construction*, 152, p. 104936. Available at: https://doi.org/10.1016/j.autcon.2023.104936.

Croce, V., Caroti, G., Piemonte, A., *et al.* (2021). ‘From survey to semantic representation for Cultural Heritage: the 3D modeling of recurring architectural elements’, *ACTA IMEKO*, 10(1), p. 98. Available at: https://doi.org/10.21014/acta_imeko.v10i1.842.

Croce, V., Caroti, G., De Luca, L., *et al.* (2021). ‘From the Semantic Point Cloud to Heritage-Building Information Modeling: A Semiautomatic Approach Exploiting Machine Learning’, *Remote Sensing*, 13(3), p. 461. Available at: https://doi.org/10.3390/rs13030461.

Croce, V. *et al.* (2023). ‘H-BIM and Artificial Intelligence: Classification of Architectural Heritage for Semi-Automatic Scan-to-BIM Reconstruction’, *Sensors*, 23(5), p. 2497. Available at: https://doi.org/10.3390/s23052497.

Czerniawski, T. *et al.* (2018). ‘6D DBSCAN-based segmentation of building point clouds for planar object classification’, *Automation in Construction*, 88, pp. 44–58. Available at: https://doi.org/10.1016/j.autcon.2017.12.029.

Delman, R. (2021). ‘The Queen’s House before Queen’s House: Margaret of Anjou and Greenwich Palace, 1447-1453’, *Royal Studies Journal*, pp. 6–25.

Haznedar, B. *et al.* (2023). ‘Implementing PointNet for point cloud segmentation in the heritage context’, *Heritage Science*, 11(1), p. 2. Available at: https://doi.org/10.1186/s40494-022-00844-w.

He, K. *et al.* (2018). ‘Mask R-CNN’. arXiv. Available at: https://doi.org/10.48550/arXiv.1703.06870.

Huang, H.S. *et al.* (2022). ‘FROM BIM TO POINTCLOUD: AUTOMATIC GENERATION OF LABELED INDOOR POINTCLOUD’, *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B5-2022, pp. 73–78. Available at: https://doi.org/10.5194/isprs-archives-XLIII-B5-2022-73-2022.

Kirillov, A. *et al.* (2023). ‘Segment Anything’. arXiv. Available at: https://doi.org/10.48550/arXiv.2304.02643.

Long, J., Shelhamer, E. and Darrell, T. (2015). ‘Fully Convolutional Networks for Semantic Segmentation’. arXiv. Available at: https://doi.org/10.48550/arXiv.1411.4038.

Pierdicca, R. *et al.* (2020). ‘Point Cloud Semantic Segmentation Using a Deep Learning Framework for Cultural Heritage’, *Remote Sensing*, 12(6), p. 1005. Available at: https://doi.org/10.3390/rs12061005.

Puerto, A. *et al.* (2024). ‘Building information modeling and complementary technologies in heritage buildings: A bibliometric analysis’, *Results in Engineering*, 22, p. 102192. Available at: https://doi.org/10.1016/j.rineng.2024.102192.

Remondino, F. *et al.* (2022). ‘AERIAL TRIANGULATION WITH LEARNING-BASED TIE POINTS’, *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2022, pp. 77–84. Available at: https://doi.org/10.5194/isprs-archives-xliii-b2-2022-77-2022.

Ronneberger, O., Fischer, P. and Brox, T. (2015). ‘U-Net: Convolutional Networks for Biomedical Image Segmentation’. arXiv. Available at: https://doi.org/10.48550/arXiv.1505.04597.

Şentürk, H.S. and Şimşek, C.F. (2024). ‘A review on HBIM modelling process from 3D point clouds by applying artificial intelligence algorithms in cultural heritage’, *Politeknik Dergisi*, pp. 1–1. Available at: https://doi.org/10.2339/politeknik.1503631.

Smith, R. (2003). ‘Core Cases in Critical CareLondon: Greenwich Medical Media’, *British Journal of Anaesthetic and Recovery Nursing*, 4(2), pp. 6–6. Available at: https://doi.org/10.1017/S1742645600001790.

Wang, C.-Y., Bochkovskiy, A. and Liao, H.-Y.M. (2022). ‘YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors’. arXiv. Available at: https://doi.org/10.48550/arXiv.2207.02696.

Zhao, J. *et al.* (2023). ‘A REVIEW OF POINT CLOUD SEGMENTATION OF ARCHITECTURAL CULTURAL HERITAGE’, *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-1/W1-2023, pp. 247–254. Available at: https://doi.org/10.5194/isprs-annals-x-1-w1-2023-247-2023.

Zhao, J. *et al.* (2025). ‘A deep learning-based study on visual quality assessment of commercial renovation of Chinese traditional building facades’, *Environmental Impact Assessment Review*, 113, p. 107862. Available at: https://doi.org/10.1016/j.eiar.2025.107862.

Zhong, D. *et al.* (2025). ‘OmniSAM: Omnidirectional Segment Anything Model for UDA in Panoramic Semantic Segmentation’. arXiv. Available at: https://doi.org/10.48550/arXiv.2503.07098.