COMBINING IMAGE AND POINT CLOUD SEGMENTATION TO IMPROVE HERITAGE UNDERSTANDING

Maarten Bassier[a], Gabriele Mazzacca[b,c], Roberto Battisti [b], Salim Malek [b], Fabio Remondino[b]

[a] Dept. of Civil Engineering – Geomatics, KU Leuven – Faculty of Engineering Technology, Ghent, Belgium
Email: maarten.bassier@kuleuven.be

[b] 3D Optical Metrology (3DOM) unit, Bruno Kessler Foundation (FBK), Trento, Italy
Email: <gmazzacca><rbattisti><smalek><remondino>@fbk.eu

[c] Dept. Mathematics, Computer Science and Physics, University of Udine, Italy

Commission II

KEY WORDS: 3D classification, heritage, deep learning, joint semantic segmentation, fusion methods

ABSTRACT:

Current 2D and 3D semantic segmentation frameworks are developed and trained on specific benchmark datasets, often rich of synthetic data, and when they are applied to complex and real-world heritage scenarios they offer much lower accuracy than expected. In this work, we present and demonstrate an early and late fusion of methods for semantic segmentation in cultural heritage applications. We rely on image datasets, point clouds and BIM models. The early fusion utilizes multi-view rendering to generate RGBD imagery of the scene. In contrast, the late fusion approach merges image-based segmentation with a Point Transformer applied to point clouds. Two scenarios are considered and inference results show that predictions are primarily influenced by whether the scene has a predominantly geometric or texture-based signature, underscoring the necessity of fusion methods.

1. INTRODUCTION

The semantic segmentation of constructions encompasses the segmentation of primary, secondary, and auxiliary building classes, as noted in the reference (Armeni et al., 2017). This segmentation is an intermediate step crucial for detecting different instances of elements within buildings, a requirement for various tasks such as scan-to-BIM procedures and building enrichment pipelines, among others (Croce et al., 2023). Prior to 2020, traditional machine learning methods, along with specific features, were the preferred choice. However, the state of the art has now completely shifted towards deep learning methods, as evident in references (Bello et al., 2020, Guo et al., 2021). These deep learning techniques generally offer improved generalization and reduce the need for feature engineering, such as radiometric feature extraction. Nonetheless, they do demand a significantly larger amount of training data to achieve similar detection rates.

Currently, most semantic segmentation approaches still primarily focus on a single modality, which could be either imagery or point cloud data (Coudron et al., 2020). This bias toward single-modality methods is largely due to benchmark datasets that predominantly promote such competitions (Armeni et al., 2017) or limitation in processing methods. However, these single-modality approaches fall short in achieving market-ready detection rates, particularly when dealing with objects of heritage that exhibit intricate geometries and textures. For example, identifying different types of columns in a heavily eroded setting can greatly benefit from both visual and geometric interpretations, even when the latter might introduce noise. Multi-modal data fusion in machine learning is a growing sector (Townend et al., 2024) and some recent works started also to introduce background knowledge into the neural network's learning pipelines (Grilli et al., 2023).

In our work, we propose a framework that integrates image and point cloud segmentation techniques for cultural heritage building elements. To achieve this, we have developed an integration pipeline that combines state-of-the-art methods for semantic segmentation of both images and point clouds. In summary, our contributions include:

1. The theoretical framework and implementation for early and late image and point cloud semantic segmentation.
2. An implementation for automated image and point cloud training sample production.
3. An empirical study on two heritage assets to compare the proposed joint semantic segmentation.

2. RELATED WORK

Heritage Semantic Segmentation - Researchers have been exploring the application of machine learning techniques for the semantic enrichment of 3D point clouds in the cultural heritage field for some time now (Fiorucci et al., 2020, Yang et al., 2023). Supervised machine learning methods have primarily focused on mapping various materials, building techniques, and deterioration phenomena. Leveraging the geometric characteristics of 3D data (Weinmann et al., 2015), these methods utilize extracted geometric features, and sometimes sensor-based features, to train machine learning algorithms to perform their tasks (Grilli et al., 2018, Grilli and Remondino, 2020, Croce et al., 2021). Despite the potential of these approaches, even the field of cultural heritage has seen an increasing change in research interest towards deep learning methods due to their noteworthy improvements in performing the semantic enrichment of 3D point clouds(Pierdicca et al., 2020, Matrone et al., 2020).

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-2-W4-2024-49-2024 | © Author(s) 2024. CC BY 4.0 License.

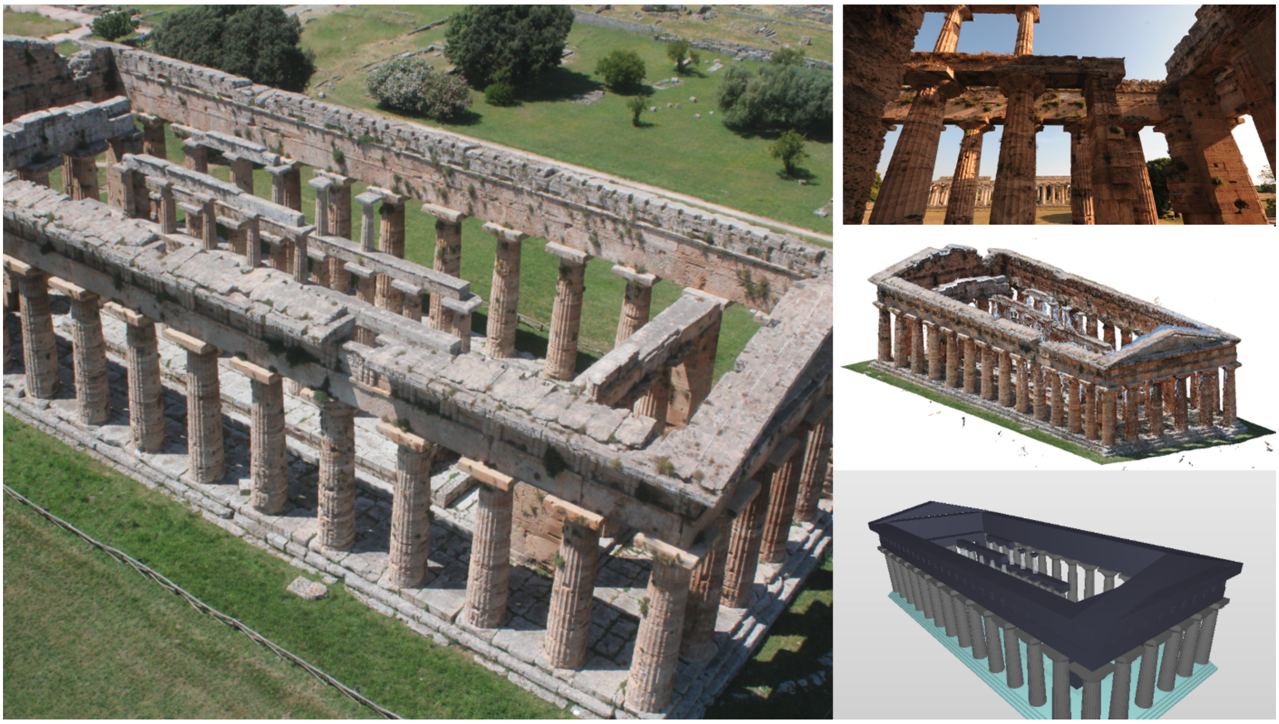

Figure 1: Overview of the project inputs: (left) hand-held and UAV images, point clouds and BIM model.

Joint point cloud and image segmentation - Joint point cloud and image segmentation is a popular topic in deep learning fusion approaches (Qi et al., 2021). Both data fusion (early) and method fusion (late) are predominantly pursued in academia. Early fusion methods, for instance, involve rendering 3D information as multi-view 2D images with an additional depth channel (RGBD), which can then be processed by standard 2D convolutions (Cui et al., 2022) (MVCNN). Alternatively, 2D images can be rendered as a 3D graph, tree, or raster point cloud representation (Lu et al., 2022). However, these 2D methods often lose some 3D geometric context and struggle with per-point label prediction. Recent advancements in MVCNN networks include ShapeConv (Cao et al., 2021) and FPS-Net (Xiao et al., 2021). On the other hand, late fusion combines the outputs of multiple networks and averages the results, for example, by integrating Point Transformers (Lu et al., 2022) with purely image-based networks. The advantage here is that each modality can be trained separately, leveraging numerous available benchmarks. However, the averaging of results is typically suboptimal and does not consider the quality of the geometric/texture signature at that location. Beyond early and late fusion, there are hybrid solutions that combine the strengths of both approaches (Zhang et al., 2021). These typically involve the use of intermediate fusion blocks that enable the parallelization of different networks, merging them at strategic points.

3. METHODOLOGY

3.1 Training data production

The successful integration of Deep Learning (DL) methods into heritage projects is fundamentally linked to the automated generation of training and testing data. In our work, we aim for a seamless transition between IFC BIM models, geolocated images produced by a photogrammetric pipeline, and the combined point cloud resulting from both photogrammetry and 3D terrestrial laser scanning (Fig. 1). To effectively train semantic segmentation models, it is crucial to amalgamate information from these three sources, thereby generating the necessary training data. First, there is the choice of the initial modality to which training labels are assigned. Generally, labelling 3D objects is more efficient than 2D formats, as images in photogrammetric pipelines typically have over 60% overlap. Among the 3D formats, the IFC model is much more efficient to label due to the limited number of elements. Additionally, IFC models already contain metadata that can be utilized for the production of training data. For instance, in the first test case (Section 4.1), the IFC model comprises only 214 elements across four types of building elements, while it has a point cloud of 56 million points and 894 24MP images.

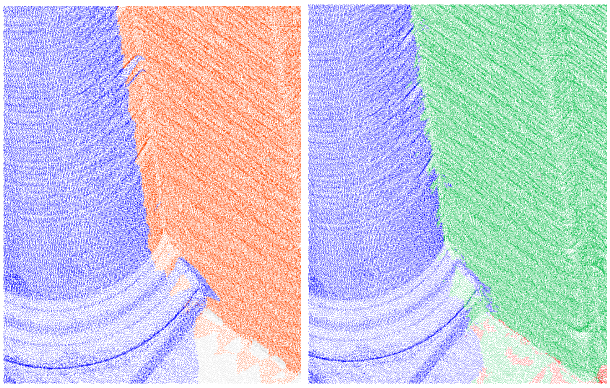

Figure 2: Overview of the transfer of 3D semantic labels from the IFC model to the 3D point cloud: without normal filtering, showing poor results near edges (left) and with normal filtering for a more nuanced segmentation (right).

3D Point Cloud Annotation - Thus in the initial phase of training data creation, the BIM information is associated with the point cloud data. The annotation of BIM information onto the point cloud data, denoted as P, relies on a nearest neighbours variant involving a uniformly sampled BIM point cloud, represented as Q. Given the substantial abstractions present in the BIM, the criterion for assigning information is determined by the difference in normals between a source point $p_i \in P$ and a set of neighbouring BIM points $Q_j \subset Q$, as expressed in Equation 1.

$$\begin{aligned} \boldsymbol{Q_j} &= \left\{ q_i \middle| p_i \in P, q_j \in Q : \|p_i - q_j\| \leq t_d \right\} \\ q_j &= \left\{ q_i \middle| p_i \in P, q_j \in Q_j : argmax_{q_j} | \overrightarrow{n(p_i)} \cdot \overrightarrow{n(q_j)} | \right\} \end{aligned} \tag{1}$$

In this context, Q represents the joint visibility point cloud, which is obtained by sampling points from the BIM objects. However, points q_j that are situated within neighbouring objects are removed, up to a specified threshold. The sets $\boldsymbol{Q_j}$ consist of points that are in close proximity to every p_i, determined by the Euclidean distance threshold t_d. To find the best fit q_j for each p_i, a maximization process is applied to the dot product between the two normals, represented as $\overrightarrow{n(p_i)}$ and $\overrightarrow{n(q_j)}$. As shown in Fig. 2, it is evident that the normal filtering improves the fit between the BIM and the point cloud annotation without significantly increasing computational complexity. Subsequently, the class information of the object that q_j belongs to is transferred to p_j as an additional point label.

2D Image Annotation - The IFC or point cloud data are used to automatically label the imagery. Operating on the full imagery has a major advantage as it has significantly higher detailing (ranging from 12 to 40 megapixels, resulting in avg. 0.002 m ground sampling distance - GSD) than the point cloud (avg. density of about 0.005 m). The training data for the image classification is automatically derived from the manually annotated point cloud. Firstly, the images are undistorted using OpenCV, utilizing the intrinsic camera matrices K for each image. Subsequently, each image is subdivided into pixel regions in accordance with the requirements of the image classification model. Next, a set of depth maps denoted as D is generated. This is achieved by performing a dense ray tracing of the photogrammetric point cloud for each image, utilizing the extrinsic camera matrices M for each image (Fig. 3 left, Equation 2).

$$\mathbf{D} = \left\{ D \middle| I \in \mathbf{I} : D = MKP \right\} \tag{2}$$

However, raycasting on the original point cloud is not ideal due to its limited density. Rays tend to pass in between points, resulting in labels for objects situated behind the initial layer of points (Fig. 3 middle). Instead, we adopt an alternative approach by generating a voxel mesh from the octree representation of the point cloud. By enhancing the voxel traversal mechanisms available in Open3D, we can create a dense mesh with the appropriate labels, making it considerably more traceable (Fig. 3 right and Fig. 4).

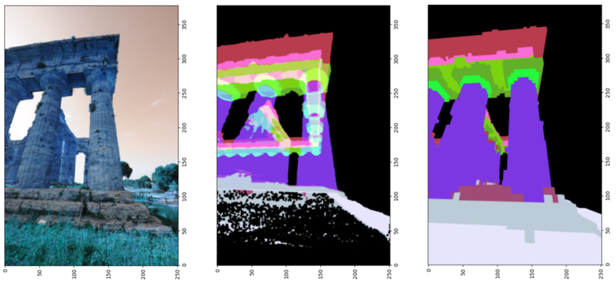

Figure 3: Overview of the image raycasting: original image (left), raycasting on the original point cloud, which is unusable due the lack of surfaces (middle) and raycasting on the voxel mesh, which does yield proper masks for image segmentation (right).

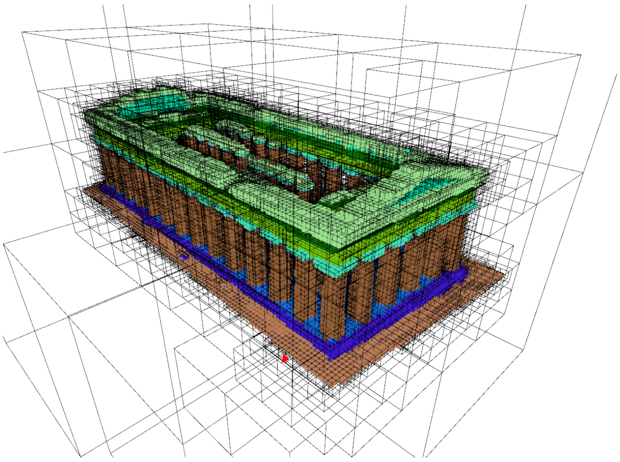

Figure 4: Voxel mesh generated from the point cloud octree.

3.2 Semantic segmentation

In the early fusion, images and point clouds are merged into RGBD images: this reduces the complexity of geometric reasoning but allows for the joint semantic segmentation of image and point cloud modalities. For late fusion, we first conduct an image-based semantic segmentation: the results of this segmentation are then associated with the point cloud data. Subsequently, the final classification is determined through a second semantic segmentation step, which is based on features extracted from the point cloud.

Geolocated Imagery - For image classification, we employ transfer learning with DeepLabV2 (Adam et al., 2021), which uses ResNet152 (He et al., 2016) as the backbone and pre-trained on the ImageNet database. The initial version of DeepLab (DeepLabV1 (Chen et al., 2018)) introduced a novel concept called atrous (dilated) convolutions. The use of atrous convolutions allowed the model to capture wider context in images without reducing their resolution. while DeepLabV2 introduced Atrous Spatial Pyramid Pooling (ASPP) that greatly improved the model's ability to handle objects of varying scales and has demonstrated robustness against various image perturbations and high-class variance, among other factors. Initially, the generated masks are divided into training, validation, and testing datasets. The sparse categorical cross-entropy loss function was used, given the multi-class semantic segmentation task. Class balancing techniques

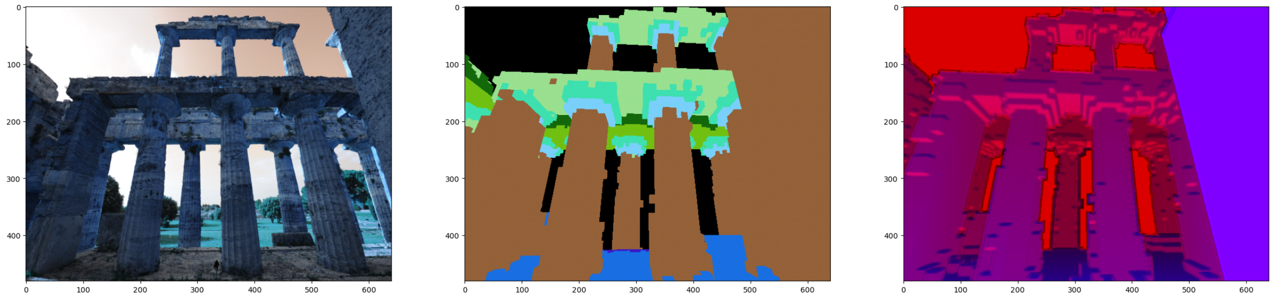

Figure 5: Overview of the early fusion modality: (left) original undistorted image, (middle) projected point cloud labels and (right) HHA imagery with depth information.

were applied to account for low class presence, and data augmentation methods, as recommended in (Shorten and Khoshgoftaar, 2019), were employed. The training process occurred in two stages. Initially, only the output layer was trained using automatically generated training data. Subsequently, the model was further fine-tuned. Out of the total 30,925,387 parameters in DeepLabV2, 30,840,427 were trained for both the building elements and the materials. Given the image segmentation, the outcomes are associated with the most suitable points in the point cloud P. By utilizing the image coordinates of the labels $\mathbf{I}$ and depth maps $\mathbf{D}$, a reference point cloud Q can be created using the same raycasting mechanism (Equation 3). As there is significant overlap in the imagery, mislabeling in $\mathbf{I}$ will result in a cluttered reference point cloud. To obtain the final result, a k-nearest neighbour evaluation between the initial point cloud and the reference cloud. The labelling Y is then obtained by the weighted average label of the project image labels, given inverse distance weights w. These image labels are then assigned as an additional feature in the point cloud semantic segmentation.

$$\boldsymbol{Q_j} = \left\{ q_j \middle| p_i \in P, q_j \in Q : argmin_{q_j} \|p_i - q_j^{(k)}\| \right\}$$

$$Y = \left\{ y \middle| Q_j \in \boldsymbol{Q_j} : argmax_{|y|} \sum_{q_j \in Q_j} w_j y(q_j) \right\} \qquad (3)$$

Point clouds - For the point cloud segmentation, a set of covariance features are computed for P, including linearity, planarity, verticality, and others as proposed in (Niemeyer et al., 2014). These features, together with the results of the 2D segmentation, are then passed to a neural network as an extra channel of input data. For the tests, we employed the Point Transformer architecture(Zhao et al., 2021), a deep learning method that relies on the self-attention operator for essential tasks in scene understanding. In the Point Transformer, the self-attention mechanism is applied locally, allowing the network to upscale its capabilities for tasks on large scenes with millions of points. The training process was conducted in a single step, with class balancing techniques applied to account for low class presence. The resulting labels, Y, can then be directly applied to the point cloud P.

RGBD - In the early fusion of image and point semantic segmentation, we project the 3D coordinate information onto the image depth channel to form RGBD imagery using the aforementioned techniques. As it is challenging to unify depth maps based on their respective depths, we opt for producing HHA imagery, as proposed in (Gupta et al., 2014). This format incorporates the depth and viewing direction into a uniform depth format, which is more comprehensible than conventional depth maps, albeit being quite computationally demanding to compute, as shown in Fig. 5.

For the semantic segmentation itself, we employ ShapeConv combined with Deeplabv3+(Chen et al., 2018) and a ResNet-101 backbone(He et al., 2016). ShapeConv is a model-agnostic convolutional layer that can be easily integrated into existing networks, focusing on jointly learning shape and base components. In the original paper, ShapeConv significantly improved the generalization and performance of the base networks on known datasets such as SiD, NYUv2-40, and SUN, as shown in Table 1.

Table 1: Baseline ShapeConv results on benchmark datasets.

Class	Mean IoU (%)	f.w. IoU (%)	Pixel Acc (%)	Mean acc (%)
SID	60.6	71.2	82.7	70.0
NYUv2-40	51.3	63.0	74.5	59.5
SUN	48.6	71.3	76.4	63.5

4. EXPERIMENTS

The early and late fusion are compared against traditional networks that process only a single modality. Specifically, two photogrammetric datasets, each with a unique signature, are selected for these tests.

4.1 Dataset I: Paestum

The first test is a photogrammetric reconstruction of the Greek Temple of Neptune (ca 25m x 60m x 15m), located in Paestum, Italy (Fiorillo et al., 2013). The dataset comprises 894 geolocated images captured by hand and UAVs, resulting in a point cloud of ca 56 million points. Although the temple is constructed entirely of one material, it features 10 different building techniques (Fig. 6). Consequently, the temple exhibits a predominantly geometric signature rather than a distinct texture signature. Each method was trained and validated on 25% of the data for 300 epochs, and inference was performed on the entire dataset. Table 2 presents the data distribution in the dataset, which is fairly unbalanced as typically happens in such datasets. Notably, the distributions are similar for the point and pixel distributions, except for classes 4 (6.9% vs 0.2%) and 7 (4.5% vs 0.1%), which are underrepresented in the image dataset, and class 3 (15.8% vs 41.2%), which is significantly over-represented. This over-representation is attributable to the large number of

Table 2: Class distribution (%) in the imagery and point cloud of Paestum.

Class	Points (%)	Pixels (%)
0. Grass	28.7	28.1
1. Crepidoma	4.9	6.6
2. Pavement	10.6	11.1
3. Shaft	15.8	41.2
4. Echinus	6.9	0.2
5. Abacus	3.7	3.9
6. Architrave	5.9	3.2
7. Frieze	4.5	0.1
8. Cornice	18.6	5.2
9. Tympanum	0.4	0.4

images (76%) taken inside the structure at eye level, primarily featuring columns.

The image segmentation is conducted as outlined in Section 3.. For the late fusion, the segmentation results are projected onto the point cloud as an additional feature. Following this, the Point Transformer network was trained again on the same partition, utilizing also the covariance features listed in Table 3. A batch size of 48,000 points was employed, with a subsampling of 0.005m. Additionally, the ShapeConv combined with Deeplabv3+ and a ResNet-101 backbone was trained on the RGB and HHA channels. Achieved results are presented in Table 4, 5 and Fig. 7, 8. Overall, while each method scores well for the more prevalent classes, some key differences are observed. Firstly, there is a notable divergence between the mIoU and the weighted mIoU, primarily attributed to class balancing. This effect is most pronounced with ShapeConv, which disproportionately favours the majority classes, resulting in a skewed performance as it neglects minority classes (4,7 and 9). Contrarily, other methods which incorporate weighted training approaches demonstrate a more balanced performance profile. Notably, the detection rates across different classes still show considerable variance, with the image segmentation networks being particularly susceptible to discrepancies in training sample sizes. Secondly, Point Transformer scores the best results (71.4% mIoU) which is expected due to the geometric nature of the dataset. On the other hand, the DeepLabV2 network, rather than improving these results, actually contributes to greater confusion in the late fusion (with a lower mIoU of 68.5%) due to its subpar classification of the less represented classes. This underscores the importance of careful integration of network results in late fusion, potentially by including the confidence levels.

Thirdly, the ShapeConv network has mixed results. It scores better than most image classes and even some late fusion classes. Nevertheless, it underperforms in representing the minority classes from the image perspective, suggesting a loss of contextual understanding when transitioning from a general to a viewpoint-specific approach, in part due to the severe unbalancing of the training data. Finally, the training efficiency for early fusion is significantly higher than its late fusion counterpart. This depends on the implementation but also the data modality (2D convolutions are faster) and the joint training of a single network with fewer parameters, which is less demanding.

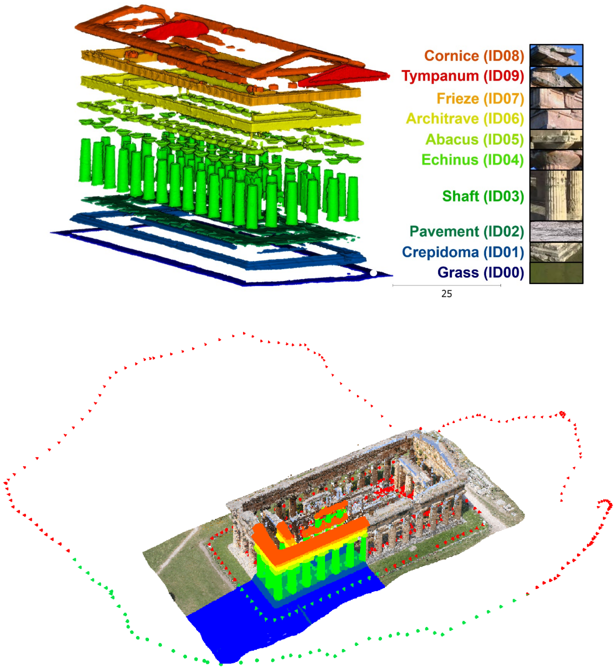

Figure 6: Paestum dataset, exploded in its 10 building technique classes (above). 25% training and validation (RGB coloured) and 75% test (highlighted) images (below).

Table 3: Covariance features.

Feature	Radius (m)
Omnivariance	1.5
Sphericity	1.5
Sphericity	2
Surface variation	0.4
Surface variation	0.7
Surface variation	1.0
Surface variation	1.5
Surface variation	2.0
Verticality	0.2
Verticality	0.6
Verticality	1.0
Verticality	1.2

Table 4: Semantic segmentation results per method.

Method	Time (s)	mIoU (%)	Weighted mIoU (%)
DeepLabV2 (RGB)	28620	44.2	69.5
Point Transformer (PCD)	12510	71.4	80.3
DeepLabV2,PT (RGB+PCD)	40710	68.5	78.4
ShapeConv (RGBD)	6784	52.5	82.9

Table 5: Average IoU per class for the 75% test area.

Class	RGB (%)	PCD (%)	RGB+PCD (%)	RGBD (%)
0. Grass	47.21	77.5	85.6	92.3
1. Crepidoma	27.1	89.5	86.8	49.0
2. Pavement	61.16	85.3	88.6	55.7
3. Shaft	79.31	90.2	90.7	88.6
4. Echinus	44.58	76.1	66.8	0.0
5. Abacus	39.42	60.8	50.7	38.9
6. Architrave	45.82	74.1	62.3	37.6
7. Frieze	38.3	56.2	53.0	0.0
8. Cornice	50.26	77.6	73.1	53.9
9. Tympanum	9.17	26.2	27.3	5.4

4.2 Dataset II: Wall of the Pecile

The second test is a photogrammetric 3D reconstruction of the Wall of the Pecile (18m x 1m x 8m), a part of the courtyard of the Roman Villa Adriana in Tivoli, Rome. The dataset comprises 54 geolocated images taken by hand, resulting in a point cloud of 2.5 million points. It includes 6 building techniques (Fig. 9). However, these classes primarily have texture signatures since the reconstruction consists of a gate at the center of a flat wall with limited geometric signatures. Therefore, it is expected that

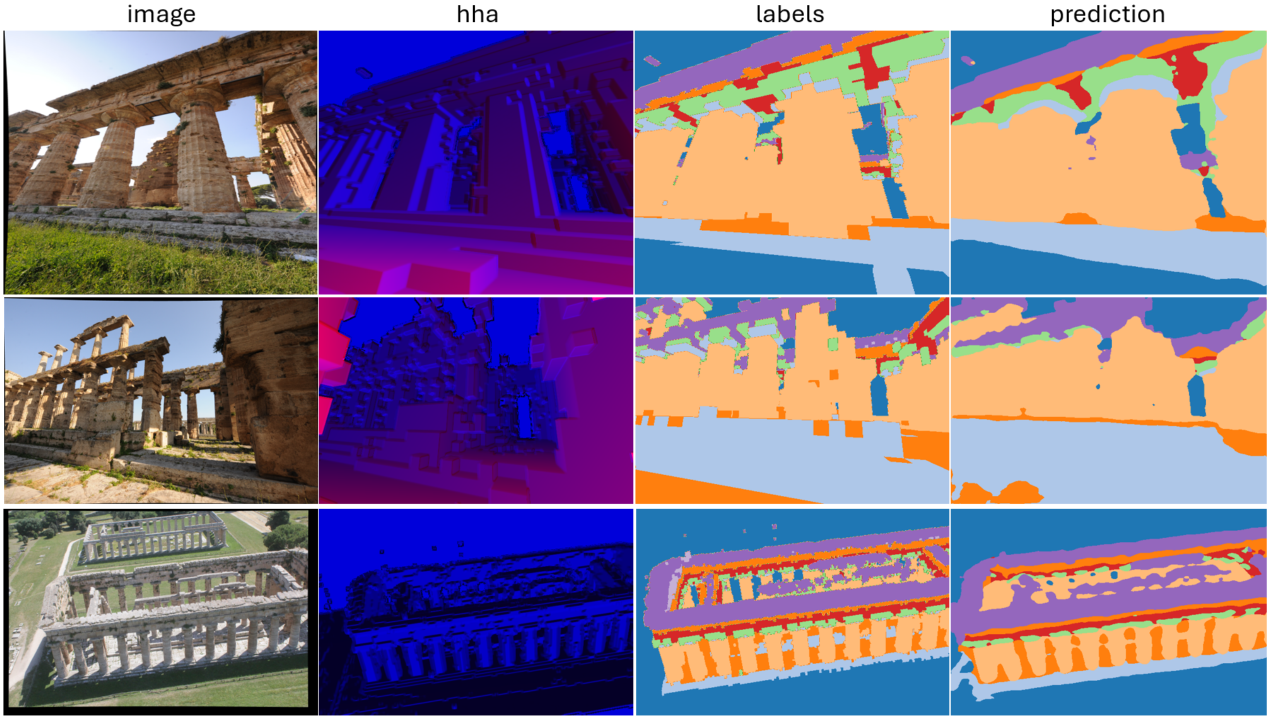

Figure 7: Semantic segmentation results of the early fusion method with ShapeConv on the Paestum dataset.

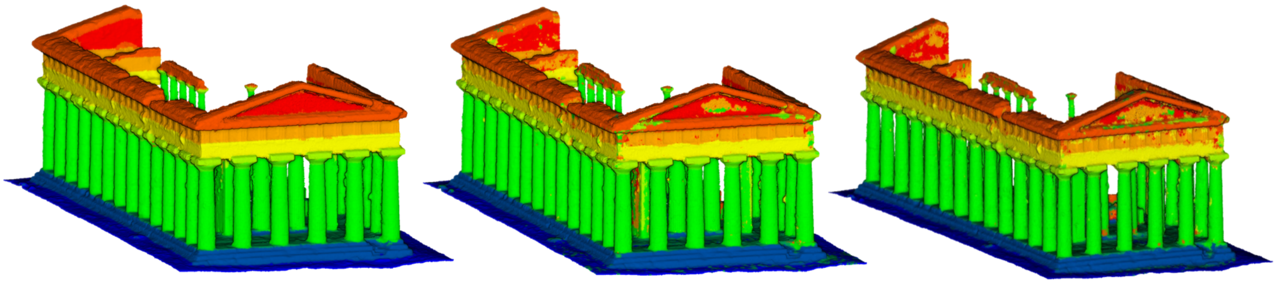

Figure 8: Semantic segmentation results for the Paestum dataset: ground truth (left), Point transformer (middle) and DeepLabV2 plus Point Transformer (right).

detection in the imagery will outperform the point cloud in semantic segmentation.

Each method was trained and validated on 50% of the data for 300 epochs, and inference was performed on the entire dataset. Table 6 again reveals a highly unbalanced dataset, with an average class balancing spread of $\sigma_C = 15\%$. Similar imbalances in class representation are observed as in Paestum, with classes at eye-level being overrepresented. However, given that the Pecile dataset is significantly smaller, these effects are more pronounced. For instance, the average difference in class balance in Paestum is 5.5%, whereas in Pecile it is 10.5%.

Table 6: Pecile class distribution (%) in the imagery and point cloud.

Class	Points (%)	Pixels (%)
0. Plaster	9.9	41.3
1. Old opus reticulatum	41.3	28.3
2. Restored opus reticulatum	8.4	6.5
3. Opus reticulatum grey	34.7	19.7
4. Old opus latericium	0.7	0.1
5. Restored opus latericium	5.0	4.1

All methods were processed analogously to those in Paestum. The Point Transformer network was trained with a batch size of 48000 points, a subsampling of 0.005m, and the features listed in Table 7. For the image segmentation using DeeplabV2 and ShapeConv, the imagery was divided into 9 tiles, thereby increasing the number of samples to 504. This partitioning incurs minimal overhead on the total calculations. The generation of HHA imagery took 423 seconds.

Results are presented in Table 8, 9 and Fig. 10. The average detection rate of the methods is 63% mIoU while the weighted mIoU is 78.9%, showing a similar trend as the Paestum dataset due to training data differences. However, the image-based methods score significantly better with Point Transformer now being the weakest performer. A significant observation here is the superior performance of image-based methods, with the Point Transformer being the less effective method. This underscores the importance of choosing networks that leverage both texture and geometric features in a scene.

Both early and late fusion techniques show comparable efficacy, with the contribution in late fusion primarily coming from the DeepLabV2. Again, it is observed that ShapeConv does not deal well with low presence classes. Despite this, the added geometry channels in ShapeConv do improve the detection rate as some of the materials have some depth-

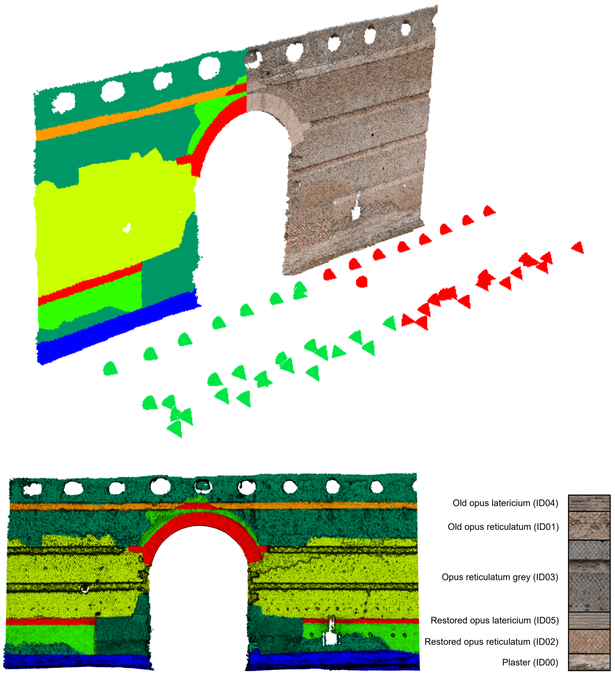
Old opus latericium (ID04)
Old opus reticulatum (ID01)
Opus reticulatum grey (ID03)
Restored opus latericium (ID05)
Restored opus reticulatum (ID02)
Plaster (ID00)

Figure 9: The Pecile dataset including 6 building technique classes, 50% training and validation images (green) and 50% test images (red).

sensitive erosion.
An interesting insight is that each network has variable performance depending on the scene, expect for the late fusion. It seems that by directly embedding image detection results into the geometric processing, simultaneously the best and worst results are filtered out, leading to a stable performance across scenes with varying texture and geometric signatures. Contrary to expectations, early fusion didn't mirror this behavior. The further imbalance in training data appeared to hamper the network's efficacy. Additionally, the limited parameter set in early fusion, as opposed to the more elaborate setup in late fusion involving multiple networks, seemed to restrict its ability to encapsulate the same level of complexity effectively.

Table 7: Pecile covariance features.

Feature	Radius (m)
Anisotropy	0.05
Gaussian curvature	0.1
Mean curvature	0.1
Normal change rate	0.05
Roughness	0.05
Roughness	0.1
Roughness	0.2

Table 8: Average Pecile semantic segmentation results per method.

Method	Time (s)	mIoU (%)	weighted mIoU (%)
DeepLabV2 (RGB)	7260	66.6	85.2
Point Transformer (PCD)	1410	51.0	60.7
DeepLabV2,PT (RGB+PCD)	1770	67.4	82.3
ShapeConv (RGBD)	6154	67.0	87.4

5. CONCLUSIONS

This work presented the adoption of early and late fusion methods for image and point cloud semantic segmentation in cultural heritage applications. It features a methodology for seamless transition between data modalities and efficient production of training data. The late fusion approach merges image-based segmentation with a Point Transformer applied to point clouds. In contrast, the early fusion utilizes multi-view rendering to generate RGBD imagery of the scene.

Table 9: Pecile IoU per class for the 50% test area.

Class	RGB (%)	PCD (%)	RGB+PCD (%)	RGBD (%)
0. Plaster	80.9	73.1	85.6	96.8
1. Old opus reticulatum	82.8	61.8	86.2	84.2
2. Restored opus reticulatum	47.9	37.6	54.0	59.1
3. Opus reticulatum grey	80.0	66.1	90.5	86.4
4. Old opus latericium	50.4	22.3	34.3	0.0
5. Restored opus latericium	54.2	45.0	54.1	69.8

Experiments on two test cases demonstrate that the detection rate is primarily influenced by whether the scene has a predominantly geometric or texture-based signature, underscoring the necessity of fusion methods. Image semantic segmentation proves to be more effective in texture-rich areas, whereas Point Transformers excel in geometrically complex scenes. The combination of both approaches yields enhanced results in both cases, a pattern also observed in early fusion. Notably, late fusion tends to be more consistent, benefiting from better-suited data modalities and the absence of training entanglement.
The study concludes that employing networks in series or parallel, as seen in late fusion, tends to be more advantageous for projects than early fusion. This is because even if only one of the networks in the series performs well, satisfactory results are achievable. An essential factor in choosing between early and late fusion methods is the scene's complexity. In highly intricate scenes, late fusion is often the better choice as each modality requires a dedicated network for precise tuning. However, in simpler scenarios like those examined in this study, the ability to generalize quickly over smaller networks makes early fusion a viable option. Future research aims to explore further the relationship between scene complexity and the choice of fusion methods.

ACKNOWLEDGEMENTS

This project has received funding from the FWO Postdoc grant (1251522N) and the Geomatics group of the Department of Civil Engineering, KU Leuven, Belgium.

REFERENCES

Adam, H., Qiao, S., Yuille, A. L., Collins, M. D., Yuan, L., Yu, Q., Wang, H., Zhu, Y., Weber, M., Cremers, D., Xie, J., Schroff, F., Kim, D., Chen, L.-C. and Leal-Taixe, L., 2021. Deeplab2: A tensorflow library for deep labeling.

Armeni, I., Sax, S., Zamir, A. R., Savarese, S., Sax, A., Zamir, A. R. and Savarese, S., 2017. Joint 2d-3d-semantic data for indoor scene understanding. arXiv:1702.01105.

Bello, S. A., Yu, S., Wang, C., Adam, J. M. and Li, J., 2020. Review: Deep learning on 3d point clouds. Remote Sensing 12, pp. 1–20.

Cao, J., Leng, H., Lischinski, D., Cohen-Or, D., Tu, C. and Li, Y., 2021. Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A. L., 2018. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE TPAMI 40, pp. 834–848.

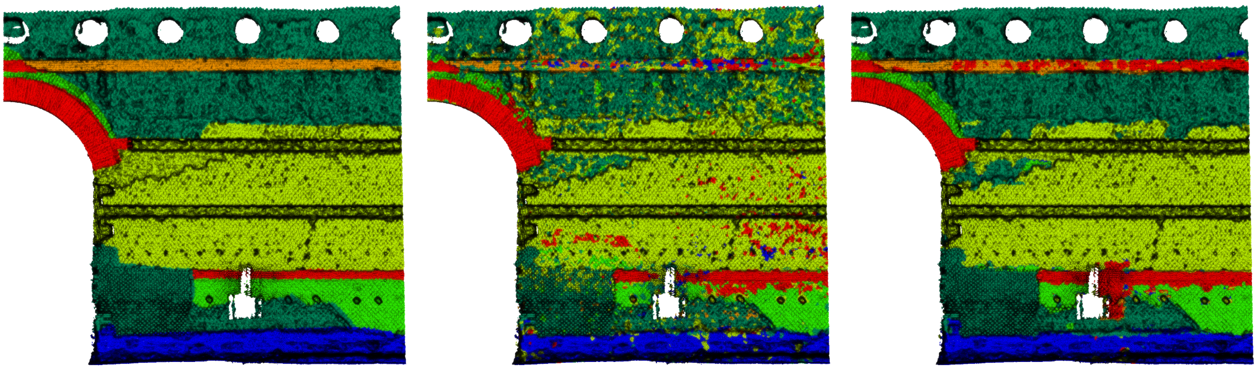

Figure 10: Semantic segmentation results for a part of the Pecile wall: ground truth (left), Point transformer (middle) and DeepLabV2 plus Point Transformer (right).

Coudron, I., Puttemans, S. and Goedemé, T., 2020. Semantic extraction of permanent structures for the reconstruction of building interiors. Sensors pp. 1–21.

Croce, V., Caroti, G., De Luca, L., Jacquot, K., Piemonte, A. and Véron, P., 2021. From the semantic point cloud to heritage-building information modeling: A semiautomatic approach exploiting machine learning. Remote Sensing 13(3), pp. 461.

Croce, V., Caroti, G., Piemonte, A., Luca, L. D. and Véron, P., 2023. H-bim and artificial intelligence: Classification of architectural heritage for semi-automatic scan-to-bim reconstruction. Sensors 23, pp. 2497.

Cui, Y., Chen, R., Chu, W., Chen, L., Tian, D., Li, Y. and Cao, D., 2022. Deep learning for image and point cloud fusion in autonomous driving: A review. IEEE Transactions on Intelligent Transportation Systems 23, pp. 722–739.

Fiorillo, F., Jiménez Fernández-Palacios, B., Remondino, F. and Barba, S., 2013. 3d surveying and modelling of the archaeological area of paestum, italy. Virtual Archaeology Review 4, pp. 55–60.

Fiorucci, M., Khoroshiltseva, M., Pontil, M., Traviglia, A., Del Bue, A. and James, S., 2020. Machine learning for cultural heritage: A survey. Pattern Recognition Letters 133, pp. 102–108.

Grilli, E. and Remondino, F., 2020. Machine learning generalisation across different 3d architectural heritage. ISPRS International Journal of Geo-Information 9(6), pp. 379.

Grilli, E., Daniele, A., Bassier, M., Remondino, F. and Serafini, L., 2023. Knowledge enhanced neural networks for point cloud semantic segmentation. Remote Sensing 15, pp. 1–26.

Grilli, E., Petrucci, G. and Remondino, F., 2018. Supervised segmentation of 3d cultural heritage. Vol. 42-2, pp. 399–406.

Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L. and Bennamoun, M., 2021. Deep learning for 3d point clouds: A survey. IEEE TPAMI 43, pp. 4338–4364.

Gupta, S., Girshick, R., Arbeláez, P. and Malik, J., 2014. Learning rich features from rgb-d images for object detection and segmentation. Lecture Notes in Computer Science 8695, pp. 345–360.

He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. IEEE Computer Society, pp. 770–778.

Lu, D., Xie, Q., Wei, M., Member, S., Gao, K., Member, S., Xu, L. and Li, J., 2022. Transformers in 3d point clouds: A survey. arXiv:2205.07417.

Matrone, F., Grilli, E., Martini, M., Paolanti, M., Pierdicca, R. and Remondino, F., 2020. Comparing machine and deep learning methods for large 3d heritage semantic segmentation. ISPRS International Journal of Geo-Information 9(9), pp. 535.

Niemeyer, J., Rottensteiner, F. and Soergel, U., 2014. Contextual classification of lidar data and building object detection in urban areas. ISPRS Journal of Photogrammetry and Remote Sensing 87, pp. 152–165.

Pierdicca, R., Paolanti, M., Matrone, F., Martini, M., Morbidoni, C., Malinverni, E. S., Frontoni, E. and Lingua, A. M., 2020. Point cloud semantic segmentation using a deep learning framework for cultural heritage. Remote Sensing 12(6), pp. 1005.

Qi, S., Ning, X., Yang, G., Zhang, L., Long, P., Cai, W. and Li, W., 2021. Review of multi-view 3d object recognition methods based on deep learning. Displays.

Shorten, C. and Khoshgoftaar, T. M., 2019. A survey on image data augmentation for deep learning. Journal of Big Data.

Townend, F., Roddy, P. J. and Goebl, P., 2024. Fusilli v1.1.0 (v1.1.0). Zenodo.

Weinmann, M., Jutzi, B., Hinz, S. and Mallet, C., 2015. Semantic point cloud interpretation based on optimal neighborhoods , relevant features and efficient classifiers. ISPRS Journal of Photogrammetry and Remote Sensing 105, pp. pp.286–304.

Xiao, A., Yang, X., Lu, S., Guan, D. and Huang, J., 2021. Fps-net: A convolutional fusion network for large-scale lidar point cloud segmentation. ISPRS Journal of Photogrammetry and Remote Sensing 176, pp. 237–249.

Yang, S., Hou, M. and Li, S., 2023. Three-dimensional point cloud semantic segmentation for cultural heritage: A comprehensive review. Remote Sensing 15(3), pp. 548.

Zhang, Y., Sidibé, D., Morel, O. and Mériaudeau, F., 2021. Deep multimodal fusion for semantic image segmentation: A survey. Image and Vision Computing 105, pp. 104042.

Zhao, H., Jiang, L., Jia, J., Torr, P. and Koltun, V., 2021. Point transformer. Proceedings ICCV pp. 16239–16248.