Complex Instance Segmentation in Point Clouds with Images and 3D Models

Zhiqi Hu, M.Sc., [1] **Yuandong Pan, Ph.D.,**[2]
Ioannis Brilakis, Ph.D., M.ASCE, [3] **and André Borrmann, Dr-Ing,** [4]

[1]Ph.D. Candidate, Department of Engineering, University of Cambridge, Cambridge, UK. ORCID: 0000-0002-0549-4500. Email: zh334@cam.ac.uk
[2]Research Associate, Department of Engineering, University of Cambridge, Cambridge, UK. ORCID: 0000-0002-5331-6901. Email: yp296@cam.ac.uk
[3]Professor, Department of Engineering, University of Cambridge, Cambridge, UK. ORCID: 0000-0003-1829-2083. Email: ib340@cam.ac.uk
[4]Professor, Chair of Computational Modeling and Simulation, Technical University of Munich, Munich, Germany. ORCID: 0000-0003-2088-7254. Email: andre.borrmann@tum.de

ABSTRACT

A geometric digital twin can help monitor the progress, control the quality, and simulate the energy of a building during its construction and operation stages. However, current methods cannot match and segment instances with a high-resolution result in real, complex environments regarding mechanical, electrical, and plumbing (MEP) objects, since the geometry of as-built MEP objects is complex and often deviates from as-designed models in terms of position, orientation, and scale. To this end, this paper proposes a hybrid method by fusing point clouds, images, and 3D models to efficiently segment complex MEP components in buildings. An experiment targeting heating terminals in a real 3-floor staircase space shows the feasibility and practicality of the proposed method.

INTRODUCTION

This research is about fusing two-dimensional (2D) and three-dimensional (3D) datasets for the segmentation of mechanical, electrical, and plumbing (MEP) components in point clouds of buildings. By 2D dataset, we refer here to images (2D matrices with red, green, and blue color modalities) or video sequences (a series of images marked with timestamps and captured sequentially from a single video camera). By 3D dataset, we refer here to point cloud data (PCD) (millions of data points represented by XYZ coordinates). By MEP components in buildings, we refer here to the most frequent object types in plumbing, heating, air conditioning, and electrical supply systems (Hu et al. 2022; Drobnyi et al. 2023). By segmentation, we refer here to match and extract a subset of points of MEP components from laser scanned PCD by mapping 2D bounding boxes from the predicted image series and iterative closest point (ICP) fine registration.

The objective of this research is to develop a framework to automatically detect, segment, and match top frequent MEP instances from PCD in Scan-vs-BIM contexts. MEP objects are major components besides structural objects in buildings. The top frequent MEP object types include round pipe segments, round elbows, light fixtures, plumbing and heating terminals, cylindrical joints, and rectangular duct segments (Drobnyi et al. 2023). Construction companies have suffered poor project performance of MEP objects due to a lack of timely progress and quality monitoring to fix any issues. A digital twin (DT) reflecting a physical building's status can be a solution to keep construction progress up to date (Hu and Brilakis 2024, Hu et al. 2022). Identifying the location of as-built MEP components and matching them with as-designed models are crucial DT construction steps. This paper will discuss the background of this problem, propose a solution, and show an experiment on heating terminals to validate the solution. A 3-floor staircase space in the Civil Engineering Building at the University of Cambridge was used for the experiment.

BACKGROUND

The relevant fundamental knowledge, the state of practice, and the state of research are reviewed in this section. Mobile and terrestrial laser scanners are widely used in collecting PCD. The raw scanned data files need to be registered to show the complete scene of a targeted 3D space. Iterative Closest Point (ICP) is widely used for fine registration between two point clouds. On the other hand, structure from motion (SFM) is widely applied for 3D point cloud reconstruction from images. Deep learning provides an efficient way for object classification.

Terrestrial laser scanners are frequently used to collect point clouds of outdoor areas and buildings (Pan et al. 2023). The accuracy of terrestrial scanners is high and thus can assist operations, such as surveying, site inspection, and building information modelling (BIM), in the construction and industrial sectors. On the other hand, mobile scanners held in the hand or installed on a vehicle are less accurate but quicker than the terrestrial scanners (Sepasgozar et al. 2018). Many companies including Faro, Leica, and Trimble produce many scanners that fulfill various use cases.

The strategy of scanning data collection is important for the quality of the point cloud result. Different scanning positions and trajectory selections can influence the captured area. A non-model-based scan planning approach has been developed to identify the best viewpoint for each scan by projecting a point cloud to generate a 2D polygon for the next best viewpoint selection (Achakir et al. 2021). The Greedy algorithm was applied to determine a suitable viewpoint for scanning optimization (Jia and Lichti 2019). However, most methods are workable in a simulation environment, and a robust method in a real and complex 3D scene with occlusions is still under research.

The main objective of PCD registration is to find the rigid transformation matrix including rotation matrix ($\boldsymbol{R}$) and translation vector ($\boldsymbol{t}$). A rigid transformation is defined as follows:

$$\boldsymbol{T}(\boldsymbol{v}) = \boldsymbol{R}(\boldsymbol{v}) + \boldsymbol{t}$$

where $\boldsymbol{T}$ represents the transformation of the input vector $\boldsymbol{v}$. One of the expansion forms of homogeneous transformation matrix $\boldsymbol{T}$ in 3D view can be represented as follows (Patil et al., 2018):

$$\boldsymbol{T} = \begin{bmatrix} cos\varphi cos\theta & cos\varphi sin\beta sin\theta - cos\beta sin\varphi & sin\varphi sin\beta + cos\varphi cos\beta sin\theta & t_x \\ sin\varphi cos\theta & cos\varphi cos\beta + sin\varphi sin\beta sin\theta & cos\beta sin\varphi sin\theta - cos\varphi sin\beta & t_y \\ -sin\theta & sin\beta cos\theta & cos\beta cos\theta & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

where $\varphi,\ \theta,\ \beta$ are the Euler angles rotating along with the principle of z-axis, y-axis, and x-axis respectively.

The common registration types include coarse registration, fine registration, and hybrid registration. A 4-plane congruent set method has been developed to deliver a ranked list of the most likely transformation matrices for registration so that the user can compare the results and manually select the best solution (Bueno et al. 2018). Kaiser et al. (2019) proposed a random sample consensus (RANSAC)-based line matching method to achieve coarse registration. Masood et al. (2020) projected 3D into 2D to do coarse registration. On the other hand, ICP algorithm is widely used for fine registration. The core idea of ICP consists of two steps: 1) It generates the correspondence between two sets of the PCD. Each point in one set matches the nearest point in the other set. 2) It computes new transformation parameters iteratively by minimizing the sum of the squares of the Euclidean distances between corresponding points (Li et al. 2020). A progressive ICP variant (Tang and Rasheed 2013) and Go-ICP (Yang et al. 2013) method were developed for fine registration efficiently. Point Feature Histograms (PFH) (Rusu et al. 2008) is also widely used for 3D Registration. Fast Point Feature Histograms (FPFH) (Rusu et al. 2009) is an improved variant of FPH that can improve the speed and reduce the computational complexity.

Photogrammetry is widely used to reconstruct 3D geometry from 2D images. Generally, the photogrammetric process consists of two steps: SFM (Schonberger and Frahm 2016) and Multi-view stereo (MVS) (Schönberger et al. 2016). The SFM process uses a set of overlapping images captured at different viewpoints as its input to reconstruct key points in 3D by detecting and matching features in photos. Subsequently, MVS takes the output of SFM process as input to create a dense point cloud by computing depth and normal information for pixels in these images. Photogrammetric methods are designed to detect positions where the color changes, making the weakly-textured surfaces hard to be reconstructed. Although the reconstruction result of the indoor environment is not as good as the outdoor environment as there are usually large amount of surfaces of white walls, all the key feature points can be found, i.e., boundary points of objects and intersection of walls. These reconstructed 3D key points can represent the location and dimension of objects in the environment.

Deep Learning is widely applied in 2D data processing. In computer vision, object detection aims to identify an object and estimate its location in an image (Szegedy et al. 2013). Variants of R-CNN (Girshick et al. 2014), including fast R-CNN (Girshick 2015) and mask R-CNN (He et al. 2017), have been proposed to improve the performance of object detection. In the AEC field, object detection networks are also applied in different domains, such as damage detection of concrete (Jiang et al. 2021), worker detection on construction sites (Jeelani et al. 2021), object recognition in buildings (Pan et al. 2022), etc. In order to solve different problems in practice, a neural network should be trained on the specific dataset. Transfer learning (Pan and Yang 2010), a process in which a neural network model is pre-trained on a related larger dataset and then re-trained on a user-specific dataset, aims to improve the network's performance.

As-built MEP components very often deviate from the as-designed model in terms of position, orientation, scale, and geometry in the real construction environment. Simply aligning the as-designed model against the as-built PCD cannot directly help locate and segment MEP object instances with complex shapes to monitor the construction progress. While deep learning models provide the possibility to efficiently segment 3D point clouds of different environments, they mainly focus more on furniture classes (e.g., tables, chairs, etc.) and structural elements (e.g., ceiling, floor, etc.). The lack of labeled MEP PCD with more various and complex shapes makes it difficult to apply these models to detect and segment complex instances. Therefore, the paper aims to develop a hybrid method by fusing point clouds, images, and 3D models to efficiently detect and segment MEP components in buildings.

METHOD OVERVIEW

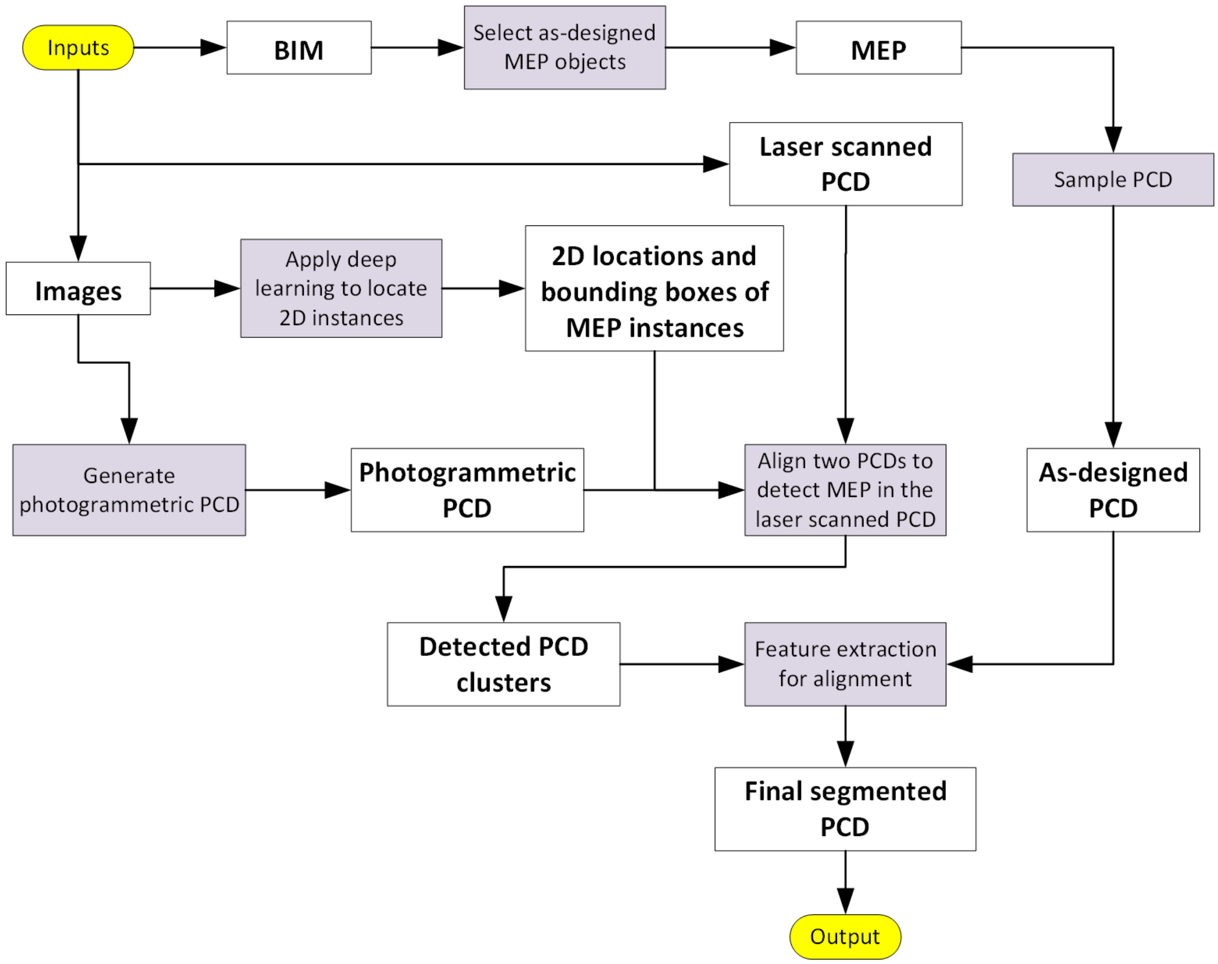

Figure 1. The framework of the proposed method

This paper proposed an automatic hybrid approach with images and as-designed models to automatically detect and segment the point clouds into instance-level point clusters. Figure 1 shows an overview of the proposed method. The inputs contain an as-designed BIM model, captured images, and a laser-scanned PCD from a building at the construction stage. First, the images are split into a training set and a test set. The training set is labeled with different MEP classes for training object classification models. Specifically, a faster region-based convolutional neural network (R-CNN) is employed to locate as-built MEP components in 2D images. Then, SFM and MVS are applied to process the test 2D dataset to generate photogrammetric PCD. After that, the scanned PCD was aligned with the photogrammetric PCD to identify the 3D location of MEP instances with the predicted label for each point from tested images. We use a bounding box to locate the instances instead of a mask to segment instances from images since it is time-saving and easier to generate bounding boxes for training datasets. Especially for different buildings at the construction stage, MEP objects have various irregular shapes, such as heaters with hollow structures. In this case, mask generation will cost time and affect the dynamic updating at the construction stage. Finally, the as-designed point cloud clusters are sampled from the 3D model to support the matching process. Features from the detected point cluster and the sampled point cluster are extracted by PFH to coarse registration and then ICP is used for fine registration as a result of high-resolution instance segmentation. Overall, the proposed hybrid method performs well in real, complex environments including: 1) the as-built point clouds are noisy with high occlusions and clutter, and 2) there are distinct discrepancies between the as-designed models and the as-built components from the perspective of position, orientation, and scale. The proposed method is capable of matching as-designed object instances in as-built PCDs with a high-resolution result.

DATA PREPARATION

A 3-floor staircase space in the Civil Engineering Building at the University of Cambridge was used to test the proposed method. The space contains MEP components including heating terminals, light fixtures, piping segments, and elbows. We use heating terminals as an example to prove the concept of the proposed method.

As for 2D dataset preparation, three videos of the ground, first, and second-floor staircases (each about 90 seconds) were captured by iPhone 13 Pro. Images were extracted from every 20 frames by MATLAB. 302 images were selected from other parts of the building as the training dataset. LabelImg (Tzutalin 2015) was used to label axis-aligned bounding boxes of MEP classes for each training image. Colmap (Schonberger & Frahm 2016; Schönberger et al. 2016) was used to generate photogrammetric PCD from 327 test images.

As for 3D dataset preparation, the equipment used to capture scanned PCD is a FARO Focus 3D X330 Terrestrial Laser Scanner (2013) (ranging error ± 2 mm at 10 m, self-levelling: accuracy 0.015° (range ± 5°)). Four scanning positions have been selected manually to ensure the captured area. The photogrammetric PCD needs to be registered against the scanned PCD. Cloudcompare was used to register two files of PCD by manually selecting 4 pair points. Figure 2 elaborates on the results of the photogrammetric PCD, the scanned PCD, and the registered outcomes.

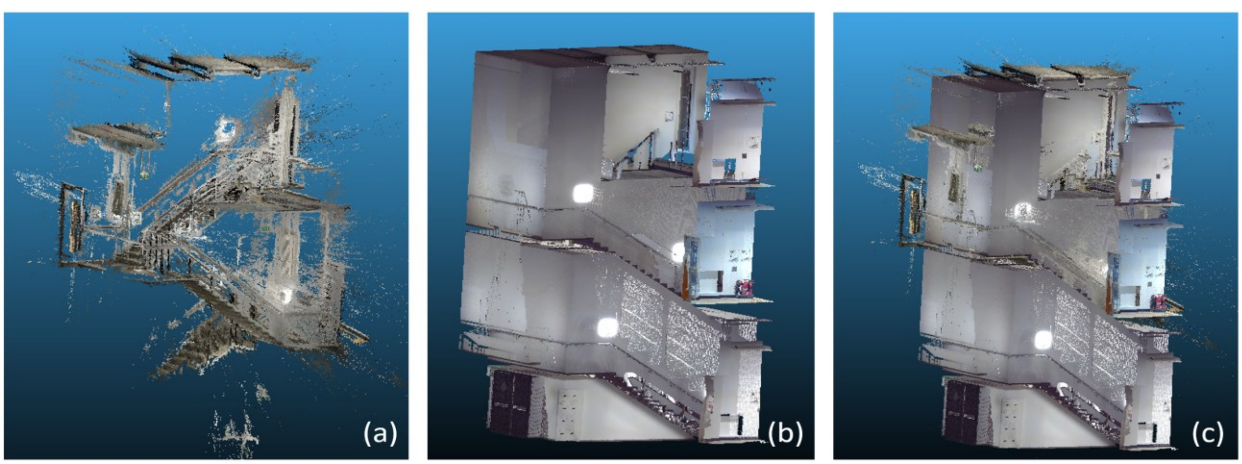

Figure 2. The results of (a) the photogrammetric PCD, (b) the scanned PCD, (c) the registered outcome of a 3-floor staircase space.

EXPERIMENTAL RESULTS

Two faster R-CNN models (Table 1) have been selected for 2D object detection. X101-FPN shows higher box average precision (AP) while R101-DC5 requires shorter train time and less train memory (Wu et al. 2019). 302 images were trained with two models separately and the result comparison was shown in Figure 3. X101-FPN was finally selected as a faster R-CNN model to detect MEP objects because of the better performance in class accuracy and less GPU consumption.

Table 1. Performance Comparison of Two Faster R-CNN Models.

Model Name	Train Time (s/iter)	Train Mem (GB)	Box AP
X101-FPN	0.638	6.7	43.0
R101-DC5	0.452	6.1	40.6

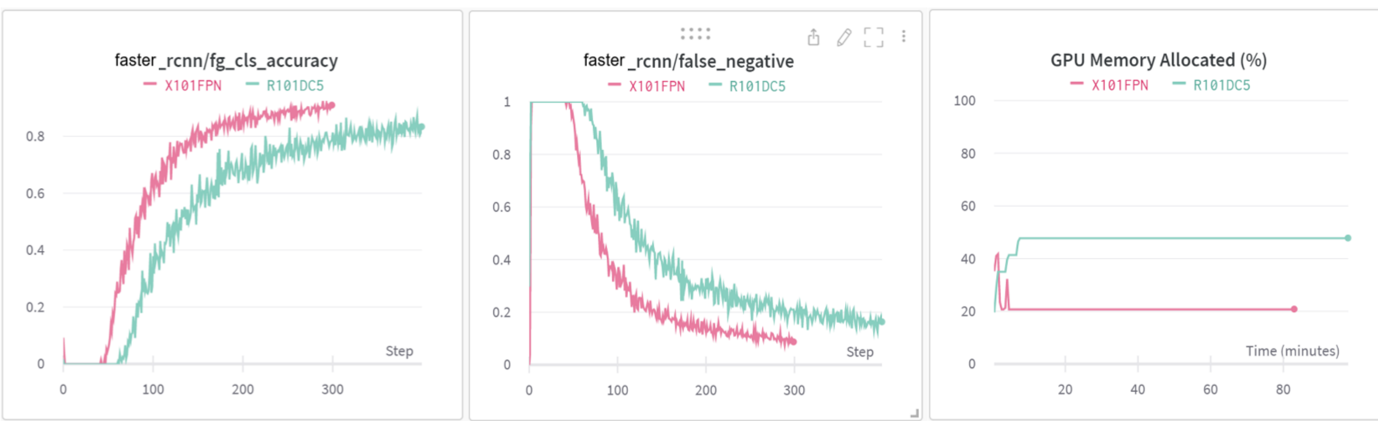

Figure 3. The training result comparison of two faster R-CNN models: X101-FPN (in pink) and R101-DC5 (in green). Models are provided by Wu et al. (2019) from Detectron2.

The intrinsic and extrinsic parameters of the camera were automatically extracted during the photogrammetry process. Therefore, the test images were automatically registered with the scanned PCD after the alignment with the photogrammetric PCD. Each pixel detected in each image was projected by the camera matrices to the 3D scanned PCD. Since the 327 test images were extracted from videos, the same object instance may exist in different images. Faster R-CNN predicted the bounding box of each object instance with its class and assigned each pixel's label to each scanned 3D point. Therefore, each point in the scanned PCD may be assigned the same label many times.

The spatial localization results of plate heating terminals are shown in Figure 4 with the prediction of different frequencies. 0.0% in Figure 4 (a1) means that no object instance was predicted in the scanned PCD so that all points (most of them are in blue) were shown in the 3D viewer. When the predicted frequency increased from 0.0% to 12.5% in (a2), two heating terminals were predicted in bright colors and located in red boxes. The frequency does not have to be very high to predict the location of MEP components in the scanned PCD since the instance can also be located when it is shown only once in a single image. The more times an identical instance showed in images, the higher the predicted frequency the point would be assigned. When the number increased to 52.9% in (a3), the viewer only showed the instances that occurred many times in images. In practice, we encourage to use a number of images for instance detection to retrieve more relevant points from PCD.

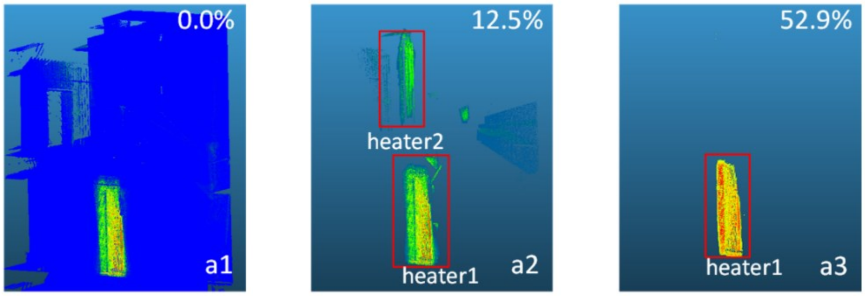

Figure 4. The spatial localization results of heating terminals (a1-a3) with the predicted frequency of points.

After locating the instances in the laser scanned PCD, we obtained coarse results of the detected heating terminals. By applying PFH for feature extraction and ICP for fine registration between the as-designed PCD and the coarse result, we can finally match and extract the corresponding points to the heating terminals. The solution can be applied for other MEP classes with various shapes. Figure 5 shows the final result of two segmented heating terminals and the corresponding ground truths. Table 2 shows a high-resolution result (average precision: 0.999, average recall: 0.876, and average intersection over union (IoU): 0.875). The instance-level point clusters can then be used to reconstruct the geometry of MEP objects and support progress monitoring and quality control during the construction and operation stages of a building.

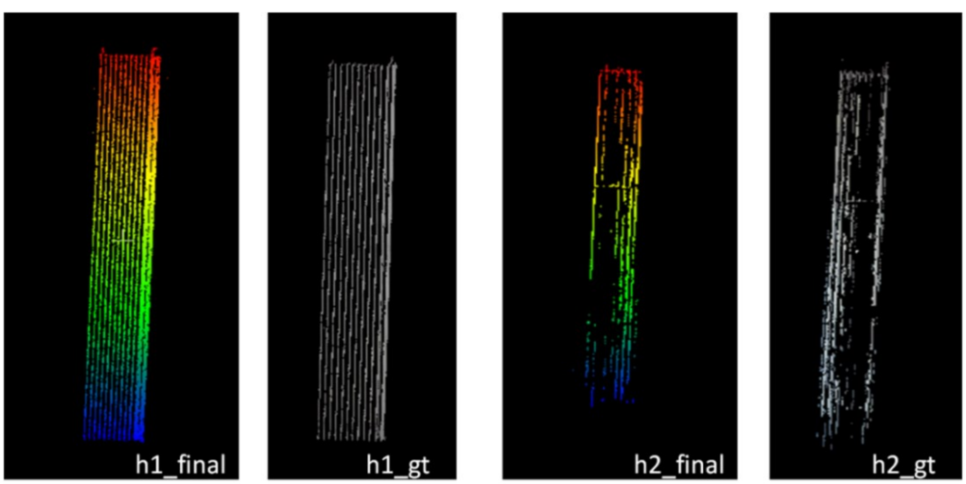

Figure 5. The visualization of final segmentation results and the related ground truth (gt) for heating terminal 1 (h1) and heating terminal 2 (h2).

Table 2. Performance Evaluation: Precision, Recall, and IoU

Metrics	Precision	Recall	IoU
h1	0.999	0.948	0.948
h2	0.998	0.803	0.802
Average	0.999	0.876	0.875

DISCUSSION

The proposed method in Figure 1 can be implemented for the localization and segmentation of MEP components in buildings in a real, complex environment. Automatic localization and segmentation of off-site as-built objects can help monitor the construction progress, and can also assist the scan-to-BIM process in the generation of DT and scan-vs-BIM to support quality control at the construction stage. The proposed solution is generic to be applied for various MEP class with different geometric properties. Further experiments will focus on more types of instances including pipes, elbows, ducts, diffusers, etc. The proposed method outperforms the state-of-the-art methods since it considers the distinct discrepancies between as-designed and as-built instances. In real-world cases, it can completely segment point cluster without missing many points and match this cluster with the corresponding as-designed model. The high-resolution performance shows better application against other advanced methods.

There are several aspects that can be improved for better solution performance. In data preparation, more training image datasets can be added from different buildings to increase the prediction's accuracy. Replacing videos by taking high-resolution images directly can also improve the prediction's accuracy. In experiments, different angles of images can influence the result of location prediction. For example, an axis-aligned bounding box may be incorrectly predicted or missed due to significant gradients of instances against axes in an image. Also, some irrelevant points can be predicted and labeled to a class because the photo was taken from a noticeable inclined angle. Therefore, taking videos or photos from the canonical view can reduce false positive or false negative results. Finally, using bounding boxes to merely localize objects can cause false positive results by predicting irrelevant points, which is less accurate than using mask segmentation. However, merely determining object location is faster than using masks in terms of generating training datasets. Also, further feature extraction can reduce the false positive results for the final outcome.

CONCLUSION

This paper proposed a hybrid method by fusing 2D images, 3D PCD, and 3D models to segment MEP components in buildings. The proposed method does not require complex 3D annotations to PCD and can deal with high-resolution segmentation when as-built instances deviate from BIMs. A case study is provided to show the feasibility and practicality of this method by the segmentation of heating terminals as complex MEP components.

In future directions, more instance segmentation will be investigated based on the proposed method. Matching mechanical and electrical instances is an essential step to facilitate quality control in the scan-vs-BIM system. Mesh reconstruction from extracted point clusters will also be investigated for geometric BIM updating and maintenance.

ACKNOWLEDGEMENT

This work is funded by European Commission's Horizon 2020 for CBIM (Cloud-based Building Information Modelling) European Training Network under agreement No.860555.

REFERENCES

Achakir, F., El Fkihi, S., & Mouaddib, E. M. (2021). Non-Model-Based approach for complete digitization by TLS or mobile scanner. ISPRS Journal of Photogrammetry and Remote Sensing, 178, 314-327.

Bueno, M., Bosché, F., González-Jorge, H., Martínez-Sánchez, J., & Arias, P. (2018). 4-Plane congruent sets for automatic registration of as-is 3D point clouds with 3D BIM models. Automation in Construction, 89, 120-134.

Drobnyi, V., Hu, Z., Fathy, Y., & Brilakis, I. (2023). Construction and Maintenance of Building Geometric Digital Twins: State of the Art Review. Sensors, 23(9), 4382.

Girshick, R. (2015). Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision, 2015 Inter, 1440–1448. https://doi.org/10.1109/ICCV.2015.169

Girshick, R., Donahue, J., Darrell, T., Malik, J., Berkeley, U. C., & Malik, J. (2014). 1043.0690. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1, 5000. https://doi.org/10.1109/CVPR.2014.81

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. 2017 IEEE International Conference on Computer Vision (ICCV), 2980–2988. https://doi.org/10.1109/ICCV.2017.322

Hu, Z., Fathy, Y., & Brilakis, I. (2022, July). Geometry updating for digital twins of buildings: a review to derive a new geometry-based object class hierarchy. In EC3 Conference 2022 (Vol. 3). University of Turin. https://doi.org/10.35490/EC3.2022.155

Hu, Z., & Brilakis, I. (2024). Matching design-intent planar, curved, and linear structural instances in point clouds. Automation in Construction, 158, 105219.

Jeelani, I., Asadi, K., Ramshankar, H., Han, K., & Albert, A. (2021). Real-time vision-based worker localization & hazard detection for construction. Automation in Construction, 121, 103448. https://doi.org/10.1016/J.AUTCON.2020.103448

Jia, F., & Lichti, D. D. (2019). A model-based design system for terrestrial laser scanning networks in complex sites. Remote Sensing, 11(15), 1749.

Jiang, Y., Pang, D., & Li, C. (2021). A deep learning approach for fast detection and classification of concrete damage. Automation in Construction, 128, 103785. https://doi.org/10.1016/J.AUTCON.2021.103785

Kaiser, T., Clemen, C., & Maas, H. G. (2019). Automated Alignment of Local Point Clouds in Digital Building Models. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 42(5/W2).

Masood, M. K., Aikala, A., Seppänen, O., & Singh, V. (2020). Multi-Building Extraction and Alignment for As-Built Point Clouds: A Case Study With Crane Cameras. Frontiers in Built Environment, 6, 581295.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), 1345-1359.
Pan, Y., Braun, A., Brilakis, I., & Borrmann, A. (2022). Enriching geometric digital twins of buildings with small objects by fusing laser scanning and AI-based image recognition. Automation in Construction, 140(February), 104375. https://doi.org/10.1016/j.autcon.2022.104375
Patil, A. K., Kumar, G. A., Kim, T. H., & Chai, Y. H. (2018). Hybrid approach for alignment of a pre-processed three-dimensional point cloud, video, and CAD model using partial point cloud in retrofitting applications. International Journal of Distributed Sensor Networks, 14(3), 1550147718766452.
Schonberger, J. L., & Frahm, J. M. (2016). Structure-from-Motion Revisited. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-Decem, 4104–4113. https://doi.org/10.1109/CVPR.2016.445
Schönberger, J. L., Zheng, E., Frahm, J. M., & Pollefeys, M. (2016). Pixelwise view selection for unstructured multi-view stereo. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14 (pp. 501-518). Springer International Publishing.
Sepasgozar, S. M., Forsythe, P., & Shirowzhan, S. (2018). Evaluation of terrestrial and mobile scanner technologies for part-built information modeling. Journal of Construction Engineering and Management, 144(12), 04018110.
Szegedy, C., Toshev, A., & Erhan, D. (2013). Deep Neural Networks for Object Detection. Advances in Neural Information Processing Systems 26 (NIPS 2013), 26.
Tang, P., & Rasheed, S. H. (2013, December). Simulation for characterizing a progressive registration algorithm aligning as-built 3D point clouds against as-designed models. In 2013 Winter Simulations Conference (WSC) (pp. 3169-3180). IEEE.
Tzutalin. LabelImg. Git code (2015). https://github.com/tzutalin/labelImg
Wu Y., Kirillov A., Massa F., Lo W.-Y., Girshick R., Detectron2, https://github.com/facebookresearch/detectron2 (2019)
Yang, J., Li, H., & Jia, Y. (2013). Go-icp: Solving 3d registration efficiently and globally optimally. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1457-1464).
Li, P., Wang, R., Wang, Y., & Tao, W. (2020). Evaluation of the ICP algorithm in 3D point cloud registration. IEEE Access, 8, 68030-68048.
Rusu, R. B., Blodow, N., & Beetz, M., and Marton Z.C. (2008). Aligning Point Cloud Views using Persistent Feature Histograms. In Proceedings of the 21st IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Nice, France
Rusu, R. B., Blodow, N., & Beetz, M. (2009, May). Fast point feature histograms (FPFH) for 3D registration. In 2009 IEEE international conference on robotics and automation (pp. 3212-3217). IEEE.
Hu, Z., & Brilakis, I. (2022, October). PriSeg: IFC-Supported Primitive Instance Geometry Segmentation with Unsupervised Clustering. In European Conference on Computer Vision (pp. 196-211). Cham: Springer Nature Switzerland.
Pan, Y., Hu, Z., & Brilakis, I. (2023). Digital Twins and Their Roles in Building Deep Renovation Life Cycle. In Disrupting Buildings: Digitalisation and the Transformation of Deep Renovation (pp. 83-96). Cham: Springer International Publishing.