Towards Automated Digital Building Model Generation from Floorplans and On-Site Images

Niklas Gard[1,2,*] and Aleixo Cambeiro Barreiro[1,*]

[1]Vision and Imaging Technologies, Fraunhofer HHI, Einsteinufer 37, Berlin, Germany

[2]Institute for Computer Science, Humboldt University of Berlin, Unter den Linden 6, Berlin, Germany

[*]Both authors contributed equally to this work.

E-mail(s): niklas.gard@hhi.fraunhofer.de, aleixo.cambeiro@hhi.fraunhofer.de

Abstract: Digitalization of buildings has become increasingly relevant for processes such as construction or refurbishing. However, the lack of availability of digital building models is a common problem. In many cases, only printed floorplans and photos of the interior of a building are available. We propose a system to automatically generate enriched digital models from this data, consisting of two AI modules: one for 3D model reconstruction from 2D plans and one for 6D localization of images taken within a building in the corresponding 3D model. Such a model can facilitate tasks such as monitoring of the condition of a building or defects of its components. We demonstrate this pipeline using a real-world building, from which a plan and pictures of a floor are available, and show that good results can be achieved, with potential to greatly reduce the human effort normally required to create digital twins of buildings.

Keywords: Deep Learning, Floorplan Analysis, Pose Estimation, Localization, BIM

Erschienen in Tagungsband 34. Forum Bauinformatik 2023, Bochum, Deutschland, *DOI:* 10.13154/294-10132

1 Introduction

A digital building model serves as a fundamental requirement for numerous applications throughout the entire life cycle of a building, encompassing essential areas such as building maintenance, refurbishment, and energy and resource management. Unfortunately, for the majority of existing buildings, the availability of such a model is scarce. This scarcity arises due to the considerable costs associated with its creation and the absence of digital planning practices during the construction phase. On the other hand, printed floorplans are readily available for nearly every building. Moreover, in many instances photos of the rooms captured, for example, during building inspections for documentation purposes or for real estate advertising are available as well. The combination of these sources of information has many potential applications, such as the visual analysis of the current state of different parts of the building. To this end, the photos must be matched to locations within the plan, which can be a difficult and time-consuming process if done manually.

In this paper, we tackle this issue with the introduction of an innovative approach that capitalizes on these existing inputs—floorplans and captured room photos—to generate a 3D model and subsequently register virtual camera positions corresponding to the photos within the reconstructed model. This registration process serves as the key to enriching the digital model with useful information. For example, reintegrating objects detected in the images back into the digital model would allow to update or correct existing information and accurately register defects or damaged components. The registration of pictures taken over time would also offer time snapshots of the actual state of the building throughout the years, leveraging the immutability for long periods of the structural design of most buildings for the localization.

2 Related Work

Floorplan Analysis

Automatic floorplan analysis is a complex task due to the variations in appearance of the different symbols and their potential overlapping with measurement annotations, text or even manual modifications of the printed documents. Early approaches addressing this problem were based on traditional computer vision techniques, such as morphological operations and handcrafted features, as in [1], [2]. Modern approaches have, however, shifted towards the use of deep-learning models due to their superior results, such as [3], [4], which constitute some of the state-of-the-art methods.

Kalervo et al. [3] improve upon [5] following a two-step approach. First, semantic segmentation and junction keypoints are extracted, creating a so-called junction layer. This information is then combined and integer programming is used to yield a so-called primitive layer, which is further refined into a vectorial output. A limitation of this method is that it only works for axis-aligned vertical and horizontal walls. Lv et al. [4] propose a different pipeline, in which a ROI for the actual building plan within the image is extracted and processed to obtain a semantic segmentation map of the structural information, to which a vectorization algorithm is applied. Moreover, symbols and text are extracted which, in combination with the length of extracted measure lines, are used to determine the scale of the plan.

The work of Cambeiro Barreiro et al. [6], which we have followed in our work, uses some of the techniques presented in the aforementioned works and proposes some modifications and new ideas, yielding state-of-the-art results on the public dataset CubiCasa5k [3]. This approach will be discussed in more detail in the following section.

Image localization

Many existing deep learning models for indoor 6D localization rely on prior scene knowledge during training. Some models establish correspondences between test images and localized images from a pre-generated image database, often created using structure-from-motion techniques [7]. Others encode the entire scene in a neural network and regress the absolute pose [8]. Acharya et al. [9] propose a visual localization approach that eliminates the need for manual image capturing for 3D reconstruction. Their method, is trained on synthetic renderings from a 3D BIM-model and achieves real-time indoor localization with approximately 2 meters accuracy. However, it is limited to the specific building used during fine-tuning and does not generalize to new buildings.

In contrast, Liu et al. [10] were the first to localize images directly within a floorplan. Their approach does not directly estimate the 3D pose but assumes a cuboidal layout and predicts the wall to which the camera is pointed. Lalaloc [11], predicts the relative panoramic camera location with respect to a grid of reference renderings. The authors focus on bridging the gap between real images and semantic renderings by constructing a robust latent space and mapping camera images in that space, similar to our approach. They extended their approach [12] to reduce rendering efforts and directly match the coded image with a coded representation of the plan. This is efficient, but also poses challenges when incorporating additional information, such as varying window heights. So far, none of the methods that rely solely on floorplans is able to directly estimate full 6D perspective camera poses.

3 Method

The input of our method is an image of a floorplan and multiple 2D perspective camera images taken within the building floor shown in the plan, and optionally prior knowledge about image to room associations. The output is a 3D model created from the floorplan and the 6D camera poses (rotation and translation) of the images in that model. We propose a modular system that can be expanded with other modules, for example to enrich the model with details from the localized image.

3.1 Data handling

The localization and reconstruction pipeline used in our study relies on a simplified 3D building representation proposed in the Structured3D (S3D) dataset [13]. It limits the semantic classes to simplified representations of walls, ceilings, floors, windows and doors. The floors are assigned to rooms and are annotated with a room type (e.g. corridor, office) and a unique identifier. This fundamental information can be extracted from all commonly available floorplans. The extracted information from floorplan analysis is converted into a human-readable JSON format, designed to accommodate future additions of more detailed information that could be derived from localized images or to be ultimately converted to an Industry Foundation Classes (IFC) model. This is exchanged between the plan analysis and localization modules and converted internally into S3D format.

3.2 Reconstruction of a 3D Model

The 3D reconstruction pipeline [6], illustrated in Figure 1, follows a series of steps to generate a 3D model from a 2D plan. First, a Faster-RCNN [14] with a ResNet [15] backbone is used for symbol detection, in order to extract windows and doors from the plan. At the same time, an FPN [16] with a

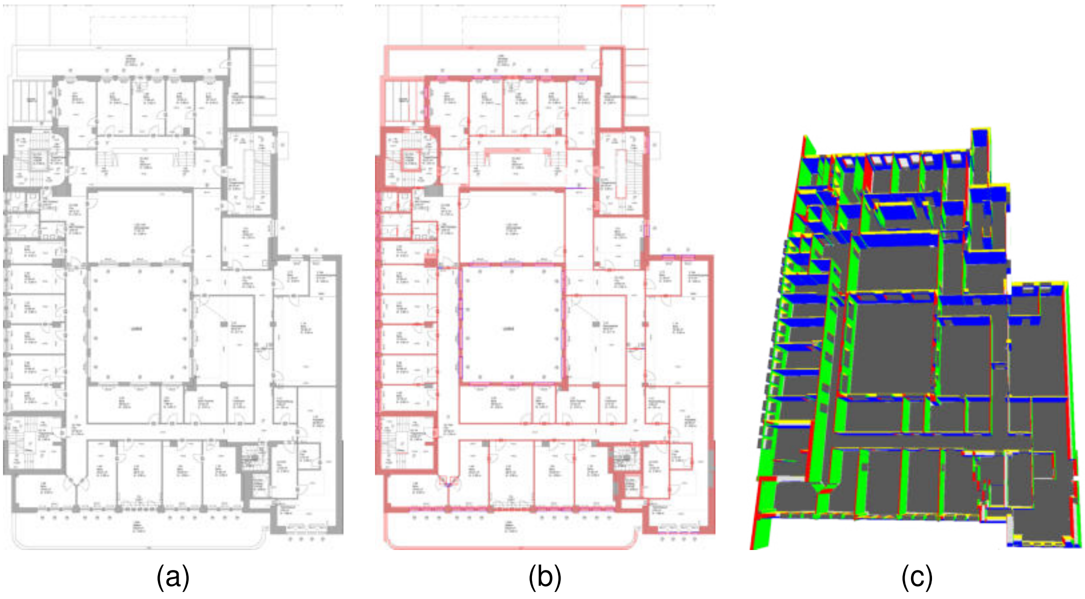

Figure 1: 3D model reconstruction pipeline: **a)** shows the input floorplan, corresponding to an office building, **b)** shows an overlay with the extracted, vectorial walls and the detected doors and windows and **c)** shows a view of the reconstructed 3D model.

ResNet [15] backbone is used to perform semantic segmentation in order to extract a pixelwise mask of the walls. A sliding-window approach ensures that the method can be applied to images of any scale, which is important due to the high resolution of many floorplan images. Data augmentation during training helps maintain the performance for inputs of different scales and alignments.

In a next step, the walls segmentation mask is processed by a vectorization algorithm, which divides it into rectangular wall segments. This algorithm first determines the main wall orientations present in the mask and then extracts the components for each of these orientations and approximates their irregular shapes with rotated rectangular boxes. This vectorial representation is then used to adjust the locations and sizes of detected doors and windows, merging all the information into a joint hierarchical structure. This data is then converted to the S3D format and a simple 3D model that can be used by the image localization module is generated using default values for information that cannot currently be extracted from the floorplan, such as height of walls, windows and doors.

3.3 Localization of the Virtual Camera

The localization algorithm follows a two-stage training inspired by LaLaLoc [11]. However, it is modified to estimate the full 6D pose of a perspective camera instead of the 3D position of a panoramic image.

First, a range of images from randomly sampled camera poses is presented to the network in a "layout" representation, consisting of semantic segmentation and normal maps. A robust latent space is constructed using an auto-encoder with a ResNet18-encoder, a small bottleneck and a decoder without skip connections. The encoder condenses the information from the layout into a vector with 256 values. In this stage, the network processes only the layout representations, not actual photos.

In the second training phase, the general task is to learn a viewport matching between a rendered panorama "layout" representation of a room and a query camera image, as well as the relative 6D pose offset between the reference panorama and the query image. Inspired by DTOID [17], the query image is treated as a template and matched with the panorama. A query image encoder generates tunable filters as a side output, which are injected into a panorama processing backbone to allow the heads of the panorama network to predict a 2D bounding box representing the camera's field of view in the panorama, via an anchor-based regression. Simultaneously, a separate network head in the panorama decoder estimates the relative pose between the perspective camera and the panoramic camera. It utilizes a Graph Neural Network (GNN) approach [18] to fuse results from nearby panorama renderings. This fusion mechanism leverages information from neighboring panoramas, enhancing pose estimation accuracy. The robust layout embedding network from the first phase provides additional guidance. The query image encoder is trained to map an image to the same embedding space to which the frozen layout encoder maps the underlying layout.

During inference, reference panorama renderings are generated for positions on a 2D grid across the floorplan (Figure 2). The query image is matched with each of the reference panoramas. From these matches, the bounding box with the highest classification accuracy is selected. The 6D pose is calculated using the GNN head with the encodings of three nearby panorama renderings.

4 Discussion and Evaluation

For our tests, we used photos taken inside a real world building and the corresponding 2D floorplan. The floorplan analysis module was trained on the CubiCasa5k [3] dataset, with around 5000 images of floorplans and the corresponding ground truth divided into training and validation sets. The plans seen in the dataset, which correspond to residential buildings, are substantially different from the test input of an office building in scale, symbols, structural layouts and style. Despite of this, it is possible to see in Figure 1 that the model has adapted reasonably well to the new data, accurately detecting most of the doors and windows and correctly vectorizing the walls to yield a faithful 3D reconstruction of the floor. Concretely, the intersection over union (IoU) of the predicted walls segmentation mask with respect to the ground truth is of 71.5%, while the IoU for a mask generated from the vectorized

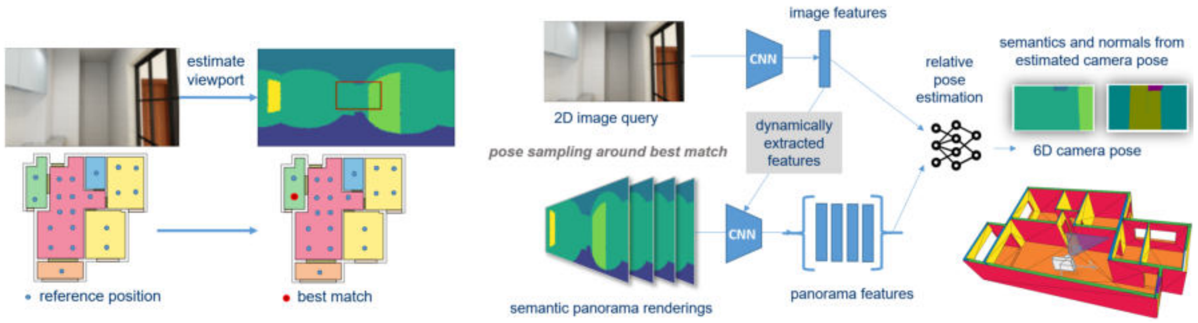
estimate viewport
reference position
best match
image features
CNN
2D image query
pose sampling around best match
dynamically extracted features
CNN
panorama features
semantic panorama renderings
relative pose estimation
semantics and normals from estimated camera pose
6D camera pose

Figure 2: Pose inference: Searching for the best matching viewport (left) and calculating the relative pose (right). Both steps use the same backbones with different heads.

walls is of 69.3%. These numbers indicate a good result, since IoU strongly penalizes deviations from the ground truth, but they also show the strong performance of the vectorization algorithm, which effectively approximates the wall segments with rectangular boxes.

The localization module is trained on 400 scenes from the S3D [13] dataset, each containing around 20 images along with the semantic 3D model. Training employs the "empty" setting, representing unfurnished rooms, and each image is presented to the network approximately 25-30 times. In an initial experiment, we evaluate on a test set of 20 previously unseen apartments from the same dataset. We consider the retrieval accuracy of the Top-1 and Top-5 results, i.e. the rate of predictions for which the estimated error of the camera pose remains within thresholds with respect to the ground truth.

Table 1: Localization evaluation for Scenes 400-419 of S3D. Percentage of correct estimates within various distance and angular error thresholds using a sample grid resolution of 1.2m $\times$ 1.2m.

	Error Threshold			
Rank	$<$ 50cm	$<$ 100cm	$<$ 50cm, 10 deg	$<$ 100cm, 10 deg
Top 1	29.4%	40.9%	27.3%	35.5%
Top 5	54.8%	71.0%	50.2%	61.9%

As shown in Table 1, in 40% of the cases, the Top-1 match is within 1 meter of the reference position. This is a promising result regarding the semantic 3D model's limited information, geometric ambiguities, and the camera's limited field of view. It suggests that both the room and the correct camera pose have been accurately estimated. Regarding the Top-5, the accuracy increases to 71%. In a real-world application, it is feasible to rely on user input to select the correct position. Additionally, incorporating prior knowledge about the room or a set of rooms can further enhance the accuracy of pose estimation.

Qualitative results using images from the reference building are shown in Fig. 3. We found that in the large floorplan with more monotonous structures a room to image association was required to correctly estimate most poses. Our tests revealed opportunities for future improvements to adress this. First,

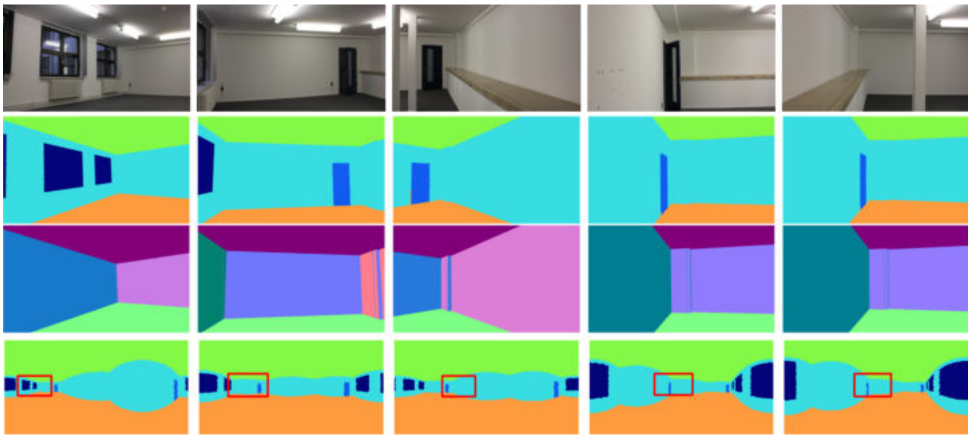

Figure 3: Visualization of pose estimation results for room images in our reference building. Room is given as prior knowledge. **2nd and 3rd row**: Rendered semantics and normals for predicted viewpoint. **Bottom row**: Highest-score bounding box in panorama. **Right column**: Failure case: normals align with the image, but a corner with a door was mistakenly selected.

extracting additional semantic information from the floor plan will make the matching more unique. Also, soft priors, such as knowledge that multiple pictures are captured in the same room, can enhance the matching process. Lastly, structural similarity between training and testing 3D representations is desirable. For instance, if the test 3D model includes freestanding columns absent in the training data, it can cause confusion for the model. Our experiments with the localization module demonstrate a proof of concept. A more detailed evaluation will be presented in future work.

5 Conclusion

In this paper, we have proposed an automated pipeline for plan analysis, 3D model reconstruction and 6D image localization aiming to simplify digitalization of buildings, leveraging the synergy of two different AI modules. We have shown that, despite having been trained on very different data, they show promising generalization capabilities, yielding in many cases good results for 6D pose estimation of pictures taken in a real world building within the 3D model reconstructed from a 2D plan of it. While the domain gap and challenging nature of the real data used in this experiment led to incorrect predictions in certain situations, our system offers a ranking of the most likely camera poses, that is much more likely to contain the correct pose that could be chosen with minimal user intervention.

Our system paves the way to the creation of enriched 3D models, potentially removing much of the manual work required in this process. Some future improvements to make it more effective could include training with data more specific to the target domain or the combination with other modules to further enrich the model. For example, a text detection module such as the one proposed by Schönfelder et al. [19] could be used to extract additional information such as walls height or room types if present in the plan. Furthermore, relevant objects such as fire extinguishers [20] or additional construction elements [21] could be detected in 2D pictures and then be placed in the reconstructed 3D model making use of the image localization.

Acknowledgements

This work has partly been funded by the German Federal Ministry of Economics Affairs and Climate Action project BIMKIT under grant number 01MK21001H and the German Federal Ministry for Digital and Transport project EConoM under grant number 19OI22009C.

References

[1] S. Ahmed, M. Liwicki, M. Weber, and A. Dengel, “Improved automatic analysis of architectural floor plans”, in *International conference on document analysis and recognition*, IEEE, 2011.

[2] L.-P. De Las Heras, S. Ahmed, M. Liwicki, E. Valveny, and G. Sánchez, “Statistical segmentation and structural recognition for floor plan interpretation: Notation invariant structural element

recognition", *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 17, no. 3, pp. 221–237, 2014.

[3] A. Kalervo, J. Ylioinas, M. Häikiö, A. Karhu, and J. Kannala, "Cubicasa5k: A dataset and an improved multi-task model for floorplan image analysis", in *Image Analysis: 21st Scandinavian Conference, SCIA 2019, Norrköping, Sweden, June 11–13, 2019, Proceedings 21*, Springer, 2019, pp. 28–40.

[4] X. Lv, S. Zhao, X. Yu, and B. Zhao, "Residential floor plan recognition and reconstruction", in *Proc. CVPR*, 2021, pp. 16 717–16 726.

[5] C. Liu, J. Wu, P. Kohli, and Y. Furukawa, "Raster-to-vector: Revisiting floorplan transformation", in *Proc. ICCV*, 2017, pp. 2195–2203.

[6] A. Cambeiro Barreiro, M. Trzeciakiewicz, A. Hilsmann, and P. Eisert, "Automatic Reconstruction of Semantic 3D Models from 2D Floor Plans", in *18th International Conference on Machine Vision Applications (MVA)*, 2023.

[7] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii, "InLoc: Indoor visual localization with dense matching and view synthesis", in *Proc. CVPR*, 2018, pp. 7199–7209.

[8] T. Naseer and W. Burgard, "Deep regression for monocular camera-based 6-dof global localization in outdoor environments", in *Proc. IROS*, IEEE, 2017, pp. 1525–1530.

[9] D. Acharya, K. Khoshelham, and S. Winter, "BIM-PoseNet: Indoor camera localisation using a 3D indoor model and deep learning from synthetic images", *ISPRS journal of photogrammetry and remote sensing*, vol. 150, pp. 245–258, 2019.

[10] C. Liu, A. Schwing, K. Kundu, R. Urtasun, and S. Fidler, "Rent3D: Floor-Plan Priors for Monocular Layout Estimation", in *Proc. CVPR*, 2015.

[11] H. Howard-Jenkins, J.-R. Ruiz-Sarmiento, and V. A. Prisacariu, "Lalaloc: Latent layout localisation in dynamic, unvisited environments", in *Proc. ICCV*, 2021, pp. 10 107–10 116.

[12] H. Howard-Jenkins and V. A. Prisacariu, "LaLaLoc++: Global Floor Plan Comprehension for Layout Localisation in Unvisited Environments", in *Proc. ECCV*, Springer, 2022, pp. 693–709.

[13] J. Zheng, J. Zhang, J. Li, R. Tang, S. Gao, and Z. Zhou, "Structured3d: A large photo-realistic dataset for structured 3d modeling", in *Proc. ECCV*, Springer, 2020, pp. 519–535.

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks", *Advances in neural information processing systems*, vol. 28, 2015.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", in *Proc. CVPR*, 2016, pp. 770–778.

[16] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection", in *Proc. CVPR*, 2017, pp. 2117–2125.

[17] J.-P. Mercier, M. Garon, P. Giguere, and J.-F. Lalonde, "Deep Template-Based Object Instance Detection", in *Proc. WACV*, Jan. 2021, pp. 1507–1516.

[18] M. Ö. Türkoğlu, E. Brachmann, K. Schindler, G. Brostow, and Á. Monszpart, “Visual Camera Re-Localization Using Graph Neural Networks and Relative Pose Supervision”, in *Proc. 3DV*, 2021.

[19] P. Schönfelder and M. König, “Deep learning-based text detection on architectural floor plan images”, in *IOP Conference Series: Earth and Environmental Science*, vol. 1101, 2022.

[20] H. Bayer and A. Aziz, “Object Detection of Fire Safety Equipment in Images and Videos using Yolov5 Neural Network”, in *Proceedings of 33. Forum Bauinformatik*, 2022.

[21] N. Gard, A. Hilsmann, and P. Eisert, “Combining Local and Global Pose Estimation for Precise Tracking of Similar Objects”, in *Proc. VISAPP*, 2022, pp. 745–756.