Indoor-Outdoor 3D Reconstruction Alignment

Andrea Cohen[1⋆], Johannes L. Schönberger[1⋆], Pablo Speciale[1], Torsten Sattler[1], Jan-Michael Frahm[2], Marc Pollefeys[1,3]

[1]ETH Zürich, [2]UNC Chapel Hill, [3]Microsoft

Abstract. Structure-from-Motion can achieve accurate reconstructions of urban scenes. However, reconstructing the inside and the outside of a building into a single model is very challenging due to the lack of visual overlap and the change of lighting conditions between the two scenes. We propose a solution to align disconnected indoor and outdoor models of the same building into a single 3D model. Our approach leverages semantic information, specifically window detections, in multiple scenes to obtain candidate matches from which an alignment hypothesis can be computed. To determine the best alignment, we propose a novel cost function that takes both the number of window matches and the intersection of the aligned models into account. We evaluate our solution on multiple challenging datasets.

1 Introduction

Recent progress in the area of 3D reconstruction enables the generation of large-scale [12] and detailed outdoor models [25], as well as accurate indoor reconstructions [9] and their floor-plans [2,19]. The resulting 3D models are useful for a wide range of applications, from virtual tourism [16,27], visualization of apartments for real estate [19], cultural heritage [7,23,29], and image-based localization [18,30], to real-time camera pose tracking on mobile devices for Augmented Reality [22], and autonomous navigation [20]. Ideally, a single joint reconstruction of the interior and exterior is desirable as it would, for example, enable a user to seamlessly enter buildings in a virtual city model rather than only exploring the outside. Similarly, a combined model would allow autonomous robots to easily transition between the indoor and outdoor world. However, state-of-the-art approaches often fail to reconstruct both parts into a single 3D model.

Obtaining a joint indoor-outdoor model is hard for multiple reasons: on the one hand, the indoor and outdoor parts of a scene typically exhibit a weak connection through a limited number of visual observations such as doorways or windows. As a result, great care must be taken when capturing data to ensure enough visual overlap for feature matching and to prevent the models from being disconnected [28]. This problem is often aggravated by the fact that there can be a strong change in illumination in transition areas. In practice, Structure-from-Motion (SfM) models disconnect quite often, even when an experienced

⋆ Equal contribution from both authors.

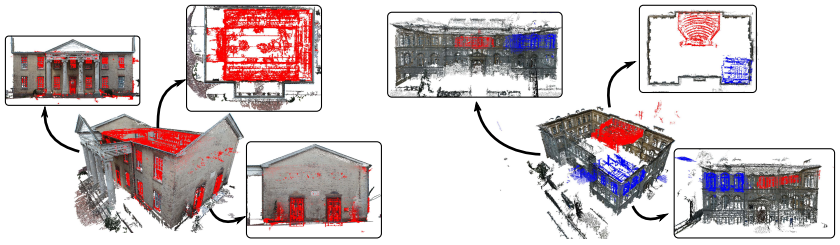

Fig. 1. The proposed method aligns disconnected Structure-from-Motion reconstructions of the inside and outside to produce a single 3D model of a building. Our approach also handles incomplete reconstructions and multiple indoor models (*c.f.* right model)

user carefully takes images of a single outdoor scene [4]. It is, thus, very hard to connect indoor and outdoor scenes through feature matches reliably for most practical applications. On the other hand, even if we capture enough imagery to visually connect indoors and outdoors, *e.g.* by recording a video sequence, the connections are usually rather weak. Consequently, it is hard to prevent drift between the two models. Additionally, indoor reconstructions are often incomplete and disconnected, *e.g.* when some rooms are not accessible. This makes the alignment problem even harder, since several indoor models have to be aligned to one or more outdoor models for which the relative scale is also unknown. For the case of incomplete models, the solution might be ambiguous even for humans without prior knowledge of the building.

In this paper, we propose an alignment algorithm that exploits scene semantics to establish correspondences between indoor and outdoor models. More precisely, we exploit the fact that the windows of a building can be seen both from the inside and the outside. Towards this goal, we apply semantic classifiers to detect windows in the indoor and outdoor scenes. A single match between an indoor and outdoor window determines an alignment hypothesis (scale, rotation, translation) between the two models. All hypotheses are inspected and grossly wrong alignments are detected and discarded using a measure of intersection of the two models. Plausible alignments are then further refined using additional window matches. Our approach is robust to noisy window detections and is able to align disconnected indoor and outdoor models (*c.f.* Fig. 1). Furthermore, our method can handle both multiple and/or incomplete indoor or outdoor models.

Concretely, we make the following contributions: we present a novel approach for aligning indoor and outdoor reconstructions of a building by detecting and aligning windows in both models. We propose a novel quality metric for the resulting alignment based on detecting intersections between the two models. We exploit multi-view redundancy to ensure robustness to noisy window detections. As a result, our proposed algorithm is able to tackle the challenging problem of joining indoor and outdoor models of a building into a single reconstruction. In addition, our method works purely on sparse point clouds and does not require

any dense geometry. We demonstrate the practical applicability of our approach on multiple challenging datasets.

2 Related Work

There is a clear trend of using higher level (semantic) information for both sparse and dense 3D reconstruction: Ceylan *et al.* [3] and Cohen *et al.* [6] actively detect and exploit symmetries and repetitions to improve the quality of SfM reconstructions. Häne *et al.* [11] and Savinov *et al.* [24] combine semantic image classification and dense 3D reconstruction, showing that jointly optimizing over the labels and the shape of the 3D model improves the results for both. All these methods use a higher level understanding of the scene to optimize the reconstruction results. In contrast, we use semantic information to enable reconstruction in the first place by aligning indoor and outdoor models that cannot be related by low-level feature matches alone.

Indoor reconstruction approaches usually exploit Manhattan world assumptions to obtain clean, dense 3D models from streams of photos [9, 29]. Given a set of panoramas as input, Cabral and Furukawa [2] determine for each pixel whether it belongs to the floor, wall, or ceiling. Given these structural classifications, they estimate a piecewise planar floor plan and create a compact, textured mesh from the generated plan. While the previous approaches operate on densely sampled images, Liu *et al.* [19] estimate the layout of each room from a sparse set of photos. Prior knowledge of the floor plan and semantic classification is then used to align the individual rooms. Recently, Ikehata *et al.* [13] showed that parsing the structure of the scene can significantly aid the indoor reconstruction process. They reason about the semantic relation between different scene parts and the structure of the rooms and use this knowledge during reconstruction.

Martin *et al.* [21] and Cohen *et al.* [4] consider the problem of aligning visually disconnected 3D models without using traditional feature matches. Martin *et al.* determine the room layout of individual 3D models by solving a jigsaw puzzle problem, utilizing annotated floor plans and the temporal flow of crowds between rooms. Cohen *et al.* reason about the spatial arrangement of individual sub-models to obtain a closed model of the outside of a single building. Their method is based on determining potential connection points between the models and detecting free-space violations using semantic information. The two approaches solely focus on indoor [21] and outdoor reconstructions [4], respectively. In contrast, our approach addresses the problem of linking previously disconnected indoor and outdoor models. In addition, we also show that indoor models can help to connect partial outdoor models and vice-versa.

Strecha *et al.* [28] reconstruct both the outside and the inside of a historic castle. In contrast to our method, which does not constrain the capture setup, Strecha *et al.* heavily constrain the capture to be able to reconstruct the whole scene as a single model. In particular, they very carefully take images with high visual overlap between indoors and outdoors to prevent the reconstruction from

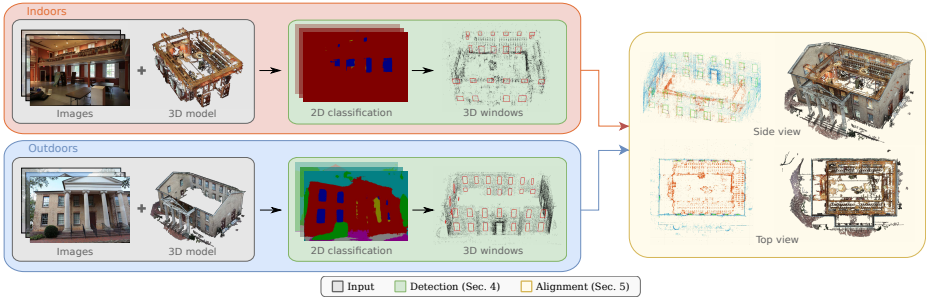

Fig. 2. Given SfM reconstructions of indoors and outdoors together with their input images, we leverage per-pixel semantic classification to detect windows in 3D. These windows are then used to compute a registration between both scenes that maximizes the number of aligned windows while avoiding that the models intersect each other

disconnecting into multiple sub-models. Often, this is impractical or even impossible. Hence, our approach is specifically designed to handle separate models.

Simultaneously to our work, Koch *et al.* [14] also developed a method to tackle the problem of indoor-outdoor model alignment using 3D line matching. 3D lines are detected using the original images and the reconstructed (separate) models. The models are then aligned using the transformation that matches the highest number of line segments. The method assumes that the 3D line segments found are mostly located on windows and doors, indirectly matching these structures between both models without explicitly using semantics, as opposed to our method. In addition, they also need to know the scale of both models in advance and they only deal with one indoor and one outdoor model. Our method overcomes these limitations. Both works are complementary, since our method's results could be used as input for [14] which would act as a refinement step.

3 Method Overview

Given separate indoor and outdoor models, we propose to align the inside and outside of a building through semantic information. Specifically, as windows are visible both from inside and outside, we use window detections to generate correspondences between the two models, which are then used to compute the alignment between the models. This approach naturally extends to room-to-room registrations by detecting and aligning doors. However, similarly to [19], we found that door detection performs poorly. In this paper, we thus focus on indoor-to-outdoor alignments via window detections.

In the following, we provide an overview of our algorithm, as illustrated in Fig. 2, before presenting algorithmic details in the next sections. As input, our method uses sparse SfM models of the indoor and outdoor scenes, as well as the images used to generate them.

Window detection. First, we apply a per-pixel classifier to detect windows in all input images. For each image, we employ a façade parsing approach on

the rectified images, similar to Cohen *et al.* [5], to obtain the 2D rectangles that most likely correspond to the actual windows seen in the photos. Next, we use the known camera poses and the sparse 3D scene points to estimate the 3D planes containing the windows. Leveraging the SfM points, we estimate 3D window positions for each image individually. We then detect overlapping 3D windows and compute consensus window positions. Using all images in the reconstruction also allows us to handle occlusions more robustly, *e.g.*, due to vegetation in front of a façade. Sec. 4 describes this process in more detail.

Model alignment. Given 3D window detections for the indoor and outdoor models, we next register the disjoint models based on window correspondences. Computing the alignment boils down to finding a similarity transformation between the models, which can be computed from three point correspondences in the general case and from two point matches if the gravity direction is known. One potential approach would be to simply obtain point correspondences by aligning the centers of gravity of the windows and apply RANSAC [8] to estimate the transformation. However, the appearance of a window can change dramatically when viewed from the inside and the outside, *e.g.*, due to illumination changes or by actually looking through the window. As such, we need to consider each pair of indoor and outdoor windows, which means RANSAC-based approaches quickly become infeasible. Consider a simple case, where 20 windows are detected for the outdoor model and 3 windows are detected for a partial indoor reconstruction of a corner room. There are 1140 (resp. 190) potential combinations to draw 3-(resp. 2)-tuples of window matches. Out of all these configurations, exactly one is correct, leading to inlier ratios below 1%.
In order to avoid the combinatorial growth in complexity, we exploit the width and height of the 3D window detections to estimate a similarity transformation from a single window correspondence. Using a single match allows us to exhaustively generate the set of all possible alignment configurations. In the previous example, there are 60 potential combinations, out of which 3 are correct. The obtained alignments are then ranked based on the fact that the indoor models must not intersect with the outdoor model by enforcing free-space constraints. Alignments that violate this constraint are discarded. Otherwise, we determine the window support of the transformation, *i.e.*, the number of correctly aligned indoor-outdoor window detections, and refine the best alignment in an iterative procedure. Sec. 5 provides details on the alignment process.

Handling ambiguities due to symmetries. Given a reconstruction of a single floor in a multi-story building, it can be impossible to determine to which floor the model belongs if the windows are symmetric between floors[1]. We thus determine the number of floors and estimate the best alignment per floor, enabling a user to choose a transformation and hence resolve the ambiguity.

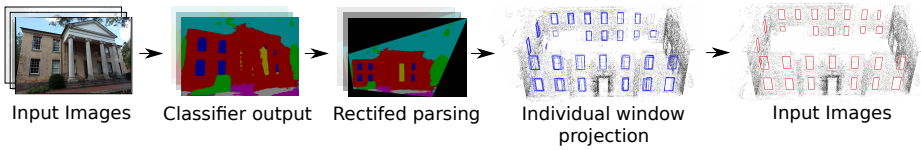

Fig. 3. The proposed 3D window detection pipeline

4 Window Detection

In this section, we describe our window extraction approach (*c.f.* Fig. 3). Our approach leverages a pre-trained per-pixel classifier to detect windows in all of the images used to reconstruct the indoor and outdoor models. The labels for image i are fed into a façade parsing algorithm to obtain a set $\mathcal{W}_{\text{2D}}(i)$ of 2D window detections, where each window $w \in \mathcal{W}_{\text{2D}}(i)$ is defined by its four corners. For each 2D window, we obtain a corresponding 3D window by projecting it onto a 3D plane estimated using the sparse SfM points. As shown in Fig. 3, these individual window projections are not necessarily consistent between images. We thus use all individual window projections to compute a consensus set $\mathcal{W}_{\text{3D}}$ of 3D windows that is consistent across all images of a model. This window detection pipeline is applied separately on each indoor and outdoor model.

Image classification. We use the supervised learning method of Ladický *et al.* [17] to obtain a pixel-wise semantic classification of the images used for reconstruction. Since we found that a classifier trained on indoor images performs poorly on photos taken on the outside and vice-versa, we train two separate classifiers. For training the indoor classifier, we use the annotated datasets provided by [19]. To train the outdoor classifier, we use the eTrims dataset [15]. The classification scores can then be used in a façade parsing algorithm to obtain the best scoring set of windows per image.

Natural frame estimation. To simplify the subsequent steps of our procedure, we align each 3D model into a canonical coordinate system. We choose the coordinate system that is aligned to the façade directions of the building. To achieve this, we determine the main axes of each model by estimating the vanishing points in each input image. The vanishing points then vote for the three coordinate directions. Next, we align the coordinate system of each 3D model with the x-y-z-axes, such that the vertical axis is aligned with z and walls are mostly aligned with the x or y direction under a Manhattan world assumption.

Image rectification and façade parsing. Following most works on indoor reconstruction [9,13,29], we use the Manhattan world assumption. This assumption is not strictly necessary, but simplifies and robustifies further processing and allows us to restrict our search for window planes to those parallel to the x-z and y-z planes. We therefore rectify all images w.r.t. x- and y-aligned planes to synthesize fronto-parallel images of the walls (*c.f.* step 3 in Fig. 3). The façade

[1] Again, detecting doors could resolve these ambiguities for the ground floor, but it would still remain for other floors.

parsing algorithm presented in Cohen *et al.* [5] is then used to extract the set $\mathcal{W}_{2D}(i)$ of 2D windows for image i by obtaining the four corner vertices of the rectangles corresponding to window detections in the rectified image. For outdoor models, this method also provides the number of floors detected per image. As discussed in Sec. 3, knowing the number of floors for a building enables our alignment approach to generate multiple plausible hypotheses if the indoor model only covers a single floor. Additionally, assuming that all windows on the same floor have the same height, façade parsing can better handle incomplete window labellings (e.g., due to occlusion). It requires < 1 second per image and provides better results compared to directly extracting windows from the semantic labels.

Individual window projection. Let $\mathcal{W}_{2D}(i) = \{w_1^i, \ldots, w_n^i\}$ be the set of 2D windows detected in the previous step for image i, and let $P^i = \{p_0^i, \ldots, p_m^i\}$ be the set of 3D SfM points that are visible in image i. We extract the subset $P'^i \subset P^i$ of points whose projections in the image fall inside any of the detected 2D windows w_j^i, $j = 1, \ldots, n$. We then use P'^i to estimate the window plane π as the best fitting plane parallel to either the x-z or y-z plane. The normal of the plane π is chosen to agree with the direction to which image i was rectified for the window extraction. All windows w_j^i are then projected onto π to obtain a set of 3D windows $\mathcal{W}_{3D}(i) = \{W_j^i\}_{j=1,\ldots,n}$ for image i.

Window grouping and consensus. Given the sets of 3D windows $\mathcal{W}_{3D}(i)$ detected for each *individual* image i, we next group the overlapping 3D windows from *all* images into clusters C. All 3D windows from the same cluster are then used to estimate a single 3D consensus window (*c.f.* the last two stages in Fig. 3).

First, we cluster all 3D windows that overlap and are on the same plane (up to a threshold computed as 20% of the average window length). To decide whether two windows W_i and W_j overlap, we intersect their areas in the common plane. We use an agglomerative clustering approach, *i.e.*, we initialize the clustering procedure by creating a separate cluster C for each window W_j^i in each image and then iteratively merge clusters. Two clusters C_s and C_t, $s \neq t$, are merged if there exist two overlapping windows $W_j^i \in C_s$ and $W_l^k \in C_t$, $i \neq k$. Once all overlapping clusters are merged, we compute a consensus window $W(C)$ for each cluster C: first, we determine the bounding box B containing all windows in the cluster, *i.e.*, $W_j^i \subseteq B$ for all $W_j^i \in C$. Next, for each image i observing a window $W_j^i \in C$, we project its per-pixel classifier scores onto B to accumulate the scores. We then compute $W(C)$ as the rectangle inside the bounding box that maximizes the sum of window scores minus wall scores in B. The computation of such a rectangle is known as the *maximum sum rectangular sub-matrix problem* and can be optimally computed using a 2D version of Kadane's algorithm [1]. The output is the set of consensus windows $\mathcal{W}_{3D} = \{W(C)\}$ for each sub-model.

5 Model Alignment

The goal of the alignment procedure is to transform the initially disjoint indoor and outdoor models into a common reference frame. Since traditional feature correspondences are not available, we instead employ window-to-window matches

to facilitate the alignment. We utilize the fact that a single window correspondence defines a similarity transformation that registers one indoor against one outdoor model. This allows us to exhaustively evaluate all potential matches rather than having to rely on appearance to establish correspondences. This is important since the appearance of a window can change quite drastically between indoors and outdoors[2] or might even be completely different, *e.g.*, due to closed shutters or partial occlusion. A natural way to define the best alignment is to find the transformation that explains the largest number of window correspondences. However, the transformation maximizing the number of inlier matches is not necessarily plausible. For example, it does not guarantee that an indoor model does not protrude from the outside of the building. In this section, we introduce and discuss a quality metric that takes both the number of inliers and the intersection between the models into account.

The input to our alignment procedure are sets $\mathcal{M}_{in}$ and $\mathcal{M}_{out}$ of axis-aligned indoor and outdoor models, respectively, as well as the consensus windows $\mathcal{W}_{\text{3D}}$ detected in the previous step. The output is a set of ranked configurations of aligned models $\mathcal{K}_s = \{(\mathcal{C}_i, e_i) \mid e_i < e_{i+1}\}$, where the energy e_i measures the cost of a configuration $\mathcal{C}_i$ and a lower energy denotes a better configuration. A configuration $\mathcal{C}_i$ relates two or more models through a set of window-to-window correspondences $\mathcal{C}_i = \{(W_a(m_j), W_b(m_k)), ... \mid j \neq k\}$. A single correspondence $(W_a(m_j), W_b(m_k))$ relates model m_j to m_k and defines a 3D similarity transformation $\mathbf{T}_{jk}$. The alignment procedure repeatedly searches for unique optimal configurations by minimizing the objective function

$$\begin{aligned} \underset{\mathcal{C}}{\text{minimize}} \quad & e = E_W(\mathcal{C}, \mathcal{W}_{\text{3D}}) + E_I(\mathcal{C}, \mathcal{M}_{in}, \mathcal{M}_{out}) \\ \text{subject to} \quad & E_I(\mathcal{C}, \mathcal{M}_{in}, \mathcal{M}_{out}) < \lambda \end{aligned} \quad . \tag{1}$$

The term E_W measures the cost of the window alignment between the models, *i.e.*, how well the estimated transformations align the windows. Likewise, the term E_I measures the cost of the model alignment in terms of the intersection of the models and λ defines the maximum intersection allowed. We solve this constrained optimization problem through exhaustive search in the space of possible configurations. For N windows in each of the M models, the number of possible configurations is $O(N^M)$. As the number of windows and models is typically relatively small, exhaustive search is feasible. A window-to-window correspondence $(W_a(m_j), W_b(m_k))$ relates the 3D consensus window W_a detected in an indoor model m_j to the 3D consensus window W_b in an outdoor model m_k. The correspondence also defines a relative 3D similarity $\mathbf{T}_{jk}$ transforming coordinates in model m_j into the coordinate frame of model m_k. Section 5.1 describes the process of establishing these correspondences and then chaining them to form a configuration $\mathcal{C}$. Section 5.2 defines the terms of E_W and E_I used to rank the set of configurations $\mathcal{K}$.

[2] We noticed that the indoor classifier sometimes splits a window into multiple parts while the outdoor classifier usually detects the whole window. This is due to the indoor images typically being taken closer to the windows, such that the frames appear larger, as well as the stronger contrast against the outdoor illumination.

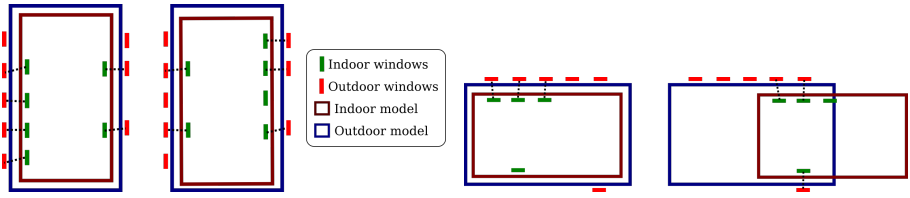

Fig. 4. (Left) Window term example. The alignment on the left has a lower cost than the one on the right. (Right) Intersection term example. Both alignments have the same E_W. The solution on the left is chosen since E_I is lower

5.1 Correspondence Search

For correspondence search, we exhaustively explore all possible configurations $\mathcal{C}_i$. We only consider window-to-window matches between indoor and outdoor models. We start by generating all unique pairwise window combinations between every unique pair of indoor and outdoor models. This initial set of combinations determines alignments between pairs of models. To handle the case of multiple indoor and outdoor models, we then generate all unique combinations of the initial set of combinations and repeat this process until all possible configurations are explored. The resulting set $\mathcal{K}$ contains the entire space of configurations aligning the models through chains of correspondences. Each correspondence in a configuration defines a relative 3D similarity that can be used to align the corresponding models into a common reference frame. For each correspondence $(W_a(m_j), W_b(m_k))$, we estimate its associated similarity transformation $\mathbf{T}_{jk}$ from the four corresponding 3D window corners in $W_a(m_j)$ and $W_b(m_k)$. To handle noisy window detections more robustly, we exploit the fact that the windows are already axis-aligned, *i.e.*, the rotations around the x- and y-axes are already fixed. Hence, we first estimate a 2D similarity transformation in the x-y plane and then independently infer the z-translation. This comes with two main benefits: first, the 2D window locations are usually less accurate than their estimated vertical plane. As a result, we obtain more robust orientation alignment around the z-axis. Second, a single window correspondence provides us with redundant observations for both the scale and z-translation estimation. To estimate the scale, we can use either the vertical or horizontal length of the window frames. For the z-translation, either the top or bottom side of the window frame. Generating these multiple possible alignments per window correspondence enables us to handle partial occlusions of windows more robustly, *e.g.*, caused by furniture or curtains. Chaining similarities using the recurrence relation $\mathbf{T}_{jkl} = \mathbf{T}_{kl} \cdot \mathbf{T}_{jk}$ enables us to transform any model's m_j coordinate system into any other model's m_k coordinate system, if they are within the same configuration $\mathcal{C}$. For each configuration, we align their contained models and windows into a single reference frame.

At this point, each $\mathbf{T}_{jk}$ is determined from a single window correspondence. However, a correct transformation chain is expected to put all corresponding windows in a configuration close to each other in 3D space. Hence, we look for

these additional window correspondences in a densification step through mutual nearest neighbor search in 3D space. For two windows $(W_a(m_j), W_b(m_k))$ to be mutual nearest neighbors, their centroids must be mutually closest in 3D space and the distance must be smaller than a fraction $\alpha = 0.25$ of their average window frame lengths. In addition, we enforce consistent orientation with a maximum angular distance of $\beta = 20°$ degrees. This densification procedure usually extends configurations by additional window correspondences. We then refine the initial alignments between models by estimating the similarity transformations from all window correspondences. The densification might lead to duplicate configurations in $\mathcal{K}$ containing the exact same correspondences. We prune these duplicates to reduce the computational cost in the following steps.

We apply the proposed correspondence search in an iterative manner, *i.e.*, we repeatedly densify the correspondences, prune duplicates, re-estimate similarities using the densified correspondences, and align the models using the refined similarities. This iterative refinement strategy terminates if the densification finds no additional correspondences.

5.2 Configuration evaluation

Given the set of unordered configurations, the next step is to determine whether they are plausible and to rank the plausible ones based on their quality. As defined in Eq. (1), we propose the energy $E_W + E_I$ to jointly model the quality of the window alignments E_W (window term) and the amount of model intersection E_I (intersection term). In the following, we define and discuss both terms.

Window term. Intuitively, a good alignment explains as many window alignments as possible, similar to inlier counting in RANSAC. Given a configuration $\mathcal{C}_i$ and the set of 3D consensus windows $\mathcal{W}_{3D}$, we define the window term as

$$E_W(\mathcal{C}_i, \mathcal{W}_{3D}) = |\mathcal{W}_{3D}| - 2 \cdot |\mathcal{C}_i| \ . \tag{2}$$

Thus, the window term counts the number windows that do not have a correspondence. A configuration with a higher number of explained window correspondences thus results in lower energy (*c.f.* Fig. 4(left)).

Intersection term. The window term reflects positive evidence for the quality of an alignment. However, it is not sufficient on its own, as illustrated in Fig. 4(right). The two configurations explain the same number of window matches, but the one to the right is clearly implausible as the indoor model intersects the outer hull of the building. We thus use a second term that determines the amount of intersection by measuring the amount of free-space violations between the aligned models. Intuitively, none of the 3D points in one reconstruction should be positioned in between a 3D point from another model and the cameras observing this second point. We thus create a 3D voxel grid for each model spanning the entire reconstruction including cameras and points, using a resolution of 200^3 voxels. A voxel is marked as free space if it is intersected by a viewing ray from one of the cameras to a sparse 3D point. The intersection ratio γ_{jk} between two aligned models m_j and m_k is then defined as the fraction of

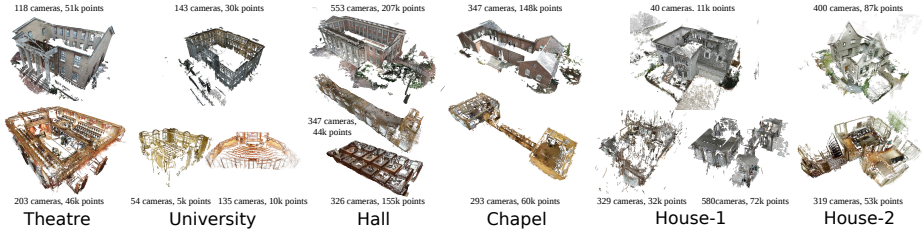

Fig. 5. Datasets used for experimental evaluation. We report the number of cameras and *sparse* points in each model. Dense point clouds are shown for visualization only

the sparse 3D points in model m_j that lie within a free-space voxel of m_k. The voxel grids can be efficiently pre-computed before the alignment procedure in the respective coordinate frames of the original models. Then, the intersection is computed by transforming the sparse points into the respective coordinate frame of the voxel grid of the other model. The energy term is defined as the maximum intersection of any combination of models in the configuration $\mathcal{C}_i$

$$E_I(\mathcal{C}_i) = \min\{1-\epsilon, \max\{\gamma_{jk} \; \forall \; m_j \in \mathcal{C}_i, m_k \in \mathcal{C}\}\}. \tag{3}$$

Here $\epsilon > 0$ is a small constant chosen to ensure that $E_I \in [0, 1)$. Ideally, no 3D point in a model should violate the free-space of another model. However, this is rarely the case in practice due to noise and outliers in the reconstruction. Thus, we allow a certain amount of intersection by setting $\lambda = 0.05$, *i.e.*, less than 5% of all points in a model are allowed to violate the free-space constraint. All configurations containing two models with an intersection ratio of λ or more are discarded (*c.f.* Eq. 1) during correspondence search.

Discussion. By definition $E_I(\mathcal{C}_i) \in [0, 1)$. Consequently, a configuration $\mathcal{C}_i$ with one more window correspondence than another configuration $\mathcal{C}_j$ will always have a lower energy. This implies that the intersection term only acts as negative evidence towards implausible configurations and it does not fully assess the quality of a configuration: scaling an indoor model such that it completely fits into the hull of a building results in no free-space violation. However, this configuration is only correct if the indoor model actually fills the whole space. If, on the other hand, the indoor model only contains part of the indoor scene, *e.g.* a single room, there is a high chance that this configuration will not have any window match, resulting in a high energy which denotes a bad configuration. alignment might be good.

6 Experimental Evaluation

In this section, we evaluate the accuracy and robustness of our proposed alignment approach. We provide both qualitative and quantitative results by showing different visualizations and comparing our estimated alignments with ground

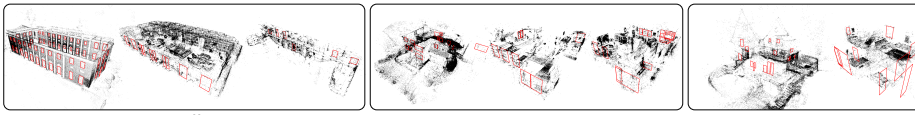

Fig. 6. Window detections obtained as described in Sec. 4

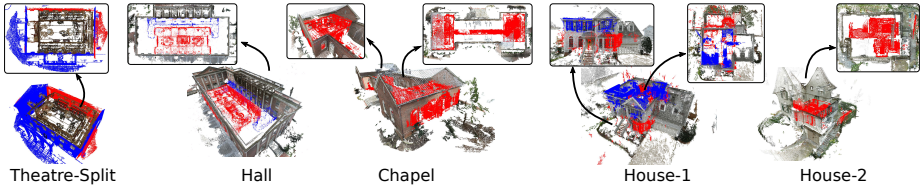

Fig. 7. Alignments between indoor and outdoor models computed by our approach

truth. In addition, we present and discuss failure cases of our approach. In the following, we first introduce the datasets.

Datasets. We collected a diverse set of six datasets (Fig. 5), spanning the possible input scenarios of our approach: (1) single indoor and single outdoor model (*Theatre*, *House-1*, *Chapel*), (2) multiple indoor models and single outdoor model (*University*, *House-2*), and (3) single indoor and multiple outdoor models (*Theatre-Split*). All buildings have multiple floors and we use a state-of-the-art SfM pipeline [25, 26] to reconstruct the models from photos taken with a single calibrated camera. Each of the datasets presents different challenges for our algorithm that we evaluate in the following sections.

Qualitative Evaluation. Figs. 1 and 7 show the best alignments produced by our proposed algorithm. We show dense models computed from the SfM output using PMVS [10] for better visualization, with the aligned models colored in red and blue. Despite the noisy window detections, we obtain the correct alignments for most datasets. These results demonstrate that our approach is able to estimate alignments that are accurate enough to pass visual inspection. There are, however, problems with the house datasets. In *House-2*, the indoor model is slightly too small. The noisy location of the detected windows also affects the alignment of the ground floor in *House-1*, where the bow-window area slightly intersects the outside of the house. This is related to the fact that the appearance of the datasets differs considerably from the training data for the outdoor model, and also the fact that a bow-window does not fall under the Manhattan-world assumption. Note that our approach correctly aligns the small, disconnected rooms in *University*, while a pure room layout-based alignment would fail.

Quantitative Evaluation. To quantitatively evaluate the alignment accuracy, we generate ground truth as follows: we manually label window corners in indoor and outdoor images for the *Theatre*, *University*, *House-2* and *Hall* datasets. The technique described in Sec. 4 is then used to obtain the 3D coordinates for

each window. Next, we manually select correspondences between indoor and outdoor windows to estimate a ground truth similarity transformation using least-squares. The absolute scale of the reconstruction is determined by measurement of the real-world window sizes. Using this ground truth alignment, we determine the quality of our alignments by calculating the positional error of the aligned sparse 3D points produced by our method. With a mean error of $\approx 0.05m$ (*Theatre*), $\approx 0.42m$ (*University*), $\approx 0.19m$ (*Hall*) and $\approx 0.54m$ (*House-2*, which results in an inaccurate alignment), the indoor and outdoor models are accurately aligned to a degree that is already difficult to notice by visual inspection. Note that the sizes of *University*, *Hall* and *House-2* are $36 \times 25m$, $40 \times 16m$ and $9 \times 9.5m$, resulting in an error of less than 1% of the dataset's size for the success cases, and only 5% for the failure case.

In addition, we also removed images from the outdoor reconstruction for *Theatre*. As a result, SfM splits this model into back and front façade models. We call this dataset *Theatre-Split*. Fig. 7 shows that our approach successfully connects the outdoor models through the indoors. We manually labelled corresponding cameras, and obtain an average camera pose error of $\approx 0.16m$ with a median of $\approx 0.05m$. Beyond the quantitative evaluation, the *Theatre-Split* dataset is a very interesting scenario, demonstrating additional applications of our method. For example, it is often impossible to create full models for individual houses in a connected building block or occlusions prevent feature matches around the corners of buildings [4]. Our approach enables the creation of full building models even in these cases.

Windows Evaluation. Fig. 6 shows the 3D window detections obtained with the approach described in Sec. 4 for a selection of datasets. Many window detections are noisy, especially indoors, where many windows are either missing (inside of *Hall*) or their shape, size, or location is inaccurate (*House-1* and *House-2*). In addition, there are a few false-detections due to noisy SfM points. Despite the large number of windows detected in some cases, our approach generates the ranked alignments for all datasets in under one minute. This can be attributed to our proposed combinatorial correspondence search scheme (Sec. 5.1). In our experiments, we were able to detect, on average, 73.9% of all indoor and 66% of all outdoor windows. Even for detection rates as low as 45%, our approach still works.

6.1 Discussion

Even though our approach is robust to noisy and missing window detections, it fails if there are no common windows between two models or if the detected number of windows is very small and their shape is too inaccurate. Possible reasons for missing or corrupt window detections include occlusion, incorrect labeling by our semantic classifier, a lack of 3D points preventing the estimation of the 3D window locations, *etc*. This results in different windows sizes for indoor and outdoor models, which in turn leads to wrong scale estimates. This is especially problematic if the number of common windows is small. *House-2* depicts one such case, in which we are not able to infer the correct scale of the interior model. If

there are enough common windows, our approach is rather robust against such cases, since it is likely that at least one of the window matches leads to a correct scale estimate. In addition, our proposed similarity estimation can handle partial occlusions. Further robustness could be gained by considering indoor-indoor and outdoor-outdoor alignments, *e.g.*, using techniques similar to [4, 21]. Another potential failure may arise in the presence of many noisy points in the reconstructions. A correct alignment could potentially violate the intersection constraint.

Similar to most vision-based reconstruction systems, our approach is vulnerable to multiple symmetry effects. First, along the vertical direction, where a room placement would be plausible on multiple floors. We obtain valid room placements on all three floors for the *University* and on two floors for the *Hall* dataset. With prior knowledge, a human could manually select the correct floor from the set of top-ranked configurations. Second, rotational symmetry, as depicted by the *Chapel* dataset. Even though the alignment shown in Fig. 7 looks visually plausible, it is actually off by a 180° rotation around the z-axis. Our approach finds the rotated alignment as the best solution due to window occlusions on one side of the outdoor model. Given an alignment computed with the proposed method, we could use an approach similar to [6] to detect symmetry planes for either the inside or outside model. The symmetry planes can then be used to hypothesize additional rotationally symmetric alignments, while the intersection constraint would rule out any invalid configurations. Last, if the task is to align a small room to a building with many rooms and windows, our approach will generate many plausible room placements. Choosing the best alignment is impossible without prior knowledge of the building layout.

7 Conclusion

We are among the first to tackle the problem of indoor-outdoor alignment. Our insight is to use semantic features (windows) to bridge the appearance gap in the alignment. This insight is potentially more broadly applicable, e.g., aerial-ground image alignment. We qualitatively and quantitatively showed the efficacy of our method on six challenging datasets. Our method handles disjoint reconstructions that might have been acquired at different times, thus giving more flexibility to the data acquisition stage for 3D reconstruction. Our results provide a valuable baseline for this difficult and important problem. In the future, we would like to explore other semantic cues such as doors, elevators, or staircases, in order to disambiguate across floors and symmetric configurations.

Acknowledgements This project was funded by the CTI Switzerland grant #17136.1 Geometric and Semantic Structuring of 3D point clouds, and the European Union's Horizon 2020 research and innovation programme under grant agreement #637221.

References

1. Bentley, J.: Programming pearls: algorithm design techniques. Comm. ACM (1984)
2. Cabral, R., Furukawa, Y.: Piecewise Planar and Compact Floorplan Reconstruction from Images. In: CVPR (2014)
3. Ceylan, D., Mitra, N.J., Zheng, Y., Pauly, M.: Coupled Structure-from-Motion and 3D Symmetry Detection for Urban Facades. ACM Trans. Graphics (2013)
4. Cohen, A., Sattler, T., Pollefeys, M.: Merging the Unmatchable: Stitching Visually Disconnected SfM Models. In: ICCV (2015)
5. Cohen, A., Schwing, A.G., Pollefeys, M.: Efficient Structured Parsing of Facades Using Dynamic Programming. In: CVPR (2014)
6. Cohen, A., Zach, C., Sinha, S., Pollefeys, M.: Discovering and Exploiting 3D Symmetries in Structure from Motion. In: CVPR (2012)
7. Cosmas, J., Itegaki, T., Green, D., Joseph, N., Gool, L.V., Zalesny, A., Vanrintel, D., Leberl, F., Grabner, M., Schindler, K., Karner, K., Gervautz, M., Hynst, S., Waelkens, M., Vergauwen, M., Pollefeys, M., Cornelis, K., Vereenooghe, T., Sablatnig, R., Kampel, M., Axell, P., Meyns, E.: Providing multimedia tools for recording, reconstruction, visualisation and database storage/access of archaeological excavations. In: VAST (2003)
8. Fischler, M., Bolles, R.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Comm. ACM 24(6), 381–395 (1981)
9. Furukawa, Y., Curless, B., Seitz, S.M., Szeliski, R.: Reconstructing Building Interiors from Images. In: ICCV (2009)
10. Furukawa, Y., Ponce, J.: Accurate, Dense, and Robust Multi-View Stereopsis. PAMI 32(8), 1362–1376 (2010)
11. Häne, C., Zach, C., Cohen, A., Angst, R., Pollefeys, M.: Joint 3D Scene Reconstruction and Class Segmentation. In: CVPR (2013)
12. Heinly, J., Schönberger, J.L., Dunn, E., Frahm, J.M.: Reconstructing the World* in Six Days *(As Captured by the Yahoo 100 Million Image Dataset). In: CVPR (2015)
13. Ikehata, S., Yan, H., Furukawa, Y.: Structured Indoor Modeling. In: ICCV (2015)
14. Koch, T., Korner, M., Fraundorfer, F.: Automatic Alignment of Indoor and Outdoor Building Models using 3D line segments. In: CVPR Workshops (2016)
15. Korč, F., Förstner, W.: eTRIMS Image Database for interpreting images of manmade scenes. Tech. Rep. TR-IGG-P-2009-01, Dept. of Photogrammetry, University of Bonn (April 2009), http://www.ipb.uni-bonn.de/projects/etrims_db/
16. Kushal, A., Self, B., Furukawa, Y., Gallup, D., Hernandez, C., Curless, B., Seitz, S.: Photo Tours. In: 3DIMPVT (2012)
17. Ladický, L., Russell, C., Kohli, P., Torr, P.: Associative Hierarchical Random Fields. PAMI 36(6), 1056–1077 (2014)
18. Li, Y., Snavely, N., Huttenlocher, D.P., Fua, P.: Worldwide Pose Estimation Using 3D Point Clouds. In: ECCV (2012)
19. Liu, C., Schwing, A.G., Kundu, K., Urtasun, R., Fidler, S.: Rent3D: Floor-Plan Priors for Monocular Layout Estimation. In: CVPR (2015)
20. Lynen, S., Sattler, T., Bosse, M., Hesch, J., Pollefeys, M., Siegwart, R.: Get Out of My Lab: Large-scale, Real-Time Visual-Inertial Localization. In: RSS (2015)
21. Martin-Brualla, R., He, Y., Russell, B.C., Seitz, S.M.: The 3D Jigsaw Puzzle: Mapping Large Indoor Spaces. In: ECCV (2014)

22. Middelberg, S., Sattler, T., Untzelmann, O., Kobbelt, L.: Scalable 6-DOF Localization on Mobile Devices. In: ECCV (2014)
23. Russell, B.C., Martin-Brualla, R., Butler, D.J., Seitz, S.M., Zettlemoyer, L.: 3D Wikipedia: Using online text to automatically label and navigate reconstructed geometry. In: SIGGRAPH Asia (2013)
24. Savinov, N., Ladicky, L., Häne, C., Pollefeys, M.: Discrete Optimization of Ray Potentials for Semantic 3D Reconstruction. In: CVPR (2015)
25. Schönberger, J.L., Radenovic, F., Chum, O., Frahm, J.M.: From Single Image Query to Detailed 3D Reconstruction. In: CVPR (2015)
26. Schönberger, J.L., Frahm, J.M.: Structure-from-Motion Revisited. In: CVPR (2016)
27. Snavely, N., Garg, R., Seitz, S.M., Szeliski, R.: Finding Paths through the World's Photos. In: SIGGRAPH (2008)
28. Strecha, C., Krull, M., Betschart, S.: The Chillon Project: Aerial / Terrestrial and Indoor Integration. Tech. rep., Pix4D (June 2014), https://pix4d.com/chillon/
29. Xiao, J., Furukawa, Y.: Reconstructing the World's Museums. In: ECCV (2012)
30. Zeisl, B., Sattler, T., Pollefeys, M.: Camera Pose Voting for Large-Scale Image-Based Localization. In: ICCV (2015)