SALVe: Semantic Alignment Verification for Floorplan Reconstruction from Sparse Panoramas

John Lambert[2⋆], Yuguang Li[1], Ivaylo Boyadzhiev[1], Lambert Wixson[1], Manjunath Narayana[1], Will Hutchcroft[1], James Hays[2], Frank Dellaert[2], and Sing Bing Kang[1]

[1] Zillow Group
[2] Georgia Institute of Technology

Abstract. We propose a new system for automatic 2D floorplan reconstruction that is enabled by *SALVe*, our novel pairwise learned alignment verifier. The inputs to our system are sparsely located 360° panoramas, whose semantic features (windows, doors, and openings) are inferred and used to hypothesize pairwise room adjacency or overlap. SALVe initializes a pose graph, which is subsequently optimized using GTSAM [16]. Once the room poses are computed, room layouts are inferred using HorizonNet [50], and the floorplan is constructed by stitching the most confident layout boundaries. We validate our system qualitatively and quantitatively as well as through ablation studies, showing that it outperforms state-of-the-art SfM systems in completeness by over 200%, without sacrificing accuracy. Our results point to the significance of our work: poses of 81% of panoramas are localized in the first 2 connected components (CCs), and 89% in the first 3 CCs.

Keywords: floorplan reconstruction; 3d reconstruction; structure from motion; extreme pose estimation

1 Introduction

Indoor geometry reconstruction enables a variety of applications that include virtual tours, architectural analysis, virtual staging, and autonomous navigation. There are solutions for image-based reconstruction based on inputs ranging from dense image capture to sparser capture using specialized imaging equipment (e.g., Matterport Pro2). For scalability of adoption, however, data bandwidth, equipment costs, and amount of labor must be considered.

We reconstruct floorplans from sparsely captured 360° panoramas, as provided by ZInD [14]. Currently, this problem is far from solved. Traditional Structure-from-Motion (SfM) [35,25] suffers from very limited reconstruction completeness [14,45]. Semantic SfM has been proposed [4,12,13], but accuracy is still limited, typically requiring a human in the loop [14].

Indoor floorplan reconstruction from unordered panoramas is a *discrete* instance of the wide-baseline SfM problem. Unlike traditional SfM, which is associated with a continuous estimation problem, for indoor residential floorplan reconstruction, discrete

⋆ Work completed during an internship at Zillow Group.

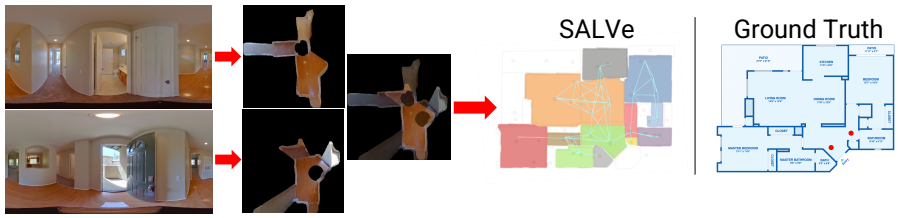

Fig. 1. A challenging wide-baseline scenario where traditional SfM systems that rely upon keypoint feature matches struggle, but where we succeed by exploiting semantic features such as doors, windows, and openings, or W/D/O). We infer layout and hypothesize plausible pairwise relative poses, which are then accepted or rejected, by feeding top-down aligned renderings into our learned *SALVe* verifier. Our global pose estimation has high completeness, leading to dramatic improvements in floorplan reconstruction (indicated by colored regions) vs. state-of-the-art systems such as OpenMVG [35] and OpenSfM [25]. For this hallway/entryway pano pair, SALVe easily validates a relative pose that was generated by grounding on a hallway opening feature.

room pieces must align at specific junction points (such as doors and walls), similar to solving a jigsaw puzzle [30]. We show how objects with repetitive structure, such as windows and doors, can be used to hypothesize room adjacency or overlap. Each hypothesis, i.e. a matched semantic element, provides a relative 2D room pose. The main innovation of our work is *SALVe*, a learned pairwise room alignment verifier. Given a room pair alignment hypothesis, *SALVe* uses the bird's eye view (BEV) of floors and ceilings to predict the likelihood score of adjacency or overlap. Our use of a discrete combinatorial proposal step, followed by a learned deep verifier, is akin to recent trends in language models, for tasks requiring multi-step reasoning [11,46], as "*Verifiers benefit both from their inherent optionality and from verification being a simpler task than generation in general.*" [11].

Once the relative poses are computed and verified, we perform global pose graph optimization using GTSAM [16]. Using the estimated poses and room layouts generated using HorizonNet [50], we construct the floorplan by stitching these layouts.

Our contributions are:

- To our knowledge, the first system for creating floorplans from unaligned panoramas with small to extremely wide baselines. These baselines can be so large that traditional SfM techniques fail.
- SALVe, a novel learning-based approach for validating discrete pairwise alignment proposals between panoramas in polynomial time.
- We show how our network verifies measurements with a high enough signal-to-noise ratio to directly apply global aggregation and optimization techniques.

2 Related Work

We briefly review approaches in floorplan reconstruction, SfM, and pose estimation under extreme baselines. While single-room layout estimation and depth estimation are

also relevant, we do not claim novelty in these areas. Good surveys of such methods can be found in [41] and [1].

Floorplan Reconstruction. Early systems require a human in the loop [15,22]. One notable manual approach is that of Farin *et al*. [22], which uses sparsely located 360° panoramas for joint floorplan and camera pose estimation.

For more automated solutions, SfM is used on densely captured perspective images [24] or 360° panoramas [5]. Both use SfM and MVS output to formulate graph optimization problems on a regular grid, through either graph cuts [24] or shortest-path problems [5], from which a rough 2D floorplan can be extracted. For sparser image inputs, semantic information such as floors, ceilings, and walls are used as additional cues [39]. Pintore *et al*. [40] cluster panoramas by room using photo-consistency at the central horizon line and plane sweeping with superpixel object masks to model clutter and floorplans in 3D. There are also methods on floorplan reconstruction from known camera poses [31,7,30,49,42] or RGBD data [31,7,30,49,42,36,28,21,20].

Structure from Motion (SfM). Much work has been done on SfM, and we refer readers to surveys such as [38]. Recently, deep learning with graph-based attention [44] or transformers [52] for deep, differentiable key point matching has been exploited to learn and match features from data. These "deep front-ends" offer a promise of less noisy input to back-end optimization [44]. Our system can be viewed as a deep verifier network (a deep front-end) that feeds measurements to global SfM [34,53]; however, instead of requiring complex outlier rejection schemes typical of global SfM [19,34,35,57,53,54], we show that outlier rejection can simply be based on predicted scores.

Semantic information has been used to overcome the limitation of keypoint matching for large baselines or scenes with little detail or repetitive textures [4,10]. Cohen *et al*. [12] first introduced a combinatorial approach for 3D model registration by aligning semantic objects such as windows [13]. More recent work [14,45] exploits this same idea to assemble floorplans from room layouts.

Extreme Pose Estimation. This refers to computing relative pose with little to no visual overlap. On localizing RGBD images, Yang *et al*. [55,56] demonstrate scan completion to a 360° image, followed by feature-based registration can be useful. Chen *et al*. [8] introduce DirectionNet to estimate a distribution of relative poses in 5 DOF space, i.e., when scale is unknown. SparsePlanes [27] uses planar surface estimation from perspective views within a single room for relative pose estimation. Other CNN-based approaches on perspective image re-localization include [29,3,18].

In concurrent work, Shabani *et al*. [45] use semantic information to generate global pose hypotheses by synthesizing Manhattan-only floorplans. The hypotheses are then scored by ConvMPN [58] and used to produce plausible room layout arrangements along with camera poses. They assume each panorama is captured in separate but connected rooms. Another key difference from our work is that their learning-based verifier is trained to evaluate the *final floorplan arrangements*, after using heuristics to enumerate many possible solutions. This is *exponential* in the number of input panoramas. Their approach is expected to produce several layout arrangements. In contrast, *SALVe* matches semantic elements between pairs of panoramas in polynomial time. Our model is then trained to verify the individual pairwise arrangements, allowing our approach

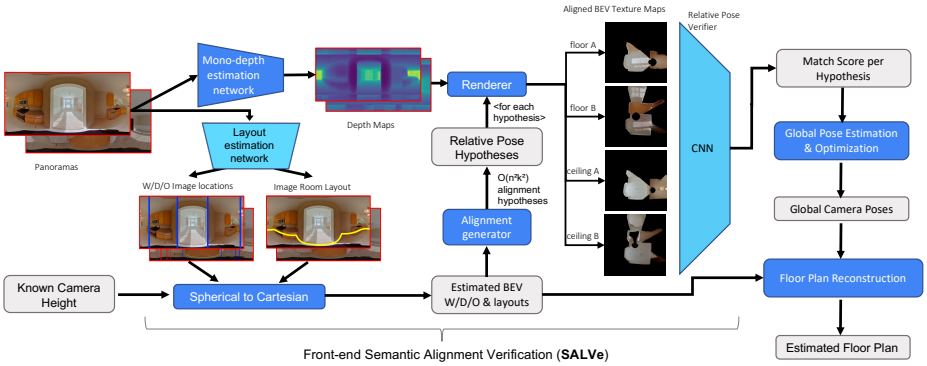

Fig. 2. Overview of our floorplan reconstruction system. "BEV" = "bird's eye view". Blue boxes are processing components, gray boxes are data. Trapezoids denote components based on deep networks; lighter blue networks are trained by us. 'Image Room Layout' represents the image coordinates of the floor-wall boundary (at each panorama column). n is the number of panoramas and k is the average number of detected windows/doors/openings per panorama. We show rendered floor and ceiling texture maps for a consistently-aligned pair of panoramas.

to be substituted as a front-end in any pose-graph optimization and producing a single reconstruction with higher reliability.

3 System Overview

We address the problem of global pose estimation of sparsely located panoramas, for the purpose of floorplan reconstruction. Formally, we define the global pose estimation problem as, given an unordered collection of n panoramas $\{\mathbf{I}_i\}$, determine poses $\{{}^w\mathbf{T}_i\}_{i=1}^n \in SE(2)$ of each panorama in global coordinate frame w. Similar to [42], we define the floorplan reconstruction problem as generating a *raster* (1) floor occupancy and (2) per-room masks.

Global pose estimation inherently relies on methods that build up global information from local signals. In our work, these local signals are estimated relative poses between pairs of panoramas. Our system for generating the floorplan from sparsely located panoramas is shown in Figure 2. The system consists of a front-end designed to hypothesize and compute relative pairwise poses, and a back-end designed to optimize global poses using these measurements.

The front-end (*SALVe*, or Semantic Alignment Verifier) first generates hypotheses of relative pose between the input pair of panoramas using their estimated room layout and detected semantic objects (specifically windows, doors, and openings, or W/D/O).[3] A hypothesis consists of pairing the same type of object across the two panoramas. Each pair of hypothesized corresponding W/D/O detections generates two relative pose hypotheses, by solving for the 2D translation that aligns their centers (on the ground plane), and the two possible rotation angles $\alpha, 180° + \alpha$ that align their extents. Each pairing allows us to compute the relative SE(2) pose.

[3] Openings are constructs that divide a large room into multiple parts [14].

A main novelty in this paper is how we test whether a hypothesis is plausible with *SALVe*. For a hypothesized relative pose, the system renders bird's-eye views of the floor and ceiling for both panoramas in the same BEV coordinate system, which produces overlapped top-down renderings. The rendering is computed with per-panorama depth distribution estimation using HoHoNet [51]. Then we use a deep CNN with a ResNet [26] backbone to generate a likelihood score that the overlapped images are a plausible match. Implausible matches are discarded, and from the remaining plausible matches we construct a pose graph. The back-end then globally optimizes the constructed pose graph using GTSAM [16]. Finally, floorplans are created by clustering the panoramas by room, extracting the most confident room layout given predicted panorama poses, and finally stitching these room layouts.

4 Approach

In this section, we detail the steps taken to generate a 2D floorplan from sparsely distributed 360° panoramas. The first step is to generate alignment hypotheses between pairs of panoramas.

4.1 Assumptions

We assume the inputs are a set of unordered 360° panoramas, captured from an indoor space. The images cover the entire space and the connecting doors between different rooms. Neighboring images may or may not have visual overlap. We assume the panoramas are in equirectangular form, i.e., their fields of view are 360° (horizontal) and 180° (vertical). The camera is assumed to be of known height and fixed orientation parallel to the floor[4], so pose is estimated in a 2D bird's-eye view (BEV) coordinate system.

4.2 Generating Alignment Hypotheses

Since our floorplan is 2D, alignment between pairs of panoramas has 3 DOFs (horizontal position and rotation). Scale is not a free parameter, assuming known, fixed camera height and a single floor plane (see [2] or our supplementary material for a derivation). To handle wide baselines, we use semantic objects (windows, doors, and openings, or W/D/O) to generate alignment hypotheses. While this is similar to the W/D/O-based room merge process in [14], we additionally make use of estimated room layout. Each room layout is estimated using a modified HorizonNet model [50]; it is trained with partial room shape geometry to predict both the floor-wall boundary with an uncertainty score and locations of W/D/O.

Each alignment hypothesis is generated with the assumption that W/D/O being aligned are in either the same room or different rooms. The outward surface normals of W/D/O are either in the same or opposite directions; we assume a window can only be

[4] We achieve this orientation assumption via pre-processing that straightens the panoramas using vanishing points [59].

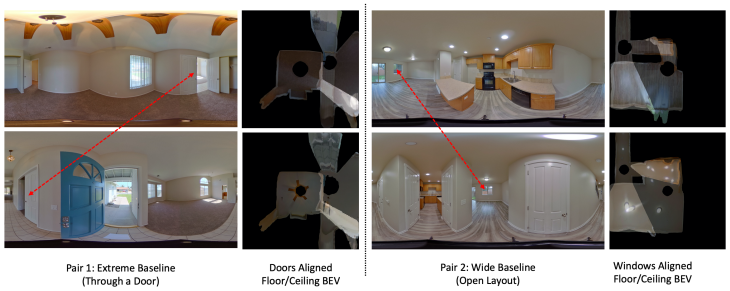

Fig. 3. Generating training samples. Orthographic BEVs of given panoramas, after semantic alignment proposal. Red arrows indicate the W/D/O, used to generate the pose proposals. ***Column 1:*** Example of extreme baseline pair. ***Column 2:*** overlaid floor **(top)** and ceiling **(bottom)**. ***Column 3:*** Example of a wide baseline pair. ***Column 4:*** overlaid floor **(top)** and ceiling **(bottom)**.

aligned in the direction of its interior normal, while a door or opening could be aligned in either direction. The hypothesis for rotation is refined using dominant axes of the two predicted room layouts.

Exhaustively listing pairs of W/D/O can produce many hypotheses for alignment verification. We halve the combinatorial complexity by ensuring that each pair of matched W/D/O have widths with a ratio within $[0.65, 1.0]$, i.e. a door that is 2 units wide cannot match to a door that is 1 units wide. Once the alignment hypotheses are found, they need to be verified.

4.3 SALVe: Semantic Alignment Verifier

While domain knowledge of indoor space such as room intersections and loop closure can be helpful in constructing the floorplan [14], visual cues can also be used to verify pairwise panorama overlap [17]. We use bird's eye views (BEVs, which are orthographic) of the floor and ceiling as visual cues for alignment verification. Given the significant variation in lighting and image quality across panoramas, traditional photometric matching techniques may not be very effective. Instead, we train a model to implicitly verify spatial overlap based on these aligned texture signals.

We extract depth using HoHoNet[51], which is used to render the BEVs. Example views can be found in Figure 3. Given an alignment hypothesis, we map the BEVs of the floor and ceiling for both panoramas to a common image coordinate system. The four stacked views are then fed into our deep-learning based pairwise alignment verification model to classify 2-view alignment. Given n panoramas, each with k W/D/O, $\mathcal{O}(n^2k^2)$ alignments are possible and thus need to be verified.

SALVe uses a ResNet [26] ConvNet architecture as the backbone for verification. Its input is a stack of 4 aligned views (2 from each panorama), with a total of 12 channels. It is trained with softmax-cross entropy over 2 classes, representing the "mismatch" and "match" classes. We generate these classes by measuring the deviation of generated relative poses (alignments from window-window, opening-opening, or door-door pairs) against the ground truth poses. Those below a certain amount of deviation are considered "matches", and all others are considered "mismatches".

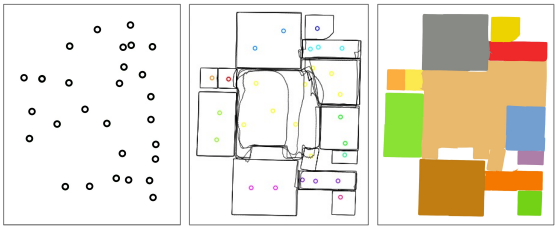

Fig. 4. An example of different stages of floorplan reconstruction: ***Left***: Estimated positions of panorama centers. ***Center***: Grouped panoramas with estimated dense room layouts. Panorama centers with the same color are part of the same group. Notice that each open space is grouped together. Distinct groups correspond largely to physical rooms separated by doors. ***Right***: The final floorplan after highest-confidence contour extraction is applied to each group. Each contour is filled with a unique color.

4.4 Global Pose Estimation and Optimization

SALVe is used to generate a set of pairwise alignments, which are used to construct a pose graph; its nodes are panoramas and edges are estimated relative poses. The pose graph has an edge between any two panoramas $\mathbf{I}_{i_1}$ and $\mathbf{I}_{i_2}$ where pairing a detection $\mathbf{d}_{k_1}^{i_1}$ with detection $\mathbf{d}_{k_2}^{i_2}$ yields a plausible (according to SALVe) alignment. A detection may participate in multiple edges e.g., pairing $(\mathbf{d}_{k_1}^{i_1}, \mathbf{d}_{k_2}^{i_2})$ may add an edge between i_1 and i_2, and pairing $(\mathbf{d}_{k_1}^{i_1}, \mathbf{d}_{k_3}^{i_3})$ may add an edge between panos i_1 and i_3. Although conflicting relative pose hypotheses are possible, in practice SALVe is a sufficiently accurate verifier that they are quite rare.

When multiple disjoint graphs result, we only consider the largest connected component. We experiment with two algorithms for global localization: spanning tree pose aggregation and pose graph optimization (PGO) with a robust noise model, detailed in the supplementary material.

4.5 Floorplan Reconstruction

Figure 4 shows the progression of floorplan reconstruction, from estimated panorama poses and room layouts to the output. There are three steps: panorama room grouping, highest confidence room contour extraction, and floorplan stitching. To refine a room layout, we first identify all the panoramas within that room; this is done using 2D IoU. Since each panorama has its own layout with local shape confidence (Section 4.2) within a room, we extract a single global layout by searching for the most confident contour points. The search is done by raycasting from panorama centers and voting for the most confident contour point along each ray. The final floorplan is found by taking the union of (stitching) all room layouts. Details are in the supplementary material.

5 Experimental Results

In this section, we explain why we use ZInD [14], provide implementation details, and describe our metrics before showing results for different global pose estimation

techniques. We also describe ablation studies that show how different types of inputs affect the results.

5.1 Use of ZInD [14]

In order to evaluate every part of our approach, as well as the entire system, we use the recently released Zillow Indoor Dataset (ZInD) [14]. ZInD has all the required components: (1) *large scale* with $67,448$ panoramas taken in $1,575$ real homes; (2) *multiple localized panoramas per-room* with 42 panoramas over 15 rooms per-home on average; (3) *layout and W/D/O* annotations including complex, non-Manhattan layouts and (4) *2D floor-plans* with 1.8 number of floors per-home on average. We use the official train, val, and test splits that contain 1260, 157, and 158 homes, and 2168, 278, 291 floors respectively. We acknowledge that in ZInD most rooms are unfurnished, but this is a frequent scenario in the domain of real estate floor plan reconstruction. While there are other real [6,61,45] and synthetic [60,48] indoor datasets, none of them have all the required components. Structured3D [60] is a synthetic dataset with only one panorama per room and doors in almost all rooms are closed (uncommon in real estate capture scenarios); these factors result in a significant change of modality.

5.2 Implementation Details

Layout and W/D/O estimation. We use a modified version of HorizonNet [50], trained to jointly predict room layout as well as 1D extents of W/D/O. We trained the joint model on ZInD and make the predictions publicly available.
Verifier supervision. We consider a pair-wise alignment to be a "match" if ground truth relative pose $(x, y, \theta) \in SE(2)$ and generated relative pose $(\hat{x}, \hat{y}, \hat{\theta}) \in SE(2)$ differ by less than 7° (θ) for doors and windows, and less than 9° for openings. A larger threshold is used for openings because there is more variation in their endpoints. We also require that $\left\|[x, y]^\top - [\hat{x}, \hat{y}]^\top\right\|_\infty < 0.35$ in normalized room coordinates (i.e., when camera height is scaled to 1).
Texture mapping, verifier data augmentation and verifier training. Details are provided in the supplementary material.

5.3 Evaluation Metrics

In order to evaluate our entire system, we measure increasing subsets of components.
Layout estimation and W/D/O detection accuracy. To evaluate the quality of the layout estimation, we report 2D IoU between the predicted and ground truth room layouts per panorama. Because we project 1D W/D/O on the predicted layout, we use 1D IoU to measure the accuracy of those semantic elements, with F1 score evaluated at a true positive 1D IoU threshold of 70%.
Relative pose classification accuracy. We report intermediate system metrics that measure the model's accuracy at discerning between correct and inaccurate alignments. We use mean accuracy over two classes, as well as precision, recall, and F1 score.
Global pose estimation accuracy and completeness. We first align an estimated pose

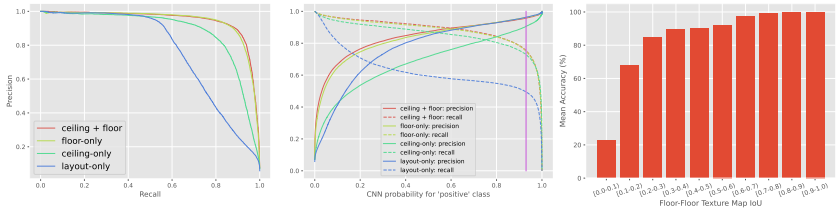

Fig. 5. Precision-recall analysis of SALVe. ***Left***: curve for SALVe under different inputs ('layout-only' refers to a model with access only to estimated room geometry, but no floor or ceiling texture). ***Center***: Comparison of confidence thresholds versus their effect on precision and recall. The purple line indicates our operating point (93% confidence). ***Right***: Classification accuracy vs. visual overlap for the GT **positive** class only from SE(2) alignments generated from predicted W/D/O's. Small visual overlap often corresponds to "extreme" baselines.

graph $\{\hat{\mathbf{T}}_i\}_{i=1}^{M}$ to a ground truth pose graph $\{\mathbf{T}_i\}_{i=1}^{N}$ where $\mathbf{T}_i \in SE(2)\ \ \forall i \in 1, \ldots, N$, by estimating a $\mathrm{Sim}(2)$ transformation between them, where $M \leq N$, since not all poses may be estimated. To reduce the influence of outliers for mostly-correct global pose estimates, we perform pose graph alignment in a RANSAC loop, with a randomly selected subset ($2/3$ of the M estimated poses) used to fit each alignment hypothesis, over 1000 hypotheses. We then measure the distance between the predicted and true i'th camera location $\|t_i - \hat{t}_i\|_2$, and difference between true and predicted i'th camera orientation $|\theta_i - \hat{\theta}_i|$. Completeness is essential to floorplan reconstruction, so we also report the percent of panoramas localized in the largest connected component.

Floorplan reconstruction accuracy and completeness. We measure the 2D IoU between a rasterized binary occupancy map of the ground truth and the predicted floorplans. This metric measures the quality of our end-to-end system, as it encapsulates the *accuracy* of our *pair-wise relative pose proposal* in combination with the *accuracy* and *completeness* of the global pose estimation and the fusion of the room layouts (see supplementary for more details).

5.4 Layout and W/D/O Estimation Accuracy

The layout estimation module used in the system yields an average of 85% IoU with ground truth shape. W/D/O detection is accurate; at a 70% 1D IoU threshold, we correctly identify W/D/O with F1 scores of 0.91, 0.89, and 0.67, respectively. Our model is the least accurate in predicting openings. As discussed in [14], there are issues with annotator error and possibly ambiguous tagging of rooms in open spaces that cover different room types, making locations of openings less clear. We speculate that these contribute to the errors, especially for openings. In the Supplement, we provide qualitative examples of the various types of failure modes of the model.

5.5 Relative Pose Classification

We first measure the performance of the SALVe "front-end". These trained models achieve 92-95% accuracy on the test split (see Supplement). We show that a larger

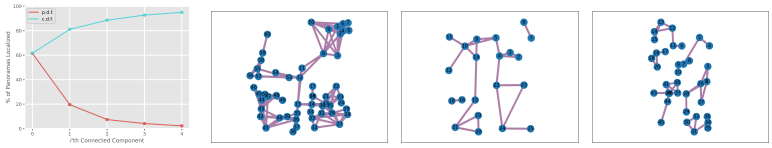

Fig. 6. ***Left:*** Distribution of localization percentage in the first 5 connected components, averaged over all test tours. ***Right:*** Topology of global pose graphs for various different homes.

capacity model than ResNet-50 (i.e. ResNet-152) further improves performance. We also note that the accuracy is limited by noisily-generated 'ground truth'. We train on 587 number of tours from ZInD, and use the official train/val/test splits.

In Figure 5, we show a PR curve, indicating the precision of the model at different recall thresholds. We choose a 93% confidence threshold as our operating point, as it maximizes precision just before a precipitous drop in recall.

How does the amount of visual overlap affect relative pose classification accuracy? More overlap yields higher accuracy for the ground truth positive class, but lower accuracy for the ground truth negative class. In Figure 5, we analyze the performance of our relative pose classification method under varying amounts of visual overlap. 100% overlap would indicate that two panoramas were captured in exactly the same position, with the scene unchanged between the two captures. On the other hand, 0% overlap would indicate that the panoramas were captured in completely different locations, i.e. in two rooms, on opposite sides of a closed door (an example of an "extreme" baseline). We use a proxy metric, IoU of the texture map generated using HoHoNet-estimated [51] monocular depth, which introduces some amount of noise.

5.6 Global Pose Estimation Results

Next, we measure performance of both the "front-end" along with some form of global aggregation ("back-end"). We compare with two baselines from state-of-the-art structure from motion systems that support optimization from 360° images.

OpenMVG [34,35]. We use the recommended setting for 360° image input, with incremental SfM using an upright SIFT feature orientation, an upright 3-point Essential matrix solver with A-Contrario RANSAC, following the planar motion model described by [2,37,9], with an angular constraint for matching.

OpenSfM [25]. Incremental SfM system that uses the Hessian-Affine interest point detector [33], SIFT feature descriptor [32], and RANSAC [23].

In Table 1, we show the results of global pose estimation on the ZInD test set. We outperform OpenMVG by 656% and OpenSfM by 257% in the median percentage of panoramas localized (their 8.7% and 22.2% vs. our 57.1%), with even lower median rotation error (our 0.17° vs. their 0.37° and 0.36°). Our median translation error is comparable (our 25 cm vs. their 12 cm and 10 cm). PGO is significantly more accurate than spanning tree when VP estimation is not employed (see Table 3). However,

Table 1. Results of global pose estimation on the ZinD test set. Two global aggregation methods are evaluated: spanning tree ('ST'), and pose graph optimization ('PGO'), with axis-alignment ('AA'). ST and PGO both use the same largest connected component of $\mathcal{G}$ as input, and thus localize an equal number of panoramas.

Method	Localization %		Tour Avg. Rotation Error (deg.)		Tour Avg. Translation Error (meters)	
	mean	median	mean	median	mean	median
OpenSfM [25]	27.62	22.22	9.52	0.36	1.88	0.12
OpenMVG [34,35]	13.94	8.70	3.84	0.37	**0.41**	**0.10**
Ours (w/ ST + AA)	**60.70**	**57.10**	**3.69**	**0.03**	0.81	0.26
Ours (w/ PGO + AA)	**60.70**	**57.10**	3.73	0.17	0.80	0.25

when using vanishing point-based dominant axis-alignment, both spanning trees and pose graph optimization on SALVe-verified measurements produce similar global aggregation results. In the left column of Figure 7, we show the topological structure of the largest component of the pose graph for a few homes.

6 Discussion

Is deep learning necessary for verification, or can heuristics be used? To verify pairwise alignment, matching texture is necessary but hard to feature engineer. Using geometry alone is insufficient (See Figure 5(a-b) and Table 2), motivating others to explore graph neural networks for the task [45]. We implemented rule-based baselines that classify BEV image pairs via FFT cross-correlation scores [43], and found they do not work well due in part to difficulty in choosing thresholds. Previous works such as LayoutLoc [14] have explored rule-based checking, but found that it only can be successful when given access to *oracle* within-room pano grouping information; estimation of such within-room grouping (i.e. adjacency) is itself one of the fundamental challenges of global pose estimation in an indoor environment.

What type of semantic object is most useful for alignment in this semantic SfM problem? Doors, but all are essential. Openings are the second-most effective object type to achieve complete localization, and windows are least effective. Among the alignments that the model predicts to be positives with confidence $\geq 97\%$, we find that 63% originate from door-door hypotheses, 24% originate from opening-opening hypotheses, and 20% originate from window-window hypotheses. While rooms in residential homes are rarely connected by a window, these window alignments can provide additional redundancy, or ground alignments in very large open spaces when doors are not visible as in Fig. 3, pair 2. In Table 2, we report global pose estimation results when only one type of semantic object is used to create the edges $\mathcal{E}$ of the relative pose graph $\mathcal{G}$.

To what extent is the pose graph shattered into multiple clusters? Typically, the first three connected components contain 61%, 20%, and 7% of all panoramas (See Figure 6a). We measure the distribution of connected components (CCs), as global pose estimation relies upon a single CC (we use the largest), and we find that often the second and third largest CCs are also large, indicating the potential for merging, e.g. combining ideas from [45] or [47]. We compute an average probability density function and cumulative density function by averaging per-floor distributions across the test set.

Table 2. Results of ablation experiments on how inputs to SALVe affect global pose estimation accuracy and completeness. Pose graph optimization and vanishing point-based axis alignment ('PGO + AA') are utilized for all entries below.

W/D/O Inputs			Raster Inputs			Localization %		Tour Avg. Rotation Error (deg.)		Tour Avg. Translation Error (meters)	
Doors	Windows	Openings	Floor Texture	Ceiling Texture	Layout	MEAN	MEDIAN	MEAN	MEDIAN	MEAN	MEDIAN
✓	✓	✓	✓	✓		60.70	57.14	3.73	0.17	0.80	0.25
✓			✓	✓		43.30	40.00	2.41	0.07	0.59	0.20
	✓		✓	✓		15.57	13.33	2.20	**0.00**	0.74	**0.11**
		✓	✓	✓		23.87	23.08	**0.66**	0.05	**0.34**	0.18
✓	✓	✓	✓			60.64	58.33	3.75	0.15	0.91	0.25
✓	✓	✓		✓		**60.93**	**64.58**	10.94	0.28	2.12	0.35
✓	✓	✓			✓	19.19	16.67	3.43	0.03	0.53	**0.11**

Is the RGB photometric signal from panoramas actually necessary, as opposed to solely using geometric context? Yes, the RGB texture is essential. In Table 2, we show that using a layout-only rasterization as input to the CNN, instead of a photometric texture map, leads to severe performance degradation.

Does floor or ceiling texture provide a more useful signal for alignment classification? Floor texture. However, using both signals jointly improves performance. In Table 2, we show the results of using as input to the network only the floor texture maps, or only the ceiling texture maps, as opposed to reasoning about both jointly.

Is a Manhattan world assumption helpful? For pose estimation, yes, but for shape estimation, no. Many rooms at critical junctures in the floorplan are non-Manhattan in shape, and 'Manhattanizing' them would be destructive when chaining together. However, room organization in a home is usually tied to three dominant, orthogonal directions. In Table 3, we show that using vanishing point estimation to align relative poses up to a 15° correction significantly improves both global pose estimation accuracy and slightly improves floorplan reconstruction accuracy. Both vanishing point relative rotation angle correction and pose graph optimization are effective means of decreasing the rotation error. In the supplementary we show how using ground truth layout (near-perfect shape) and W/D/O locations affects performance, as an upper-bound on performance of the first module in our system.

6.1 Floorplan Reconstruction Results

Next, we compare performance of the entire floorplan reconstruction system. In Table 4, we demonstrate that compared to traditional SfM with oracle room layout and oracle

Table 3. Comparison of results with and without axis-alignment ('AA') of relative poses (via vanishing angles) before global aggregation. The amount of panoramas localized is unaffected, as adjacency is maintained during the correction. For this comparison, 'oracle' layouts are used to isolate the effect of pose error. With vanishing point (VP) information, the difference between PGO and Spanning Tree is not statistically significant (1 cm and 0.04° error on average).

Method	Tour Avg. Rotation Error (deg.)		Tour Avg. Translation Error (meters)		Floorplan IoU	
	MEAN	MEDIAN	MEAN	MEDIAN	MEAN	MEDIAN
Spanning Tree	5.41	1.92	0.86	0.33	0.55	0.52
Spanning Tree + AA	**3.69**	**0.03**	0.81	0.26	**0.56**	0.52
PGO	4.93	1.53	0.81	0.29	**0.56**	0.52
PGO + AA	3.73	0.17	**0.80**	**0.25**	**0.56**	**0.53**

Fig. 7. Qualitative comparison of floorplan results. ***Column 1:*** OpenSfM. ***Column 2:*** OpenMVG. ***Column 3:*** Ours. ***Column 4:*** Ground truth floorplan. All results are superimposed on the ground truth floorplan. Colored regions indicated the reconstruction result; at times, the baselines localize no panos. Our floorplan recall is significantly better than the state-of-the-art. Each row corresponds to a single floor of a different home. Colored lines represent W/D/O objects – doors, openings and windows. The multiple cyan edges in the overlaid graph correspond to verified W/D/O alignment hypotheses. For an open layout, a successful case often involves edges from panoramas in many different rooms to a single pano. These examples are intended to offer an even-handed selection of reconstructions that indicate both good performance as well as areas for improvements. Rows 1 and 6 illustrate good reconstructions. Row 2 illustrates a more challenging case with only 1-2 panos in most rooms. Rows 3-5 are more challenging as they include bottlenecks in the actual physical layout, which is critical in joining connected components.

Table 4. Floorplan reconstruction results against the ground truth manually annotated floorplan. Floorplan 2D IoU is measured in the bird's eye view. The IoU is measured on the largest connected component. 'AA' represents axis-alignment.

METHOD	GLOBAL POSES		LAYOUT		FLOORPLAN IOU	
	ORACLE	ESTIMATED	ORACLE	ESTIMATED	MEAN	MEDIAN
OPENSFM		✓	✓		0.29	0.26
OPENMVG		✓	✓		0.16	0.07
OURS	✓			✓	**0.94**	**0.95**
OURS (PGO + AA)		✓	✓		0.56	0.53
OURS (PGO + AA)		✓		✓	0.49	0.45

scale, our end-to-end system is able to produce more accurate floorplans with estimated room layouts (our 0.49 mean IoU vs. OpenSfM's 0.29 and OpenMVG's 0.16). The 0.56 mean IoU score using our estimated global poses and oracle layout primarily reflects the completeness of our final floorplan. With oracle pose and estimated room layouts, the 0.94 mean IoU reflects the accuracy of our layout estimation and stitching stages. This baseline has significantly larger IoU in part because the 'oracle' poses are provided for *all* panoramas (see the Supplement for comparison visualizations).

Qualitative Results. Fig. 7 provides qualitative results for a number of different homes. For floors of some homes, our method produces nearly complete reconstructions, while for others, the results are more sparse. As shown by the third column of Fig. 7, the topology of the pose graph directly affects the completeness of the reconstruction; when multiple large connected components appear, the reconstruction is shattered apart. For several homes, OpenMVG and OpenSfM fail to converge, localizing no panoramas.

7 Conclusion

We present a new system for automatic 2D floorplan reconstruction from sparse, unordered panoramas. This work represents a breakthrough in the completeness of reconstructed floorplans, with over two times more coverage than previous systems [25,35], without sacrificing accuracy. We demonstrate how *SALVe*, our novel pairwise learned alignment verifier, capitalizes on the mature field of semantic detection of features (W/D/O) to handle a tractable number of alignment hypotheses and generate high-quality results. A human annotator may use it to accelerate labeling by automatically generating the majority of necessary decisions before making the final choices about glueing connected components. Fig. 7 only illustrates the largest CC; other CCs are also generated, but not shown (Fig. 6, a CDF of 89% for the first 3 CCs).

Limitations. Because the number of pairwise alignments is combinatorial in the number of W/D/O, the runtime of the current system is limited, although we have not heavily optimized it. As ZInD [14] contains only unfurnished homes, our system has not yet been evaluated in a furnished home regime, due to dataset availability. Camera localization completeness is still in the 55-60% range. With future improvements to each part of the system, especially omnidirectional depth estimation and layout estimation, we expect floorplan reconstruction performance to continue to improve.

References

1. Albanis, G., Zioulis, N., Drakoulis, P., Gkitsas, V., Sterzentsenko, V., Álvarez, F., Zarpalas, D., Daras, P.: Pano3D: A holistic benchmark and a solid baseline for 360° depth estimation. CVPR Workshops (2021)
2. Aly, M., Bouguet, J.Y.: Street view goes indoors: Automatic pose estimation from uncalibrated unordered spherical panoramas. In: 2012 IEEE Workshop on the Applications of Computer Vision (WACV). pp. 1–8 (2012)
3. Balntas, V., Li, S., Prisacariu, V.: Relocnet: Continuous metric learning relocalisation using neural nets. In: ECCV (2018)
4. Bao, S.Y., Savarese, S.: Semantic structure from motion. In: CVPR (2011)
5. Cabral, R., Furukawa, Y.: Piecewise planar and compact floorplan reconstruction from images. In: CVPR (2014)
6. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from RGB-D data in indoor environments. International Conference on 3D Vision (3DV) (2017)
7. Chen, J., Liu, C., Wu, J., Furukawa, Y.: Floor-SP: Inverse CAD for floorplans by sequential room-wise shortest path. In: ICCV (2019)
8. Chen, K., Snavely, N., Makadia, A.: Wide-baseline relative camera pose estimation with directional learning. In: CVPR (2021)
9. Choi, S., Kim, J.H.: Fast and reliable minimal relative pose estimation under planar motion. Image and Vision Computing **69**, 103–112 (2018)
10. Choudhary, S., Trevor, A.J., Christensen, H.I., Dellaert, F.: SLAM with object discovery, modeling and mapping. In: 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 1018–1025. IEEE (2014)
11. Cobbe, K., Kosaraju, V., Bavarian, M., Hilton, J., Nakano, R., Hesse, C., Schulman, J.: Training verifiers to solve math word problems. ArXiv:2110.14168 (2021)
12. Cohen, A., Sattler, T., Pollefeys, M.: Merging the unmatchable: Stitching visually disconnected SfM models. In: ICCV (2015)
13. Cohen, A., Schönberger, J.L., Speciale, P., Sattler, T., Frahm, J., Pollefeys, M.: Indoor-outdoor 3D reconstruction alignment. In: ECCV (2016)
14. Cruz, S., Hutchcroft, W., Li, Y., Khosravan, N., Boyadzhiev, I., Kang, S.B.: Zillow indoor dataset: Annotated floor plans with 360deg panoramas and 3D room layouts. In: CVPR (2021)
15. Debevec, P.E., Taylor, C.J., Malik, J.: Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. In: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques. SIGGRAPH '96 (1996)
16. Dellaert, F.: Factor graphs and GTSAM: A hands-on introduction. Tech. rep., Georgia Institute of Technology (2012)
17. Dellaert, F., Burgard, W., Fox, D., Thrun, S.: Using the condensation algorithm for robust, vision-based mobile robot localization. In: CVPR (1999)
18. Ding, M., Wang, Z., Sun, J., Shi, J., Luo, P.: CamNet: Coarse-to-fine retrieval for camera re-localization. In: ICCV (2019)
19. Enqvist, O., Kahl, F., Olsson, C.: Non-sequential structure from motion. In: ICCV Workshops (2011)
20. Fang, H., Lafarge, F., Pan, C., Huang, H.: Floorplan generation from 3D point clouds: a space partitioning approach. ISPRS Journal of Photogrammetry and Remote Sensing **175** (2021)
21. Fang, H., Pan, C., Huang, H.: Structure-aware indoor scene reconstruction via two levels of abstraction. ISPRS Journal of Photogrammetry and Remote Sensing **178**, 155–170 (2021)

22. Farin, D., Effelsberg, W., de With, P.H.: Floor-plan reconstruction from panoramic images. In: Proceedings of the 15th ACM international conference on Multimedia (2007)
23. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6) (1981)
24. Furukawa, Y., Curless, B., Seitz, S.M., Szeliski, R.: Reconstructing building interiors from images. In: ICCV (2009)
25. Gargallo, P., Kuang, Y., et al.: OpenSfM (2016)
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
27. Jin, L., Qian, S., Owens, A., Fouhey, D.F.: Planar surface reconstruction from sparse views. In: ICCV (2021)
28. Kim, Y.M., Dolson, J., Sokolsky, M., Koltun, V., Thrun, S.: Interactive acquisition of residential floor plans. In: ICRA (2012)
29. Laskar, Z., Melekhov, I., Kalia, S., Kannala, J.: Camera relocalization by computing pairwise relative poses using convolutional neural network. In: ICCV Workshops (2017)
30. Lin, C., Li, C., Wang, W.: Floorplan-jigsaw: Jointly estimating scene layout and aligning partial scans. In: ICCV (2019)
31. Liu, C., Wu, J., Furukawa, Y.: FloorNet: A unified framework for floorplan reconstruction from 3D scans. In: ECCV (2018)
32. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60** (2004)
33. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. Int. J. Comput. Vision **60**(1) (Oct 2004)
34. Moulon, P., Monasse, P., Marlet, R.: Global fusion of relative motions for robust, accurate and scalable structure from motion. In: ICCV (2013)
35. Moulon, P., Monasse, P., Perrot, R., Marlet, R.: OpenMVG: Open multiple view geometry. In: International Workshop on Reproducible Research in Pattern Recognition. Springer (2016)
36. Okorn, B., Xiong, X., Akinci, B., Huber, D.: Toward automated modeling of floor plans. In: 3D DPVT (2010)
37. Oskarsson, M.: Two-view orthographic epipolar geometry: Minimal and optimal solvers. Journal of Mathematical Imaging and Vision **60**(2) (2018)
38. Ozyesil, O., Voroninski, V., Basri, R., Singer, A.: A survey of structure from motion. Acta Numerica **26** (May 2017)
39. Pintore, G., Ganovelli, F., Pintus, R., Scopigno, R., Gobbetti, E.: 3D floor plan recovery from overlapping spherical images. Computational visual media **4**(4) (2018)
40. Pintore, G., Ganovelli, F., Villanueva, A.J., Gobbetti, E.: Automatic modeling of cluttered multi-room floor plans from panoramic images. Comput. Graph. Forum **38**(7) (2019)
41. Pintore, G., Mura, C., Ganovelli, F., Fuentes-Perez, L., Pajarola, R., Gobbetti, E.: State-of-the-art in automatic 3D reconstruction of structured indoor environments. Computer Graphics Forum **39**(2) (2020)
42. Purushwalkam, S., Garí, S.V.A., Ithapu, V.K., Schissler, C., Robinson, P., Gupta, A., Grauman, K.: Audio-visual floorplan reconstruction. In: ICCV (2021)
43. Reddy, B., Chatterji, B.: An fft-based technique for translation, rotation, and scale-invariant image registration. IEEE Transactions on Image Processing **5**(8), 1266–1271 (1996)
44. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperGlue: Learning feature matching with graph neural networks. In: CVPR (2020)
45. Shabani, M.A., Song, W., Odamaki, M., Fujiki, H., Furukawa, Y.: Extreme structure from motion for indoor panoramas without visual overlaps. In: ICCV (2021)

46. Shen, J., Yin, Y., Li, L., Shang, L., Zhang, M., Liu, Q.: Generate & Rank: A multi-task framework for math word problems. In: Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics (2021)
47. Son, K., Moreno, D., Hays, J., Cooper, D.B.: Solving small-piece jigsaw puzzles by growing consensus. In: CVPR (2016)
48. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic scene completion from a single depth image. In: CVPR (2017)
49. Stekovic, S., Rad, M., Fraundorfer, F., Lepetit, V.: Montefloor: Extending MCTS for reconstructing accurate large-scale floor plans. In: ICCV (2021)
50. Sun, C., Hsiao, C.W., Sun, M., Chen, H.T.: Horizonnet: Learning room layout with 1D representation and pano stretch data augmentation. In: CVPR (2019)
51. Sun, C., Sun, M., Chen, H.T.: Hohonet: 360 indoor holistic understanding with latent horizontal features. In: CVPR (2021)
52. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loftr: Detector-free local feature matching with transformers. In: CVPR (2021)
53. Sweeney, C., Hollerer, T., Turk, M.: Theia: A fast and scalable structure-from-motion library. In: Proceedings of the 23rd ACM international conference on Multimedia (2015)
54. Wilson, K., Snavely, N.: Robust global translations with 1DSfM. In: ECCV (2014)
55. Yang, Z., Pan, J.Z., Luo, L., Zhou, X., Grauman, K., Huang, Q.: Extreme relative pose estimation for RGB-D scans via scene completion. In: CVPR (2019)
56. Yang, Z., Yan, S., Huang, Q.: Extreme relative pose network under hybrid representations. In: CVPR (2020)
57. Zach, C., Klopschitz, M., Pollefeys, M.: Disambiguating visual relations using loop constraints. In: CVPR (2010)
58. Zhang, F., Nauata, N., Furukawa, Y.: Conv-MPN: Convolutional message passing neural network for structured outdoor architecture reconstruction. In: CVPR (2020)
59. Zhang, Y., Song, S., Tan, P., Xiao, J.: Panocontext: A whole-room 3D context model for panoramic scene understanding. In: ECCV (2014)
60. Zheng, J., Zhang, J., Li, J., Tang, R., Gao, S., Zhou, Z.: Structured3d: A large photo-realistic dataset for structured 3D modeling. In: ECCV (2020)
61. Zou, C., Su, J.W., Peng, C.H., Colburn, A., Shan, Q., Wonka, P., Chu, H.K., Hoiem, D.: Manhattan room layout reconstruction from a single 360° image: A comparative study of state-of-the-art methods. International Journal of Computer Vision **129**(5) (May 2021)