Extreme Structure from Motion for Indoor Panoramas without Visual Overlaps

Mohammad Amin Shabani
Simon Fraser University
mshabani@sfu.ca

Weilian Song
Simon Fraser University
weilians@sfu.ca

Makoto Odamaki
Ricoh Company, Ltd.
makoto.odamaki@jp.ricoh.com

Hirochika Fujiki
Ricoh Company, Ltd.
hirochika.fujiki@jp.ricoh.com

Yasutaka Furukawa
Simon Fraser University
furukawa@sfu.ca

Figure 1. The paper introduces an extreme Structure from Motion problem for indoor panoramas that have little to no visual overlaps. Our approach learns to evaluate the realism of room/door/window arrangements in the top-down semantic space and solve for the camera poses.

Abstract

This paper proposes an extreme Structure from Motion (SfM) algorithm for residential indoor panoramas that have little to no visual overlaps. Only a single panorama is present in a room for many cases, making the task infeasible for existing SfM algorithms. Our idea is to learn to evaluate the realism of room/door/window arrangements in the top-down semantic space. After using heuristics to enumerate possible arrangements based on door detections, we evaluate their realism scores, pick the most realistic arrangement, and return the corresponding camera poses. We evaluate the proposed approach on a dataset of 1029 panorama images with 286 houses. Our qualitative and quantitative evaluations show that an existing SfM approach completely fails for most of the houses. The proposed approach achieves the mean positional error of less than 1.0 meter for 47% of the houses and even 78% when considering the top five reconstructions. We will share the code and data in https://github.com/aminshabani/extreme-indoor-sfm.

1. Introduction

The emergence of consumer-grade panorama cameras is making a revolution in the real-estate industry. With only a few hundred dollars per unit, increasingly more number of real-estate agents and home owners utilize the cameras to snap panoramas, enabling house renters or buyers to browse through full-360 interior views with the flick of a finger. In particular, the THETA series from RICOH is collecting 100 million panoramas for real-estate applications.

This incredible market growth comes from mass-consumer crowd sourcing, whereas the operation must be simple. Given lengthy instructions on how to 1) use a camera, 2) set up a mono/tri-pod, and 3) use a smartphone app to verify shootings, create annotations, and moderate contents, users are simply asked to take a picture in the middle of each room. Therefore, panorama images have little to no visual overlaps, making the pose estimation infeasible for existing techniques. A robust panorama alignment algorithm will enable a plethora of applications such as automated floorplan generation, accurate price prediction, and verification of building-codes.

This paper proposes an extreme Structure from Motion (SfM) problem for indoor panoramas with little to no visual overlaps and provides a novel compelling solution to the problem. Our key idea is to learn the arrangement of rooms, doors, and windows, and solve for camera parameters that maximize the realism of their arrangement. For example, a rest-room and a shower-room are often adjacent and nearby

the entrance. Bedrooms are connected to a living room, and a balcony is typically on the opposite side from an entrance.

Concretely, given a set of panoramas, we use standard techniques to apply Manhattan-rectification, infer a room layout, detect doors/windows with their types, and classify a room type for each panorama. Inferred semantic information is re-projected into a Nadir (i.e., top-down) view as a semantic image. We generate arrangement candidates by aligning Nadir semantic images based on the door detections. Finally, a convolutional message passing neural network learns to score the generated arrangements, where we output the one with the highest score as the reconstruction.

We use 1029 panoramas for 286 houses from a production pipeline. A standard SfM approach fails to align even two panoramas for most of the houses [12]. The proposed system reconstructs compelling arrangements, concretely the mean positional error being less than 1.0 meter in the top five reconstructions for 78% of the test houses.

The contribution of the paper is three fold: 1) A new extreme indoor SfM problem with the new dataset for the exploding market; 2) One-of-a-kind SfM algorithm which learns to evaluate the arrangement of semantic information; and 3) State-of-the-art performance where existing techniques fail. We will share the code, models, and data.

2. Related Work

The paper tackles a pose estimation problem from indoor photographs. We study related works in Structure from Motion (SfM), extreme pose estimation with minimal visual overlaps, and indoor digital scanning.

Structure from Motion: Feature matching, geometric verification, and reconstruction has been the golden standard for camera pose estimation, known as the SfM pipeline [14]. Successful SfM system has been presented even for Internet photo collections in a massive scale [15, 1]. SfM also has been used for floorplan reconstruction [3] or reconstructing a single 3d model of a building [5]. Deep neural networks further robustify the feature matching process for wide-baseline scenarios [20]. Nonetheless, these techniques require ample visual overlaps with well textured surfaces among input images, incapable of handling our problem where images have little to no visual overlaps.

Extremal pose estimation: Priors on standard room shapes have been exploited for the alignment of perspective images [19] or partial SLAM reconstructions with minimal data overlaps in a single room [8]. In contrast, this paper seeks to align images from different rooms by exploiting the regularities of room arrangement at a house-scale. A sitemap was utilized for the registration of SfM reconstructions without any data overlap [11] via heuristics. Our problem does not have a map (i.e., floorplan).

Indoor digital scanning: Image-based indoor 3D reconstruction made great progress nearly a decade ago [6], but was not robust enough for production. The advent of consumer-grade depth sensors made a breakthrough in the indoor 3D scanning via RGBD videos [13, 10]. However, the operation was too complicated for non-experts to use as a production system. Panorama RGBD cameras have been successful in industry for 3D indoor scanning, where Matterport is a good example [2]. Their operations are much simpler than those of RGBD videos. However, the system is still cumbersome for mass consumers and has suffered from slow adoption in the real-estate business. This paper proposes a novel SfM algorithm, namely a sparse set of panoramas with little to no visual overlaps, which has been exploding in the past five years.

3. Dataset and Problem Definition

A dataset contains 1029 panoramas and 286 apartments/houses from a production pipeline. RICOH THETA camera series are used for the data acquisition. The number of panoramas per house ranges from 2 to 7, in particular, $44/91/91/58/2$ houses contain $2/3/4/5/7$ panoramas, respectively. Panorama images are rescaled to 1024×512.

Annotations: Each panorama is associated with the following set of information/annotations (See Fig. 2).

- A floorplan image of the apartment/house.
- Manhattan rectification parameters estimated by the official HorizonNet [16] code package.
- A Manhattan room layout. We seek to identify the extent of a current room instead of estimating the entire visible floor region through doorways. As seen in Fig. 4, our layout annotations do not include spaces behind doors.
- A camera pose (a 2D position and a heading) with respect to the floorplan. The positions are calculated in the unit of meters by assuming that the room height is 3.2 meters. [1]
- Window and door instance segmentation. Each instance is associated with a window/door type label. [2]
- A room type label where a panorama is taken.
- A set of room-to-room connections via doors.

Problem definition: The input is a set of panoramas from a single house/apartment. The output is 2D relative camera poses, that is, a 2D position and a heading angle per panorama. We assume that cameras are placed at a constant height from a flat floor and the gravity rectification is successful [3]. An SfM reconstruction is defined up to a sim-

[1] Assuming a fixed camera height, we convert a room layout into a 2D segmentation mask in a top-down view and manually place it in the floorplan while adjusting the overall scale of the masks. See the supplementary document for the details of our annotation system.

[2] Window-door types are Door/Glass-door/Frame/Window/Kitchen-counter/Closet. Room types are Balcony/Closet/Western-style/Japanese-style/Dining-room/Kitchen/Corridor/Washroom/Bathroom/Toilet.

[3] Production data acquisition usually requires a monopod or a tripod to avoid hand-shakes, making the camera height roughly constant. Together

Figure 2. Extreme indoor panorama dataset consisting of 1029 panoramas and 286 apartments/houses. Our annotations include a floorplan image, a Manhattan room layout, door/window detections with types, room-type classifications, room-to-room connections, and camera poses with respect to the floorplan image.

ilarity transformation, and the reconstructed camera centers are registered to the ground truth by minimizing the sum of squared distances (i.e., aligning the center of mass and using SVD to solve for a scaled rotation). Our reconstruction is metric, which we align by a rigid transformation.

Metric: The performance is evaluated by the rate of success, where a method is defined to be a success if top-K reconstructions contain a solution, whose mean positional error is below δ meters. We vary K from 1 to 10. The threshold δ is set to 0.2, 0.6, or 1.0 meter.

4. Extreme Indoor Structure from Motion

Extreme Indoor SfM is challenging even for trained human annotators, where large-scale data/annotation collection is not easy. We extract the architectural semantic information from each panorama and re-project into a "Nadir semantic representation", which cuts the flow of raw pixel information and avoids network over-fitting. After generating arrangement candidates by aligning doors in the semantic images, a convolutional message passing neural network learns to evaluate the realism of the arrangements. The section explains the Nadir semantic image representation, the semantic image construction, the arrangement generation, and the arrangement evaluation (See Fig. 3).

4.1. Nadir semantic image

A nadir semantic image is a 16-channel image in the nadir (i.e., top-down) view, representing the room shape, the room type, the door/window locations, and the door/window types. The image is of resolution 256× 256 (1 pixel = 4 cm). The panorama center is at the center of the image and the left border of the (Manhattan-rectified) panorama maps to the image x-axis.

The first 10 channels are segmentation masks for the 10 room types. Suppose a room type is "Kitchen" which corresponds to the first channel, the room shape is given as a segmentation mask in the first channel. The remaining 9 channels become 0. There are 6 door/window types, and the remaining six channels are their segmentation masks.

4.2. Semantic image construction

A room layout can be reprojected to the Nadir image as a polygonal shape, where the scale comes from the assumption that the room height is 3.2 meters as in the dataset preparation. Bounding boxes of the doors and windows are assumed to be on the room walls, and reprojected to the Nadir image. At times, a bounding box overlaps with two edges of a room polygon, in which case we clip the bounding box at the corner and keeps the longer side. PIL Image library [4] is used to draw an image, where doors/windows are drawn with a thickness of 3 pixels.

During training, we use the ground truth to generate a semantic nadir image for each panorama. During testing, we apply standard techniques for Manhattan-rectification [16], room layout estimation [16], door/window detection and type classification [18], and room type classification [17]. The room and door/window types are probabilistic in this case, where the room and door/window segmentation masks store the probability scores instead of being binary. See Sect. 5.1 for the architecture details and Fig. 5 for the samples of estimated nadir semantic images.

4.3. Arrangement generation

We use door detections to align panoramas and create arrangement candidates. [4] Given two doors from two panora-

with the accelerometer in an IMU yielding the gravity, it suffices to estimate only the horizontal position and heading angle.

[4] Door/window type classification is not reliable and we treat all the door/window detections (except for "Window") as the same doors. The full door/window types are utilized by baseline methods in the experiments.

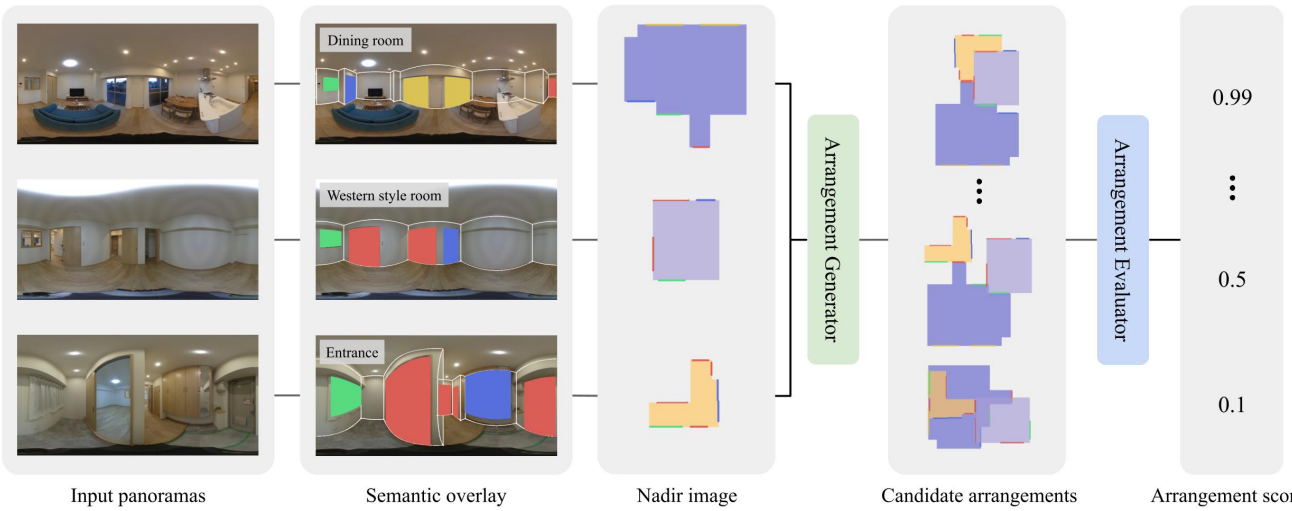

Figure 3. Overview of our system. For each panorama, we estimate a room layout and detect doors/windows by standard CNNs. After re-projecting the information into a nadir (i.e., top-down) view, we generate arrangement candidates based on the door connections. Finally, a convolutional message passing network is used to evaluate the realism of each arrangement.

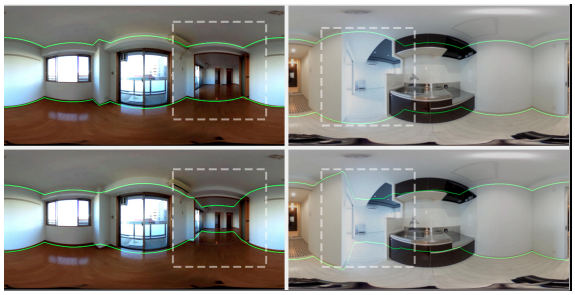

Figure 4. (Top) Our room layout annotations capture only the extent of the current room. (Bottom) A state-of-the-art layout estimation algorithm HorizonNet [16] seeks to cover the entire visible floor region through doorways and open-spaces instead.

mas, we register their nadir semantic images by making the doors parallel and aligning their centers, while keeping the rooms on the opposite side. Candidate generation is exhaustive, using DFS to enumerate all possible arrangements while enforcing that each door is used at most once.

4.4. Arrangement evaluation

Given an arrangement of Nadir semantic images, a convolutional message passing network (ConvMPN) [21] learns to evaluate its realism score. ConvMPN is a variant of a graph neural network, whose input is a graph of nadir semantic images, which are fully connected in the relational graph. We could use the room-to-room connections to define a different relational graph, but this made no difference in the experiments. Starting from a resolution of 256×256, we iterate convolutional message passing and max-pooling to shrink the feature resolutions, pool features from all the nodes, and use a FC layer to predict the realism score. See

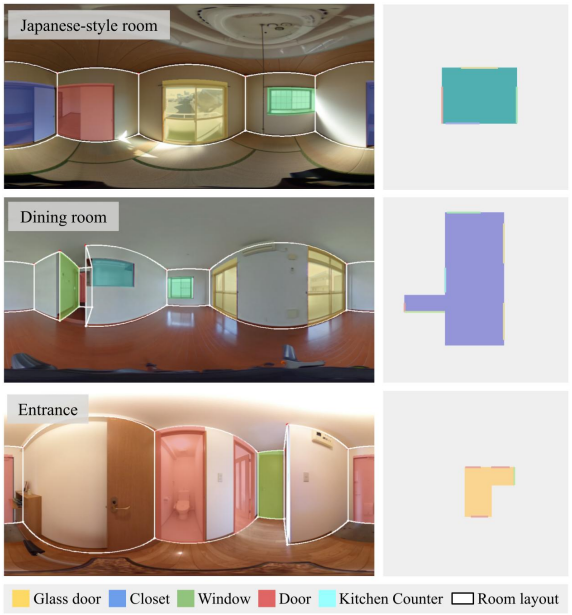

Figure 5. Left: Panoramas with the inferred room type, the room layout, and the detected doors/windows (plus type and segmentation). Right: Constructed nadir semantic images.

the supplementary document for the architecture details.

Training data generation: We use the ground-truth to generate positive samples instead of directly using the GT arrangements, in which door masks are not exactly aligned and the network might cheat in distinguishing positive samples. Concretely, given the GT room-to-room connections via doors, we exhaustively enumerate all the possible panorama connections with a tree topology (i.e., minimal required connections) and use the same algorithm as in

Sect. 4.3 to produce positive arrangement samples.

Negative samples are the arrangements made by the arrangement generation process in Sect. 4.3, excluding the positive ones. The challenge is the data imbalance, where negative samples (1,968,679) are a lot more than the positive ones (353). Similar in spirit to hard negative mining, we found that the following two heuristic filters are effective in subsampling easy negatives and focusing on hard samples.

• **Overlap filter** removes arrangements where rooms have significant overlaps. We define the "overlap-ratio" of a panorama to be the ratio of pixels in its room mask that overlap with another panorama. For example, if a mask is completely enclosed by another, the overlap-ratio becomes 1.0. The filter rejects an arrangement if the average overlap-ratio is above 0.1, while ignoring "Dining-room" and "Kitchen", which tend to overlap with each other. At test time, a room type is set to the one with the highest probability, and the same goes for the door/window types next.

• **Door-type filter** removes arrangements where doors of different types are matched even at a single connection.

By using these two filters, we divide the negative arrangements into mutually exclusive three groups: Hard (passing both filters), Intermediate (passing only the overlap filter), and Easy (the rest). From 240 training units, the process generates 353 positive, 34035 hard-negative, 1,025,699 intermediate-negative, and 908,592 easy-negative samples.

During training, we form a batch of size 32 by randomly sampling (4, 8, 16, and 4) samples from the four groups (positive, hard-negative, intermediate-negative, and easy-negative), respectively. 32 training samples do not fit in GPU memory and we process a sample one by one, while accumulating gradients over the 32 samples before updating the network parameters. Lastly, we set the regression target of the samples in these 4 groups as (1, 0, -1, -1) with a mean squared error loss instead of the binary classifier loss.

5. Implementation details

We use a workstation with 2.20 GHz Xeon (40 CPU cores) and dual NVIDIA GTX 1080 Ti GPUs. The training takes roughly 11 hours for 340k iterations with a batch of 32 samples. At test time, the semantic image construction, the arrangement generation, and the arrangement evaluation takes less than 1 min, 3 mins, and 20 seconds for a typical house/apartment. Test time execution slows down exponentially as the number of panorama grows. Our biggest house with 7 panoramas takes 1 hour for processing. The section explains the semantic image construction networks and the competing methods in our comparative evaluation.

5.1. Semantic image construction networks

We use standard CNN architectures for the implementation of panorama preprocessing networks in Sect. 4.2.

Layout estimation: We downloaded pretrained HorizonNet [16] which was trained on 18362 panoramas from Structured3D dataset [22]. We fine-tune the network on our layout annotations for 300 epochs with a batch size of 8, which improves the 3D IoU score [16] from 74.42 to 88.12.

Door detection and type classification: We use an instance segmentation network from Detectron2 [18], in particular, Feature Pyramid Networks [9] with ResNet-101 [7] as the backbone. We fine-tune the pretrained model for 250k iterations with a learning rate of 0.0025 and the batch-size of 2, which improved the average precision to 52.498 at AP50.

Room type classification: We design a CNN encoder with 8 convolutional layers of 8 channels, each followed by a group normalization [17] of 4 groups, ReLU as an activation function, and MaxPool for downsampling. The input is a 256×256 panorama image (rescaled to the 1:1 aspect ratio). The trained network achieves 68.51% accuracy, where most confusions occur between Dining Hall and Kitchen or Western-Style-Room and Japense-Style-Room.

5.2. Competing methods

We compare against four competing methods. Note that recent extreme pose estimation algorithms [19, 8] assume that images belong to the same room and cannot be used.

• **Overlap and door-type filters** are heuristic arrangement filters used for sampling training data. We build a baseline by adding a random selection after the filters: randomly picking an arrangement as the answer.

• **Overlap filter** is the same as above except that only one filter is used before the random selection.

• **Retrieval-baseline** looks at the nadir semantic images of a given panorama set, finds the most similar set from the training data, and returns the corresponding GT arrangement as the answer. Architectural design follows certain layout principles, but this baseline is to verify that such a simple approach fails. Concretely, considering a nadir semantic image as a 16 channel image, we define the distance of semantic images as their L1 norm, while considering the four-fold rotational ambiguities and using the best case. Given two sets of panoramas, we solve a bipartite matching and uses the sum of the distances as their inverse similarity, which is used to retrieve the most similar panorama set from the training data. We ignore training samples whose number of panoramas is less than that of the query.

• **SfM** is a traditional SfM pipeline designed for panorama images from an open-source OpenMVG library [12].

6. Experimental results

We randomly split our data into 240 training and 46 testing houses/apartments. 46 test houses contain 154 panoramas. In particular, 8/15/22/1 houses contain 2/3/4/5 panora-

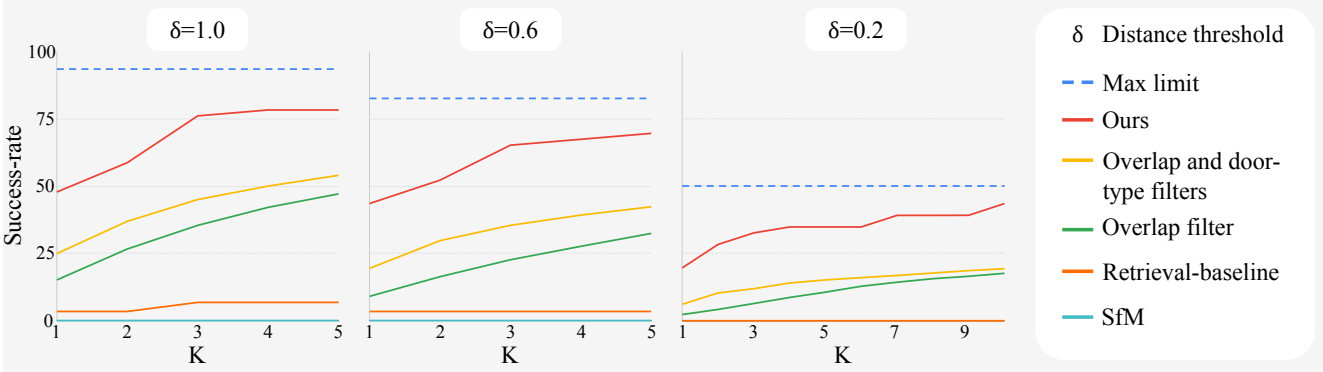

Figure 6. Quantitative evaluations. We compare against four competing methods with the "success-rate" metric in Sect. 3. For example, the left plot shows the rate in which top-K arrangements contain a successful reconstruction (i.e., mean positional error less than 1.0 meter).

Table 1. Ablation study on the importance of the room-shape (R_{shape}), the room-type (R_{type}), and the door/window (DW) information. The table shows the mean average precision over the values of K=1∼5 at three different distance thresholds (δ). The second row is our system, where predictions (Pred) are used for all the information. The first row is a case where the ground-truth (GT) information are used instead. In the last three rows, we drop the room-type and/or door/window information (denoted as 'x').

R_{shape}	R_{type}	DW	$\delta = 0.2$	$\delta = 0.6$	$\delta = 1.0$
GT	GT	GT	50.4	64.8	70.9
Pred	Pred	Pred	30.0	59.6	67.8
Pred	x	Pred	21.3	60.0	64.3
Pred	Pred	x	12.6	26.5	34.7
Pred	x	x	4.7	14.4	24.5

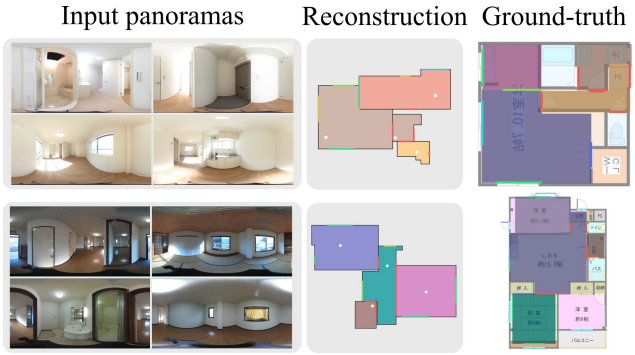

Figure 7. Failure examples. Our algorithm relies strongly on pre-processing networks and is not able to recover from mistakes by the room layout estimation and/or door/window detections.

mas, respectively. At total, door/window detectors find 279 Doors, 70 Glass doors, 53 Frames, 69 Windows, 29 Kitchen counters, and 59 Closets. The average number of arrangements per unit is 3512. Overlap-filter reduces the number to 1938, and Door-type filter further reduces the number to 187, which are used in our baseline methods.

Quantitative evaluations: Figure 6 is our main result, comparing our approach against the four competing methods based on the success-rate metric (See Sect. 3). The blue dashed line (*Max Limit*) shows the maximum possible success-rate by the oracle arrangement evaluator, given the constructed semantic nadir images, which contain errors. For example, Max limit is at 82.6 when δ=0.6, indicating that none of the generated arrangements have a mean positional error less than 0.6 for 17.4% of the test samples.

SfM completely fail for every single example and has 0 success-rate for every entry. It manages to align 2 panoramas for a few testing cases, but was never able to align all. The heuristic filters do some reasonable jobs, in fact, much better than Retrieval-baseline, demonstrating the challenges of our problem. Our method outperforms all the other competing methods with significant margins.

Table 1 provides an ablation study on the importance of different information components in the nadir semantic representation. The first row presents a scenario where the semantic image construction is perfect (i.e., no errors in the panorama pre-processing networks). In the bottom three rows, we drop the room-type information (i.e., treating all the room types to be the same) and the door/window-information (i.e., setting the door/window segmentation masks to be empty). For each setup, we retrain our evaluators with the modified data representation. The fourth entry shows that the door/window information is critical in identifying the correct arrangement as expected.

Qualitative evaluations: Figure 8 provides qualitative comparisons against the competing methods. The left column reveals the extreme nature of our problem, where there exists little to no visual overlaps and the task appears infeasible even for humans. Given the difficulty, baseline results are impressive. Retrieval baseline does not utilize the inferred room shape/size information and often makes gross errors. Our approach consistently outperforms the others.

Figure 9 shows the top 5 reconstructions by our method for more examples. The top example has a panorama at a balcony, which is challenging for a standard SfM system,

Figure 8. Qualitative comparisons against the three competing methods. We show the top-2 reconstructions from each method based on their scoring functions. Room colors indicate their types

because 1) the majority of the image sees the outdoor space; and 2) the house interior is under-exposed due to the limited dynamic range. On the contrary. this is one of the easiest panoramas for our system, as the room type classification is trivial and a balcony should be connected to a bright door/window in a living room, which can be easily detected. Our approach learns the architectural rules of layouts and often finds an accurate arrangement in the top 3, which look realistic even when they are incorrect.

Discussions and future work: This paper introduces a new extreme SfM problem for indoor panoramas and proposes a unique SfM algorithm, which learns to evaluate the arrangement of rooms/doors/windows without solving a correspondence problem. Our algorithm makes significant improvements over the current state-of-the-art, which completely fails for every single example.

Our solution is still far from perfect (See Fig. 7). First, the running time is exponential in the number of panoramas.

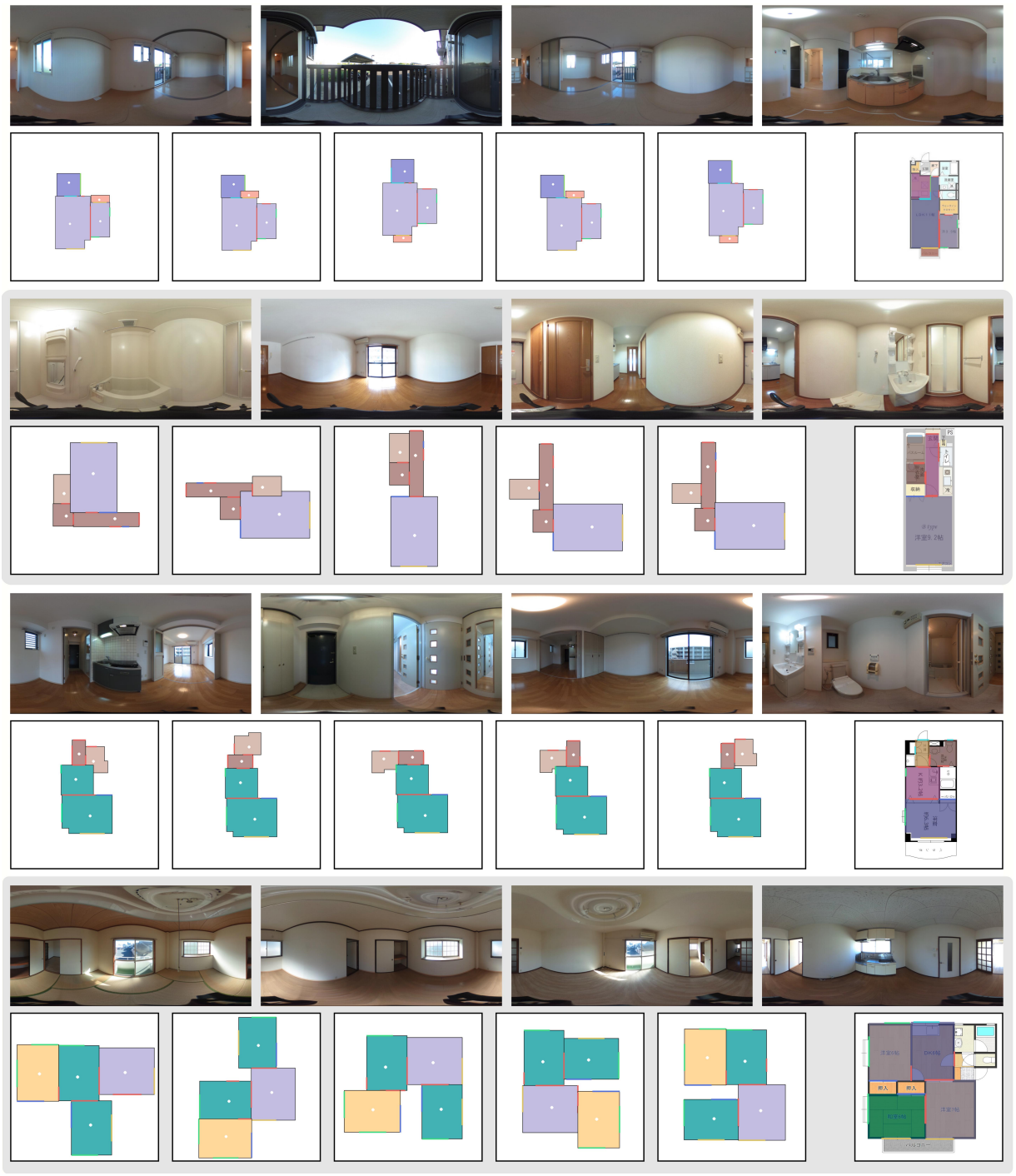

Figure 9. Qualitative evaluations. Top-5 reconstructions by our method against the ground-truth arrangement.

Second, the algorithm highly depends on the door detection, in particular, is not able to recover from missing doors. Our future work is the development of more scalable algorithm that does not require hard door/window detections. Please refer to the supplementary for system/architecture details, more results on panorama-preprocessing networks, more reconstruction results, and intermediate visualization revealing what the arrangement evaluators learned.

Acknowledgement: This research is partially supported by NSERC Discovery Grants with Accelerator Supplements and DND/NSERC Discovery Grant Supplement.

References

[1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM*, 54(10):105–112, 2011. 2

[2] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1534–1543, 2016. 2

[3] Ricardo Cabral and Yasutaka Furukawa. Piecewise planar and compact floorplan reconstruction from images. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 628–635. IEEE, 2014. 2

[4] Alex Clark. Pillow (pil fork) documentation. *Readthedocs. https://buildmedia. readthedocs. org/media/pdf/pillow/latest/pillow. pdf*, 2015. 3

[5] Andrea Cohen, Johannes L. Schönberger, Pablo Speciale, Torsten Sattler, Jan-Michael Frahm, and Marc Pollefeys. Indoor-outdoor 3d reconstruction alignment. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 285–300, Cham, 2016. Springer International Publishing. 2

[6] Yasutaka Furukawa, Brian Curless, Steven M Seitz, and Richard Szeliski. Reconstructing building interiors from images. In *2009 IEEE 12th International Conference on Computer Vision*, pages 80–87. IEEE, 2009. 2

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[8] Cheng Lin, Changjian Li, and Wenping Wang. Floorplan-jigsaw: Jointly estimating scene layout and aligning partial scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5674–5683, 2019. 2, 5

[9] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017. 5

[10] Chen Liu, Jiaye Wu, and Yasutaka Furukawa. Floornet: A unified framework for floorplan reconstruction from 3d scans. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–217, 2018. 2

[11] Ricardo Martin-Brualla, Yanling He, Bryan C Russell, and Steven M Seitz. The 3d jigsaw puzzle: Mapping large indoor spaces. In *European Conference on Computer Vision*, pages 1–16. Springer, 2014. 2

[12] Pierre Moulon, Pascal Monasse, Romuald Perrot, and Renaud Marlet. Openmvg: Open multiple view geometry. In *International Workshop on Reproducible Research in Pattern Recognition*, pages 60–74. Springer, 2016. 2, 5

[13] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE international symposium on mixed and augmented reality*, pages 127–136. IEEE, 2011. 2

[14] GF Page. Multiple view geometry in computer vision, by richard hartley and andrew zisserman, cup, cambridge, uk, 2003, vi+ 560 pp., isbn 0-521-54051-8.(paperback£ 44.95), 2005. 2

[15] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. *ACM siggraph*, pages 835–846, 2006. 2

[16] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1047–1056, 2019. 2, 3, 4, 5

[17] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 3, 5

[18] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 3, 5

[19] Zhenpei Yang, Siming Yan, and Qixing Huang. Extreme relative pose network under hybrid representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2455–2464, 2020. 2, 5

[20] Kwang Moo Yi, Eduard Trulls, Yuki Ono, Vincent Lepetit, Mathieu Salzmann, and Pascal Fua. Learning to find good correspondences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2666–2674, 2018. 2

[21] Fuyang Zhang, Nelson Nauata, and Yasutaka Furukawa. Conv-mpn: Convolutional message passing neural network for structured outdoor architecture reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2798–2807, 2020. 4

[22] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Proceedings of The European Conference on Computer Vision (ECCV)*, 2020. 5