# CoVisPose: Co-Visibility Pose Transformer for Wide-Baseline Relative Pose Estimation in 360° Indoor Panoramas

Will Hutchcroft[1], Yuguang Li[1], Ivaylo Boyadzhiev[1], Zhiqiang Wan[1], Haiyan Wang[2], and Sing Bing Kang[1]

[1] Zillow Group {willhu,yuguangl,ivaylob,zhiqiangw,singbingk}@zillow.com
[2] The City College of New York hwang005@citymail.cuny.edu

**Abstract.** We present *CoVisPose*, a new end-to-end supervised learning method for relative camera pose estimation in wide baseline 360° indoor panoramas. To address the challenges of occlusion, perspective changes, and textureless or repetitive regions, we generate rich representations for direct pose regression by jointly learning dense *bidirectional* visual overlap, correspondence, and layout geometry. We estimate three image column-wise quantities: *co-visibility* (the probability that a given column's image content is seen in the other panorama), *angular correspondence* (angular matching of columns across panoramas), and *floor layout* (the vertical floor-wall boundary angle). We learn these dense outputs by applying a transformer over the image-column feature sequences, which cover the full 360° field-of-view (FoV) from both panoramas. The resultant rich representation supports learning robust relative poses with an efficient 1D convolutional decoder. In addition to learned direct pose regression with scale, our network also supports pose estimation through a RANSAC-based rigid registration of the predicted corresponding layout boundary points. Our method is robust to extremely wide baselines with very low visual overlap, as well as significant occlusions. We improve upon the SOTA by a large margin, as demonstrated on a large-scale dataset of real homes, ZInD.

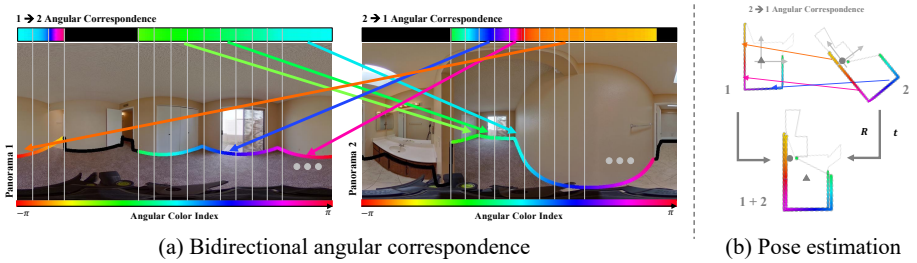
**Keywords:** Indoor, 360° panorama, indoor, pose estimation, camera localization, structure-from-motion, layout

## 1 Introduction

With the increasing affordability of 360° capture devices, omnidirectional imagery has become an important capture modality for indoor environments[3]. The large-FoV provides an immersive experience as well as comprehensive context for indoor scene understanding; this enables applications such as AR/VR, autonomous navigation, virtual tours, room layout estimation, and floor plan reconstruction. The omnidirectional information allows sparser capture while maintaining geometric context. Concurrent with the rise of deep learning, these advantages have inspired

[3] https://www.ricoh360.com/tours/

a rapid increase in research focused on the spherical domain, including layout estimation [45,46,36,54,58,55], depth estimation [58,53,46], semantic segmentation [59,19], object detection [30], and network design [44,29].




(a) Bidirectional angular correspondence (b) Pose estimation

Fig. 1: Our system (a) establishes bidirectional column-wise (angular) correspondence while computing (b) pose estimation between two panoramas that visually overlap.

Commercially, for practical and economic reasons, panoramic captures for virtual tours and floor plan reconstruction typically result in sparse coverage, with wide baselines between panoramas [43,12,37]. In particular, the indoor environments of real homes in ZInD[12] have extensive featureless regions and strong visual similarity between rooms, which present challenges to classical feature-based Structure-from-Motion (SfM) approaches. To alleviate that, other methods require denser RGB [1,6] or RGB-D [38,8] captures. However, those are at the expense of more time (to capture) or investment in specialized hardware.

In this paper, we propose a new method to estimate relative pose for a pair of 360° indoor panoramas under a wide range of baselines *(small to extreme)*. To address the challenges associated with operating in the spherical domain, we take inspiration from the horizontal (1D column-wise) representation for layout estimation [45,46]. We jointly learn to regress relative pose alongside estimation of image-column-wise representations of visual overlap (co-visibility), correspondence, and layout geometry. To learn the highly non-local associative tasks of co-visibility and correspondence, we apply a transformer to image-column feature sequences, allowing the network to attend globally across the full 360° context from both panoramas. Further, jointly learning to estimate layout geometry provides the network with a strong prior for the indoor environment. By providing dense supervision, we reduce ambiguity and guide the transformer to learn rich representations for robust pose regression, even in the presence of minimal visual overlap and wide baselines.

Our contributions are:

- Novel representation for relative pose estimation between two upright 360° cameras, which factors the auxiliary tasks of visual overlap, correspondence, and layout estimation as image-column-wise quantities.
- Transformer-based architecture that operates over the image-column feature sequences across both panoramas, applying inter and intra-image column attention.

- Support for both end-to-end direct pose regression as well as a post-processing step using iterative robust model fitting, e.g., RANSAC [61], enabled through the densely predicted corresponding layout boundary points.
- Co-visibility score generated by our co-visibility representation, which gives a strong measure of trust in a given predicted relative pose. Using this measure, we demonstrate strong pose precision and recall performance for pose estimation in unordered panoramas.
- Direct estimation of two-view SfM with scale. To our knowledge, this is the first end-to-end learning approach to estimate the pose of upright panoramas with respect to predicted layout geometry.
- SOTA performance on a challenging real world dataset [12]. We achieve a 78% and 85% decrease in median rotation and translation error, respectively, over a SOTA deep feature matching approach, while at the same time successfully estimating a pose for 35% more panorama pairs.

## 2 Related Work

In this section, we briefly review representative methods in the related areas of pose estimation and room layout estimation.

**Two-View Pose Estimation** Classical methods for relative pose estimation (RPE) first extract and match image features like SIFT [31], which are then used to derive the relative camera motion by estimating the fundamental or essential matrix, for uncalibrated or calibrated cameras, respectively [62]. These methods are generally well behaved and robust when the camera motions are small and the scene texture is amenable to extraction and matching of features; however, common failure modes include repetitive or limited texture, as well as large camera motion between views. Recent works have attempted to address these challenges using deep learning. Many works have focused on modeling those components of the classical pipeline which are especially susceptible to failure, such as feature detection [18,14], correspondence estimation [57,40,47], and model fitting [4,39].

Of the recent works for deep feature detection, description, and correspondence estimation, a combination of [14] and [40] has proven particularly effective for relative pose estimation. The system of [14] replaces hand-crafted interest-point detection and descriptors with learned counterparts by attaching two separate decoder branches to a shared CNN encoder. The method of [40] accepts two detection sets and learns the feature matching step with a Graph Neural Network (GNN). Its attentional GNN variant combines both inter- and intra-image attention in order to reason about both appearance and spatial cues. However, there is no feedback from matching; correspondence estimation is subject to the input detection quality with no global reasoning.

LoFTR [47] learns to perform both steps in a detector-free approach and directly outputs dense correspondences. Their system also leverage inter- and intra-image attention with a transformer applied in a coarse-to-fine approach. This global reasoning improves matching for regions with limited texture or

repetitive patterns. This method sets a new state-of-the-art on multiple benchmarks, including indoor relative pose estimation in the perspective imagery of ScanNet. Nevertheless, this method only focuses on one portion of the SfM pipeline, and may still be subject to difficulties with model fitting.

In contrast, other techniques directly learn the mapping function between images and pose [9,32,50], demonstrating improved performance for classically challenging cases, including wide baselines. Similar to our work, many techniques [32,20,9] first apply a feature extractor in a Siamese configuration, sharing weights across the input pair. Melekhov et al. [32] use the extracted feature representations directly as input to two fully-connected layers to regress the relative pose. En et al. [20] propose multiple variants, including relative pose computation from two absolute pose estimates, with a similar fully-connected regressor displaying the best overall performance. Notably, they estimate the full translation vector and report errors in meters. Chen et al. [9] formulate a directional parameterization of the relative pose. They stack multiple decoder blocks on top of a Siamese encoder and estimate discrete distributions over the sphere. Their best performing architecture follows a two-stage approach which first derotates the image before estimating the translation component, making it robust to wide baselines.

Such direct regression methods have also been proposed for the task of absolute pose estimation (APE), with the aim of learning the camera-to-scene transformation directly [26,52,25,5]. One subproblem of APE is the visual relocalization task, which aims to localize one or more target images against scene images of known pose. This problem may be framed as retrieval-then-RPE, wherein nearest neighbors are retrieved, with the scene pose subsequently determined through RPE. Such methods have demonstrated strong ability to generalize [2,28] as the pose regressor is not tied to a scene-specific coordinate frame. Laskar et al. [28] first train a Siamese architecture to regress relative pose, and then use the learned feature representation for database retrieval of neighboring panoramas. Balntas et al.'s work [2] is in a similar spirit with ours in that they estimate a camera frustrum overlap, analogous to our co-visibility. From the same feature embeddings, they additionally regress the relative pose, which strengthens the retrieval performance while serving to localize to scene coordinates.

Two-view pose estimation with given priors (e.g., gravity-aligned vertical direction [27,17,16]) or constrained motions (e.g., planar camera motion [21,11,1]) is an active area of research. This has many practical applications such as robotics [34] and virtual walkthroughs [12,1], and are enabled by the availability and robust integration of low-cost IMU sensors [22] and improved algorithms for upright camera corrections [24,13,23]. In our work, we assume a planar motion model for the spherical cameras, i.e., all the cameras lie on the same plane with a fixed height and a gravity-aligned vertical direction. These practical constraints are used in commercial applications, e.g., application of Street View technology indoors [1]. Those assumptions reduce the general 6-DoF two-view spherical geometry [48] to a 3-DoF problem solved in the 2D plane [1].

**Layout Estimation** HorizonNet [45] introduced the horizontal representation for layout estimation in 360° indoor panoramas, significantly improving SOTA

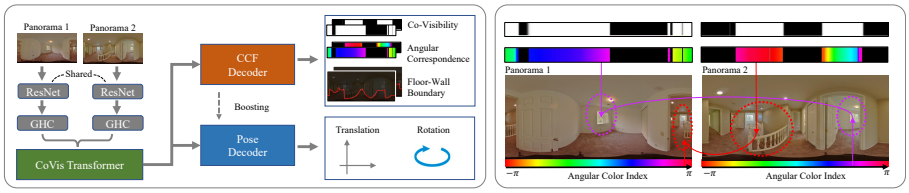
by factoring layout into output vectors over the image columns. By applying an LSTM to the image column features produced through a height compression module (HCM), they estimate both the floor and ceiling boundary contours as well as the corner probability score. HorizonNet further exploits the upright camera assumption to post-process Manhattan layouts. HoHoNet [46] improves the HCM efficiency, and demonstrates the ability to predict per-pixel modalities by applying an inverse discrete cosine transform to decompress the latent feature representation.

Our work is inspired by the success of the horizontal representation in the indoor domain. To learn rich pose representations, we frame correspondence and visual overlap estimation as column-wise prediction, and additionally estimate the layout contour from both views, to provide a strong prior for the indoor environment. Similar to [47], we leverage a transformer to attend to inter and intra-image relationships; however, our transformer is applied over image column feature sequences, analogous to sequence processing in NLP, to efficiently aggregate the full 360° range. Moreover, we do not require full depth maps for training; only the sparse wall layout geometry is necessary. As has been demonstrated in the perspective domain [9], direct pose regression provides advantages for wide-baselines, occlusion, textureless regions, and other classically challenging cases. The rich representations learned by our CoVis transformer bring these benefits, for the first time, to the spherical domain.

## 3 Overview

Our CoVisPose architecture is shown in Fig. 2 (a). The inputs to our system are a pair of 360° panoramas in equirectangular projection, captured in an indoor space. We assume the camera is upright, with a fixed height for each home. The orientation with the gravity vector is imposed via straightening as a pre-processing step [60]. Each panorama may or may not have visual overlap with neighboring panoramas. Further, we assume Atlanta world layouts [42], where the walls are upright and orthogonal to the floor. As seen in Fig. 2 (a), we adopt the feature extractor from HorizonNet [45] in a Siamese configuration, with shared weights between the branches. Each branch consists of a ResNet50 backbone followed by an HCM to produce a feature sequence over the image columns. The feature sequences from each image are then summed with fixed positional encodings and per-image segment embeddings with learnable weights, concatenated length-wise, and passed as input sequence to the CoVis transformer. The output embedding sequence from the transformer is then passed as input to two decoders which are trained jointly.

The Co-visibility, angular Correspondence and Floor-wall boundary (CCF) decoder is a single fully connected layer which maps the transformer embedding space to the per-column outputs. In Fig. 2 (b), we illustrate the CCF decoder's output representation for an example image pair. In the predicted co-visibility probability vector, note the large gap in estimated co-visibility in panorama 1's view of the interior of the room, spanning from the right edge of the window to the doorway. This represents the section of floor-wall boundary not being




(a) Architecture of CoVisPose Network (b) Illustration of CCF Decoder's output

Fig. 2: CoVisPose system.(a) The architecture consists of CoVis transformer, the Co-visibility, angular Correspondence and Floor-wall boundary (CCF) decoder, and pose decoder. (b) An example to illustrate the CCF decoder's outputs.

visible to panorama 2. We highlight two particular angular correspondences. The highlighted correspondence from panorama 1 estimates that the center of the window is visible at location $\approx 90°$ in panorama 2, while the correspondence from panorama 2 estimates that the banister is visible at location $\approx 180°$ in panorama 1.

Our pose decoder consists of a simple 6 layer 1D CNN, applied separately to the output transformer embedding sequence from each image. The outputs are then once again concatenated and passed to a fully-connected layer to regress the relative pose. In addition to this direct pose regression, we also demonstrate recovery of relative pose through a RANSAC procedure applied directly to the correspondence network's outputs. Projected into the floor plane, the floor-wall boundary contours form a pair of 2D point sets. For those boundary points with predicted co-visibility, the correspondence angles then suggest corresponding point pairs between the two sets. The pose may subsequently be recovered by rigid registration, which we estimate with a RANSAC iteration.

## 4 Method

In this section, we provide details on the components of CoVisPose, how training is done, and how RANSAC is used for relative pose estimation. We also describe how ZInD [12] is processed for evaluation. We first define the architecture outputs: relative pose, co-visibility, angular correspondence, and floor-wall boundary.

**Relative Pose.** Given an equirectangular image pair, $(I_1, I_2) \in \mathbb{R}^{3\times H\times W}$, we estimate the relative pose $\mathbf{P}_{2,1}$ of panorama $I_2$ w.r.t. the local coordinate system of $I_1$ centered at the origin. Under the upright camera, camera-axis-aligned walls, and orthogonal floor orientation assumptions, the camera pose may be simplified to planar motion with a single rotation angle about the camera axis, i.e., a translation vector $\mathbf{t} \in \mathbb{R}^2$ and a rotation matrix $R \in SO(2)$. Therefore, the pose $\mathbf{P}_{2,1} \in SE(2)$. For direct regression, we represent the pose by five parameters, estimating the unit rotation and translation vectors $\mathbf{r}$ and $\mathbf{t}$, as well as the translation scale $s$. The network is trained to estimate the translation scale normalized by the camera height. This decoupled representation allows both rotation and translation to be framed as directional estimation.

**Co-Visibility, Angular Correspondence, and Floor-Wall Boundary.** Under the wall-floor geometry assumptions, wall geometry may be represented by a 1D contour, the position of which is defined for a given image column $i$ as a vertical angle $\phi_i \in [0, \pi/2]$ as in [45,47]. For relative pose estimation, we extend this column-wise vector representation to two additional quantities: (1) The co-visibility vector, i.e., a binary value $p_{c,i}$ with value 1 if column $i$'s floor-wall boundary is visible to the other panorama and 0 otherwise, and (2) the angular correspondence, i.e., the horizontal angle $\alpha_i \in [-\pi, \pi]$ at which column $i$'s floor-wall boundary is visible in the other panorama's FoV (defined only if $p_{c,i} = 1$). For a given panorama pair, these quantities are defined bidirectionally.

### 4.1 CoVisPose Network Architecture

Our architecture consists of the feature extractor, CoVis transformer, Co-Visibility, Angular Correspondence, Floor-Wall Boundary (CCF) decoder, and pose decoder.

**Feature Extractor.** We leverage the single image feature extractor from [45], a ResNet50 backbone followed by an HCM. For each panorama, this produces a feature sequence over the (downsampled) column-space. For an input image $I_k \in \mathbb{R}^{3\times512\times1024}$, we obtain the features $f_k \in \mathbb{R}^{256\times1024}$.

**CoVis Transformer.** While LSTM and CNN architectures have been applied successfully for per-column prediction of layout [45,47], the local inductive biases of these architectures [56,3] makes them ill-suited for the inherently non-local tasks of co-visibility and correspondence estimation across pairs of 360° panoramas. On the contrary, transformers [51] update embeddings globally and in parallel.

Inspired by [15,51], we add fixed sinusoidal positional encodings and learnable per-image segment embeddings. Both are crucial for the angular correspondence task; they provide the permutation invariant transformer the necessary information to distinguish both relative position of image columns and image membership while attending globally to both intra and inter-image column relationships. We then concatenate the updated column-wise feature sequences length-wise to form the input to the transformer, $F_{in} = (\hat{f}_1, \hat{f}_2) \in \mathbb{R}^{512\times1024}$. Our transformer consists of 6 encoder layers, each with 8 heads of internal self-attention, followed by a feed-forward layer of dimension 2048. The output embeddings are of the same dimensionality as the input, $F_{tr} \in \mathbb{R}^{512\times1024}$. Those serve as input to two decoders, to estimate the per-column outputs, and regress the relative pose.

**Co-Visibility, Angular Correspondence, and Floor-Wall Boundary Decoder.** We apply a single fully connected layer to map the transformer embeddings $F_{tr}$ to the column-wise outputs $F_{out} \in \mathbb{R}^{512\times12}$. Each vector of the output sequence predicts the values for 4 image columns, and thus may be reshaped to three column-wise vectors for each image, $[\boldsymbol{\phi}^k, \boldsymbol{\alpha}^k, \mathbf{p}^k] \in \mathbb{R}^{1024}$, $k \in [1, 2]$.

**Pose Decoder.** To regress the relative pose from the rich column-wise co-visibility, correspondence, and wall depth information embedded in $F_{tr}$, we separately apply a 6-layer 1-D CNN to each image's embedding sequence. Each layer consists of a convolution with kernel size 3, followed by batch normalization

and a ReLU non-linearity. Each conv layer reduces the feature dimension by half. We then once again concatenate the feature sequences, before flattening the sequence and applying a fully-connected layer to map to the five-dimensional pose output representation, $(\mathbf{r}, \mathbf{t}, s)$. The scale parameter is made non-negative by applying a ReLU.

### 4.2 Training

The model is implemented in PyTorch; we train in mixed precision on 4 NVIDIA Tesla V100 for 30 epochs with a learning rate of .0001. We select the best model by the lowest validation loss sum. During training we apply random rotational augmentation, as well as randomly swapping the panorama pair order, inverting the relative pose for regression.

**Loss Functions.** We apply the L1 loss for the angular floor-wall boundary and correspondence outputs, and the binary cross entropy (BCE) loss for the co-visibility probability. Being both angular quantities in radians, the floor-wall boundary and correspondence losses are naturally of similar magnitude. To equilibrate the magnitude of the BCE loss, we apply a scaling parameter $\beta_c$. The total column-wise output loss is

$$L_{covis} = ||\boldsymbol{\phi}^k - \hat{\boldsymbol{\phi}}^k||_1 + ||\boldsymbol{\alpha}^k - \hat{\boldsymbol{\alpha}}^k||_1 + \beta_c \cdot BCE(\mathbf{p}_c^k, \hat{\mathbf{p}}_c^k), k \in [1, 2]. \quad (1)$$

For regressing the relative pose, we normalize the estimated rotation and translation direction to be unit vectors, and multiply the translation direction by the estimated scale, $s$, to produce the final estimated translation $\mathbf{t}_s$. To learn the pose parameters we minimize mean-squared error over both vectors. We similarly scale the magnitude of both the rotation and translation loss functions as we find these losses to have an overall stronger effect than the per-column loss functions above:

$$L_{pose} = \beta_r \cdot ||\mathbf{r} - \hat{\mathbf{r}}||_2^2 + \beta_t \cdot ||\mathbf{t}_s - \hat{\mathbf{t}}_s||_2^2. \quad (2)$$

We did not carefully tune the loss scaling parameters as we did not find the optimization to be particularly sensitive to these values. In our experiments, we used $\beta_c = .25, \beta_r = 3 \times 10^{-3}, \beta_t = 6 \times 10^{-2}$.

**Positive/Negative Sampling.** For constructing the dataset for training and testing, we form "positive" training examples by retaining all panorama pairs from ZInD which have >= 10% co-visibility score. We further sample "negative" examples with zero co-visibility with a probability of .1. These settings result in an overall ratio of positives to negatives of approximately 2.5.

### 4.3 Relative Pose Estimation by RANSAC

In addition to learning direct pose regression, the CCF decoder outputs support alignment by rigid registration. Projected into the floor plane using the assumptions of upright camera and orthogonal floor plane, the floor-wall boundary points from both images form two 2D point sets. For those boundary points whose co-visibility probability is high, the predicted correspondence angle can be used to

sample a point on the neighboring panoramas floor-wall boundary contour. In this way, a set of corresponding points can be determined bidirectionally using the predictions from both images in the pair. Fig. 2 (a) illustrates bidirectional correspondence estimation, while (b) illustrates the resultant connectivity between floor-wall boundary points generated by the estimated correspondence, which can then be used to obtain pose by rigid registration.

It is possible to directly solve for the rotation and translation using singular value decomposition on the corresponding point sets [49]; however, this method is sensitive to noise. To compensate, we instead apply RANSAC. At each iteration, we randomly select two pairs of corresponding points from the set predicted by the model. A candidate rotation and translation is then determined from the point pair by a 2-point minimal solver. A Hungarian algorithm-based assignment to determine inliers can be used; however, given the quality of the correspondences predicted by the model, we find greedy assignment to be faster with minimal performance loss. After the RANSAC loop, given the alignment candidate with the highest number of inliers, we use the inlier assignment to do an SVD refit in order to determine the final rotation and translation.

### 4.4 ZInD Pre-Processing

ZInD's complete geometry allows computation of visual overlap and floor-wall boundary angular correspondence by comparing points on the visible wall layout for panorama pairs that share the same space; however, with no signal on whether the doors are open or closed, these quantities cannot be confidently extended across doorways. To allow extension of our method to whole floors without the burden of fully labeling all doors, a sample of 5K doors were labeled. We then bootstrap off of these annotations by training a classification network consisting of one convolutional and one fully connected classification layer, stacked on top of pretrained mid-level depth, normal, and room layout features from [41]. This classifier was then used to label the remainder of the dataset in a semi-supervised manner. See the supplementary for more details and examples of this step.

## 5 Results

In this section, we describe how ZInD is used for evaluation, the evaluation metrics, and comparisons with a few baselines. We also describe results of our ablation study.

### 5.1 Dataset

We split out dataset with mined positives and negatives into train, test, and validation sets according to the publicly released ZInD split. We evaluate our method on the test set; the statistics on number of examples by visual overlap and baseline distance can be found in the supplementary. Our dataset contains a wide range of co-visibility, with a large fraction of examples subject to extremely low visual overlap; 32% of positive examples have less than 25% co-visibility. Further, our dataset contains extremely wide baselines between panoramas; 36% of positive pairs have a baseline distance of more than 4 meters.

### 5.2 Baselines

We compare our method to multiple baselines for relative pose estimation.

**SIFT+OpenMVG.** We run 2-view SfM from OpenMVG [33] with the three-point upright relative pose solver [35] on the full panorama image pair. It assumes an upright camera and solves for the horizontal rotation and 3 DoF translation. Note that the recovered translation is not up-to-scale with the predicted room layout geometry from CoVisPose.

**LoFTR+OpenMVG.** We split each panorama image into perspective crops with a combination of horizontal angle at [0°, 60°, 120°, 180°, 240°, 300°] and vertical angle [-30°, 0°, 30°]. The crops are projected with a 90° horizontal field-of-view, and a resolution of 640 by 640 pixels. We use the LoFTR [47] feature matcher as trained in the original paper, exhaustively run on combinations of crops from a panorama pair, and project putative feature matches back to spherical space. Then we solve relative camera pose using openMVG with settings similar to the SIFT+OpenMVG baseline.

**LayoutLoc.** We run LayoutLoc [12] on panorama pairs; it applies a semantic-based camera alignment based on predicted room layout and wall features. It solves the panorama relative poses with the same scale as predicted room layout geometry. The success of a localization is determined by thresholding the estimated camera pose confidence score.

**DirectionNet.** DirectionNet [10] is a recent supervised learning approach, achieving SOTA performance on the challenging task of wide-baseline relative pose estimation for perspective, limited FoV cameras in indoor scenes. DirectionNet is a representative of the line of work focused on end-to-end camera pose estimation without explicit feature correspondences. We adapted and re-trained their released code-base[4] on our domain (pairs of 360° spherical cameras) using the same training set and data augmentations as described above. More details on the training protocol are provided in the supplementary.

### 5.3 Evaluation Metrics

For relative pose error, we report the absolute error in the predicted rotation and translation angles for all methods. For those methods which produce scale, we additionally report the translation error in meters. For all of these quantities, we report the mean and median errors in Table 1. For rotation and translation angle errors, we also report % of total samples which have error less than 2.5°, and for translation errors, % of total samples with error less than .5 meters.

As optimization-based models may fail given insufficient input correspondences, or lack of consistency in the set, to better understand this, we report the success rate in %. Further, as many direct regression methods may not come with a measure of confidence in a given pose, to demonstrate the strength of our co-visibility output, we compute true positive/false positive rate curves as function

[4] `https://arthurchen0518.github.io/DirectionNet`

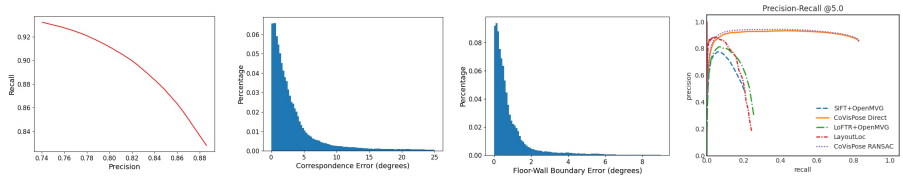
Table 1: Relative pose statistics by co-visibility. We report the mean ("Mn") and median ("Med") angular rotation and translation errors in degrees, as well as the fraction of test set pairs for which the angular error was less than 2.5°. In addition, as our method additionally estimates two-view scale, we report the mean and median translation distance error in meters, as well as the fraction of pairs for which the translation error was less than .5 meters. Highlights: 1st, 2nd and 3rd best results.

| Co-Vis.% | Method | Success (% ↑) | Rotation | | | Translation angle | | | Translation vector | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Mn(°↓) | Med(°↓) | 2.5°(% ↑) | Mn(°↓) | Med(°↓) | 2.5°(% ↑) | Mn(m.↓) | Med(m.↓) | .5m.(% ↑) |
| 75 - 100 | SIFT-OpenMVG [31] | 69.17% | 12.50 | 0.58 | 53.10 | 17.31 | 1.62 | 41.68 | - | - | - |
| | LoFTR-OpenMVG [47] | 89.12% | 15.71 | 0.90 | 60.71 | 19.25 | 2.29 | 46.53 | - | - | - |
| | DirectionNet [10] | **100.00%** | 16.35 | 1.48 | 69.28 | 18.43 | 5.53 | 25.12 | - | - | - |
| | LayoutLoc [12] | 78.69% | 13.13 | **0.00** | 70.19 | 14.86 | 1.46 | 51.12 | 0.64 | 0.11 | 63.96 |
| | CoVisPose-Ransac | 99.73% | **1.20** | 0.53 | 96.51 | **2.86** | **0.91** | **84.09** | **0.10** | **0.07** | 98.51 |
| | CoVisPose-Direct | **100.00%** | 1.27 | 0.56 | **97.87** | 3.38 | 1.15 | 78.50 | 0.12 | 0.09 | **98.87** |
| 50 - 75 | SIFT-OpenMVG [31] | 47.88% | 22.01 | 0.83 | 31.79 | 25.01 | 2.22 | 25.23 | - | - | - |
| | LoFTR-OpenMVG [47] | 71.36% | 24.54 | 1.84 | 38.25 | 26.53 | 4.13 | 27.93 | - | - | - |
| | DirectionNet [10] | **100.00%** | 27.44 | 2.02 | 57.52 | 24.06 | 6.46 | 22.18 | - | - | - |
| | LayoutLoc [12] | 60.84% | 41.64 | **0.00** | 40.17 | 38.57 | 4.26 | 26.18 | 1.86 | 0.53 | 30.13 |
| | CoVisPose-Ransac | 99.22% | **1.45** | 0.67 | 92.36 | **1.92** | **0.89** | **83.46** | **0.16** | **0.08** | 94.93 |
| | CoVisPose-Direct | **100.00%** | 1.48 | 0.73 | **94.71** | 2.13 | 1.09 | 81.55 | **0.16** | 0.10 | **96.93** |
| 25 - 50 | SIFT-OpenMVG [31] | 26.58% | 41.51 | 7.94 | 11.05 | 45.18 | 14.68 | 8.13 | - | - | - |
| | LoFTR-OpenMVG [47] | 52.26% | 40.58 | 12.86 | 16.36 | 43.07 | 18.25 | 12.18 | - | - | - |
| | DirectionNet [10] | **100.00%** | 38.38 | 2.75 | 47.58 | 29.88 | 9.04 | 17.56 | - | - | - |
| | LayoutLoc [12] | 49.85% | 77.39 | 90.00 | 18.57 | 63.40 | 50.52 | 8.27 | 3.56 | 3.15 | 8.94 |
| | CoVisPose-Ransac | 96.42% | **2.51** | **0.98** | 80.02 | **2.19** | **1.00** | **77.49** | **0.24** | **0.12** | 88.06 |
| | CoVisPose-Direct | **100.00%** | 3.47 | 1.03 | **83.89** | 3.00 | 1.24 | 75.57 | 0.28 | 0.14 | **92.45** |
| 10 - 25 | SIFT-OpenMVG [31] | 16.55% | 68.70 | 61.78 | 2.46 | 72.02 | 64.10 | 1.63 | - | - | - |
| | LoFTR-OpenMVG [47] | 39.76% | 59.32 | 37.75 | 5.16 | 63.71 | 50.83 | 3.95 | - | - | - |
| | DirectionNet [10] | **100.00%** | 53.37 | 5.97 | 37.85 | 42.47 | 18.94 | 9.53 | - | - | - |
| | LayoutLoc [12] | 46.59% | 91.30 | 90.00 | 11.85 | 77.21 | 70.11 | 2.19 | 5.04 | 4.71 | 1.64 |
| | CoVisPose-Ransac | 88.46% | **6.18** | **1.78** | 54.36 | **4.82** | **1.59** | **57.66** | **0.56** | **0.22** | 67.64 |
| | CoVisPose-Direct | **100.00%** | 8.90 | 2.10 | **55.87** | 6.79 | 2.13 | 56.15 | 0.72 | 0.27 | **73.56** |

of a threshold applied to a method-specific measure of confidence. For the methods which solve for the essential matrix given correspondences, we use the number of inliers of the fit. For CoVisPose, we use a threshold on the estimated co-visibility score [12] [55] between the pair. For LayoutLoc we use the pose confidence score.

### 5.4 Relative Pose Estimation Accuracy

We report error metrics stratified by the GT co-visibility score in Table 1. First, we observe that our CoVisPose achieves the best performance in almost all cases across the range of visual overlap, for both rotation and translation errors. Similar to [9,7], our empirical studies suggest that feature-based approaches, e.g., classic methods like SIFT [31] and learned ones like LoFTR [47], are competitive in the high-overlap regime, where point features can be matched robustly. However, those become less reliable as the visual overlap decreases. LayoutLoc [12] can be seen as a learned, semantic feature-based approach, which outperforms the point-based base-lines (SIFT and LoFTR) in this modality. Similar to them, its performance quickly drops for wide to extreme baselines. Note that LayoutLoc achieves 0° median rotation error for high visual overlap cases. This is due to rooms in ZInD being aligned by the computed vanishing angle which LayoutLoc aligns with as a final step. This results in zero rotation error when the geometric




(a) (b) (c) (d)

Fig. 3: (a) Precision and recall of per-column co-visibility. (b) Histogram of angular correspondence error, in degrees. (c) Histogram of floor-wall boundary error, in degrees. (d) Relative pose precision and recall curves for CoVisPose and baseline methods. True positives defined by maximum of rotation and translation errors less than 5°.

alignment is successful. Similar to the trends reported in Chen et al. [9], DirectionNet (that we trained end-to-end) performs substantially better than the feature-based approaches in the mid-to-low overlap regimes.

CoVisPose RANSAC shows a small advantage over CoVisPose Direct; when accurate correspondences are available, a robust iterative fitting is often capable of finding a more accurate pose. However, CoVisPose Direct inference only requires 50 ms, whereas RANSAC iteration and inlier assignment typically requires 5-30 secs to fully exploit the dense correspondences. Moreover, direct regression, like DirectionNet, returns a valid pose 100% of the time. Under very low visual overlap this may be an advantage as a fitting algorithm may have few available correspondences as input. This 100% success rate is potentially both a positive and negative; without a reliable measure of confidence it is difficult to know whether or not a directly regressed pose should be trusted. CoVisPose's co-visibility vector provides such a measure, allowing the method to be run in an unordered set of panoramas, without a separate retrieval module.

**Precision and Recall.** We demonstrate the estimated co-visibility as a measure of pose confidence by computing precision and recall over the entire test set, including negative examples with no visual overlap, in Fig. 3(d). CoVisPose demonstrates strong precision and recall, with similarly high accuracy for both direct regression and RANSAC poses, correctly rejecting negative examples through accurate estimated co-visibility (see Fig.3(a)(b)(c) for an in-depth analysis). For computing the curves, true positives are defined as poses with the maximum of rotation and translation angle errors less than 5°. Returning a confident pose for a pair with zero visual overlap is considered a false positive for this analysis.

### 5.5 Per-column prediction accuracy analysis

We report per-column prediction errors for co-visibility, angular correspondence and floor-wall boundary estimation over all image columns from test images in Fig. 3(a)(b)(c). We see that CoVisPose produces less than 5° of error in angular correspondence for more than 73.1% of all image columns and less than 2.5° error in the floor-wall boundary position for more than 83.1% of all columns.


Fig. 4: **Qualitative evaluation on ZInD.** (a) SIFT feature point inliers generated by 2-view SfM with OpenMVG, (b) LoFTR feature point inliers generated by the same solver as (a), (c) CoViSPose prediction, (d) predicted floor-wall boundaries for **Panorama 1** and **Panorama 2** aligned by CoViSPose direct regression. (e) Pose errors for each method. Color scheme corresponds to maximum of rotation and translation errors: Failure to recover pose or error $> 10°$, error in $[2.5°, 10°]$, error $< 2.5°$.

Table 2: Ablation study on direct pose regression performance by removing dense column-wise outputs.

| Method | Rotation | | | Translation angle | | | Translation vector | | |
|---|---|---|---|---|---|---|---|---|---|
| | Mn(°↓) | Med(°↓) | 2.5°(% ↑) | Mn(°↓) | Med(°↓) | 2.5°(% ↑) | Mn(m.↓) | Med(m.↓) | .5m.(% ↑) |
| CoVisPose Boundary | 16.35 | 6.07 | 24.84 | 14.15 | 5.76 | 26.02 | 1.01 | .48 | 51.89 |
| CoVisPose Boundary+CoVis | 4.94 | 1.24 | 74.34 | 4.83 | 1.57 | 66.45 | .41 | .16 | 86.13 |
| CoVisPose Boundary+CoVis+AC | **4.33** | **.97** | **80.17** | **4.14** | **1.39** | **71.13** | **.36** | **.14** | **88.72** |

### 5.6 Qualitative Results

Fig. 4 shows results from CoVisPose and baseline methods on ZInD panoramas. We arrange result rows by ground truth (GT) co-visibility score, in decreasing order. Common sources of baseline failure include lack of texture for feature matching, wide-baselines, and low visual overlap. (Failure examples are in the supplemental.) On the contrary, we see that CoVisPose produces accurate poses for the vast majority of cases, including for panorama pairs with wide to extreme baselines. In column (c), we further see the high spatial precision in geometry estimation and alignment; in most cases the predicted room contours align well. In row 4 we show a case where the competing direct regression method, DirectionNet, shows better performance. We further examine limitations and failure cases in the supplementary.

### 5.7 Ablation Study

We demonstrate the impact of jointly training our pose decoder alongside dense column-wise outputs by training three CoVisPose variants. We found it difficult to tune the model to converge without *any* column-wise outputs. When compared with just the boundary output, we find that the performance increases dramatically when adding the co-visibility mask (CoVis), and increases further with addition of the angular correspondence (AC). While the boundary information offers a strong geometry prior, we hypothesize that co-visibility increases performance by providing a strong signal on inter-view association, which the angular correspondence further refines. We also note that the gap between the training and validation errors is markedly reduced as column-wise outputs are added, indicating increased generalization. Details can be found in the supple-mentary.

## 6 Conclusion

We present a novel end-to-end learning approach for relative pose estimation in wide baseline 360° indoor panoramas. We have shown how jointly learning dense column-wise representations of visual overlap, correspondence, and layout geometry increases the feasibility and accuracy of direct pose regression. This representation yields accurate poses through a RANSAC-based approach applied to the densely predicted corresponding boundary points. Further, our co-visibility vector provides a strong measure of pose confidence. We set a new SOTA for this task, improving upon existing methods, including classical SfM, deep feature matching-based SfM, and direct pose regression.

## References


1. 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. IEEE (2012)
2. Balntas, V., Li, S., Prisacariu, V.: Relocnet: Continuous metric learning relocalisation using neural nets. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
3. Bello, I., Zoph, B., Vaswani, A., Shlens, J., Le, Q.V.: Attention augmented convolutional networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
4. Brachmann, E., Krull, A., Nowozin, S., Shotton, J., Michel, F., Gumhold, S., Rother, C.: Dsac — differentiable ransac for camera localization. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2492–2500 (2017). https://doi.org/10.1109/CVPR.2017.267
5. Brahmbhatt, S., Gu, J., Kim, K., Hays, J., Kautz, J.: Mapnet: Geometry-aware learning of maps for camera localization. CoRR **abs/1712.03342** (2017), `http://arxiv.org/abs/1712.03342`
6. Cabral, R., Furukawa, Y.: Piecewise planar and compact floorplan reconstruction from images. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 628–635. IEEE (2014)
7. Cai, R., Hariharan, B., Snavely, N., Averbuch-Elor, H.: Extreme rotation estimation using dense correlation volumes. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
8. 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. arXiv preprint arXiv:1709.06158 (2017)
9. Chen, K., Snavely, N., Makadia, A.: Wide-baseline relative camera pose estimation with directional learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3258–3268 (June 2021)
10. Chen, K., Snavely, N., Makadia, A.: Wide-baseline relative camera pose estimation with directional learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3258–3268 (2021)
11. Choi, S., Kim, J.H.: Fast and reliable minimal relative pose estimation under planar motion. Image and Vision Computing **69**, 103–112 (2018)
12. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2133–2143 (June 2021)
13. Davidson, B., Alvi, M.S., Henriques, J.F.: 360 camera alignment via segmentation. In: European Conference on Computer Vision. pp. 579–595. Springer (2020)
14. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. CoRR **abs/1712.07629** (2017), `http://arxiv.org/abs/1712.07629`
15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, `https://aclanthology.org/N19-1423`

16. Ding, Y., Barath, D., Kukelova, Z.: Homography-based egomotion estimation using gravity and sift features. In: Proceedings of the Asian Conference on Computer Vision (2020)
17. Ding, Y., Barath, D., Yang, J., Kong, H., Kukelova, Z.: Globally optimal relative pose estimation with gravity prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 394–403 (2021)
18. Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T.: D2-net: A trainable cnn for joint description and detection of local features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
19. Eder, M., Shvets, M., Lim, J., Frahm, J.M.: Tangent images for mitigating spherical distortion. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
20. En, S., Lechervy, A., Jurie, F.: Rpnet: an end-to-end network for relative camera pose estimation. In: ECCV Workshops (2018)
21. Guan, B., Zhao, J., Li, Z., Sun, F., Fraundorfer, F.: Minimal solutions for relative pose with a single affine correspondence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
22. Herath, S., Yan, H., Furukawa, Y.: Ronin: Robust neural inertial navigation in the wild: Benchmark, evaluations, & new methods. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 3146–3152. IEEE (2020)
23. Jung, J., Lee, J.Y., Kim, B., Lee, S.: Upright adjustment of 360 spherical panoramas. In: 2017 IEEE Virtual Reality (VR). pp. 251–252. IEEE (2017)
24. Jung, R., Lee, A.S.J., Ashtari, A., Bazin, J.C.: Deep360up: A deep learning-based approach for automatic vr image upright adjustment. In: 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). pp. 1–8. IEEE (2019)
25. Kendall, A., Cipolla, R.: Geometric loss functions for camera pose regression with deep learning. CoRR **abs/1704.00390** (2017), http://arxiv.org/abs/1704.00390
26. Kendall, A., Grimes, M.K., Cipolla, R.: Posenet: A convolutional network for real-time 6-dof camera relocalization. 2015 IEEE International Conference on Computer Vision (ICCV) pp. 2938–2946 (2015)
27. Kukelova, Z., Bujnak, M., Pajdla, T.: Closed-form solutions to minimal absolute pose problems with known vertical direction. In: Asian Conference on Computer Vision. pp. 216–229. Springer (2010)
28. Laskar, Z., Melekhov, I., Kalia, S., Kannala, J.: Camera relocalization by computing pairwise relative poses using convolutional neural network. CoRR **abs/1707.09733** (2017), http://arxiv.org/abs/1707.09733
29. Lee, Y., Jeong, J., Yun, J., Cho, W., Yoon, K.J.: Spherephd: Applying cnns on a spherical polyhedron representation of 360deg images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
30. Li, J., Su, J., Xia, C., Tian, Y.: Distortion-adaptive salient object detection in 360<inline-formula><tex-math notation="latex">$^\circ$</tex-math></inline-formula> omnidirectional images. IEEE Journal of Selected Topics in Signal Processing **14**(1), 38–48 (2020). https://doi.org/10.1109/JSTSP.2019.2957982
31. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2), 91–110 (2004)
32. Melekhov, I., Kannala, J., Rahtu, E.: Relative camera pose estimation using convolutional neural networks. CoRR **abs/1702.01381** (2017), http://arxiv.org/abs/1702.01381

33. Moulon, P., Monasse, P., Perrot, R., Marlet, R.: Openmvg: Open multiple view geometry. In: International Workshop on Reproducible Research in Pattern Recognition. pp. 60–74. Springer (2016)
34. Ortin, D., Montiel, J.M.M.: Indoor robot motion based on monocular images. Robotica **19**(3), 331–342 (2001)
35. Oskarsson, M.: Two-view orthographic epipolar geometry: Minimal and optimal solvers. Journal of Mathematical Imaging and Vision **60**(2), 163–173 (2018)
36. Pintore, G., Agus, M., Gobbetti, E.: AtlantaNet: Inferring the 3D indoor layout from a single 360 image beyond the Manhattan world assumption. In: Proc. ECCV (August 2020), http://vic.crs4.it/vic/cgi-bin/bib-page.cgi?id='Pintore:2020:AI3'
37. Pintore, G., Ganovelli, F., Pintus, R., Scopigno, R., Gobbetti, E.: 3d floor plan recovery from overlapping spherical images. Computational visual media **4**(4), 367–383 (2018)
38. Ramakrishnan, S.K., Gokaslan, A., Wijmans, E., Maksymets, O., Clegg, A., Turner, J.M., Undersander, E., Galuba, W., Westbury, A., Chang, A.X., Savva, M., Zhao, Y., Batra, D.: Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021), https://openreview.net/forum?id=-v4OuqNs5P
39. Ranftl, R., Koltun, V.: Deep fundamental matrix estimation. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
40. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
41. Sax, A., Emi, B., Zamir, A.R., Guibas, L.J., Savarese, S., Malik, J.: Mid-level visual representations improve generalization and sample efficiency for learning visuomotor policies. (2018)
42. Schindler, G., Dellaert, F.: Atlanta world: An expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex man-made environments. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004. vol. 1, pp. I–I. IEEE (2004)
43. Shabani, M.A., Song, W., Odamaki, M., Fujiki, H., Furukawa, Y.: Extreme structure from motion for indoor panoramas without visual overlaps. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5703–5711 (October 2021)
44. Su, Y.C., Grauman, K.: Learning spherical convolution for fast features from 360°imagery. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper/2017/file/0c74b7f78409a4022a2c4c5a5ca3ee19-Paper.pdf
45. Sun, C., Hsiao, C.W., Sun, M., Chen, H.T.: Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
46. Sun, C., Sun, M., Chen, H.T.: Hohonet: 360 indoor holistic understanding with latent horizontal features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2573–2582 (June 2021)

47. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loftr: Detector-free local feature matching with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8922–8931 (June 2021)
48. Torii, A., Imiya, A., Ohnishi, N.: Two-and three-view geometry for spherical cameras. In: Proceedings of the sixth workshop on omnidirectional vision, camera networks and non-classical cameras. pp. 81–88. Citeseer (2005)
49. Umeyama, S.: Least-squares estimation of transformation parameters between two point patterns. IEEE Transactions on Pattern Analysis & Machine Intelligence **13**(04), 376–380 (apr 1991). https://doi.org/10.1109/34.88573
50. Ummenhofer, B., Zhou, H., Uhrig, J., Mayer, N., Ilg, E., Dosovitskiy, A., Brox, T.: Demon: Depth and motion network for learning monocular stereo. In: CVPR. pp. 5622–5631. IEEE Computer Society (2017), http://dblp.uni-trier.de/db/conf/cvpr/cvpr2017.html#UmmenhoferZUMID17
51. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
52. Walch, F., Hazirbas, C., Leal-Taixé, L., Sattler, T., Hilsenbeck, S., Cremers, D.: Image-based localization using lstms for structured feature correlation. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 627–637 (2017)
53. Wang, F.E., Yeh, Y.H., Sun, M., Chiu, W.C., Tsai, Y.H.: Bifuse: Monocular 360 depth estimation via bi-projection fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
54. Wang, F.E., Yeh, Y.H., Sun, M., Chiu, W.C., Tsai, Y.H.: Led2-net: Monocular 360deg layout estimation via differentiable depth rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12956–12965 (June 2021)
55. Wang, H., Hutchcroft, W., Li, Y., Wan, Z., Boyadzhiev, I., Kang, S.B.: Psmnet: Position-aware stereo merging network for room layout estimation (in press). In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
56. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7794–7803 (2018). https://doi.org/10.1109/CVPR.2018.00813
57. Yi, K.M., Trulls, E., Ono, Y., Lepetit, V., Salzmann, M., Fua, P.: Learning to find good correspondences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
58. Zeng, W., Karaoglu, S., Gevers, T.: Joint 3d layout and depth prediction from a single indoor panorama image. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 666–682. Springer International Publishing, Cham (2020)
59. Zhang, C., Liwicki, S., Smith, W., Cipolla, R.: Orientation-aware semantic segmentation on icosahedron spheres. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
60. Zhang, Y., Song, S., Tan, P., Xiao, J.: Panocontext: A whole-room 3d context model for panoramic scene understanding. In: European conference on computer vision. pp. 668–686. Springer (2014)

61. Zhang, Z., Deriche, R., Faugeras, O., Luong, Q.T.: A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. Artificial intelligence **78**(1-2), 87–119 (1995)
62. Özyeşil, O., Voroninski, V., Basri, R., Singer, A.: A survey of structure from motion. Acta Numerica **26**, 305–364 (2017). https://doi.org/10.1017/S096249291700006X