Supplementary: 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

1 CoVisPose: Terminology

To clarify terminology, the *co-visibility score* (%), which measures visual overlap ($\in [0, 100\%]$) between two panoramas, is derived from the column-wise *co-visibility vectors* by taking the mean over the image columns for each panorama and then averaging over the pair. We stratify our results as a function of the ground truth (GT) co-visibility score [2] [5] to understand robustness as a function of visual overlap. We estimate the column-wise co-visibility vector to improve direct pose regression accuracy and to filter high quality correspondences. We compute the estimated co-visibility score as a measure of pose confidence (applied for both CoVisPose-Direct and CoVisPose-RANSAC), with the reasonable assumption that higher visual overlap leads to more accurate poses on average. This also serves to exclude poses between panoramas that do not share visual overlap. As direct pose regression methods typically return an estimated pose regardless of input, in absence of a measure like the co-visibility score, they would need to rely on a separate retrieval model to determine likelihood of pose validity.

2 ZInD Preprocessing

Extension of our method to predict layout, co-visibility, correspondence and pose through doorways, requires information on whether or not doors are open or closed, which is not contained in ZInD. For these cases, we illustrate our pipeline for data creation in Fig. 1. We first (1.) extract and label door crops for a subset of the dataset (5K examples). We then (2.) train a classifier as explained in the main paper. For data creation, we then (3.) deploy the trained classifier to label doors. When an open door is encountered, we (4.) collect the adjacent layouts. We represent the open door between the two layouts by carving the doorway region of the two layout polygons and bridging the gap with two line segments. From this combined representation, we then compute a visibility map for each pano, which then support computation of the co-visibility map between the pair. (5.) The visibility and co-visibility maps are then projected to image space. We compute the co-visibility mask by binning the co-visibility map angularly along the horizontal FoV. For co-visible regions, we additionally check

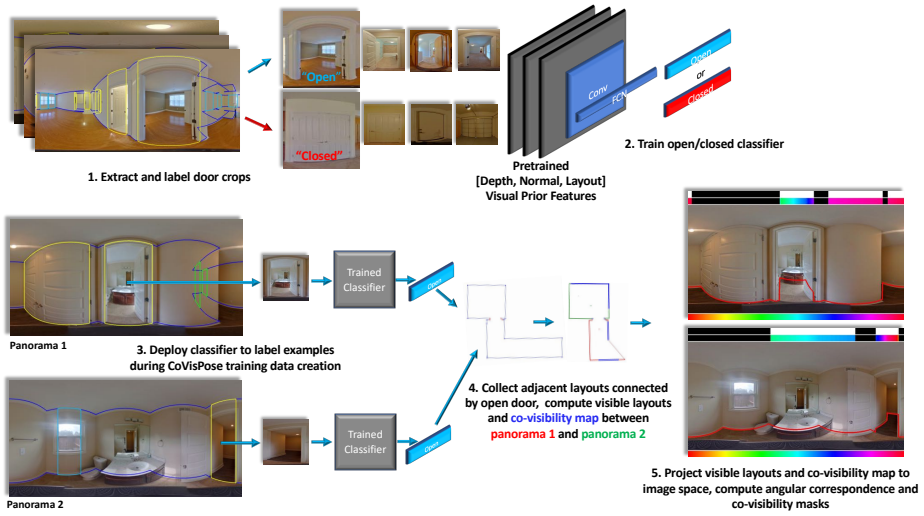


Fig. 1: CoVisPose data creation through open doors.

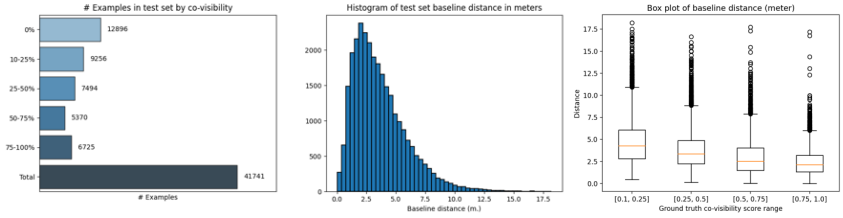
the angular line-of-sight of the co-visible points in each image to determine the correspondence.

Note that in the example shown, the door is open in both panoramas on either side, but this need not be the case. It is also possible to compute these quantities for examples in which the door is only open on one side, in one panorama. In these cases, the co-visible region is entirely on one side of the door; however, our model is capable of estimating reliable poses nonetheless, albeit with less usable signal.

On a validation set set aside from the original human-annotated door crops, the accuracy of the classifier was $\approx 95\%$. This means that there are undoubtedly mistakes made during data creation; however, the doors which matter most, those with a clear view between spaces that are closer to the pano's foreground, are also the easiest to label correctly, which diminishes the impact of this semi-supervised labeling method on the CoVisPose training data. Another limitation associated with doorways is the lack of information about doors that are partially open, or open doors which generally block the panorama's view. The latter is a general limitation of ZInD, as noted in the original paper.

3 Test Set Statistics

To demonstrate the difficulty and scope of our dataset, we examine the test set statistics for co-visibility and baseline distance in Fig. 2 (a) and (b). We see that our dataset contains a large amount of examples with low visual overlap, and is biased towards lower visual overlap, as opposed to higher. Further we see that our dataset has a large fraction of examples with baseline distances of 3 or more meters. Both of these properties ensure that our dataset serves to demonstrate the competency of our method under these challenging conditions. In Fig. 2 (c),



(a) **(b)** **(c)**

Fig. 2: (a) Test set statistics by GT co-visibility score. (b) Test set baseline distance for panorama pairs with $> 10\%$ GT co-visibility score, in meters. The mean and median are 3.6 and 3.1 meters, respectively. (c) Box plots of baseline distance (meters) at different levels of co-visibility.

Table 1: Relative difference between training and validation mean rotation and translation errors, at epoch 30.

Method	Gap Between Training and Validation Mean Error	
	Rotation(°↓)	Translation (m.↓)
CoVisPose Boundary	12.03	0.59
CoVisPose Boundary+CoVis	3.01	0.22
CoVisPose Boundary+CoVis+AC	2.61	0.18

we additionally share boxplots of baseline distance by co-visibility band. While these metrics are highly correlated, we note that even the high co-visibility band contains examples with extreme baseline distance.

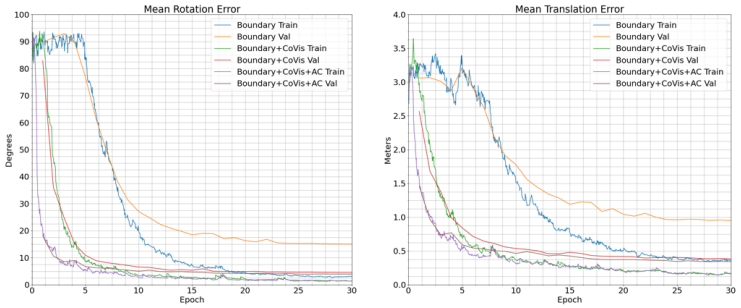
4 Joint Training of Pose and CCF Decoders Improves Pose Estimation Generalization

As shown in the training and validation error curves in Fig. 3, the addition of the dense column-wise outputs increases pose decoder generalization, as well as rate of convergence. With only the column-wise floor-wall boundary output, the gap between the training and validation curves for mean rotation and translation error is considerable. For the training curves, we reduce noise by smoothing with an exponential moving average. We share this gap, the relative difference between the training and validation error for mean rotation and translation, at epoch 30 in Table 1. Upon addition of the column-wise co-visibility output, this gap decreases substantially, in addition to decreases in the errors themselves (as seen in the ablation study). Additionally adding the angular correspondence output (AC) further decreases the train-val gap as well as the absolute errors.

5 DirectionNet Training Details

We adapted their released code[3] base for our domain, a *pair of upright 360 panoramas* with *small to extreme* baselines. We trained the best performing

[3] https://github.com/google-research/google-research/tree/master/direction_net



(a) **(b)**

Fig. 3: Training and validation curves for mean rotation and translation error demonstrate the increased generalization in direct pose regression by jointly training the CCF and pose decoders. (a) Mean rotation error over epochs, in degrees. (b) Mean translation error over epochs, in meters.

configuration reported in their paper [1], which is the 9D rotation (with SVD orthogonalization) followed by derotation (that we adapted for the 360 domain) before the 3D translation network is trained to infer up-to-scale translation direction. The training process (described below) took around 1 week on two Quadro RTX 6000 GPUs with 24GB each.

Rotation 9D: We trained the 9D rotation network on the same training set as our CoVisPose using the same data augmentation (uniform yaw rotation) as discussed in the main paper. Assuming an upright pair of cameras, the 3×3 rotation matrices, fed into that module, can be represented as a yaw-only 3D rotations around the y-axes, with an underlying rank of 1. We observed that this bias is quickly and successfully learned by their underling over-parameterized 9D representation. We used the same hyper-parameters as proposed in their paper with panoramas resized to 256×256 and a batch size of 20. We trained the rotation network for 1M steps, which resulted in around 80 epochs, i.e. passes of the whole training set of 248725 (positive) pairs.

Derotation: In their paper, they propose to train the rotation and translation networks separately, so we followed this two-stage training regime from their code-base. While this step can result in "empty" pixels in their original use-case (wide-baseline perspective, limited FoV, cameras), it is actually very well suited for our domain of 360 cameras, where the full 360 FoV allows for complete and lossless de-rotation of the target image to match the orientation of the source image.

Translation 3D: We first trained the translation network for around 40 epochs using GT rotations with noise. This allowed us to train the rotation and translation networks in parallel. We then fixed the rotation network and fine-tuned the

Table 2: Relative pose statistics for 1274 kitchen and bathroom pairs. We report statistics with highlighted performance ranking, as in Table 1 of the main paper.

Method	Success (% ↑)	Rotation			Translation angle			Translation vector		
		Mn(°↓)	Med(°↓)	2.5°(% ↑)	Mn(°↓)	Med(°↓)	2.5°(% ↑)	Mn(m.↓)	Med(m.↓)	.5m.(% ↑)
SIFT-OpenMVG [3]	65.93%	11.18	0.49	52.98	14.11	1.31	44.27	-	-	-
LoFTR-OpenMVG [4]	82.26%	13.62	0.75	58.01	16.25	1.98	46.31	-	-	-
DirectionNet [1]	**100.00%**	30.05	2.24	52.83	23.69	7.39	20.57	-	-	-
LayoutLoc [2]	67.03%	30.46	**0.00**	50.39	29.09	2.70	32.18	1.11	0.15	43.80
CoVisPose-Ransac	99.76%	1.76	0.72	90.19	3.08	**1.12**	**77.94**	**0.11**	**0.07**	97.96
CoVisPose-Direct	**100.00%**	**1.54**	0.69	**93.88**	**2.97**	1.38	73.39	0.12	0.09	**98.59**

translation network with predicted rotation (with noise) for around 40 more epochs. Similar to the discussion in their supplementary materials, we found the rotation noise (both with GT as well as predicted rotations) to be *the key* for generalization of the translation network. We used the same hyper-parameters as in their paper.

6 Additional Results and Analysis

We evaluate our method on ZInD, which, to our knowledge, is the only large-scale public dataset consisting of floor plans with full layout annotations containing multiple panoramas per space. While ZInD homes are typically unstaged, kitchens and bathrooms allow us to demonstrate robustness to clutter and occlusion through fixtures and cabinetry in section 6.1. On the contrary, though empty rooms to some extent simplify the recovery of layout geometry, the lack of interior features present highly repetitive textures and indistinct regions, which increases the difficulty of co-visibility and correspondence estimation. To demonstrate this challenge, we examine difficult negative examples in section 6.2, and further explore our model's false positives in section 6.3.

6.1 Robustness to Occlusion: Kitchens and Bathrooms

To demonstrate our model's robustness to occlusion, we collect all panorama pairs from the test set which have bathroom or kitchen labels. With this set we evaluate our method against the baselines in Table 2. Additionally, we share qualitative examples from this set in Fig. 4. Despite the significant increase in objects which occlude the floor-wall boundary, and thus challenge geometry estimation, our model performance is high, with accuracy numbers similar to those seen in the general data distribution.

6.2 Difficult Negative Examples

In many cases, differentiating between two highly similar rooms (negative examples) requires learning global information in order to correctly predict zero visual overlap. Rooms with similar windows, fixtures, closet doors, ceiling fans, etc. present stiff challenges. Especially for feature matching-based approaches, these cases may result in matches that are consistent enough to estimate an erroneous

essential matrix with high inliers. Here, we define false positives (FP) as those negative examples for which a pose was successfully estimated for the feature-matching based approaches, and for which CoVisPose predicts a co-visibility score of >= 10%. With this definition, out of the 12896 negative examples in the test set, we compute an overall FP rate of 9.36%, 52.66%, and 2.26%, for SIFT-OpenMVG, LoFTR-OpenMVG, and CoVisPose, respectively. We note that the LoFTR-based baseline results in more cases of estimated poses with a low number of inliers, compared to SIFT. In practice, this FP rate can be reduced by setting a higher threshold on the number of inliers in order to increase precision; however, this comes at the cost of recall as seen in our precision/recall curve. By comparison, CoVisPose is robust to false positives down to very low visual overlap.

We demonstrate this challenge, as well as our high performance on these cases, in Fig. 5. Both CoVisPose-Direct and CoVisPose-RANSAC use the same estimated co-visibility score and thus will have the same performance for negative examples. In particular, see row 2, where the strong similarities between the two bathrooms' vanities and fixtures prompt many erroneous matches which in turn support estimation of a pose with many inliers. In row three we share a difficult example of two bedrooms with nearly identical windows, closets and doors, carpet, and paint. For both of these cases, CoVisPose correctly predicts zero co-visibility, successfully differentiating the spaces.

6.3 Limitations and Failure Cases

As shown in section 6.2, ZInD contains many challenging examples with extremely similar looking rooms that give even a human observer pause. We share a selection of false positive examples for which our model struggled to correctly estimate zero visual overlap. Additionally, we share false negative examples, where our model failed to estimate a pose at a co-visibility threshold of >= 10%. See Fig. 6 for both cases. For false positives, note the first and second example pairs in the second row. In the first result, the bathrooms are virtually identical other then the differing lighting fixtures above the vanities. In the second result, the two garages are also nearly identical, and the model fails to pick up on the important differences in the rear walls, with one garage containing only partial storage and a rear door. On false negatives, causes include extreme wide baselines coupled with low visual overlap, unaccounted for occlusions such as open doors which block the panorama's view, and even exposure issues such as that seen on the left side of the last row.

6.4 Additional Qualitative Results

We present additional examples sorted by baseline distance in Fig 7, 8. For shorter baselines, the feature matching-based methods have a higher chance of estimating an accurate pose; however, we find that under large baseline distances, these methods commonly fail. Conversely, CoVisPose demonstrates robust performance over the entire range.

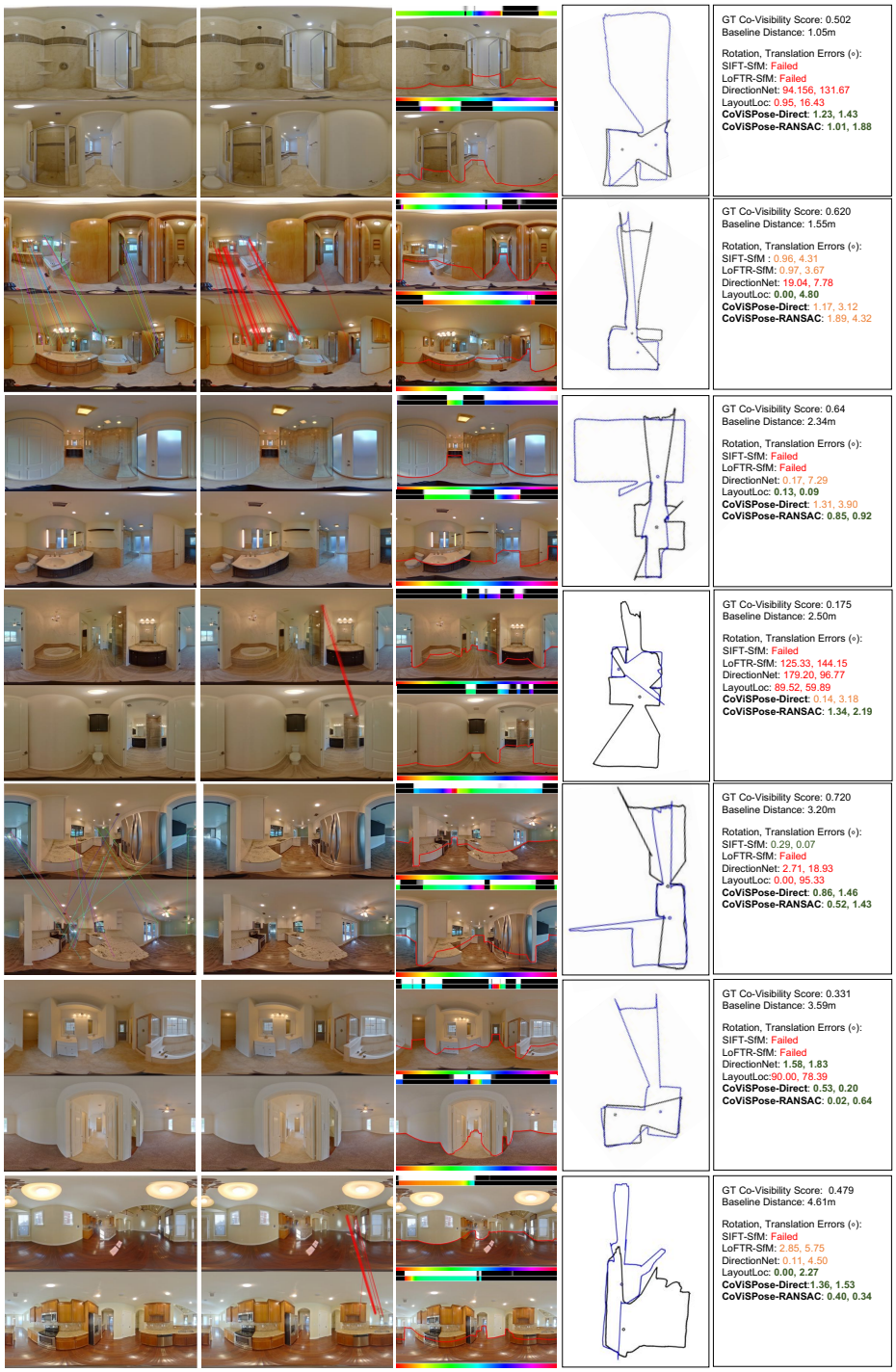


Fig. 4: Qualitative results from panorama pairs captured in kitchens and bathrooms.

Fig. 5: We show examples of the difficult negative cases contained in our dataset; rooms which look highly similar, but that CoVisPose has learned to differentiate by correctly predicting zero co-visibility.

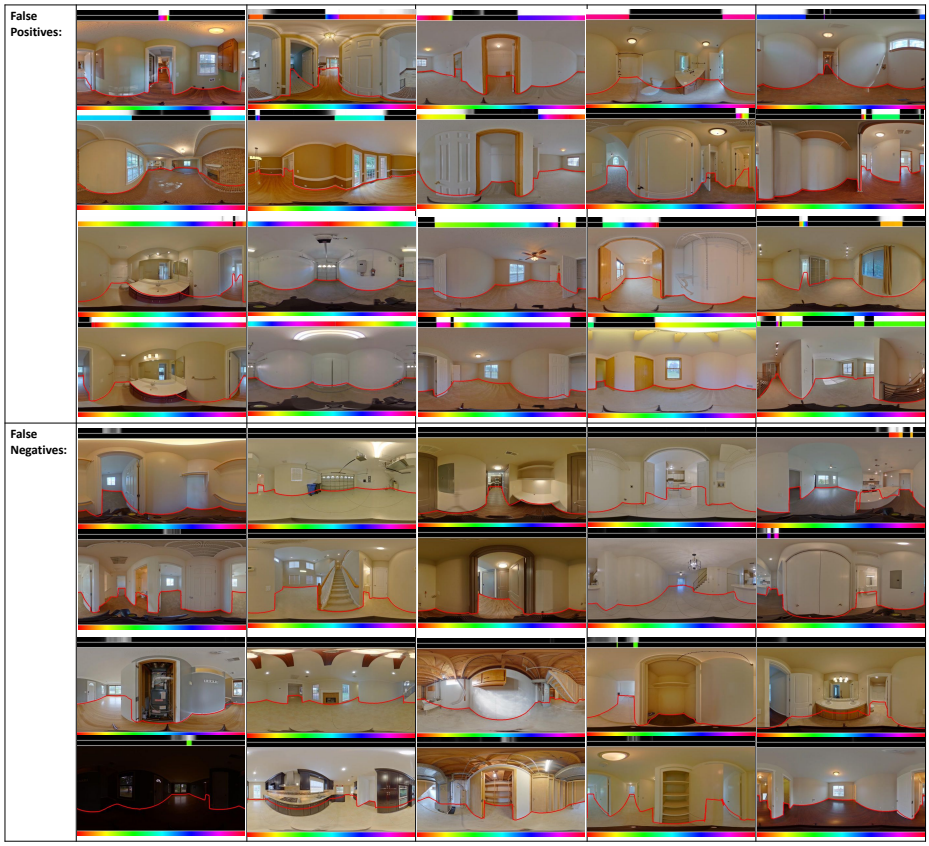


Fig. 6: Example failure cases from CoVisPose direct regression and CoVisPose RANSAC. Here, false positives are negative panorama pairs with an estimated co-visibility score of >= 10%. False negatives are positive panorama pairs with an estimated co-visibility score of < 10%.

Fig. 7: Qualitative results arranged similar to Fig. 4 of the main paper. The examples are sorted by baseline distance, with baselines less than 3 meters.

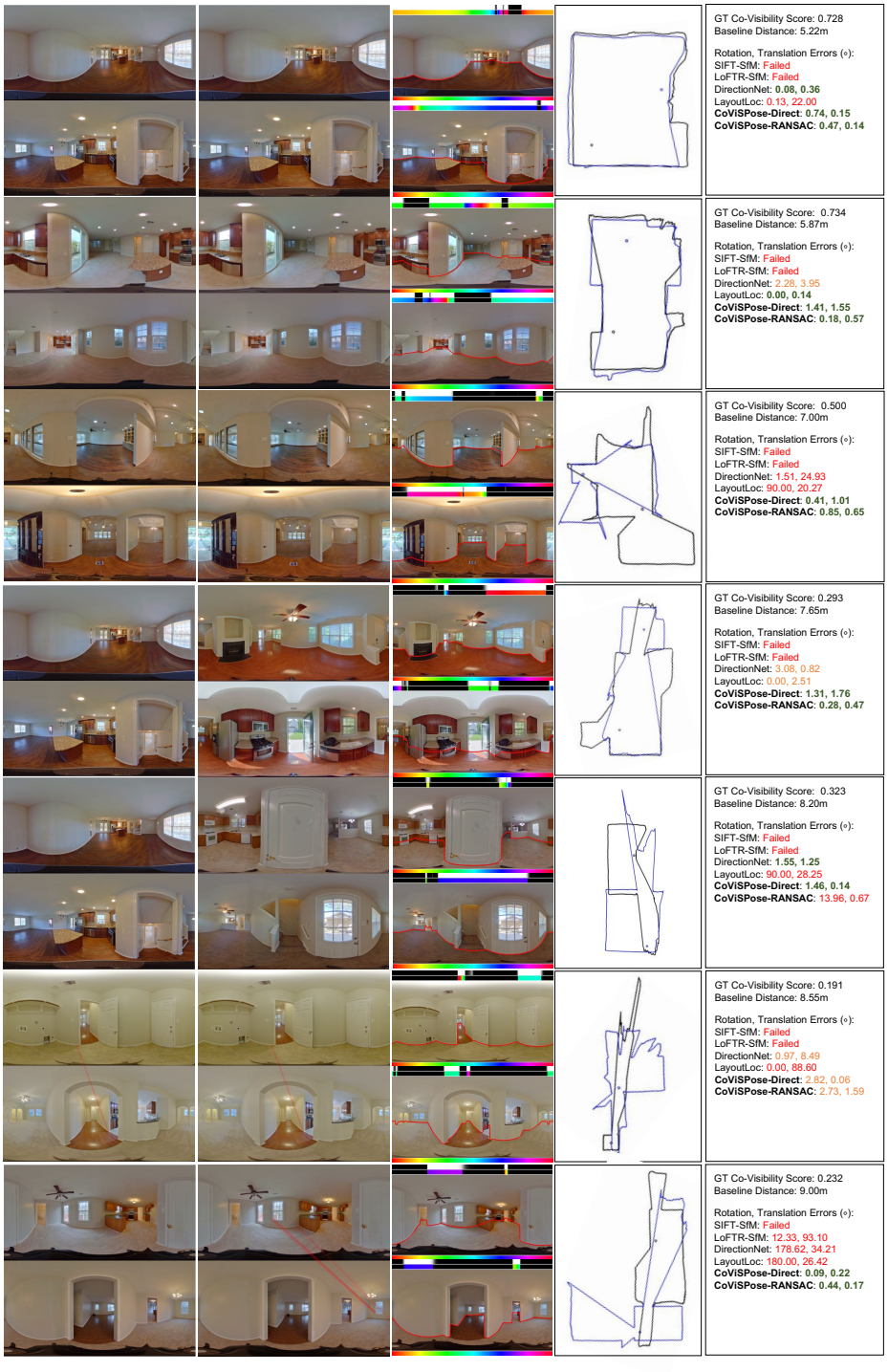


Fig. 8: Qualitative results for panorama pairs with baselines greater than 5 meters. Examples sorted by baseline distance.

References

1. 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)
2. 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)
3. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2), 91–110 (2004)
4. 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)
5. 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)