Wide-Baseline Relative Camera Pose Estimation with Directional Learning

Supplementary Material

Kefan Chen*
Google Research
chenkefan950518@gmail.com

Noah Snavely
Google Research
snavely@google.com

Ameesh Makadia
Google Research
makadia@google.com

A. Spherical Padding

We propose using spherical padding in our decoder network to reflect the correct topology on a spherical representation (See Fig 1 for our motivation).

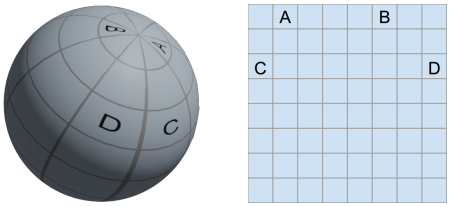


Figure 1. Discrete distributions on the sphere (left) are represented internally as equirectangular grids (right). Although pixels A and B are adjacent on the sphere, as are C and D, they are not adjacent in the grid. Our spherical padding (shown in Fig. 2b, page 4) corrects for this.

B. Dataset Generation

Since large-scale wide-baseline stereo datasets are difficult to acquire, we create our datasets from corpora of panoramic scene captures by taking pairs of panoramas that observe overlapping parts of scenes, sampling camera look-at directions for each panorama using a heuristic that ensures image overlap, and projecting the panoramas to perspective views with a given field of view. Figure 2(a) illustrates this process. The look-at direction, ℓ_1, for the first camera is uniformly sampled from a band around the equator, which is bounded in colatitude $\theta \in [-\frac{\pi}{4}, \frac{\pi}{4}]$ and azimuth angle $\phi \in [0, 2\pi)$. The look-at direction, ℓ_2, for the second camera is uniformly sampled from a circular cone centered at the direction ℓ_1 so that the magnitude of rotations are uniformly distributed in each of our datasets. The limit in latitude prevents the cameras from looking only at the ceiling or floor, which are relatively textureless for many scenes. The aperture of the cone can be adjusted to vary the amount of overlap between image pairs while maintaining variability in the relative camera orientations. Each camera rotation matrix is then constructed from the appropriate look-at vector and the world up vector.

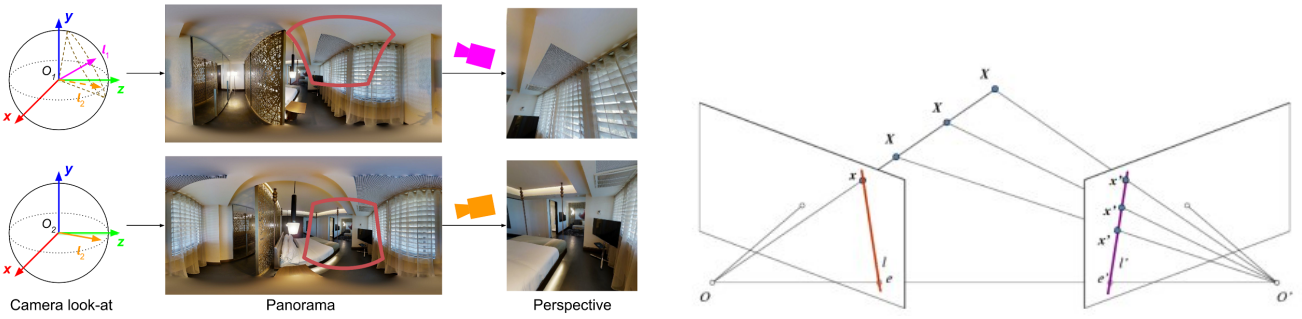
In Figure 3(b) in the main paper, we show results on the Matterport-B test set grouped by overlap percentage between the input images. Matterport3D panoramas contain depth channels which allows us to calculate the overlap percentage between the input image pair as

$$O(I_0, I_1) = \min\left(\frac{|I_0 \cap I_1|}{|I_0|}, \frac{|I_0 \cap I_1|}{|I_1|}\right) \tag{1}$$

C. Training details

We implemented our model in Tensorflow [1]. The model was trained asynchronously on 40 Tesla P100 GPUs. A single DirectionNet has approximately 9M parameters. Our full model DirectionNet-9D/6D, which consists of two DirectionNets, contains in total 18M trainable parameters. Each net was trained for 3M steps.

*Work done while Kefan was a member of the Google AI Residency program (g.co/airesidency).



(a) Data generation (b) Epipolar geometry

Figure 2. (a) We randomly sample perspective images from pairs of panoramas by picking the look-at directions l_1 and l_2 of the source and target cameras based on a heuristic (left). The red boundaries overlaid on the spherical images (center) match the output perspective views (right). (b) Any point x in one image plane corresponds to a ray shooting from its optical center o, which represents all possible 3D locations of x in the world. The projection of this ray into the second image plane forms a line called the *epipolar line*, shown in purple in the figure. The 2D point in the second image corresponding to x must lie on this line.

Rotation Perturbation. To improve the robustness of DirectionNet-T to rotation estimation errors, we apply data augmentation to its input by perturbing the rotations used for derotation. Given $R \in SO(3)$, we perturb it by randomly sampling three unit vectors no further than 15° away from the component vectors of R and projecting the result back onto $SO(3)$. We perturb the estimated rotation from DirectionNet-R before derotating the input images for DirectionNet-T. This perturbation is critical to performance. Without it, the translation error is 4° worse on InteriorNet, and much worse when the rotation range is large (MatterportB).

D. Relative Pose Baselines

We now provide additional details of the baselines including the ones not in the main paper.

- **DirectionNet-Quat** directly generates a probability distribution over the half-hypersphere in S^3. In this case, the spherical decoder consists of 3D upsampling and 3D convolutional layers. Since the output is on a hypersphere, the discretization requires much higher resolution ($O(N^3)$) compared with our model ($O(N^2)$). DirectionNet-Quat generates output at 32^3. We believe the limited resolution is partly responsible for the poor performance compared to DirectionNet-9D and -6D.

- **Bin&Delta** [6] adopts the Bin-Delta hybrid model that consists of a classification network which gives a coarse estimation of the rotation and a regression network that refines the estimate. The rotation space is discretized K-Means clustering on the training data (we use $K = 200$). We use the same encoder as ours and DirectionNet-T for the translation.

- **Spherical regression** [4] uses a novel spherical exponential activation on the n-sphere to improve the stability of gradients during training. The final outputs of the model are the absolute values of the coordinates of a unit vector in $\mathbb{R}^n$, along with n classification outputs for their signs. We use the same encoder as ours followed by separate two-layer prediction networks (one for a quaternion representation of rotation and one for translation).

- **6D** regression uses the same image encoder as ours, followed by two fully connected layers with leaky ReLU and dropout, to produce a 6D continuous representation for the rotation and 3D for the translation. The 6D output is mapped to a rotation matrix with a partial Gram-Schmidt procedure; see [12] for details. This approach uses a continuous representation for 3D rotation, and consequently facilitates training.

- The **quaternion** regression baseline is implemented using same Siamese network as described in [7] without the spatial pyramid pooling layer, followed by fully connected layers to produce a 4D quaternion and a 3D translation. We normalize the quaternion and the translation during training, and use the same loss as suggested in the paper (L2). In our experiment, we weight the quaternion loss with $\beta = 10$, as in the original paper.

		InteriorNet-A						InteriorNet-B					
		R			t			R			t		
		mean (°)	med (°)	rank	mean (°)	med (°)	rank	mean (°)	med (°)	rank	mean (°)	med (°)	rank
DirectionNet	9D	**2.87**	**1.53**	**2.30**	**12.36**	**7.40**	**2.59**	3.88	**2.20**	**2.42**	**16.36**	**9.72**	**2.60**
	6D	2.90	1.68	2.36	12.48	7.53	2.90	**3.81**	2.25	2.44	16.67	10.05	2.77
	9D-Single	3.93	2.61	3.29	18.17	12.56	4.71	4.84	3.24	3.46	26.56	19.51	4.85
	Quat.	23.88	24.53	8.54	32.85	26.92	6.16	22.11	22.43	8.55	39.45	32.05	5.50
Regression	Bin&Delta	8.79	6.59	6.10	19.53	13.45	3.87	6.73	4.57	5.21	31.87	22.61	4.16
	Spherical	19.76	15.52	8.21	31.17	23.47	5.97	11.36	8.82	6.25	44.20	35.34	6.23
	6D	4.86	3.33	3.77	30.94	22.66	5.95	6.03	3.95	3.99	41.29	34.41	6.04
	Quat.	11.14	9.64	6.27	33.14	26.23	6.31	13.94	11.64	6.70	45.44	38.96	6.39
SIFT	LMedS	29.55	7.01	7.76	37.91	19.25	8.28	30.46	7.64	7.92	41.83	24.50	5.58
	RANSAC	16.69	8.21	7.49	45.51	30.12	8.85	18.75	10.52	7.10	52.46	43.56	8.85

Table 1. **Quantitative results on the InteriorNet datasets.** We report the mean and median angular error in degrees, as well as mean rank of each method over all test pairs. Rotation (R) and translation (t) shown separately.

- **SIFT+LMedS** is a classic technique for recovering the essential matrix from correspondences. Local features are detected in images with SIFT, and subsequently matched across images. These feature matches are filtered with Lowe's proposed distance ratio test. Given the remaining putative correspondences, least median of squares (LMedS) is used to robustly estimate the essential matrix, from which we can recover the rotation and normalized translation direction. We use the OpenCV implementation for all of these steps.
- **SIFT+RANSAC** is the same as SIFT+LMedS, with RANSAC instead of LMedS.
- **SuperGlue** [50] uses CNN and graph neural network to extract and match local features from images. **D2-Net** [9] trains a single CNN as a dense feature descriptor and a feature detector. Both methods require correspondence/depth supervision from real data, which is not available in our Matterport datasets. We ran pretrained outdoor SuperGlue (training code not available) and D2-Net with RANSAC.

Additional baselines. The following baselines are not presented in the main paper due to the limit of space. For reference, the mean rotation error of our DirectionNet-9D tested on Matterport-A is $3.96°$, on Matterport-B is $13.60°$.

- **vM** [10] provides a probabilistic formulation for the 2D pose by estimating parameters of von Mises distribution on a circle (S^1). We adapt this method to estimate the 3D rotation by producing three von Mises distributions representing the Euler angles. However, the training is hindered by the singularities known in the Euler angles representation[12]. The mean rotation error tested on Matterport-A is $20.28°$ (5x worse than DirectionNet-9D) and the training diverges on Matterport-B.
- 3D-RCNN[3] uses a classification-regression hybrid model for 2D pose estimation by uniformly discretize the 2D circle into bins. This can be directly adapted to 3D rotation by estimating the three Euler angles. Due to the discontinuity of the Euler angles representation[12], the performance is poor compared with the similar hybrid **Bin&Delta** model. The mean rotation error tested on Matterport-A is $18.61°$ and the error on Matterport-B is $32.33°$.
- [9] combines probabilistic regression and an ensemble of the **quaternion** regression uisng a multi-headed network called HydraNet. The mean rotation error tested on Matterport-A is $9.38°$ and the error on Matterport-B is $16.09°$.
- PoseNet[2] relocalizes images in known scenes; we consider relative pose in scenes never seen during training. PoseNet regresses to a 3D position and quaternion, and this is similar to the **quaternion** regression baseline.

E. Additional Results and Discussion

DirectionNet consistently outperforms regression methods, showing the potential value in a fully convolutional model that avoids fully-connected regression layers and discontinuous parameterizations of pose. We show more comprehensive results to compare our model DirectionNet-9D with the baselines. Note that the spherical regression baseline generally has a higher error in rotation compared with the 6D regression method. Even though the spherical exponential activation does improve training the regression model, the 6D continuous rotation representation is still preferable to quaternions. Figure 3

	R mean (°)	R med (°)	T mean (°)	T med (°)
DirectionNet-9D (20%)	10.50	9.21	26.74	15.67
DirectionNet-9D (100%)	**9.19**	6.31	**19.36**	**11.71**
Regression 6D (100%)	13.44	12.74	22.53	16.68
SuperGlue (outdoor)	16.35	11.53	24.24	17.15
D2-Net	24.07	**5.18**	34.36	14.05

Table 2. Generalization to KITTI [15].

and Figure 4 compare the error histogram distribution of our model with the best two regression baselines, the Bin&Delta and the 6D Regression. Figure 5 compare the error histogram distribution of ours with SIFT+LMedS. Note that SIFT+LMeds has a higher mode close to 0 degree error compared with other baselines. With accurate correspondences, feature-based methods will usually outperform deep learning techniques.

To visualize results of the different methods, we select a few points detected by SIFT in image I_1 and draw their corresponding *epipolar lines* in image I_0 as determined by the estimated relative pose.[1] Figure 2(b) illustrates the epipolar geometry. We show additional qualitative results to compare our primary model DirectionNet-9D with baselines representative of regression models and the classic method, see Fig. 6 and 7) In Fig. 8, we highlight scenarios where our method struggles, such as repeating or complex texture, scenes with few objects and minimal texture, or extreme motion between images.

Ablation study on loss terms. We study the effects of the loss terms by training the DirectionNet-9D on Matterport-A. The mean rotation error is 4.68° without the spread loss, 4.85° without direction loss, and 14.66° without distribution loss, compared with 3.96° with all losses. The distribution loss which provides the direct supervision on the output distribution plays the key role in the training, because we provide the prior knowledge on the distribution by generating the ground truth from von Mises-Fisher distribution on 2-sphere which resembles the spherical normal distribution. This shows evidence that distributional learning with dense supervision is advantageous to direct regression [11, 5]. Alternatively, the distribution loss could use the KL divergence but we found MSE performs better in our experiments.

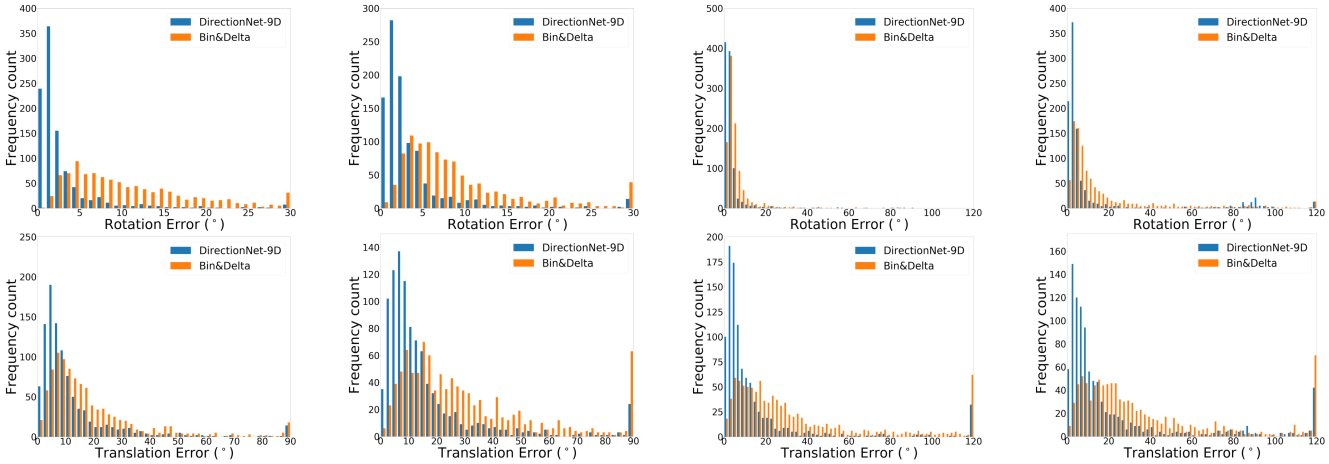
Multimodal distribution on high uncertainty scenarios. In rare scenarios, our model gives higher uncertainty and produces multimodal or even antipodal distributions. Based on our observations, this usually happens in certain scenes, for example, the scene structure exhibits some symmetry or repetitive textures and causes ambiguity in the direction of the motion from two images. (See Figure 9 and Figure 10 for more examples.)

Outdoor scenes. We used KITTI odometry [15] dataset (sequence 0-8 for train, 9-10 for eval) and sampled image pairs with a min rotation of 15° and translation of 10m (36K train pairs, 1K test pairs, mean translation $\sim$18m). Table 2 shows generalization from MatterportA to KITTI (we cropped Matterport images to approximate the KITTI FoV). This is a hard generalization task as the distribution of relative poses in KITTI is extremely different from Matterport, yet fine-tuning with just 20% of data is on par with the local feature baselines, and strong results after retraining with 100% of the data indicates DirectionNet is also effective outdoors.

RANSAC vs. LMedS. We use the OpenCV library (findEssentialMat() and recoverPose()) to implement both baselines by solving the essential matrix using the 5-point algorithm [8] from which we recover the pose. In the main paper, we showed that LMedS performs better than RANSAC in terms of errors in translation and median errors in rotation, but RANSAC has much lower mean errors in rotation on all datasets. Note that due to the nature of indoor images, a large portion of the feature correspondences may be co-planar (e.g. features on a wall or a floor). For RANSAC, we use the default parameters (threshold equals 1.0 and the confidence equals 0.999). Figure 11 shows that the design choice of robust fitting method doesn't make a big overall difference in our experiments.

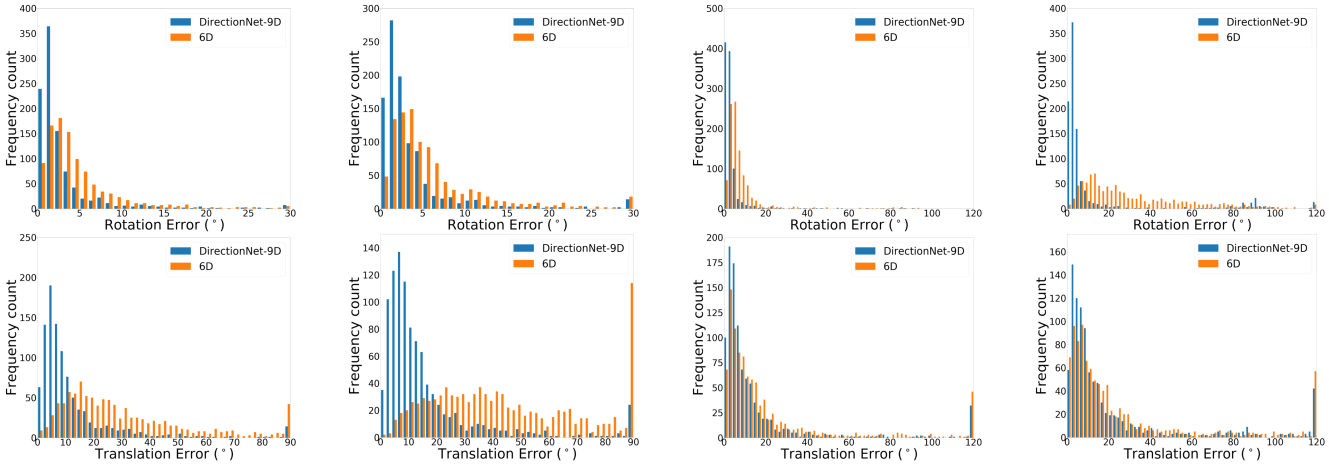
Runtime performance. DirectionNet-Single inference takes under 0.02 seconds with a TESLA P100.

[1] Note, since we do not have ground truth point correspondences between images in our datasets, we cannot draw matching points on the two images for visualization.



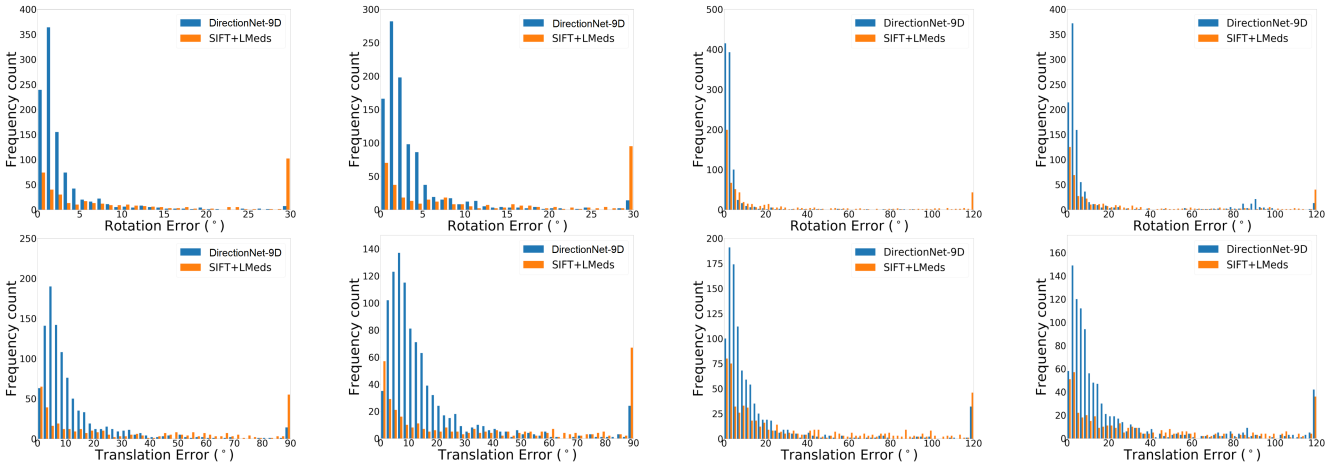
(a) InteriorNet-A (b) InteriorNet-B (c) Matterport-A (d) Matterport-B

Figure 3. **Error histograms DirectionNet-9D vs. Bin&Delta.** Top: rotation, bottom: translation.



(a) InteriorNet-A (b) InteriorNet-B (c) Matterport-A (d) Matterport-B

Figure 4. **Error histograms DirectionNet-9D vs. 6D Regression.** Top: rotation, bottom: translation.



(a) InteriorNet-A (b) InteriorNet-B (c) Matterport-A (d) Matterport-B

Figure 5. **Error histograms DirectionNet-9D vs. SIFT+LMedS.** Top: rotation, bottom: translation.

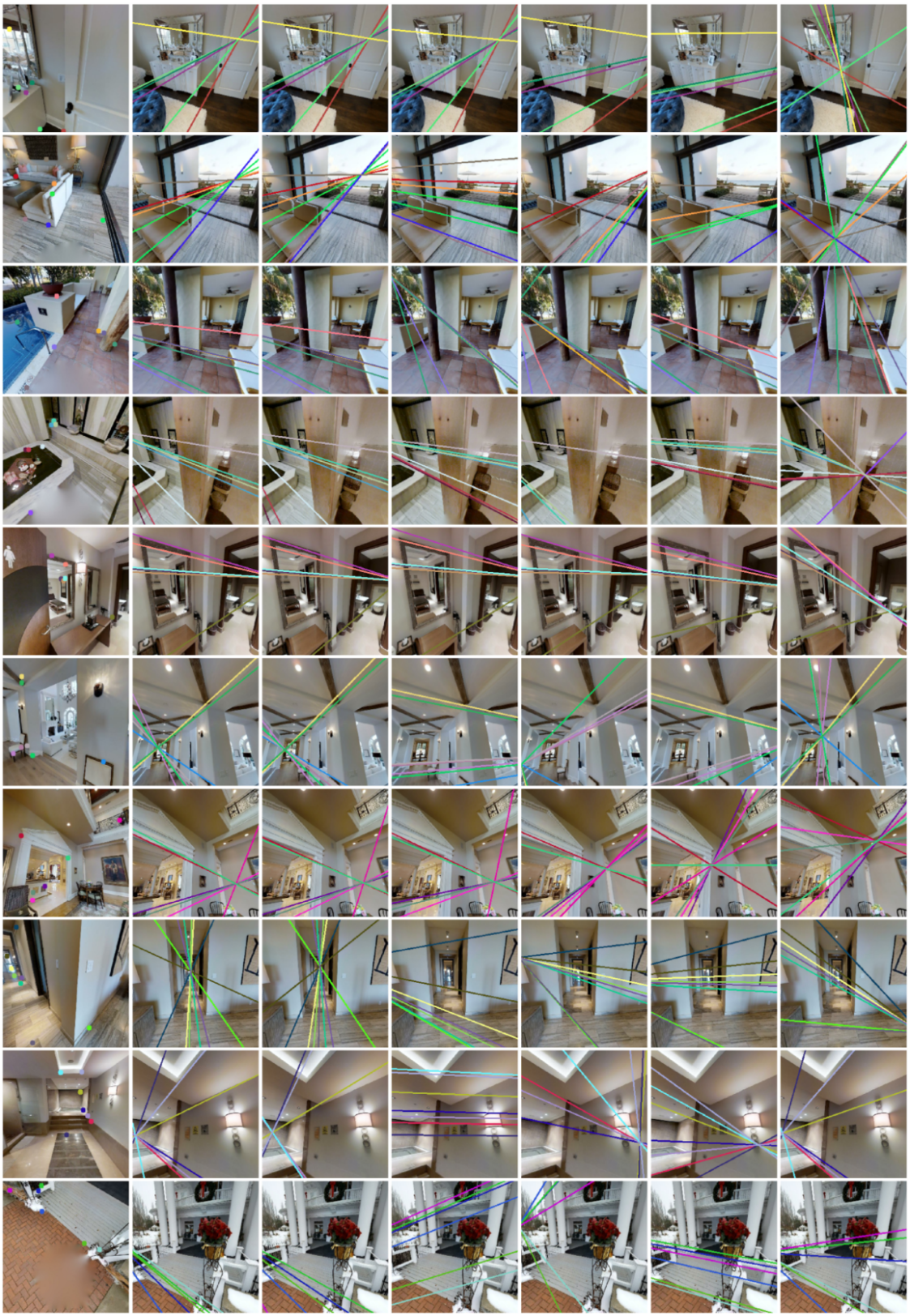


Figure 6. **Additional qualitative results on Matterport-B.**

Figure 7. **Additional qualitative results on InteriorNet-A.**

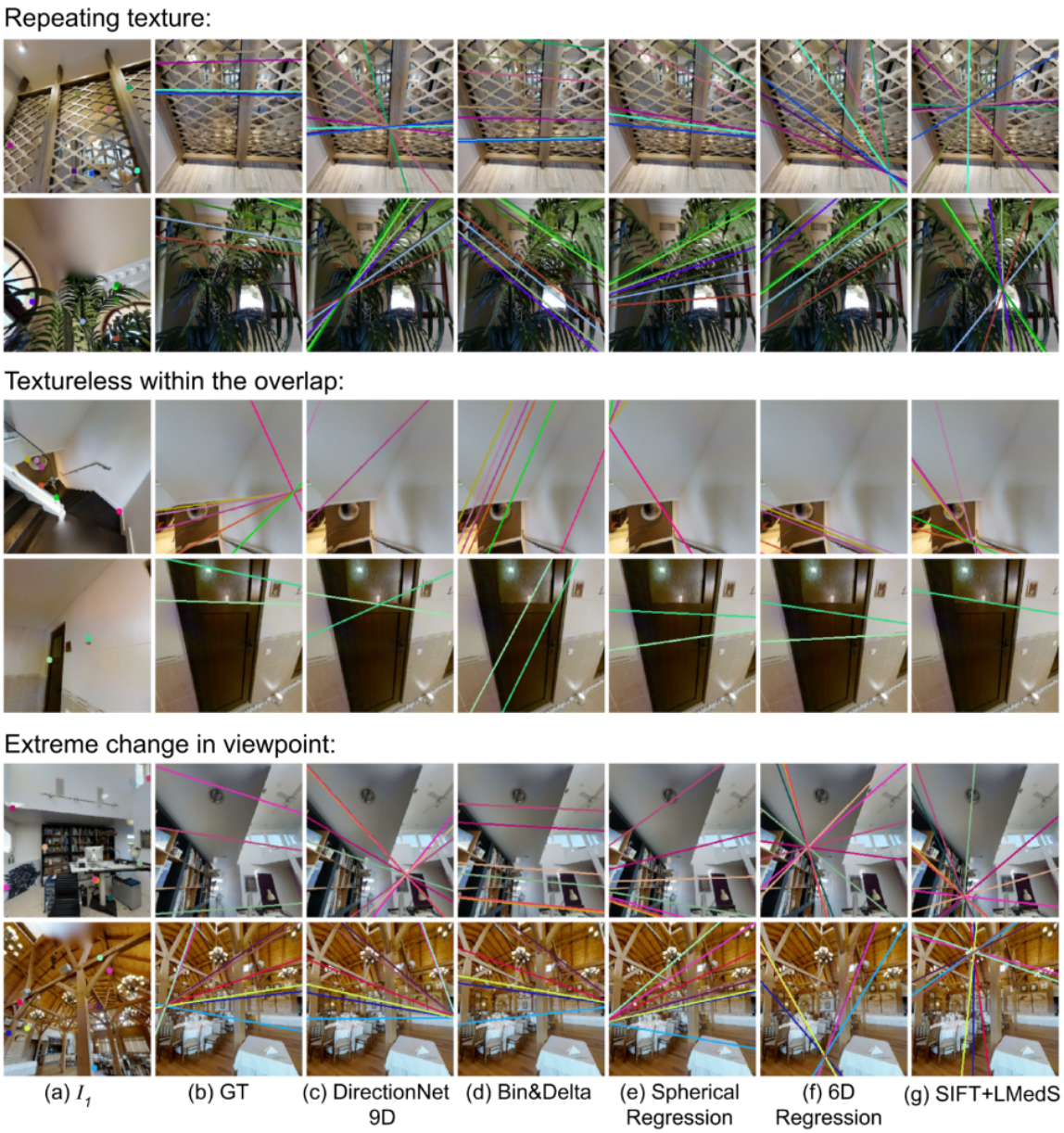


Figure 8. **Failure cases.** Our method could fail in cases such as repeating or complex textures, large textureless area or space with few objects, and extremely large motion.

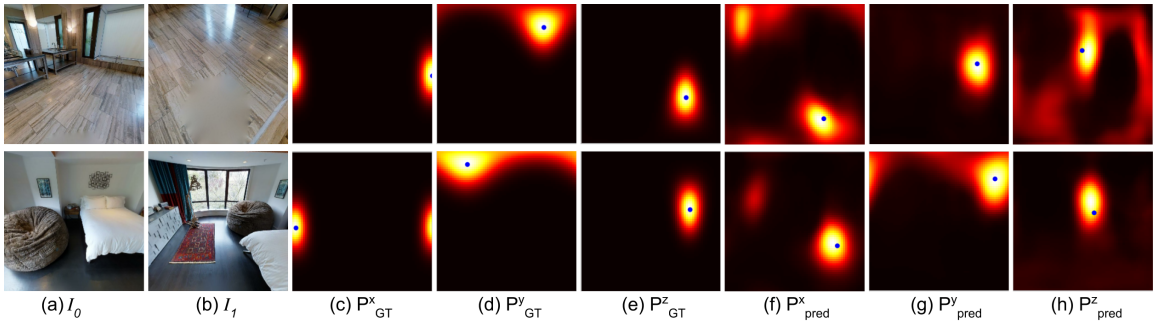


Figure 9. **Multimodal prediction on the rotation.** This figure shows two cases in Matterport-B when the DirectionNet-R fails and gives very high uncertainty in the presence of repeating texture or extreme motion. I_0 and I_1 are input images. $P^x{}_{GT}$, $P^y{}_{GT}$, and $P^z{}_{GT}$ are the ground truth distributions corresponds to v_x, v_y, and v_z respectively. $P^x{}_{pred}$, $P^y{}_{pred}$, and $P^z{}_{pred}$ are the predictions corresponds to the three directions. The spherical distributions are illustrated as equirectangular heatmaps and the blue dot shows where the spherical expectation locates.

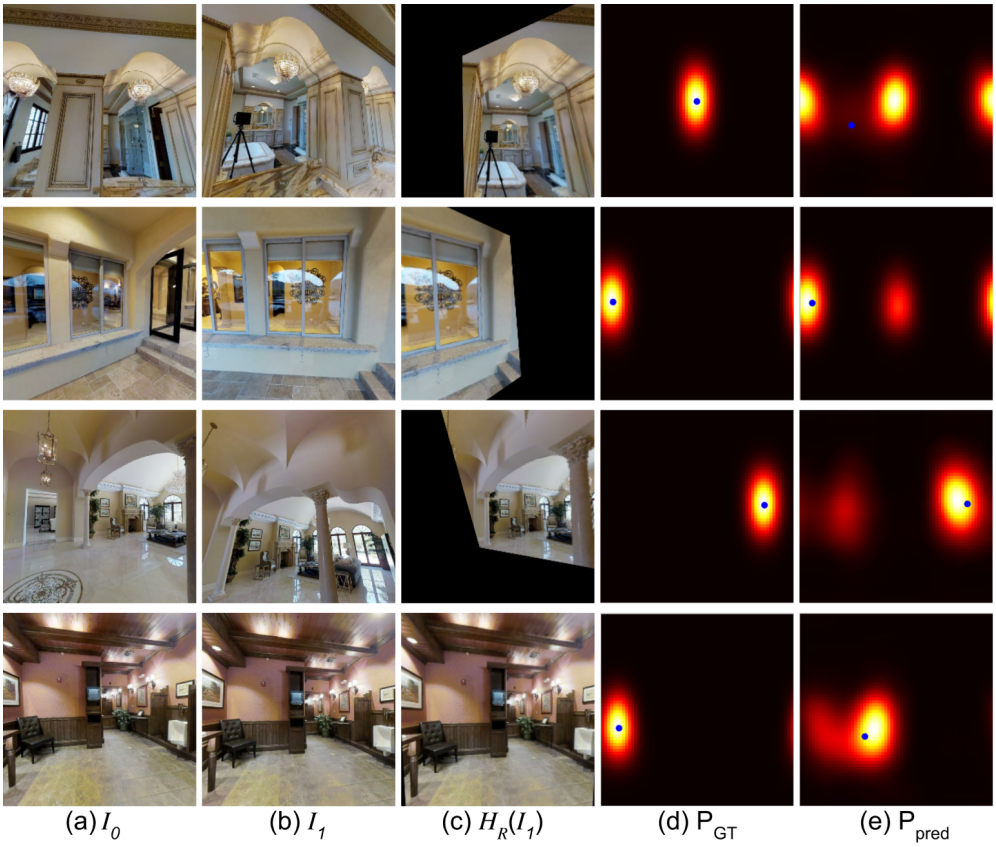
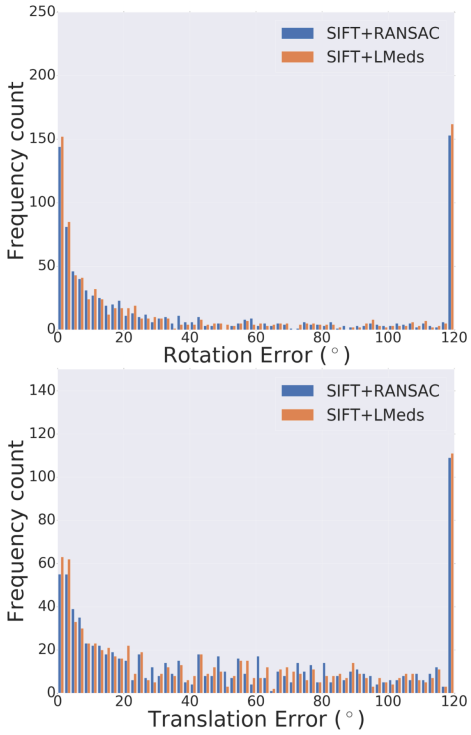


Figure 10. **Multimodal prediction on the translation.** This figure shows examples in Matterport-B when the DirectionNet-T produces multimodal distributions. I_0 and I_1 are original images and $H_R(I_1)$ is the input derotated image. P_{GT} is the ground truth distribution of the translation and P_{pred} is the prediction. The arcade scene in the first example has a symmetry, so it is ambiguous whether the camera is moving left or right. Thus, our model produces an antipodal distribution with almost equal uncertainty. The second example has similar symmetry because of the two identical glass windows, but we can figure out the motion from some inconspicuous clues such as the objects through the window. Thus, the model produces two modes but it is much more certain toward the correct one and the predicted direction is very close to the truth. The last two examples demonstrate another difficult scenario for the model to figure out the translation direction when the translation amount is tiny (3rd: large R and small T, 4th: tiny R and small T). The 3rd example is less ambiguous than the 4th, so the network gives high certainty at the correct mode. Note that even though the model is highly uncertain in the 4th case, the network still manages to produce a secondary mode at the correct location.

(a) Visualizations of RANSAC vs. LMedS



(b) Error histogram on Matterport-A.

Figure 11. (a) We visualize a few examples and compare RANSAC with LMedS in different scenes. Note that the feature detected by SIFT often occurs co-planar due to the nature of the indoor scenes. (b) The error histogram shows that RANSAC and LMedS has similar performance on our dataset.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *IEEE International Conference on Computer Vision (ICCV)*, page 2938–2946, 2015.

[3] A. Kundu, Y. Li, and J. M. Rehg. 3D-RCNN: Instance-level 3D object reconstruction via render-and-compare. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3559–3568, 2018.

[4] Shuai Liao, Efstratios Gavves, and Cees G. M. Snoek. Spherical regression: Learning viewpoints, surface normals and 3D rotations on n-spheres. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[5] Diogo C. Luvizon, Hedi Tabia, and David Picard. Human pose regression by combining indirect part detection and contextual information. *Computers & Graphics*, 85:15–22, Dec 2019.

[6] Siddharth Mahendran, Haider Ali, and René Vidal. A mixed classification-regression framework for 3D pose estimation from 2d images. *The British Machine Vision Conference (BMVC)*, 2018.

[7] Iaroslav Melekhov, Juha Ylioinas, Juho Kannala, and Esa Rahtu. Relative camera pose estimation using convolutional neural networks. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, 2017.

[8] D. Nister. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):756–770, 2004.

[9] Valentin Peretroukhin, Brandon Wagstaff, and and Jonathan Kelly. Deep Probabilistic Regression of Elements of SO(3) using Quaternion Averaging and Uncertainty Injection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[10] Sergey Prokudin, Peter Gehler, and Sebastian Nowozin. Deep directional statistics: Pose estimation with uncertainty quantification. In *European Conference on Computer Vision (ECCV)*, September 2018.

[11] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *European Conference on Computer Vision (ECCV)*, 2018.

[12] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.