6D Camera Relocalization in Ambiguous Scenes via Continuous Multimodal Inference Supplementary Material

Mai Bui[1] Tolga Birdal[2] Haowen Deng[1,3] Shadi Albarqouni[1,4]
Leonidas Guibas[2] Slobodan Ilic[1,3] Nassir Navab[1,5]

[1] Technical University of Munich [2] Stanford University
[3] Siemens AG [4] ETH Zurich [5] Johns Hopkins University

Abstract. This document supplements our paper **6D Camera Relocalization in Ambiguous Scenes via Continuous Multimodal Inference**. In particular we present the following: (1) technical details on network architecture, training and modeling of translations, (2) more evaluations on synthetic data, (3) the used dataset, (4) error metrics, (5) additional quantitative (ablation) studies on the backbone network, multiple hypotheses training, uncertainty estimation, effect of the number of hypotheses on computational time, rotation parameterization and different means of assembling the Bingham matrix from the network output, (6) qualitative results on our real dataset.

1 Modeling translations

As described in the main paper we model translations using mixture density networks [3]. In more detail, for a sample input image $\mathbf{X} \in \mathbb{R}^{W \times H \times 3}$, we obtain a predicted translation $\hat{\mathbf{t}} \in \mathbb{R}^{c=3}$ from a neural network with parameters $\mathbf{\Gamma}$. This prediction is set to the most likely value of a multivariate Gaussian distribution with covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_c^2 \end{bmatrix}_{c \times c}, \tag{1}$$

where $\boldsymbol{\sigma}^2$ is predicted by our model. As a result our model for a unimodal Gaussian is defined as:

$$p_{\mathbf{\Gamma}}(\mathbf{t} \mid \mathbf{X}) = \frac{\exp(-\frac{1}{2}(\mathbf{t} - \hat{\mathbf{t}})^\top \mathbf{\Sigma}^{-1}(\mathbf{t} - \hat{\mathbf{t}}))}{(2\pi)^{c/2}|\mathbf{\Sigma}|^{1/2}}, \tag{2}$$

where $c = 3$ and both $\hat{\mathbf{t}}$ as well as $\mathbf{\Sigma}$ are trained by maximizing its log-likelihood.

Similar to forming a Bingham Mixture Model, we can equally compute a Gaussian Mixture Model with K components and corresponding weights $\pi(\mathbf{X}, \mathbf{\Gamma})$, such that $\sum_{j=1}^{K} \pi_j(\mathbf{X}, \mathbf{\Gamma}) = 1$, to obtain a multi-modal solution. Again both $\hat{\mathbf{t}}$

Table 1: Layer specifications. We report the dimensionality of the input feature vector, N_{in}, resulting output feature vector, N_{out}, whether or not batch normalization (BN) is used and the activation function for each layer.

	N_{in}	N_{out}	BN	activation
$\mathbf{q}$	2048	K*4	no	none
$\mathbf{t}$	2048	K*3	no	none
$\mathbf{\Lambda}$	2048	K*3	no	softplus
$\mathbf{\Sigma}$	2048	K*3	no	softplus
$\boldsymbol{\pi}$	2048	1024	yes	ReLU
	1024	512	yes	ReLU
	512	K	yes	ReLU / softmax

and $\mathbf{\Sigma}$ as well as $\pi(\mathbf{X}, \mathbf{\Gamma})$ are learned by the network and trained by maximizing the log-likelihood of the mixture model. Note that, in this case, the components of $\hat{\mathbf{t}}$ are assumed to be statistically independent within each distribution component. However, it has been shown that any density function can be approximated up to a certain error by a multivariate Gaussian mixture model with underlying kernel function as defined in Eq (2) [3, 12].

2 Network and training details

We resize the input images to a height of 256 pixels and use random crops of size 224×224 for training. For testing we use the central crop of the image. As described in the main paper we use a ResNet-34 [7] as our backbone network, which was pretrained on ImageNet [15], and remove the final classification layers. Fully-connected layers are then appended as specified in Tab. 1, where we output K camera pose hypotheses, $\mathbf{q}$ and $\mathbf{t}$, corresponding distribution parameters, $\mathbf{\Lambda}$ and $\mathbf{\Sigma}$, as well as shared mixture weights $\pi(\mathbf{X}, \mathbf{\Gamma})$. We use the *softplus* activation function to ensure positivity of the Bingham concentration parameters and Gaussian variances. To satisfy the convention, the Bingham concentration parameters are then negated. In case of our single component and Bingham-MDN models we use a softmax activation function, such that $\sum_{j=1}^{K} \pi_j(\mathbf{X}, \mathbf{\Gamma}) = 1$ holds true. In our MHP version, we first apply a ReLU activation function, that, during training, is passed to a cross-entropy loss function. Once trained, we again apply a softmax on the final weights to form a valid mixture model. Our single component model equals setting $K = 1$, whereas for mixture model predictions, we use $K = 50$ pose hypotheses. During training, we follow a projected ADAM optimization [9] with an exponential learning rate decay and train each model for 300 epochs and a batch size of 20 images. For all models we train with an initial learning rate of $1e^{-4}$.

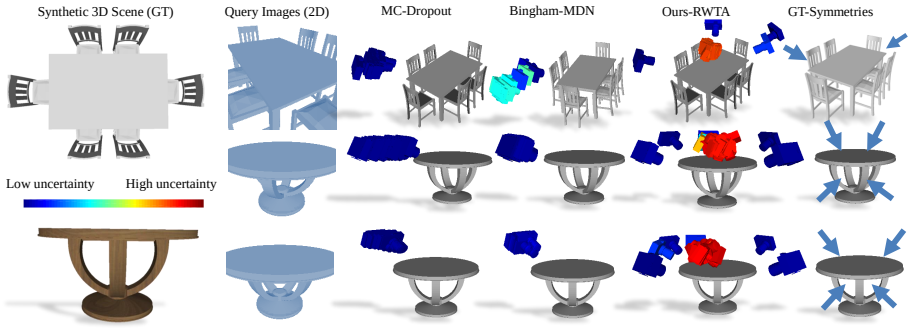

Fig. 1: Additional qualitative results of our synthetically created dataset. If available, camera poses are colored by their uncertainty.

3 Evaluation on synthetic scenes

We now show in Fig. 1 different query images and localization results on the synthetic scenes provided in the paper. The superiority of our approach is consistent across different viewpoints. We also provide additional quantitative and qualitative results on our synthetic dataset. For this aim, we render the objects/scenes from the predicted camera poses of our methods in Fig. 2. There, we show the most certain predictions sorted according to the entropy of the resulting Bingham and Gaussian distributions.

Last but not least, we compute the intersection over union (IoU) with the renderings obtained from the ground truth camera poses. Considering the hypothesis with the highest weight as the single best prediction, on average our Bingham-MDN reaches an IoU of 0.62, whereas our MHP distribution model, *Ours-RWTA*, achieves 0.88.

4 Details on the acquisition of real ambiguous dataset

Besides our synthetically created dataset, we captured a highly ambiguous real dataset, consisting of five scenes using Google Tango [11]. Fig. 3 shows ground truth training and testing camera trajectories, plotted with Open3D [18], as well as example batch images we acquired for our ambiguous scene dataset. The resolution of the captured RGB images is 540×960 and the spatial extent of our scenes can be found in Tab. 2. Further, for each image in the *Blue Chairs* and *Meeting Table* scenes, we obtain a ground truth estimate by training an autoencoder on reconstructing the input images and using the resulting feature descriptors to obtain the nearest neighbor camera poses. Then we cluster the resulting camera poses using a Riemannian Mean Shift algorithm [16] and use the centroids of the resulting clusters as "ground truth" modes. We visually verify the results. The autoencoder we use to compute said features contains a ResNet-34 encoder, followed by subsequent deconvolutions with batch normalization and

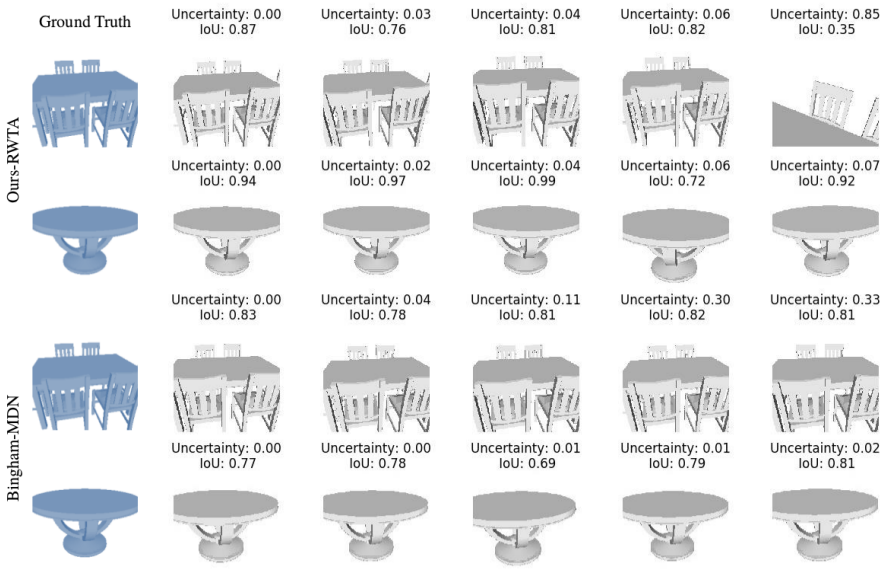

Fig. 2: Renderings of the top five camera pose hypotheses according to their uncertainty values for our Bingham-MDN and MHP version, Ours-RWTA. Further we show the corresponding ground truth query images as well as the intersection over union of the ground truth and predicted renderings.

Table 2: Spatial Extent of our scenes in meters.

Blue Chairs	Meeting Table	Staircase	Staircase Ext.	Seminar Room
$5 \times 4.6 \times 1.3$	$4.3 \times 5.8 \times 1.4$	$4.9 \times 4.4 \times 5.1$	$5.6 \times 5.2 \times 16.6$	$5.3 \times 7.8 \times 2.6$

ReLU activation as the decoder. It is trained with an l_2 reconstruction loss for 300 epochs using the Adam Optimizer [9] with a learning rate of $1e^{-3}$ and a batch size of 20 images. Examples of the obtained ground truth modes can be found in Fig. 4 (left).

5 Error Metrics

Given a ground truth camera pose, consisting of a rotation, represented by a quaternion $\mathbf{q}$, and its translation, $\mathbf{t}$, we evaluate the performance of our models with respect to the accuracy of the predicted camera poses by computing the recall of ours and the baseline models. We consider a camera pose estimate to be correct if both rotation and translation are below a pre-defined threshold and compute the angular error between GT, $\mathbf{q}$, and predicted quaternion, $\hat{\mathbf{q}}$, as

$$d_q(\mathbf{q}, \hat{\mathbf{q}}) = 2 \arccos(|\mathbf{q} \circ \hat{\mathbf{q}}|). \tag{3}$$

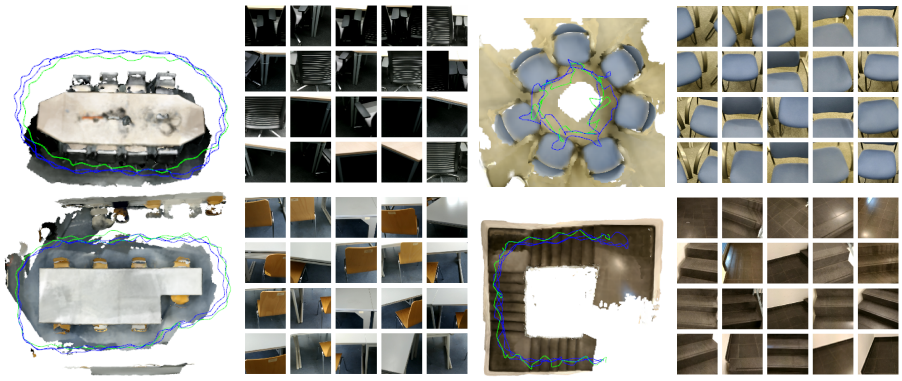

Fig. 3: Ground truth training (blue) and test (green) trajectories of our ambiguous scenes and example RGB images.

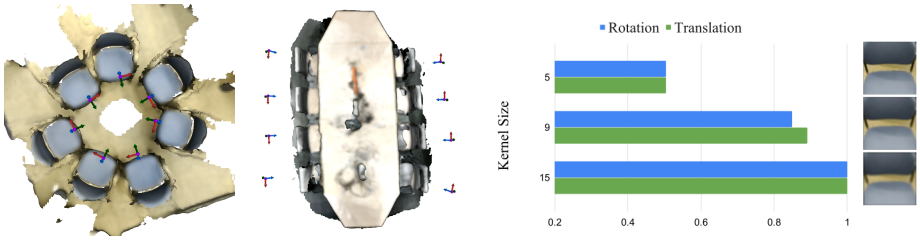

Fig. 4: **(left)** Estimated ground truth modes in *Blue Chairs* and *Meeting Table* scenes, which we use to evaluate our model's mode detection performance and diversity of predictions. **(right)** Change in uncertainty prediction in the presence of increasing image blur. For varying kernel sizes of a Gaussian filter used to blur the input images, we compute the average uncertainty over all images obtained from the predictions of our model. Reported here are the normalized values.

For translations we use the norm of the difference between GT $\mathbf{t}$, and predicted translation $\hat{\mathbf{t}}$: $d_t(\mathbf{t}, \hat{\mathbf{t}}) = \|\mathbf{t} - \hat{\mathbf{t}}\|_2$ to compute the error in position of the camera.

6 Ablation Studies

6.1 Multiple hypothesis estimation

Recently, [10] proposed EWTA, an evolving version of WTA, to alleviate the collapse problems of the RWTA training schemes proposed in [14]. Updating the top k hypotheses instead of only the best one, EWTA increases the number of hypotheses that are actually used during training, resulting in fewer wrong mode predictions that do not match the actual distribution. We evaluated the different versions of MHP training schemes for our particular application for which the

Table 3: Comparison between different MHP variants, RWTA [14] and EWTA [10], averaged over all scenes of our ambiguous real dataset.

Threshold	EWTA (k=50)	EWTA (k=25)	RWTA (k=1, used)
10° / 0.1m	0.12	0.18	**0.20**
15° / 0.2m	0.34	0.40	**0.56**
20° / 0.3m	0.47	0.51	**0.68**

Table 4: Mean ratio of correct poses for different backbone networks on all scenes.

	Threshold	PoseNet	Unimodal	Bingham-MDN	MC-Dropout	Ours-RWTA
ResNet-34	10° / 0.1m	0.15	0.12	0.08	0.15	**0.20**
	15° / 0.2m	0.46	0.39	0.28	0.39	**0.56**
	20° / 0.3m	0.60	0.53	0.37	0.54	**0.68**
ResNet-18	10° / 0.1m	0.15	0.16	0.09	0.15	**0.19**
	15° / 0.2m	0.47	0.42	0.29	0.39	**0.52**
	20° / 0.3m	0.60	0.54	0.39	0.54	**0.66**
ResNet-50	10° / 0.1m	**0.20**	0.15	0.10	0.15	**0.20**
	15° / 0.2m	0.49	0.36	0.30	0.40	**0.55**
	20° / 0.3m	0.62	0.53	0.38	0.53	**0.69**
Inception-v3	10° / 0.1m	0.11	0.10	0.11	0.08	**0.18**
	15° / 0.2m	0.38	0.33	0.38	0.31	**0.49**
	20° / 0.3m	0.55	0.53	0.52	0.49	**0.63**

results can be found in Tab. 3. As it is not straightforward how k should be chosen in EWTA, we 1) start with $k = K$, where K is the number of hypotheses and gradually decrease k until $k = 1$ (as proposed in [10]) and 2) start with the best half hypotheses, i.e. $k = 0.5 \cdot K$. We set $K = 50$ in our experiments. In our setting, we have found this parameter to strongly influence the accuracy of our model. Meanwhile, the wrong predictions are showing very high uncertainty so that, if desired, they can easily be removed. Therefore, we chose to remain with RWTA to train our models. This implicitly admits $k = 1$. Note, however, that these conclusions were drawn from experimental results on our datasets, such that the optimal choice of training scheme remains application dependant.

6.2 Backbone network

To evaluate the effect of different network architectures on our model, we change the backbone network of ours and the SoTA baseline methods. As most of the recent SoTA image based relocalization methods [1,4,13] use a version of ResNet, we compare between ResNet variants: ResNet-18, ResNet-34 and ResNet-50 and Inception-v3 [17]. All the networks are initialized from an ImageNet [6] pre-trained model. We report our findings in Tab. 4. Naturally all methods are slightly dependant on the features that serve as input to the final pose regression

layers. However, regardless of the backbone network used, Ours-RWTA shows, on average, superior performance over the baseline methods.

6.3 Uncertainty evaluation

Due to fast camera movements, motion blur easily arises in camera localization applications and is one factor that can lead to poor localization performance. As a first step in handling such problems, additional information in the form of uncertainty predictions could aid in detecting such events. Therefore, to evaluate how our model performs in the presence of noise, we use our single component model, i.e. $K = 1$, trained on the original input images, and blur the RGB images to evaluate the change in uncertainty prediction of the model. Ideally, with increasing image blur, we would expect our model to be less certain in its predictions. To ablate on this, we apply a Gaussian Filter to the input images, with varying kernel sizes, and report the change in uncertainty prediction in Fig. 4 (right) on the blurred images. We use the entropy over each image to obtain a measure of uncertainty and compute the mean over our dataset images. For visualization, we show the normalized values. An increase in uncertainty could be clearly observed with growing kernel size and thus highly blurred images.

6.4 Number of Hypotheses and Computational Times

Incorporating our method into an existing regression model, simply leads to a change in the last fully-connected layers of the network. We extend the last layer to output an additional $(K-1)\cdot 4$ and $(K-1)\cdot 3$ parameters for predicting the camera pose, as well as overall $6\cdot K$ for uncertainty prediction of both rotation and translation. Further, we incorporate extra layers for the mixture coefficients as described in Tab. 1. We run our model on a 8GB NVIDIA GeForce GTX 1080 graphics card and report the inference time of our network with respect to K in Tab. 5. In comparison to a direct regression method our model with $K = 50$ incurs a negligible computational overhead around $1ms$.

Further, we evaluate the effect of hyperparameter K, i.e. the number of hypothesis to be regressed, for our proposed method. Based on the results, which are summarized in Fig. 5, we suspect the optimal number of hypotheses to be dependant on the spatial extent of the scene and on the ambiguities contained in them. However, due to the increased complexity of the model as well as instability issues during training, we observed a drop in performance with high increase of the number of hypotheses.

Table 5: Inference time of our method with respect to the number of hypotheses.

PoseNet	$K = 1$	$K = 50$	$K = 200$	$K = 500$
$7.23ms$	$7.27ms$	$8.11ms$	$8.19ms$	$8.74ms$

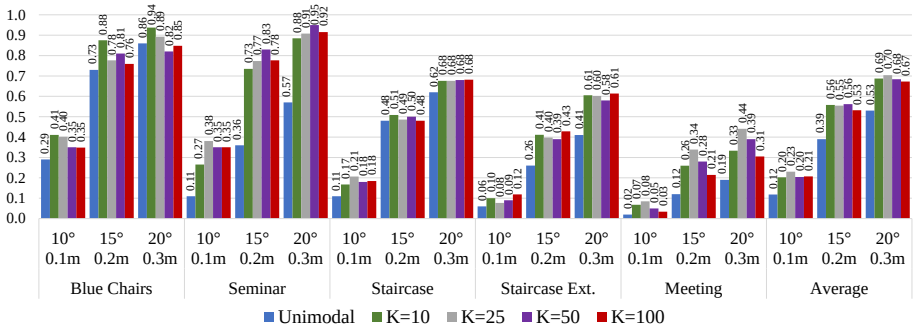

Fig. 5: Influence of the number of hypotheses, i.e. parameter K, on the performance of our method, Ours-RWTA.

6.5 Rotation parameterization

The best choice of rotation parameterization for training deep learning models is an open question. PoseNet [8] proposed to use quaternions due to the ease of normalization. The ambiguities can be resolved by mapping the predictions to one hemisphere. MapNet [4] further showed improvements in using the axis angle representation. Recently it has been shown that any representation with four or less degrees of freedom suffers from discontinuities in mapping to $SO(3)$. This might harm the performance of deep learning models. Instead, [19] proposed a continuous 6D or 5D representation. We ablate in this context by mapping all predictions to the proposed 6D representation and model them using a GMM, similar to a MDN, but treating rotation and translation separately. Therefore for each camera pose, in total we have $9 \cdot 2$ parameters to regress, plus mixture coefficients. Tab. 6 shows our results, where 'Geo + L1' refers to a direct regression using the geodesic loss proposed in [19] and an l_1 loss on the translation. When using the proposed 6D representation, we found either improvements or similar performance to their quaternion counterparts. However, overall our 9D-Ours-RWTA remains the most promising model. In terms of Oracle Error MC-Dropout sometimes outperforms our method. This comes from the fact that MC-Dropout mostly predicts multiple hypothesis around one mode, which if this mode is relatively close to the ground truth one, results in a high Oracle. 9D-Ours-RWTA predicts diverse hypothesis, but not multiple versions of the same mode. However it shows much better performance in predicting the correct mode than MC-Dropout.

6.6 Ablation studies for constructing V

Alternatively to the method proposed in the main paper, Gram-Schmidt can be used to compute an orthonormal matrix $\mathbf{V}$ from a given matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$, where the column vectors $\mathbf{v_i}$ of $\mathbf{V}$ are computed from the column vectors $\mathbf{m_i}$ as

Table 6: Ratio of correct poses when using the continuous 6D representation of [19] to model rotations instead of a Bingham distribution on the quaternion.

	Threshold	Geo+L1	Uni.	MDN	MC-Dropout	9D-Ours-RWTA	MC-Dropout Oracle	9D-Ours-RWTA Oracle
Blue Chairs	10° / 0.1m	0.41	**0.48**	0.01	0.26	0.38	**0.58**	0.44
	15° / 0.2m	**0.90**	0.89	0.14	0.83	0.81	0.91	**0.96**
	20° / 0.3m	**0.96**	0.92	0.23	0.91	0.84	0.94	**1.0**
Meeting Table	10° / 0.1m	0.03	0.03	0.02	0.02	**0.06**	**0.09**	0.08
	15° / 0.2m	0.16	0.16	0.11	0.13	**0.29**	0.24	**0.47**
	20° / 0.3m	0.22	0.23	0.14	0.21	**0.38**	0.32	**0.71**
Staircase	10° / 0.1m	0.17	**0.19**	0.12	0.12	0.18	**0.27**	0.21
	15° / 0.2m	0.46	**0.51**	0.36	0.36	0.44	**0.56**	0.54
	20° / 0.3m	0.62	**0.67**	0.47	0.56	0.56	0.70	**0.71**
Staircase Extended	10° / 0.1m	0.07	0.01	0.01	0.04	**0.08**	**0.15**	0.09
	15° / 0.2m	0.30	0.06	0.09	0.18	**0.35**	**0.40**	0.38
	20° / 0.3m	0.48	0.13	0.14	0.36	**0.55**	0.59	**0.62**
Seminar Room	10° / 0.1m	**0.34**	0.24	0.30	0.21	**0.34**	**0.45**	0.34
	15° / 0.2m	0.74	0.63	0.65	0.65	**0.76**	**0.84**	0.77
	20° / 0.3m	0.84	0.79	0.76	0.82	**0.88**	**0.9**	0.88
Average	10° / 0.1m	0.20	0.19	0.09	0.13	**0.21**	**0.31**	0.23
	15° / 0.2m	0.51	0.45	0.27	0.43	**0.53**	0.59	**0.62**
	20° / 0.3m	0.63	0.55	0.35	0.57	**0.64**	0.69	**0.78**

follows

$$\hat{\mathbf{v}}_\mathbf{i} = \mathbf{m}_i - \sum_{k=1}^{i-1} \langle \mathbf{v}_k, \mathbf{m}_i \rangle \cdot \mathbf{v}_k \ \text{, where } \mathbf{v}_i = \frac{\hat{\mathbf{v}}_i}{\|\hat{\mathbf{v}}_i\|}. \tag{4}$$

Note that in the GS procedure, we predict 16 values for $\mathbf{V}$ and use GS to project onto the orthonormal matrices. Yet the degrees of freedom of $\mathbf{V}$ is much less. For instance the matrix scheme of [2] uses only four. As an ablation, we propose another way to construct $\mathbf{V}$ using the *Cayley transform* [5] as follows: Given a vector $\mathbf{q}$ (not necessarily with unit norm), we compute $\mathbf{V}$ as

$$\mathbf{V} = (\mathbf{I}_{d\times d} - \mathbf{S})^{-1}(\mathbf{I}_{d\times d} + \mathbf{S}), \tag{5}$$

where $\mathbf{I}_{d\times d}$ is the identity matrix and

$$\mathbf{S}(\mathbf{q}) \triangleq \begin{bmatrix} 0 & -q_1 & q_4 & -q_3 \\ q_1 & 0 & q_3 & q_2 \\ -q_4 & -q_3 & 0 & -q_1 \\ q_3 & -q_2 & q_1 & 0 \end{bmatrix} \tag{6}$$

a skew-symmetric matrix parameterized by $\mathbf{q}$. We compare between the proposed method used in the paper and these two alternatives, GS orthonormalization and the construction using skew-symmetric matrices. The results can be found in Tab. 7. In comparison to GS the remaining methods only require four parameters to be estimated instead of the 16 entries of the matrix $\mathbf{V}$. For

Table 7: Ratio of correct poses for several thresholds of Gram-Schmidt (G), Skew-Symmetric (S) and Birdal *et al.* [2] (B) methods to construct $\mathbf{V}$.

		Unimodal	Bingham-MDN	Ours-RWTA
	Threshold	G / S / B	G / S / B	G / S / B
Blue Chairs (A)	10° / 0.1m	0.24 / 0.23 / 0.29	0.04 / 0.17 / 0.24	0.30 / 0.12 / 0.35
	15° / 0.2m	0.63 / 0.58 / 0.73	0.15 / 0.49 / 0.75	0.73 / 0.39 / 0.81
	20° / 0.3m	0.76 / 0.73 / 0.86	0.18 / 0.59 / 0.80	0.79 / 0.43 / 0.82
Meeting Table (B)	10° / 0.1m	0.02 / 0.07 / 0.02	0.04 / 0.01 / 0.01	0.04 / 0.09 / 0.05
	15° / 0.2m	0.16 / 0.20 / 0.12	0.18 / 0.14 / 0.07	0.12 / 0.23 / 0.28
	20° / 0.3m	0.24 / 0.25 / 0.19	0.21 / 0.24 / 0.10	0.18 / 0.27 / 0.39
Staircase (C)	10° / 0.1m	0.17 / 0.16 / 0.11	0.21 / 0.16 / 0.04	0.17 / 0.14 / 0.18
	15° / 0.2m	0.46 / 0.51 / 0.62	0.43 / 0.37 / 0.15	0.46 / 0.42 / 0.50
	20° / 0.3m	0.62 / 0.64 / 0.62	0.60 / 0.49 / 0.25	0.60 / 0.62 / 0.68
Staircase Extended (D)	10° / 0.1m	0.04 / 0.04 / 0.06	0.04 / 0.07 / 0.06	0.05 / 0.06 / 0.09
	15° / 0.2m	0.16 / 0.16 / 0.26	0.19 / 0.29 / 0.21	0.23 / 0.26 / 0.39
	20° / 0.3m	0.27 / 0.27 / 0.41	0.31 / 0.41 / 0.32	0.34 / 0.36 / 0.58
Seminar Room (E)	10° / 0.1m	0.27 / 0.33 / 0.06	0.30 / 0.35 / 0.06	0.15 / 0.28 / 0.35
	15° / 0.2m	0.69 / 0.69 / 0.23	0.56 / 0.59 / 0.23	0.47 / 0.70 / 0.83
	20° / 0.3m	0.82 / 0.80 / 0.40	0.64 / 0.70 / 0.40	0.58 / 0.79 / 0.95
Average	10° / 0.1m	0.15 / **0.16** / 0.11	**0.13** / **0.13** / 0.08	0.14 / 0.14 / **0.20**
	15° / 0.2m	0.42 / **0.43** / 0.36	0.30 / **0.38** / 0.28	0.40 / 0.40 / **0.56**
	20° / 0.3m	**0.54** / **0.54** / 0.50	0.39 / **0.49** / 0.37	0.50 / 0.49 / **0.68**

our unimodal as well as multimodal MDN we found the Skew-Symmetric construction to outperform both Gram-Schmidt (GS) and the employed method of Birdal *et al.* [2]. However, for our method, *Ours-RWTA*, the latter [2] performs the best. Additionally it achieves overall the best performance in comparison to the remaining methods and constructions.

Expressive power of $\mathbf{V}$ In this section as well as in the main paper we have presented a variety of ways to establish $\mathbf{V}(\mathbf{q})$. These methods range from regressing 16 parameters (full flexibility) to regressing only 4 (less flexibility but also a smaller parameter space). This raises an interesting trade-off on the expressiveness of $\mathbf{V}$ and the performance of the neural network *i.e.* how many parameters would *suffice* to capture all the necessary Bingham distributions? This remains to be an open question as, like many others, Birdal *et al.* [2] (our choice of construction) did not provide an analysis on the extent of sufficiency. Nevertheless, we would like to point out that once $\mathbf{V}$ is chosen to be a particular frame, it can explain other orthogonal bases as a linear combination. This introduces certain degree of expressive power (albeit quantized), which we have empirically found to be sufficient compared to other potentially over-parameterized schemes. Besides, the method of Birdal *et al.* [2] is computationally the cheapest.

7 Additional qualitative results

Further, we provide more qualitative results from different query images on all scenes of our ambiguous dataset in Fig. 6 and Fig. 7.

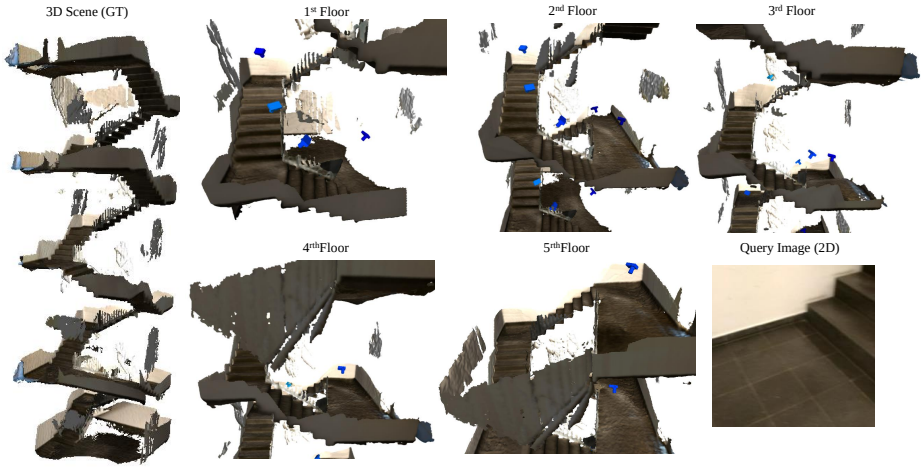

Fig. 6: Qualitative results of our model, Ours-RWTA, on the *Staircase Extended* scene.

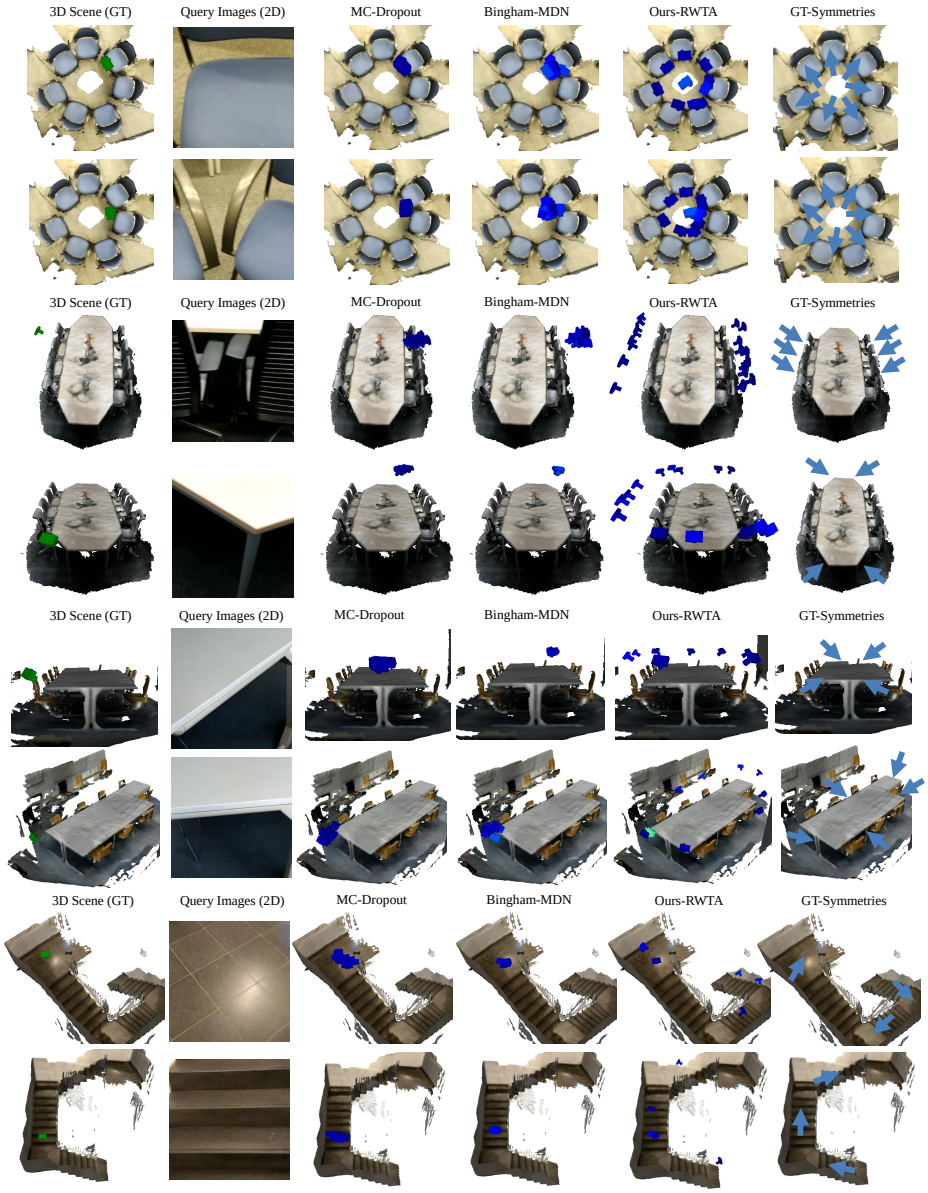

Fig. 7: Additional qualitative results of our ambiguous scenes dataset. We show the ground truth camera pose, query images and resulting camera pose predictions. Both MC-Dropout and our Bingham-MDN suffer from mode collapse, whereas our MHP-based model, Ours-RWTA, predicts diverse hypotheses covering all possible modes.

References

1. Balntas, V., Li, S., Prisacariu, V.: Relocnet: Continuous metric learning relocalisation using neural nets. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 751–767 (2018) 6
2. Birdal, T., Simsekli, U., Eken, M.O., Ilic, S.: Bayesian pose graph optimization via bingham distributions and tempered geodesic mcmc. In: Advances in Neural Information Processing Systems. pp. 308–319 (2018) 9, 10
3. Bishop, C.M.: Mixture density networks (1994) 1, 2
4. Brahmbhatt, S., Gu, J., Kim, K., Hays, J., Kautz, J.: Geometry-aware learning of maps for camera localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2616–2625 (2018) 6, 8
5. Cayley, A.: Sur quelques propriétés des déterminants gauches. Journal für die reine und angewandte Mathematik **32**, 119–123 (1846) 9
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 6
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 2
8. Kendall, A., Grimes, M., Cipolla, R.: Posenet: A convolutional network for real-time 6-dof camera relocalization. In: Proceedings of the IEEE international conference on computer vision. pp. 2938–2946 (2015) 8
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 2, 4
10. Makansi, O., Ilg, E., Cicek, O., Brox, T.: Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7144–7153 (2019) 5, 6
11. Marder-Eppstein, E.: Project tango. pp. 25–25 (07 2016) 3
12. McLachlan, G.J., Basford, K.E.: Mixture models: Inference and applications to clustering, vol. 84. M. Dekker New York (1988) 2
13. Peretroukhin, V., Wagstaff, B., Giamou, M., Kelly, J.: Probabilistic regression of rotations using quaternion averaging and a deep multi-headed network. arXiv preprint arXiv:1904.03182 (2019) 6
14. Rupprecht, C., Laina, I., DiPietro, R., Baust, M., Tombari, F., Navab, N., Hager, G.D.: Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3591–3600 (2017) 5, 6
15. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y 2
16. Subbarao, R., Meer, P.: Nonlinear mean shift over riemannian manifolds. International journal of computer vision **84**(1), 1 (2009) 3
17. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016) 6
18. Zhou, Q.Y., Park, J., Koltun, V.: Open3D: A modern library for 3D data processing. arXiv:1801.09847 (2018) 3

19. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5745–5753 (2019) 8, 9