Supplementary Material
Fusing the Old with the New: Learning Relative Camera Pose with Geometry-Guided Uncertainty

Bingbing Zhuang[1] Manmohan Chandraker[1,2]
[1]NEC Labs America [2]University of California, San Diego

This supplementary material contains (1) additional comparison with the aleatoric uncertainty learning [2, 4, 5], (2) additional analyses on uncertainty, generalization and self-attention, and (3) additional details and discussions on our framework.

S1. Comparison with Aleatoric Uncertainty

Here, we provide more details and analyses on the aleatoric uncertainty learing method in our ablation study. In contrast to our method which guides the uncertainty learning by the geometric uncertainty, the aleatoric uncertainty is directly learned from data, by minimizing the following loss,

$$L(\hat{\boldsymbol{\theta}}_d, \boldsymbol{\sigma}_d^2) = \sum_i \frac{(\hat{\theta}_{i,d} - \bar{\theta}_i^*)^2}{\sigma_{i,d}^2} + log(\sigma_{i,d}^2), \tag{S1}$$

$$\begin{aligned} \hat{\boldsymbol{\theta}}_d =& \{\alpha_d, \beta_d, \phi_{y,d}, \phi_{p,d}, \phi_{r,d}\}, \\ 1/\boldsymbol{\sigma}_d^2 =& \{1/\sigma_{\alpha,d}^2, 1/\sigma_{\beta,d}^2, 1/\sigma_{y,d}^2, 1/\sigma_{p,d}^2, 1/\sigma_{r,d}^2\}, \end{aligned} \tag{S2}$$

where $(\hat{\theta}_{i,d}, \sigma_{i,d}^2)$ indicate the network predicted mean and variance of the underlying Gaussian of the motion parameter i. $\bar{\theta}_i^*$ is the ground truth defined in the circular nearest neighbor sense, as described in Sec. 3.2.2 of the main paper. We note that the uncertainty learned in such a way is not geometry-aware as the geometric solution and uncertainty are not supplied in any way during training. Thus, the uncertainty so obtained is not optimized to be fused with the geometric solution, and may not match to the geometric uncertainty in terms of numerical range. We present the results of this method ("Aleatoric Uncertainty w/o fusion") in Tab. S1. In addition, we fuse the DNN prediction with the geometric solution as a post-processing step, using the learned uncertainty; this is denoted as "Aleatoric Uncertainty w/ fusion". It can be seen that the results are overall inferior to those from UA-Fusion. To elucidate its behavior, following the study in Sec. 4.3 of the main paper, we plot the error distributions in Fig. S3 and the uncertainty distributions in Fig. S4. We observe in Fig. S3 that the fused solutions ("DNN-Fusion") are mostly equivalent to the geometric solutions ("5pt&BA"), except for those cases with highest geometric uncertainties. This is because the learned uncertainties are in general far higher than the geometric uncertainties, as can be seen in Fig. S4. This holds even for the translation estimation.

S2. Further Analyses

Geometric Uncertainty vs. DNN Uncertainty As mentioned in the main paper, the smoothed curves provide clearer visualizations of the overall trend, but may give an impression that the DNN does not have any impact at the rotation part. To clarify this, we provide more detailed analyses on the uncertainty of each individual test sample. Specifically, we plot the uncertainty for each test sample without smoothing in Fig. S5, with y-axis in log scale for better visualization. As can be seen, this clearly reveals the cases where the DNN rotation uncertainties are lower (higher inverse variances) than the geometric rotation uncertainties, and hence do make a difference to the fused solution.

Generalization to ScanNet Here, we analyze the behavior of our network on ScanNet [1] to study its generalization across different datasets. As mentioned in the main paper, we use the outdoor model of SuperGlue [6] to prevent potential overlapping between our test set and SuperGlue's training set. Following the study in Sec. 4.3 of the main paper, we compare the geometric error/uncertainty and DNN error/uncertainty in Fig. S6 and Fig. S7. Similar to the analyses in the preceding paragraph, we plot the uncertainty for each test sample without smoothing in Fig. S8, with y-axis in log scale for better visualization. Overall, we observe that the network behaves similarly as on the DeMoN datasets. In cases with larger geometric uncertainties (i.e. lower inverse variances), the fused solutions surpass the geometric solutions significantly in terms of accuracy. Conversely, in

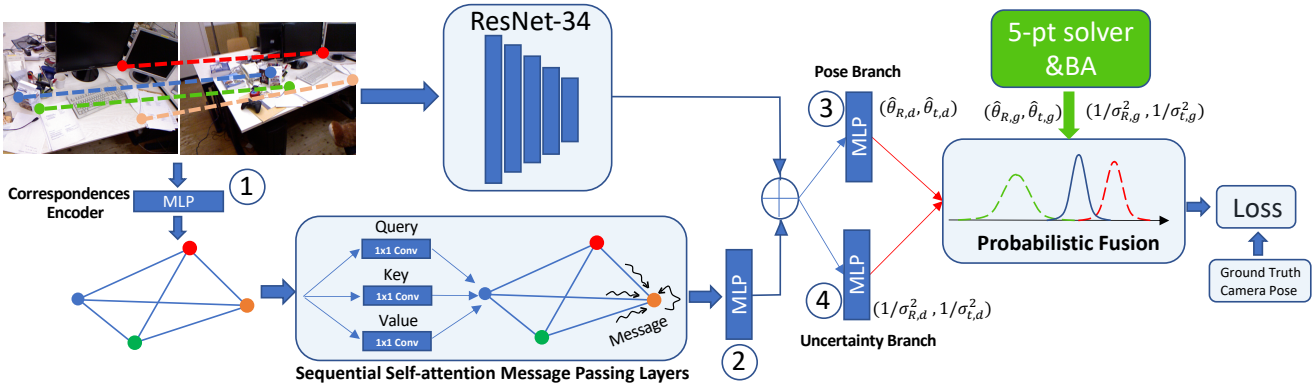

Figure S1. Overview of our geometric-DNN pose fusion network, which fuse the geometric solution with the DNN prediction in training.

	MVS		Scenes11		RGB-D		Sun3D		All	
	Rot.	Tran.	Rot.	Tran.	Rot.	Tran.	Rot.	Tran.	Rot.	Tran.
Aleatoric Uncertainty w/ fusion	4.875	6.749	0.878	5.548	4.882	13.140	3.121	12.510	3.027	8.801
Aleatoric Uncertainty w/o fusion	2.890	4.954	0.375	2.956	1.813	13.650	1.273	11.750	1.371	7.694
UA-Fusion	2.502	4.506	0.388	3.001	1.480	10.520	1.340	11.830	1.246	6.888

Table S1. Comparison with the aleatoric uncertainty learning method, including the plain DNN prediction and the post-processing fusion with the geometric solution.

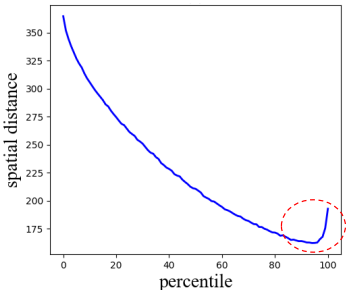

Figure S2. Study of spatial distances (pixels) against attentions.

cases with lower geometric uncertainties, the fused solutions mostly keep the geometric solutions. This corroborates the intuition that motivates our work. Furthermore, we observe that the uncertainty distributions bear close resemblance to those obtained from the DeMoN datasets in the main paper, demonstrating the generalization capability of the network.

Self-attention We have discussed in the main paper the empirically observed correlation between the attention and the spatial distance between different pairs of correspondences, as shown in Fig. S2. We reckon that, the increasing attention on the pair of correspondences with decreasing spatial distance is due to the increasing difficulty to extract additional pose-related information from two points closer to each other. However, we would also like to note here that such trend is not incessant – it may not be beneficial to attend to the relationship between two pairs of correspondences that are extremely close to each other; in the extreme case, the two pairs of correspondences merge to one and there is no any extra information to extract. We surmise that this is the reason causing the noticeable uptick in Fig. S2 as highlighted by the red circle.

S3. Further Implementation Details

Network Parameters In what follows, we discuss the detailed parameters of the MLPs in different parts of UA-fusion. For clarity, we provide our overall architecture in Fig. S1 for reference. The correspondence encoder ("1" in Fig. S1) consists of an MLP with layer size [32, 64, 128, 128]. The MLP for computing updated feature in message passing (Eq. (6) in the main paper) is of layer size [128, 128]. The feature output by the four sequential self-attention message passing layers are then passed through a single-layer MLP ("2" in Fig. S1) with layer size 128. This yields the geometric feature, which, after average pooling, is concatenated with the 512-dimension appearance feature from ResNet-34, and passed to the pose ("3" in Fig. S1) and uncertainty ("4" in Fig. S1) prediction branches. The pose branch is an MLP with layer size [512, 256, 6], predicting $\{t_{x,d}, t_{y,d}, t_{z,d}, \phi_{y,d}, \phi_{p,d}, \phi_{r,d}\}$. We then convert $\{t_{x,d}, t_{y,d}, t_{z,d}\}$ to $\{\alpha_d, \beta_d\}$; note that we do not directly regress $\{\alpha_d, \beta_d\}$ due to their discontinuity. The uncertainty

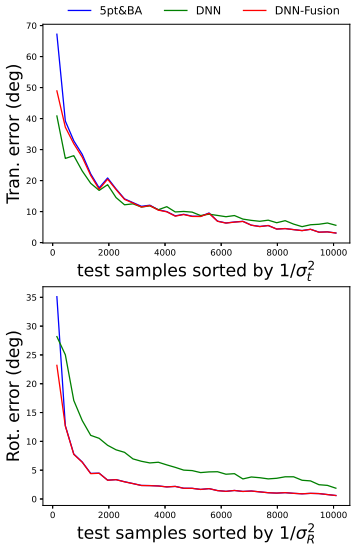

Figure S3. Error comparison between geometric and DNN predictions by learning aleatoric uncertainty.

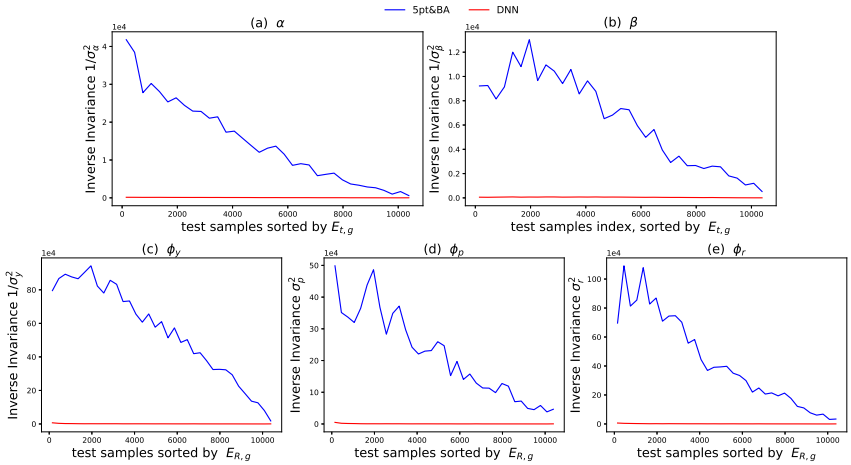

Figure S4. Comparison between the geometric and DNN aleatoric uncertainty for each parameter in $\{\boldsymbol{\theta_t}, \boldsymbol{\theta_R}\}$.

branch is an MLP with layer size [512, 256, 5], predicting the inverse variance $(1/\sigma^2_{\alpha,d}, 1/\sigma^2_{\beta,d}, 1/\sigma^2_{y,d}, 1/\sigma^2_{p,d}, 1/\sigma^2_{r,d})$. In an MLP, all the layers except for the last one are followed by ReLu and batch normalization.

Implementation details Our network is implemented with Pytorch using Adam [3] optimizer with a 10^{-4} learning rate. The input images are always resized to 192x256. We adopt an iterative training scheme; we first train the feature extraction network and the pose branch, then with the feature extraction network fixed, we alternatively train the pose and uncertainty branch till convergence. We train on batches formed by 36/24/16 images with 256/512/768 pairs of correspondences. Since the feature matcher may return varying number of correspondences, we achieve this by randomly repeat or discard some correspondences. In addition to keypoint locations, attempts were also made to provide the descriptors and scores as network inputs, which yield similar performances, so are omitted.

We would also like to mention here that the block sparsity of the Jacobian is leveraged for efficient computation, as typically done in bundle adjustment [7]. The geometric inverse variance is set to 0, i.e. infinitely large uncertainty, if the geometric method completely fails.

References

[1] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 1

[2] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NIPS*, 2017. 1

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[4] Maria Klodt and Andrea Vedaldi. Supervising the new with the old: learning sfm from sfm. In *ECCV*, 2018. 1

[5] Tristan Laidlow, Jan Czarnowski, and Stefan Leutenegger. Deepfusion: real-time dense 3d reconstruction for monocular slam using single-view depth and gradient predictions. In *ICRA*, 2019. 1

[6] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 1

[7] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *International workshop on vision algorithms*. Springer, 1999. 3

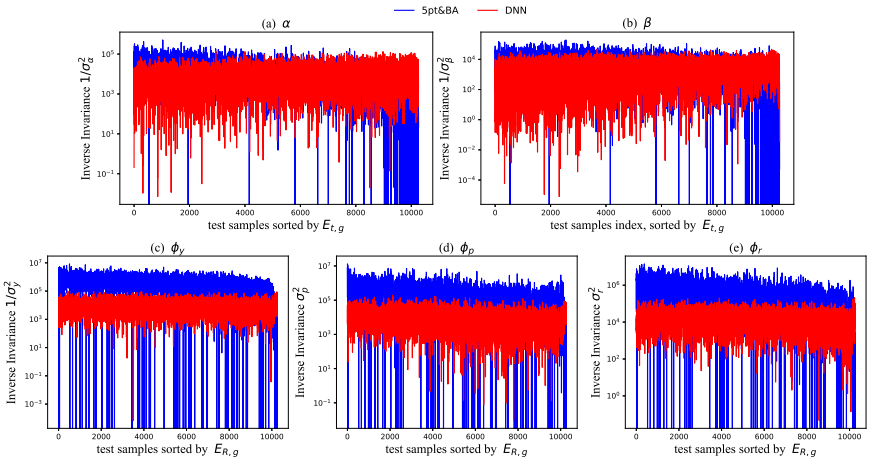

Figure S5. Visualization of the geometric and DNN uncertainty for each test sample on DeMoN.

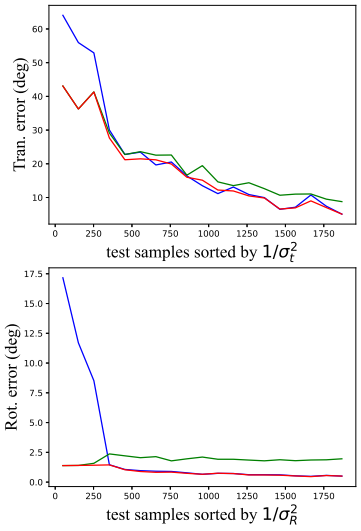

Figure S6. Error comparison between geometric and DNN predictions on ScanNet.

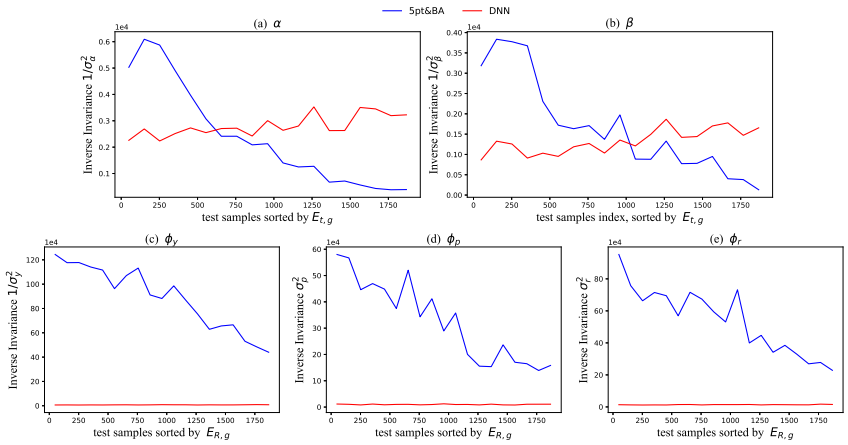

Figure S7. Comparison between the geometric and DNN uncertainty for each parameter in $\{\boldsymbol{\theta_t}, \boldsymbol{\theta_R}\}$ on ScanNet.

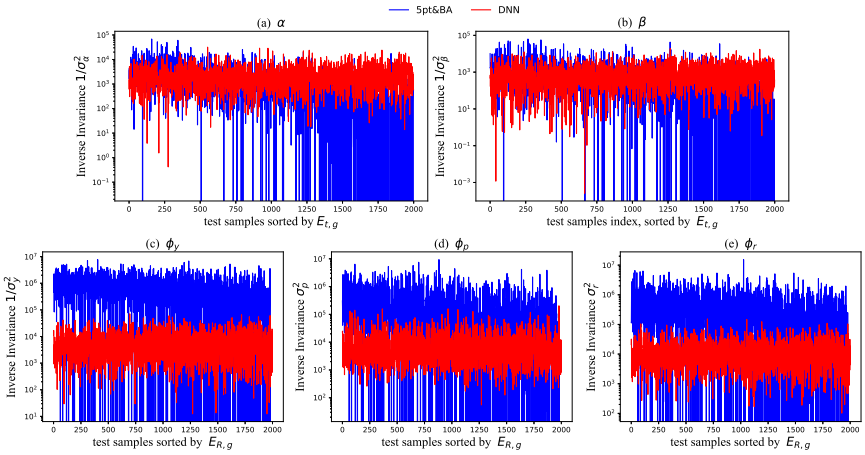

Figure S8. Visualization of the geometric and DNN uncertainty for each test sample on ScanNet.