Equalizing Eq. 3 and Eq. 5, we have
\[ \text{Recalling Eq. 7 of the main paper, we already have} \]
\[ \text{(Conversely, if Eq. 2 holds and)} \]
\[ \text{expressed as} \]
\[ \text{θ} \]
\[ \text{latent state} \]

2. Supplementary Derivation of the Bayesian Formulation

This section supplements the derivation of the distributions 8 & 9 in the main paper.

Let us denote the bivariate Gaussian distribution of the latent state θ_t and the innovation e_t conditional on I_{t-1} as
\[ \left[ \begin{array}{c} \theta_t \\ e_t \end{array} \right] | I_{t-1} \sim \mathcal{N} \left[ \begin{array}{c} \mu_1 \\ \mu_2 \end{array} , \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{bmatrix} \right] , \quad (1) \]
where \( \Sigma_{12} = \Sigma_{21}^T \). Based on the multivariate statistics theorems [1], the conditional distribution of \( \theta_t \) given \( e_t \) is expressed as \( \theta_t | e_t, I_{t-1} \sim \mathcal{N} \left( \theta_t, \Sigma_t \right) \)
\[ \mathcal{N}(\mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (e_t - \mu_2), \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}) , \quad (2) \]
and similarly, \( (e_t | \theta_t, I_{t-1}) \sim \mathcal{N}(e_t, \Sigma_t) \)
\[ \mathcal{N}(\mu_2 + \Sigma_{21} \Sigma_{11}^{-1} (\theta_t - \mu_1), \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}) . \quad (3) \]
Conversely, if Eq. 2 holds and \( \theta_t | I_{t-1} \sim \mathcal{N}(\mu_1, \Sigma_{11}) \), Eq. 1 will also hold according to [1]. Since we have had \( \theta_t | I_{t-1} \sim \mathcal{N}(\theta_t, \Sigma_t) \) in Eq. 4 of the main paper, we can note that
\[ \mu_1 = \tilde{\theta}_t , \quad \text{and} \quad \Sigma_{11} = R_t , \quad (4) \]
Recalling Eq. 7 of the main paper, we already have
\[ (e_t | \theta_t, I_{t-1}) \sim \mathcal{N}(\theta_t - \tilde{\theta}_t , V_t) . \quad (5) \]
Equalizing Eq. 3 and Eq. 5, we have
\[ \mu_2 = 0 , \]
\[ \Sigma_{12} = \Sigma_{21} = R_t , \]
\[ \Sigma_{22} = V_t + R_t . \quad (6) \]

1. Full Network Architecture

As a supplement to the main paper, we detail the parameters of the layers of SCoordNet and OFlowNet used for training 7scenes in Table 5 at the end of the supplementary material.

Figure 1: (a) The confusion matrix of 19 scenes given by our uncertainty predictions. The redder a block (i, j), the more likely it is that the images of the j-th scene belong to the i-th scene. (b) The CDFs of scene coordinate errors given by SCoordNet and OFlowNet with or without uncertainty modeling.

Substituting the variables of Eq. 4 & 6 into Eq. 1 & 2, we have reached the distributions 8 & 9 in the main paper.

3. Ablation Study on the Uncertainty Modeling

The uncertainty modeling, which helps to quantify the measurement and process noise, is an indispensable component of KFNet. In this section, we conduct ablation studies on it.

First, we run the trained KFNet of each scene from 7scenes and 12scenes over the test images of each scene exhaustively and visualize the median uncertainties as the confusion matrix in Fig. 1(a). The uncertainties between the same scene in the main diagonal are much lower than those between different scenes. It indicates that meaningful uncertainties are learned which can be used for scene recognition. Second, we qualitatively compare SCoordNet and OFlowNet against their counterparts which are trained with L2 loss without uncertainty modeling. The cumulative distribution functions (CDFs) of scene coordinate errors tested on 7scenes and 12scenes are shown in Fig. 1(b). The uncer-
Table 1: The parameters of 7-th to 12-th layers of SCoordNet w.r.t. different downsample rates and receptive fields. The number before comma is kernel size, while the one after comma is stride.

| Downsample Rate | Receptive field | Layers (kernel, stride) |
|-----------------|----------------|------------------------|
| 8               | 29             | L7: 1.2, 1.1, 1 | L8: 1, 1, 1               |
| 8               | 45             | L9: 1, 1, 1               |
| 8               | 61             | L10: 1, 1, 1               |
| 8               | 93             | L11: 1, 1, 1               |
| 8               | 125            | L12: 1, 1, 1               |
| 8               | 157            | 1, 1, 1               |
| 8               | 189            | 1, 1, 1               |
| 8               | 221            | 1, 1, 1               |
| 4               | 93             | 1, 1, 1               |
| 8               | 93             | 1, 1, 1               |
| 16              | 93             | 1, 1, 1               |
| 32              | 93             | 1, 1, 1               |

Table 2: The performance of SCoordNet w.r.t. the receptive field. The pose accuracy means the percentage of poses with rotation and translation errors less than 5° and 5cm, respectively.

| Receptive field | Relocalization accuracy | Mapping accuracy |
|-----------------|-------------------------|------------------|
|                 | pose error | pose accuracy | mean | stddev |
| 29              | 0.025m, 0.87° | 89.9%   | 29.6cm | 32.3 |
| 45              | 0.023m, 0.88° | 93.4%   | 24.6cm | 29.2 |
| 61              | 0.023m, 0.84° | 90.0%   | 17.3cm | 23.1 |
| 93              | 0.024m, 0.91° | 92.9%   | 11.5cm | 16.4 |
| 125             | 0.026m, 0.95° | 88.3%   | 11.7cm | 16.1 |
| 157             | 0.026m, 0.97° | 86.6%   | 10.3cm | 15.0 |
| 189             | 0.030m, 1.07° | 81.0%   | 10.3cm | 13.9 |
| 221             | 0.031m, 1.22° | 71.8%   | 9.5cm  | 12.9 |

Figure 2: The cumulative distribution function of scene coordinate errors w.r.t. different receptive field R. A smaller R generally has a denser distribution of errors smaller than 2cm as well as larger than 20cm. The more predictions with errors smaller than 2cm contribute to the accuracy of pose determination, while the larger number of outlier predictions with errors larger than 20cm hamper the robustness of relocalization.

4. Ablation Study on the Receptive Field

The receptive field, denoted as R, is an essential factor of Convolutional Neural Network (CNN) design. In our case, it determines how many image observations around a pixel are exposed and used for scene coordinate prediction. Here, we would like to evaluate the impact of R on the performance of SCoordNet. SCoordNet presented in the main paper has R = 93. We change the kernel size of 7-th to 12-th layers of SCoordNet to adjust the receptive field to 29, 45, 61, 125, 157, 189, 221, as shown in Table 1. Due to the time limitations, the evaluation only runs on heads of 7scenes dataset [7]. As reported in Table 2, the mean of scene coordinate errors grows up as the receptive field R decreases. We illustrate the CDF of scene coordinate errors in Fig. 2. It is noteworthy that a smaller R results in more outlier predictions which cause a larger mean of scene coordinate errors. However, a larger mean of scene coordinate error does not necessarily lead to a decrease in relocalization accuracy. For example, a receptive field of 61 has worse mapping accuracy than the larger receptive fields, but it achieves the smaller pose error and the better pose accuracy than them. As we can see from Fig. 2, a smaller receptive field has a larger portion of precise scene coordinate predictions, especially those with errors smaller than 2cm. These predictions are crucial to the accuracy of pose determination, as the outlier predictions are generally filtered by RANSAC. Nevertheless, when we further reduce R from 61 to 45 and then 29, a drop of relocalization accuracy is observed. It is because, as R decreases, the growing number of outlier predictions deteriorates the robustness of pose computation. A receptive field between 45 and 93 is a good choice that respects the trade-off between precision and robustness.

5. Ablation Study on the Downsample Rate

Due to the cost of dense predictions over full-resolution images, we predict scene coordinates for the images downsized by a factor of 8 in the main paper, following previous works [2]. In this section, we intend to explore how the downsample rate affects the trade-off between accuracy and efficiency over SCoordNet. As reported in Table 1, we change the kernel size and strides of 7-th to 12-th layers to adjust the downsample rate to 4, 8, 16 and 32 with the same receptive field of 93. The mean accuracy and the average time taken to localize frames of heads are reported in Table 3. As intuitively expected, the larger downsample rate generally leads to a drop of relocalization and mapping accuracy, as well as an increasing speed. For example, the
Table 3: The performance of SCoordNet w.r.t. the downsample rate. The pose accuracy means the percentage of poses with rotation and translation errors less than $5^\circ$ and 5cm, respectively.

downsample rate 4 and 8 have a comparable performance, while the downsample rate 8 outperforms 16 by a large margin. However, on the upside, a larger downsample rate is appealing due to the higher efficiency which scales quadratically with the downsample rate. For real-time applications, a downsample rate of 32 allows for a low latency of 34ms per frame with a frequency of about 30 Hz\(^1\).

6. Running Time of KFNet Subsystems

Table 4 reports the mean running time per frame (of size $640 \times 480$) of the measurement, process and filtering systems and NIS test, on a NVIDIA GTX 1080 Ti. Since the measurement and process systems are independent and can run in parallel, the total time per frame is 157.18 ms, which means KFNet only causes an extra overhead of 0.58 ms compared to the one-shot SCoordNet. Besides, our KFNet is 3 times faster than the state-of-the-art one-shot relocalization system DSAC++ [2].

| Modules    | KFNet       | DSAC++      |
|------------|-------------|-------------|
| Time (ms)  | Measurement | Process     | Filtering  | NIS | Total  | -       |
| 156.60     | 51.23       | 0.29        | 0.29       | 157.18 | 486.07 |

Table 4: Running time of the subsystems of KFNet.

7. Mapping Visualization

As a supplement of Fig. 5 in the main paper, we visualize the point clouds of 7scenes [7], 12scenes [8] and Cambridge [4] predicted by DSAC++ [2] and our KFNet-filtered in Fig. 3.

---

\(^1\)All the experiments of this work run on a machine with a 8-core Intel i7-4770K, a 32GB memory and a NVIDIA GTX 1080 Ti graphics card.
Figure 3: Point clouds of all the scenes predicted by DSAC++ [2] and our KFNet-filtered. Zoom in for better view.
| Input   | Layer                        | Output   | Output Size   |
|---------|------------------------------|----------|---------------|
| $I_t$   | Conv+ReLU, $K=3x3$, $S=1$, $F=64$ | conv1a   | $H \times W \times 64$ |
| conv1a  | Conv+ReLU, $K=3x3$, $S=1$, $F=64$ | conv1b   | $H \times W \times 64$ |
| conv1b  | Conv+ReLU, $K=3x3$, $S=2$, $F=256$ | conv2a   | $H/2 \times W/2 \times 256$ |
| conv2a  | Conv+ReLU, $K=3x3$, $S=1$, $F=256$ | conv2b   | $H/2 \times W/2 \times 256$ |
| conv2b  | Conv+ReLU, $K=3x3$, $S=2$, $F=512$ | conv3a   | $H/4 \times W/4 \times 512$ |
| conv3a  | Conv+ReLU, $K=3x3$, $S=1$, $F=512$ | conv3b   | $H/4 \times W/4 \times 512$ |
| conv3b  | Conv+ReLU, $K=3x3$, $S=2$, $F=1024$ | conv4a   | $H/8 \times W/8 \times 1024$ |
| conv4a  | Conv+ReLU, $K=3x3$, $S=1$, $F=1024$ | conv4b   | $H/8 \times W/8 \times 1024$ |
| conv4b  | Conv+ReLU, $K=3x3$, $S=1$, $F=512$ | conv5    | $H/8 \times W/8 \times 512$ |
| conv5   | Conv+ReLU, $K=3x3$, $S=1$, $F=256$ | conv6    | $H/8 \times W/8 \times 256$ |
| conv6   | Conv+ReLU, $K=1x1$, $S=1$, $F=128$ | conv7    | $H/8 \times W/8 \times 128$ |
| conv7   | Conv+Exp, $K=1x1$, $F=3$ | $z_t$    | $H/8 \times W/8 \times 3$ |

| SCoordNet |                           |          |               |
|-----------|---------------------------|----------|---------------|
| $I_{t-1}$ | Conv+ReLU, $K=3x3$, $S=1$, $F=16$ | feat1   | $2 \times H \times W \times 16$ |
| feat1    | Conv+ReLU, $K=3x3$, $S=2$, $F=32$ | feat2   | $2 \times H/2 \times W/2 \times 32$ |
| feat2    | Conv+ReLU, $K=3x3$, $S=1$, $F=32$ | feat3   | $2 \times H/2 \times W/2 \times 32$ |
| feat3    | Conv+ReLU, $K=3x3$, $S=2$, $F=64$ | feat4   | $2 \times H/4 \times W/4 \times 64$ |
| feat4    | Conv+ReLU, $K=3x3$, $S=1$, $F=64$ | feat5   | $2 \times H/4 \times W/4 \times 64$ |
| feat5    | Conv+ReLU, $K=3x3$, $S=2$, $F=128$ | feat6   | $2 \times H/8 \times W/8 \times 128$ |
| feat6    | Conv, $K=3x3$, $S=1$, $F=32$ | $F_{t-1} \parallel F_t$ | $2 \times H/8 \times W/8 \times 32$ |
| $F_{t-1}$ | Cost Volume Constructor | vol1     | $H/8 \times W/8 \times w \times w \times 32$ |
| vol1     | Reshape                   | vol2     | $N \times w \times w \times 32$ |
| vol2     | Conv+ReLU, $K=3x3$, $S=1$, $F=32$ | vol3     | $N \times w \times w \times 32$ |
| vol3     | Conv+ReLU, $K=3x3$, $S=2$, $F=32$ | vol4     | $N \times w/2 \times w/2 \times 32$ |
| vol4     | Conv+ReLU, $K=3x3$, $S=1$, $F=32$ | vol5     | $N \times w/2 \times w/2 \times 32$ |
| vol5     | Conv+ReLU, $K=3x3$, $S=2$, $F=64$ | vol6     | $N \times w/4 \times w/4 \times 64$ |
| vol6     | Conv+ReLU, $K=3x3$, $S=1$, $F=64$ | vol7     | $N \times w/4 \times w/4 \times 64$ |
| vol7     | Conv+ReLU, $K=3x3$, $S=2$, $F=128$ | vol8     | $N \times w/8 \times w/8 \times 128$ |
| vol8     | Conv+ReLU, $K=3x3$, $S=1$, $F=128$ | vol9     | $N \times w/8 \times w/8 \times 128$ |
| vol9     | Deconv+ReLU, $K=3x3$, $S=2$, $F=64$ | vol10    | $N \times w/4 \times w/4 \times 64$ |
| vol10    | Deconv+ReLU, $K=3x3$, $S=1$, $F=64$ | vol11    | $N \times w/4 \times w/4 \times 64$ |
| vol11    | Deconv+ReLU, $K=3x3$, $S=2$, $F=32$ | vol12    | $N \times w/2 \times w/2 \times 32$ |
| vol12    | Deconv+ReLU, $K=3x3$, $S=1$, $F=32$ | vol13    | $N \times w/2 \times w/2 \times 32$ |
| vol13    | Deconv+ReLU, $K=3x3$, $S=2$, $F=16$ | vol14    | $N \times w \times w \times 16$ |
| vol14    | Deconv+ReLU, $K=3x3$, $S=1$, $F=16$ | vol15    | $N \times w \times w \times 16$ |
| vol15    | Conv, $K=3x3$, $S=1$, $F=1$ | confidence | $N \times w \times w \times 1$ |
| confidence | Spatial Softmax [3] | flow1    | $N \times 2$ |
| flow1    | Reshape                   | flow2    | $H/8 \times W/8 \times 2$ |
| flow2, $\theta_{t-1} \parallel \Sigma_{t-1}$ | Flow-guided Warping [9, 10, 5, 6] | $\hat{\theta}_t \parallel \Sigma_t$ | $H/8 \times W/8 \times 4$ |
| vol9     | Reshape                   | fc1      | $N \times 2w^d$ |
| fc1      | FC+ReLU, $F=64$ | fc2      | $N \times 64$ |
| fc2      | FC+ReLU, $F=32$ | fc3      | $N \times 32$ |
| fc3      | FC+Exp, $F=1$ | fc4      | $N \times 1$ |
| fc4      | Reshape                   | $W_t$    | $H/8 \times W/8 \times 1$ |

Table 5: The full architecture of the proposed SCoordNet and OFlowNet. "||" denotes concatenation along i-th dimension.
References

[1] TW Anderson. An introduction to multivariate statistical analysis. 1984.

[2] Eric Brachmann and Carsten Rother. Learning less is more-6d camera localization via 3d surface regression. In CVPR, 2018.

[3] Chelsea Finn, Xin Yu Tan, Yan Duan, Trevor Darrell, Sergey Levine, and Pieter Abbeel. Learning visual feature spaces for robotic manipulation with deep spatial autoencoders. 2015.

[4] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In ICCV, 2015.

[5] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly supervised action localization by sparse temporal pooling network. In CVPR, 2018.

[6] Tomas Pfister, James Charles, and Andrew Zisserman. Flowing convnets for human pose estimation in videos. In ICCV, 2015.

[7] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In CVPR, 2013.

[8] Julien Valentin, Angela Dai, Matthias Nießner, Pushmeet Kohli, Philip Torr, Shahram Izadi, and Cem Keskin. Learning to navigate the energy landscape. In 3DV, 2016.

[9] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. In ICCV, 2017.

[10] Zheng Zhu, Wei Wu, Wei Zou, and Junjie Yan. End-to-end flow correlation tracking with spatial-temporal attention. In CVPR, 2018.