Supplementary Material of KFNet: Learning Temporal Camera Relocalization using Kalman Filtering

Lei Zhou[1] Zixin Luo[1] Tianwei Shen[1] Jiahui Zhang[2]
Mingmin Zhen[1] Yao Yao[1] Tian Fang[3] Long Quan[1]
[1]Hong Kong University of Science and Technology [2]Tsinghua University [3]Everest Innovation Technology
[1]{lzhouai,zluoag,tshenaa,mzhen,yyaoag,quan}@cse.ust.hk
[2]jiahui-z15@mails.tsinghua.edu.cn [3]fangtian@altizure.com

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.

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 $\boldsymbol{\theta}_t$ and the innovation $\mathbf{e}_t$ conditional on $\mathcal{I}_{t-1}$ as

$$\left[\begin{pmatrix}\boldsymbol{\theta}_t\\ \mathbf{e}_t\end{pmatrix}\middle|\,\mathcal{I}_{t-1}\right] \sim \mathcal{N}\left[\begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix}\right], \quad (1)$$

where $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}{}^T$. Based on the multivariate statistics theorems [1], the conditional distribution of $\boldsymbol{\theta}_t$ given $\mathbf{e}_t$ is expressed as $(\boldsymbol{\theta}_t|\mathbf{e}_t, \mathcal{I}_{t-1}) \sim$

$$\mathcal{N}(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{e}_t - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}), \quad (2)$$

and similarly, $(\mathbf{e}_t|\boldsymbol{\theta}_t, \mathcal{I}_{t-1}) \sim$

$$\mathcal{N}(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{\theta}_t - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}). \quad (3)$$

Conversely, if Eq. 2 holds and $(\boldsymbol{\theta}_t|\mathcal{I}_{t-1}) \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$, Eq. 1 will also hold according to [1]. Since we have had $(\boldsymbol{\theta}_t|\mathcal{I}_{t-1}) \sim \mathcal{N}(\hat{\boldsymbol{\theta}}_t^-, \mathbf{R}_t)$ in Eq. 4 of the main paper, we can note that

$$\boldsymbol{\mu}_1 = \hat{\boldsymbol{\theta}}_t^-, \quad \text{and} \quad \boldsymbol{\Sigma}_{11} = \mathbf{R}_t. \quad (4)$$

Recalling Eq. 7 of the main paper, we already have

$$(\mathbf{e}_t|\boldsymbol{\theta}_t, \mathcal{I}_{t-1}) \sim \mathcal{N}(\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_t^-, \mathbf{V}_t). \quad (5)$$

Equalizing Eq. 3 and Eq. 5, we have

$$\begin{aligned}\boldsymbol{\mu}_2 &= \mathbf{0},\\ \boldsymbol{\Sigma}_{12} &= \boldsymbol{\Sigma}_{21} = \mathbf{R}_t,\\ \boldsymbol{\Sigma}_{22} &= \mathbf{V}_t + \mathbf{R}_t.\end{aligned} \quad (6)$$

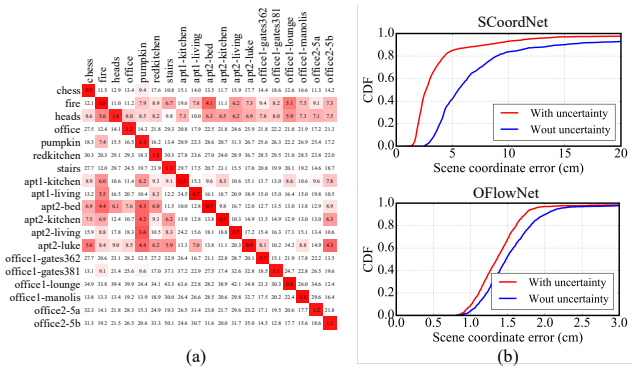

Figure 1: (a) The confusion matrix of 19 scenes given by our uncertainty predictions. The redder a block (i, j), the more likely it is that the images of the j-th scene belong to the i-th scene. (b) The CDFs of scene coordinate errors given by SCoordNet and OFlowNet with or without uncertainty modeling.

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-

Downsample Rate	Receptive field	Layers (kernel, stride)					
		L7	L8	L9	L10	L11	L12
8	29	1, 2	1, 1	1, 1	1, 1	1, 1	1, 1
8	45	3, 2	1, 1	1, 1	1, 1	1, 1	1, 1
8	61	3, 2	3, 1	1, 1	1, 1	1, 1	1, 1
8	93	3, 2	3, 1	3, 1	3, 1	1, 1	1, 1
8	125	3, 2	3, 1	3, 1	3, 1	3, 1	3, 1
8	157	3, 2	3, 1	5, 1	5, 1	3, 1	3, 1
8	189	3, 2	3, 1	5, 1	5, 1	5, 1	5, 1
8	221	3, 2	3, 1	7, 1	7, 1	5, 1	5, 1
4	93	3, 1	3, 1	5, 1	5, 1	3, 1	3, 1
8	93	3, 2	3, 1	3, 1	3, 1	1, 1	1, 1
16	93	3, 2	3, 1	3, 2	1, 1	1, 1	1, 1
32	93	3, 2	3, 1	3, 2	1, 1	1, 2	1, 1

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.

Receptive field	Relocalization accuracy		Mapping accuracy	
	pose error	pose accuracy	mean	stddev
29	0.025m, 0.87°	87.9%	29.6cm	32.3
45	0.023m, 0.88°	93.4%	24.4cm	29.2
61	**0.023m, 0.84°**	**94.0%**	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**

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.

tainty modeling leads to more accurate predictions for both SCoordNet and OFlowNet. We attribute the improvements to the fact that the uncertainties apply auto-weighting to the loss term of each pixel as in Eqs. 10 & 14 of the main paper, which prevents the learning from getting stuck in the hard or infeasible examples like the boundary pixels for SCoordNet and the occluded pixels for OFlowNet (see Fig. 2 of the main paper).

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

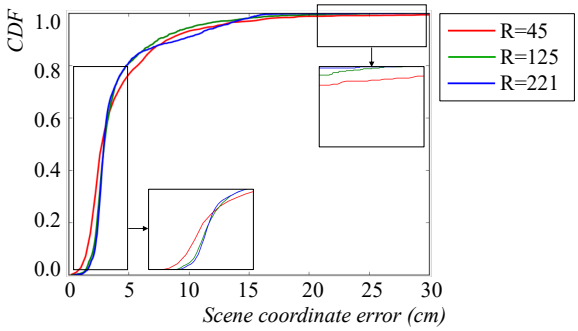

Figure 2: The cumulative distribution function of scene coordinate errors w.r.t. different receptive field R. A smaller R generally has a denser distribution of errors smaller than 2cm as well as larger than 20cm. The more predictions with errors smaller than 2cm contribute to the accuracy of pose determination,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

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

Downsample rate	Relocalization accuracy		Mapping accuracy		Time
	pose error	pose accuracy	mean	stddev	
4	**0.024m**, 0.97°	**93.6%**	**11.2cm**	17.3	1.34s
8	0.024m, **0.91°**	92.9%	11.5cm	**16.4**	0.20s
16	0.025m, 0.92°	89.1%	16.3cm	20.5	0.11s
32	0.029m, 1.06°	79.6%	20.7cm	20.7	**0.034s**

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 $5°$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 640×480) 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].

	KFNet					DSAC++
Modules	Measurement	Process	Filtering	NIS	Total	-
Time (ms)	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.

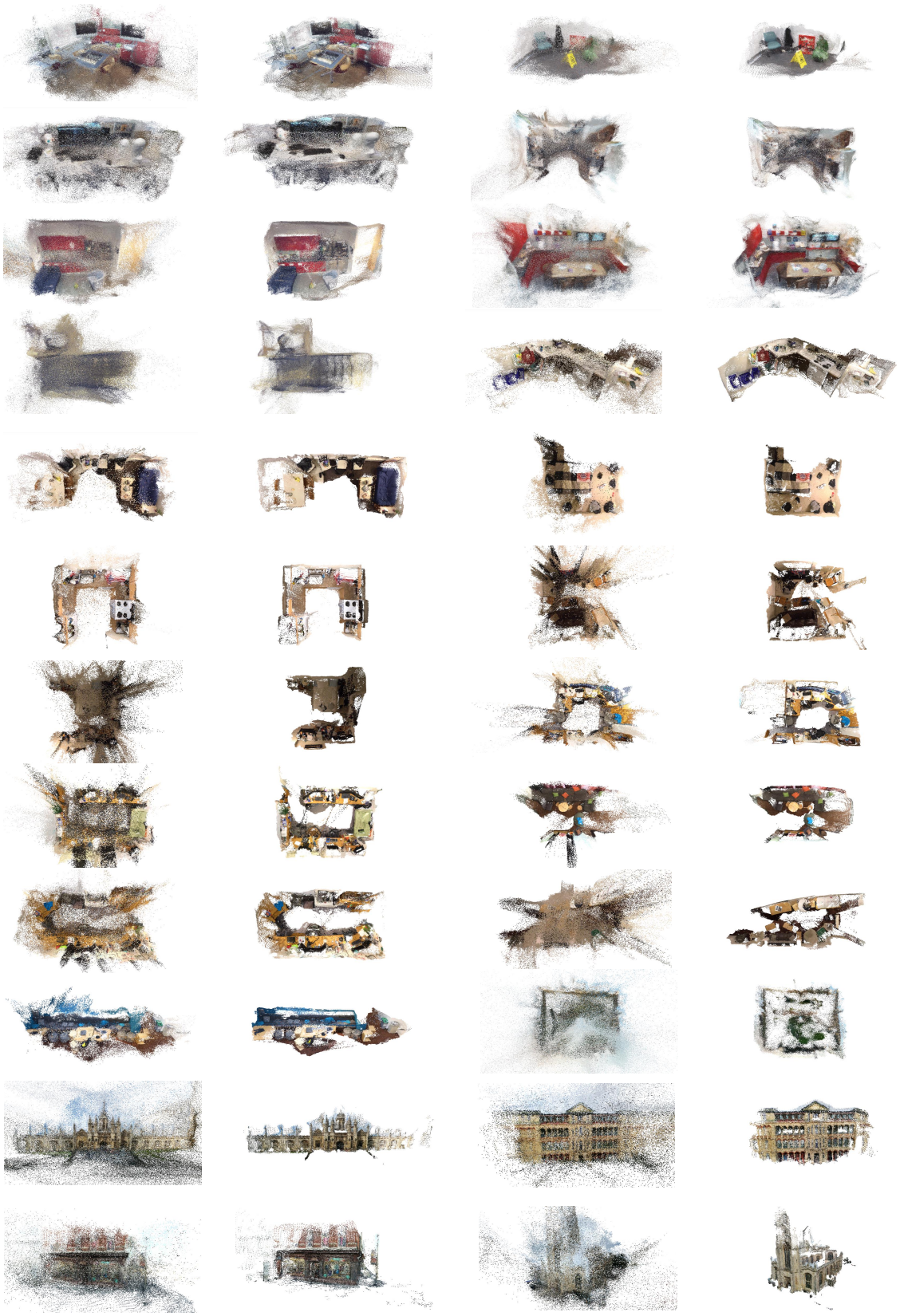

Figure 3: Point clouds of all the scenes predicted by DSAC++ [2] and our KFNet-filtered. Zoom in for better view.

Input	Layer	Output	Output Size
SCoordNet			
$\mathbf{I}_t$	Conv+ReLU, K=3x3, S=1, F=64	conv1a	$\mathrm{H} \times \mathrm{W} \times 64$
conv1a	Conv+ReLU, K=3x3, S=1, F=64	conv1b	$\mathrm{H} \times \mathrm{W} \times 64$
conv1b	Conv+ReLU, K=3x3, S=2, F=256	conv2a	$\mathrm{H}/2 \times \mathrm{W}/2 \times 256$
conv2a	Conv+ReLU, K=3x3, S=1, F=256	conv2b	$\mathrm{H}/2 \times \mathrm{W}/2 \times 256$
conv2b	Conv+ReLU, K=3x3, S=2, F=512	conv3a	$\mathrm{H}/4 \times \mathrm{W}/4 \times 512$
conv3a	Conv+ReLU, K=3x3, S=1, F=512	conv3b	$\mathrm{H}/4 \times \mathrm{W}/4 \times 512$
conv3b	Conv+ReLU, K=3x3, S=2, F=1024	conv4a	$\mathrm{H}/8 \times \mathrm{W}/8 \times 1024$
conv4a	Conv+ReLU, K=3x3, S=1, F=1024	conv4b	$\mathrm{H}/8 \times \mathrm{W}/8 \times 1024$
conv4b	Conv+ReLU, K=3x3, S=1, F=512	conv5	$\mathrm{H}/8 \times \mathrm{W}/8 \times 512$
conv5	Conv+ReLU, K=3x3, S=1, F=256	conv6	$\mathrm{H}/8 \times \mathrm{W}/8 \times 256$
conv6	Conv+ReLU, K=1x1, S=1, F=128	conv7	$\mathrm{H}/8 \times \mathrm{W}/8 \times 128$
conv7	Conv, K=1x1, S=1, F=3	$\mathbf{z}_t$	$\mathrm{H}/8 \times \mathrm{W}/8 \times 3$
conv7	Conv+Exp, K=1x1, S=1, F=1	$\mathbf{V}_t$	$\mathrm{H}/8 \times \mathrm{W}/8 \times 1$
OFlowNet			
$\mathbf{I}_{t-1} \Vert_0 \mathbf{I}_t$	Conv+ReLU, K=3x3, S=1, F=16	feat1	$2 \times \mathrm{H} \times \mathrm{W} \times 16$
feat1	Conv+ReLU, K=3x3, S=2, F=32	feat2	$2 \times \mathrm{H}/2 \times \mathrm{W}/2 \times 32$
feat2	Conv+ReLU, K=3x3, S=1, F=32	feat3	$2 \times \mathrm{H}/2 \times \mathrm{W}/2 \times 32$
feat3	Conv+ReLU, K=3x3, S=2, F=64	feat4	$2 \times \mathrm{H}/4 \times \mathrm{W}/4 \times 64$
feat4	Conv+ReLU, K=3x3, S=1, F=64	feat5	$2 \times \mathrm{H}/4 \times \mathrm{W}/4 \times 64$
feat5	Conv+ReLU, K=3x3, S=2, F=128	feat6	$2 \times \mathrm{H}/8 \times \mathrm{W}/8 \times 128$
feat6	Conv, K=3x3, S=1, F=32	$\mathbf{F}_{t-1} \Vert_0 \mathbf{F}_t$	$2 \times \mathrm{H}/8 \times \mathrm{W}/8 \times 32$
$\mathbf{F}_{t-1} \Vert_0 \mathbf{F}_t$	Cost Volume Constructor	vol1	$\mathrm{H}/8 \times \mathrm{W}/8 \times \mathrm{w} \times \mathrm{w} \times 32$
vol1	Reshape	vol2	$\mathrm{N} \times \mathrm{w} \times \mathrm{w} \times 32 \;\; (\mathrm{N} = \mathrm{HW}/64)$
vol2	Conv+ReLU, K=3x3, S=1, F=32	vol3	$\mathrm{N} \times \mathrm{w} \times \mathrm{w} \times 32$
vol3	Conv+ReLU, K=3x3, S=2, F=32	vol4	$\mathrm{N} \times \mathrm{w}/2 \times \mathrm{w}/2 \times 32$
vol4	Conv+ReLU, K=3x3, S=1, F=32	vol5	$\mathrm{N} \times \mathrm{w}/2 \times \mathrm{w}/2 \times 32$
vol5	Conv+ReLU, K=3x3, S=2, F=64	vol6	$\mathrm{N} \times \mathrm{w}/4 \times \mathrm{w}/4 \times 64$
vol6	Conv+ReLU, K=3x3, S=1, F=64	vol7	$\mathrm{N} \times \mathrm{w}/4 \times \mathrm{w}/4 \times 64$
vol7	Conv+ReLU, K=3x3, S=2, F=128	vol8	$\mathrm{N} \times \mathrm{w}/8 \times \mathrm{w}/8 \times 128$
vol8	Conv+ReLU, K=3x3, S=1, F=128	vol9	$\mathrm{N} \times \mathrm{w}/8 \times \mathrm{w}/8 \times 128$
vol9	Deconv+ReLU, K=3x3, S=2, F=64	vol10	$\mathrm{N} \times \mathrm{w}/4 \times \mathrm{w}/4 \times 64$
vol10 $\Vert_3$ vol7	Conv+ReLU, K=3x3, S=1, F=64	vol11	$\mathrm{N} \times \mathrm{w}/4 \times \mathrm{w}/4 \times 64$
vol11	Deconv+ReLU, K=3x3, S=2, F=32	vol12	$\mathrm{N} \times \mathrm{w}/2 \times \mathrm{w}/2 \times 32$
vol12 $\Vert_3$ vol5	Conv+ReLU, K=3x3, S=1, F=32	vol13	$\mathrm{N} \times \mathrm{w}/2 \times \mathrm{w}/2 \times 32$
vol13	Deconv+ReLU, K=3x3, S=2, F=16	vol14	$\mathrm{N} \times \mathrm{w} \times \mathrm{w} \times 16$
vol14 $\Vert_3$ vol3	Conv+ReLU, K=3x3, S=1, F=16	vol15	$\mathrm{N} \times \mathrm{w} \times \mathrm{w} \times 16$
vol15	Conv, K=3x3, S=1, F=1	confidence	$\mathrm{N} \times \mathrm{w} \times \mathrm{w} \times 1$
confidence	Spatial Softmax [3]	flow1	$\mathrm{N} \times 2$
flow1	Reshape	flow2	$\mathrm{H}/8 \times \mathrm{W}/8 \times 2$
flow2, $\hat{\boldsymbol{\theta}}_{t-1} \Vert_3 \boldsymbol{\Sigma}_{t-1}$	Flow-guided Warping [9, 10, 5, 6]	$\hat{\boldsymbol{\theta}}_t^- \Vert_3 \boldsymbol{\Sigma}_t^-$	$\mathrm{H}/8 \times \mathrm{W}/8 \times 4$
vol9	Reshape	fc1	$\mathrm{N} \times 2\mathrm{w}^2$
fc1	FC+ReLU, F=64	fc2	$\mathrm{N} \times 64$
fc2	FC+ReLU, F=32	fc3	$\mathrm{N} \times 32$
fc3	FC+Exp, F=1	fc4	$\mathrm{N} \times 1$
fc4	Reshape	$\mathbf{W}_t$	$\mathrm{H}/8 \times \mathrm{W}/8 \times 1$

Table 5: The full architecture of the proposed SCoordNet and OFlowNet. "$\Vert_i$" 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.