FaVoR: Features via Voxel Rendering for Camera Relocalization

Vincenzo Polizzi[1] Marco Cannici[2] Davide Scaramuzza[2] Jonathan Kelly[1]
[1] University of Toronto, [2] University of Zurich
{vincenzo.polizzi, jonathan.kelly}@robotics.utias.utoronto.ca, cannici@ifi.uzh.ch

Supplementary Material

We report on the results obtained by FaVoR at various iterations of the PnP-RANSAC scheme in in Appendix A. In Appendix B, we discuss the tradeoff between the voxel resolution and both the rendering capabilities and matching performance of our system. We also report the similarity responses for the Cambridge Landmarks [4] dataset, discussing the evidence of a lack of accuracy in the landmark triangulation in Appendix C. Finally, we provide more details on our training losses and our landmark triangulation method in Appendices D and E, respectively.

A. Extended Analysis of FaVoR Performance and Error Computation

In this section, we report the pose estimation errors obtained with different feature extractors at various iterations of the iterative PnP-RANSAC scheme. Specifically, we report the values for Alike-t, Alike-s, Alike-n, Alike-l [8] and SuperPoint [2] with 64, 94, 128, 128, and 256 channels descriptors, respectively.

Table 2 for the 7-Scenes [5] and Table 3 for the Cambrdige [4] datasets give the median pose estimate at the 1st, 2nd, and 3rd iterations of PnP-RANSAC, and the respective average number of inlier points per image (used to compute the pose estimate). The tables also report the success rates of the PnP-RANSAC iterative scheme at the various iterations, i.e., the ratio between the number of successful estimates and the total number of queries. The data shows a clear trend. Namely, the average number of inliers per image increases as the estimated camera pose converges towards the true query image pose (i.e., as the pose estimate error decreases). Furthermore, although there is a difference in matching performance between the various Alike networks (as reported in the Alike [8] manuscript), our descriptor representation effectively 'flattens' these differences, enabling robust matching despite view changes.

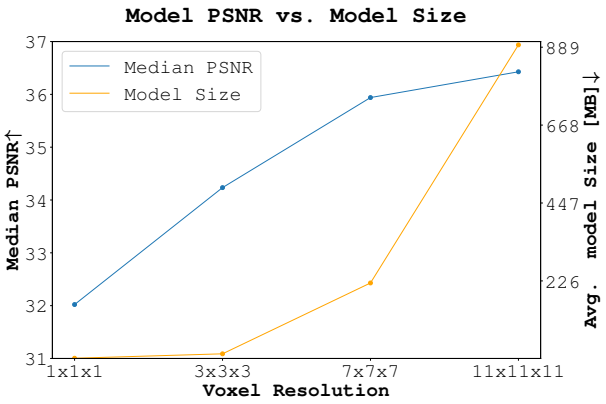

Figure 1. **PSNR and model size versus grid resolution.** We report the median peak signal-to-noise ratio (PSNR) and the average checkpoint size for FaVoR $_{\text{Alike-l}}$ at different grid resolutions of the voxel representation.

B. PSNR versus Voxel Resolution

The number of sub-voxels in each voxel representing a landmark determines the grid resolution, $R \times R \times R$. The grid resolution directly impacts the rendering quality of the descriptor patches, increasing or decreasing the peak signal-to-noise ratio (PSNR) values. The PSNR calculation is given by

$$\text{PSNR} = 10 \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right), \tag{1}$$

where MAX = 2 is the maximum span of the descriptor values, i.e, in the range (-1, 1), and MSE is the mean squared error between the ground-truth patch and the rendered patch. A grid resolution of $R \times R \times R$ implies that the grid contains $R \cdot R \cdot R$ nodes, where each node (vertex) encodes C channels (equal to the number of channels of the descriptor). The chosen grid resolution impacts the overall model size.

The plot in Figure 1 shows the median PSNR values at different grid resolutions and the corresponding model size on the chess scene of the 7-Scenes dataset [5]. The graph

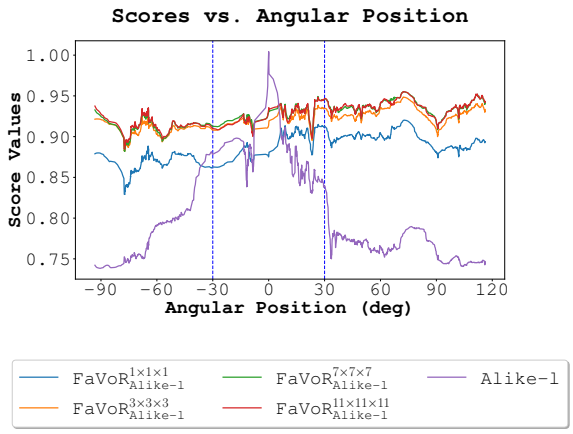

Figure 2. **Similarity response scores versus grid resolution at different view angles.** We compare the different grid resolutions' capacity to provide high score similarity score results at different view angles. Higher scores lead to better matching meaning that the rendered descriptors properly match the appearance of the ones extracted by Alike-l [8].

shows that beyond a certain grid resolution, the improvement in terms of PSNR is decreasing while, in contrast, the model size grows exponentially. Therefore, as a tradeoff between model size and *good* rendering capabilities of our representation, we choose a grid resolution of $3 \times 3 \times 3$ for our model. Also, this resolution choice is supported by the median score values, obtained as described in the manuscript in Section 4.4, reported in Figure 2. Figure 2 shows the median score values obtained in the chess scene of the 7-Scenes dataset [5] with Alike-l [8], at different grid resolutions. We note that grid resolutions greater than $1 \times 1 \times 1$ yield a similar score response; hence, we choose the lower resolution to save on memory.

C. Score Response

In Figure 3 we show the response score map obtained by convolving the dense descriptor map extracted using Alike-l [8] with two descriptors rendered using FaVoR $_{\text{Alike-l}}$. In particular, we draw a red circle centred at the projection of the triangulated landmarks on the camera plane. We add another circle (in blue) that is centred at the coordinates of the pixel with the strongest similarity response. Both the circles should be concentric to provide an accurate pose estimate. However, we notice that, most of the time, the circles do not have the same centre for the samples obtained from the Cambridge Landmarks dataset [4]. This misalignment may be due to an imprecise triangulation of the landmarks, given the depth uncertainty for the large scenes of the Cambridge Landmarks dataset [4].

D. Training and Losses

To train our voxel representation of the descriptor patches, we used two main losses, the squared L2 norm loss and the cosine similarity loss. We begin by training our model using the squared L2 norm and cosine similarity losses to enforce that the direction and norm of the rendered descriptors match the ground truth. The cosine similarity loss is calculated as

$$\mathrm{L}_{\cos}\left(\hat{\mathbf{d}}_{ij}^{uv}, \mathbf{d}_{ij}^{uv}\right) = 1 - \frac{\hat{\mathbf{d}}_{ij}^{uv} \cdot \mathbf{d}_{ij}^{uv}}{||\hat{\mathbf{d}}_{ij}^{uv}|| \cdot ||\mathbf{d}_{ij}^{uv}||}, \tag{2}$$

where $\hat{\mathbf{d}}_{ij}^{uv}$ is the rendered descriptor and $\mathbf{d}_{ij}^{uv}$ is the one extracted by $\mathcal{F}$ from the patch $\mathbf{P}_{ij}$. For the density grid, we use the cross-entropy loss, as in [6]. Also, to ensure a smooth representation of the descriptor patch, for the last 500 epochs, we introduce a total variation (TV) regularization term in the loss computation on the density and the descriptor parameters as described in [6]. The complete loss function for the voxel optimization is given by

$$\begin{aligned}\text{Loss}\left(\hat{\mathbf{d}}_{ij}^{uv}, \mathbf{d}_{ij}^{uv}\right) = \\ ||\hat{\mathbf{d}}_{ij}^{uv} - \mathbf{d}_{ij}^{uv}||_2^2 + \mathrm{L}_{\cos}\left(\hat{\mathbf{d}}_{ij}^{uv}, \mathbf{d}_{ij}^{uv}\right) + \text{TV}.\end{aligned} \tag{3}$$

During training, we choose a learning rate that depends on the visibility of each sub-voxel. In particular, we follow the same approach proposed by Sun *et al*. [6] where subvoxels visible from fewer views have a lower learning rate. Choosing the learning rate according to the visibility of the voxels allows training to *focus* more on the parts of the voxels that represent portions that have been observed from several views, and hence are more reliable than those with fewer observations. Figure 4 shows a descriptor patch, the corresponding ground truth extracted using Alike-l [8], both visualized using principal component analysis (PCA) to map the descriptor space to RGB colors, and the L2 norm between the two patches in the descriptor space.

E. Landrmark Triangulation

Our method requires landmarks positions to locate and train the associated voxels. Vision-based localization systems, such as visual-inertial odometry or visual simultaneous localization and mapping, already provide a landmark's position in 3D space. Hence, we opt for a simple-to-implement multi-view triangulation approach, since triangulation is not the main focus of our work. Given a track containing N poses $\mathbf{T}_i$, with $i = 1 \dots N$, and hence N keypoints $\mathbf{k}_{ij}$ corresponding to the projection onto each camera plane of the landmark $\boldsymbol{\ell}_j$, we wish to find the 3D coordinates ${}^{\mathcal{W}}\ell_j^x, {}^{\mathcal{W}}\ell_j^y, {}^{\mathcal{W}}\ell_j^z$ in the world frame $\mathcal{W}$ of $\boldsymbol{\ell}_j$. An initial estimate of the coordinates $\boldsymbol{\ell}_j$ can be determined by two-view

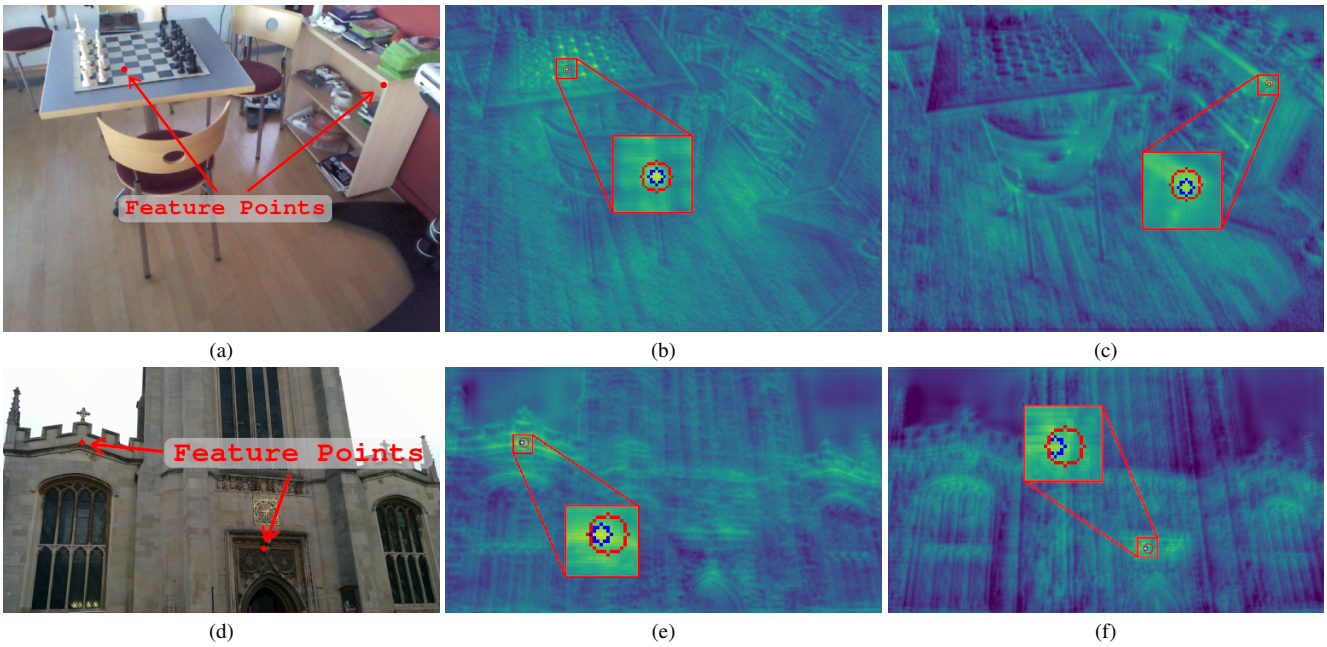

Figure 3. **Visualization of similarity response.** We render a feature tracked during training using the Alike-l descriptor from an unseen view on the 7-Scenes [5] and the Cambridge Landmarks [4] datasets. On the left, a) and d) display the ground truth positions of the rendered feature points, obtained by projecting the triangulated landmarks on the camera plane, in red. At the same time, b), e) and c), f) show the similarity response between the rendered features and the target image dense descriptor map. The yellow color indicates a strong response, concentrated around the feature positions shown in a), demonstrating the effectiveness of our descriptor rendering approach. The small circle in blue is the circle center at the highest score response, the red circle is centered at the project landmark position.

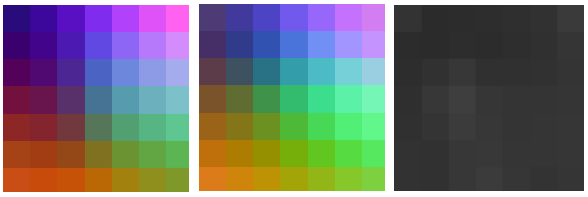

Figure 4. **Visualization of rendered vs ground truth descriptor patch.** We report the rendered descriptor patch a) and the corresponding ground-truth b) for Alike-l [8], compressed from 128 channels to 3 using PCA for visualization purposes. The patch in c) represents the normalized difference between the rendered and the ground truth, darker is better.

triangulation methods, such as the direct linear transform. Following [1], we choose an anchor pose $\mathbf{T}_a$ from among the poses in the track, and express $\boldsymbol{\ell}_j$ in camera coordinates for $\mathbf{T}_a$. We define $\mathbf{T}_a \in \mathrm{SE}(3)$, and in general any $\mathbf{T}_i$, in terms of a rotation matrix $\mathbf{R}_a \in \mathrm{SO}(3)$ and a translation vector $\mathbf{t}_a \in \mathbb{R}^3$:

$$\begin{bmatrix} {}^{\mathcal{A}}\ell_j^x \\ {}^{\mathcal{A}}\ell_j^y \\ {}^{\mathcal{A}}\ell_j^z \end{bmatrix} = \mathbf{R}_a^T \begin{bmatrix} {}^{\mathcal{W}}\ell_j^x \\ {}^{\mathcal{W}}\ell_j^y \\ {}^{\mathcal{W}}\ell_j^z \end{bmatrix} - \mathbf{R}_a^T \mathbf{t}_a. \tag{4}$$

From Equation (4), we can write the landmark coordinates in the world frame as a function of the anchor pose coordinates:

$$\begin{bmatrix} {}^{\mathcal{W}}\ell_j^x \\ {}^{\mathcal{W}}\ell_j^y \\ {}^{\mathcal{W}}\ell_j^z \end{bmatrix} = \mathbf{R}_a \begin{bmatrix} {}^{\mathcal{A}}\ell_j^x \\ {}^{\mathcal{A}}\ell_j^y \\ {}^{\mathcal{A}}\ell_j^z \end{bmatrix} + \mathbf{t}_a. \tag{5}$$

Hence, each time we need to determine the landmark coordinates $\boldsymbol{\ell}_j$ in camera frame $\mathcal{T}_i$, associated with the pose $\mathbf{T}_i$ in the track, we can write:

$$\begin{bmatrix} {}^{\mathcal{T}_i}\ell_j^x \\ {}^{\mathcal{T}_i}\ell_j^y \\ {}^{\mathcal{T}_i}\ell_j^z \end{bmatrix} = \mathbf{R}_i^T \left(\mathbf{R}_a \begin{bmatrix} {}^{\mathcal{A}}\ell_j^x \\ {}^{\mathcal{A}}\ell_j^y \\ {}^{\mathcal{A}}\ell_j^z \end{bmatrix} + \mathbf{t}_a \right) - \mathbf{R}_i^T \mathbf{t}_i \tag{6}$$

$$= \mathbf{R}_i^T \mathbf{R}_a \begin{bmatrix} {}^{\mathcal{A}}\ell_j^x \\ {}^{\mathcal{A}}\ell_j^y \\ {}^{\mathcal{A}}\ell_j^z \end{bmatrix} + \mathbf{R}_i^T (\mathbf{t}_a - \mathbf{t}_i) \tag{7}$$

To improve the numerical stability of the optimization process, we represent $\left[{}^{\mathcal{A}}\ell_j^x, {}^{\mathcal{A}}\ell_j^y, {}^{\mathcal{A}}\ell_j^z \right]^T$ using the inverse

depth parametrization,

$$\alpha_j = \frac{{}^{\mathcal{A}}\ell_j^x}{{}^{\mathcal{A}}\ell_j^z}, \quad \beta_j = \frac{{}^{\mathcal{A}}\ell_j^y}{{}^{\mathcal{A}}\ell_j^z}, \quad \rho_j = \frac{1}{{}^{\mathcal{A}}\ell_j^z} \tag{8}$$

We can then rewrite Equation (7) as

$$\begin{bmatrix} {}^{\mathcal{T}_i}\ell_j^x \\ {}^{\mathcal{T}_i}\ell_j^y \\ {}^{\mathcal{T}_i}\ell_j^z \end{bmatrix} = \frac{1}{\rho_j} \left(\mathbf{R}_i^T \mathbf{R}_a \begin{bmatrix} \alpha_j \\ \beta_j \\ 1 \end{bmatrix} + \rho_j \, \mathbf{R}_i^T (\mathbf{t}_a - \mathbf{t}_i) \right) \tag{9}$$

In turn, the camera measurement model is

$$\hat{\mathbf{z}}_{ij} = \frac{1}{{}^{\mathcal{T}_i}\ell_j^z} \begin{bmatrix} {}^{\mathcal{T}_i}\ell_j^x, {}^{\mathcal{T}_i}\ell_j^y \end{bmatrix}^T, \tag{10}$$

where $\hat{\mathbf{z}}_{ij}$ are the normalized image plane coordinates of ${}^{\mathcal{T}_i}\ell_j$. The predicted measurement can be determined by transforming $\mathbf{k}_{ij}$ into camera coordinates to obtain $\mathbf{z}_{ij}$. This involves back-projecting the keypoint coordinates $\mathbf{k}_{ij}$ from the image plane to the camera frame, followed by normalization,

$$\begin{bmatrix} x_{\mathbf{k}_{ij}} \\ y_{\mathbf{k}_{ij}} \\ z_{\mathbf{k}_{ij}} \end{bmatrix} = \mathbf{K}^{-1} \begin{bmatrix} \mathbf{k}_{ij} \\ 1 \end{bmatrix} \tag{11}$$

$$\mathbf{z}_{ij} = \frac{1}{z_{\mathbf{k}_{ij}}} \begin{bmatrix} x_{\mathbf{k}_{ij}}, y_{\mathbf{k}_{ij}} \end{bmatrix}^T, \tag{12}$$

where $\mathbf{K}$ is the intrinsic camera calibration matrix.

Finally, we find α_j, β_j, and ρ_j via Levenberg-Marquardt optimization,

$$\mathbf{e}_{ij} = \mathbf{z}_{ij} - \hat{\mathbf{z}}_{ij}, \tag{13}$$

$$u_{ij} = \sqrt{\mathbf{e}_{ij}^T \mathbf{e}_{ij}}, \tag{14}$$

$$\rho(u) = \frac{1}{2} \frac{c^2 u^2}{c^2 + u^2}, \tag{15}$$

$$\alpha_j^*, \beta_j^*, \rho_j^* = \underset{\alpha_j, \beta_j, \rho_j}{\operatorname{argmin}} \sum_{i \in \mathcal{S}_j} \rho(u_{ij}(\mathbf{e}_{ij}(\hat{\mathbf{z}}_{ij}(\boldsymbol{\ell}_{ij}(\alpha_j, \beta_j, \rho_j))))), \tag{16}$$

where $\rho(u)$ is a robust cost function [3] parameterized by c, used to prevent outliers from negatively impacting the estimate of the landmark coordinates.

F. Robustness to Pose Initialization Error

In Table 1 we report the 6-DoF median localization errors for the 7-Scenes [5] dataset using two pose initialization methods: the first frame of the test sequence (Constant) and DenseVLAD [7] (Retrieval). We perform the evaluation using FaVoR coupled with Alike-l [8]. The 'first frame' initialization choice is equivalent to adding increased noise to the starting guess, with increasing error as the target pose moves far away from the initial pose (at the first frame). However, this approach does ensure reproducibility and provides a consistent baseline for fair comparisons with future work, offering a reliable measure of our method's robustness. Our results indicate that after three iterations of the Render+PnP-RANSAC paradigm, our method converges to a low localization error, even when starting from less accurate poses than those provided by DenseVLAD [7].

References

[1] Jeff Delaune, David S. Bayard, and Roland Brockers. Range-visual-inertial odometry: Scale observability without excitation. *IEEE Robot. Autom. Lett.*, 6(2):2421–2428, 2021. 3

[2] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *IEEE Conf. Comput. Vis. Pattern Recog. Workshops (CVPRW)*, pages 224–236, 2018. 1

[3] Donald Geman and Stuart Geman. Bayesian image analysis. In *IEEE Trans. Syst. Man, Cybern. A, Syst., Humans*, pages 301–319. Springer, 1986. 4

[4] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Int. Conf. Comput. Vis. (ICCV)*, pages 2938–2946, 2015. 1, 2, 3, 10

[5] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew W. Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 2930–2937, 2013. 1, 2, 3, 4, 5, 8

[6] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022. 2

[7] Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 1808–1817, 2015. 4

[8] Xiaoming Zhao, Xingming Wu, Jinyu Miao, Weihai Chen, Peter C. Y. Chen, and Zhengguo Li. Alike: Accurate and lightweight keypoint detection and descriptor extraction. *IEEE Trans. Robot.*, March 2022. 1, 2, 3, 4

Scene	Method	Prior Err.	Iter 1	Iter 2	Iter 3
chess	Retrieval	21.9, 12.13	0.8, 0.21	0.7, 0.19	0.7, 0.18
	Constant	147.6, 29.94	1.0, 0.28	0.7, 0.19	0.7, 0.18
fire	Retrieval	34.4, 13.2	0.8, 0.3	0.9, 0.4	0.8, 0.3
	Constant	96.6, 35.6	1.2, 0.5	0.9, 0.4	0.8, 0.3
heads	Retrieval	15.8, 15.0	0.7, 0.5	0.7, 0.4	0.7, 0.4
	Constant	45.7, 37.8	28.7, 17.2	2.5, 1.5	0.5, 0.4
office	Retrieval	28.6, 11.1	1.7, 0.4	1.7, 0.4	1.6, 0.4
	Constant	113.8, 67.4	158.3, 57.4	10.2, 2.5	1.7, 0.4
pumpkin	Retrieval	31.4, 10.8	1.4, 0.3	1.4, 0.3	1.3, 0.3
	Constant	137.0, 49.7	4.1, 1.1	1.5, 0.3	1.2, 0.2
redkitchen	Retrieval	29.4, 12.0	1.2, 0.3	1.2, 0.3	1.2, 0.2
	Constant	192.4, 39.3	7.8, 1.8	1.8, 0.4	1.1, 0.2
stairs	Retrieval	26.2, 15.8	3.8, 1.0	3.0, 0.8	2.7, 0.8
	Constant	178.9, 16.3	231.4, 34.2	106.8, 16.2	4.7, 1.32

Table 1. **Different pose initialization priors for 7-Scenes dataset [5]**. We report the 6-DoF median pose errors (cm, deg) obtained with FaVoR $_{\text{Alike-l}}$ for different pose initialization methods. The results show that FaVoR is robust to different initial poses as it converges to small localization errors after iterating the Render+PnP-RANSAC scheme.

Scene	Iteration	Feature Extractor	Initial pose error (cm, deg)	Estimated pose error (cm, deg)	Avg. N. of inliers	Success rate (%)
chess	1st	alike-l	21.92, 12.13	0.77, 0.21	66	100.00
	2nd		-	0.74, 0.19	74	100.00
	3rd		-	0.72, 0.18	74	100.00
	1st	alike-n	21.92, 12.13	0.73, 0.21	64	100.00
	2nd		-	0.68, 0.18	71	100.00
	3rd		-	0.64, 0.17	71	100.00
	1st	alike-s	21.92, 12.13	0.82, 0.26	117	100.00
	2nd		-	0.76, 0.22	122	100.00
	3rd		-	0.71, 0.20	122	100.00
	1st	alike-t	21.92, 12.13	1.01, 0.29	66	100.00
	2nd		-	0.94, 0.27	73	100.00
	3rd		-	0.89, 0.25	73	100.00
	1st	SuperPoint	0.22, 12.13	0.68, 0.19	88	100.00
	2nd		-	0.67, 0.17	99	100.00
	3rd		-	0.64, 0.16	99	100.00
fire	1st	alike-l	34.37, 13.23	0.82, 0.34	73	100.00
	2nd		-	0.88, 0.36	72	100.00
	3rd		-	0.83, 0.34	72	99.85
	1st	alike-n	34.37, 13.23	0.86, 0.37	66	100.00
	2nd		-	0.93, 0.37	66	100.00
	3rd		-	0.89, 0.35	66	99.60
	1st	alike-s	34.37, 13.23	1.22, 0.47	172	100.00
	2nd		-	1.64, 0.60	162	100.00
	3rd		-	1.52, 0.56	163	99.70
	1st	alike-t	34.37, 13.23	1.17, 0.47	59	100.00
	2nd		-	1.37, 0.52	55	100.00
	3rd		-	1.30, 0.49	55	99.55
	1st	SuperPoint	34.37, 13.23	1.02, 0.39	67	100.00
	2nd		-	1.04, 0.38	67	100.00
	3rd		-	0.98, 0.36	67	98.65
heads	1st	alike-l	15.77, 14.97	0.73, 0.46	46	100.00
	2nd		-	0.67, 0.41	53	100.00
	3rd		-	0.66, 0.40	53	94.50
	1st	alike-n	15.77, 14.97	1.08, 0.59	38	100.00
	2nd		-	0.97, 0.53	43	100.00
	3rd		-	0.96, 0.56	43	91.30
	1st	alike-s	15.77, 14.97	0.70, 0.43	81	100.00
	2nd		-	0.62, 0.37	92	100.00
	3rd		-	0.59, 0.36	92	99.20
	1st	alike-t	15.77, 14.97	0.89, 0.52	52	100.00
	2nd		-	0.81, 0.48	59	100.00
	3rd		-	0.76, 0.44	59	98.90
	1st	SuperPoint	15.77, 14.97	0.62, 0.39	76	100.00
	2nd		-	0.54, 0.34	87	100.00
	3rd		-	0.52, 0.32	88	99.20

Continue on next page

office	1st	alike-l	28.58, 11.06	1.69, 0.43	36	100.00
	2nd		-	1.68, 0.41	39	100.00
	3rd		-	1.63, 0.39	40	99.25
	1st	alike-n	28.58, 11.06	1.77, 0.47	35	100.00
	2nd		-	1.74, 0.45	37	100.00
	3rd		-	1.69, 0.42	37	97.78
	1st	alike-s	28.58, 11.06	1.63, 0.45	69	100.00
	2nd		-	1.56, 0.41	72	100.00
	3rd		-	1.55, 0.40	72	99.98
	1st	alike-t	28.58, 11.06	2.57, 0.68	37	100.00
	2nd		-	2.26, 0.61	42	100.00
	3rd		-	2.21, 0.58	42	99.15
	1st	SuperPoint	28.58, 11.06	1.75, 0.43	65	100.00
	2nd		-	1.71, 0.41	72	100.00
	3rd		-	1.64, 0.37	72	99.75
pumpkin	1st	alike-l	31.38, 10.81	1.39, 0.30	69	100.00
	2nd		-	1.39, 0.29	76	100.00
	3rd		-	1.31, 0.28	76	98.65
	1st	alike-n	31.38, 10.81	1.58, 0.36	70	100.00
	2nd		-	1.53, 0.34	76	100.00
	3rd		-	1.46, 0.31	76	93.45
	1st	alike-s	31.38, 10.81	1.38, 0.31	118	100.00
	2nd		-	1.38, 0.29	121	100.00
	3rd		-	1.34, 0.28	122	99.25
	1st	alike-t	31.38, 10.81	1.87, 0.43	81	100.00
	2nd		-	1.70, 0.39	90	100.00
	3rd		-	1.67, 0.37	91	96.40
	1st	SuperPoint	31.38, 10.81	1.50, 0.33	110	100.00
	2nd		-	1.51, 0.31	120	100.00
	3rd		-	1.45, 0.29	120	99.05
redkitchen	1st	alike-l	29.38, 11.97	1.23, 0.30	45	100.00
	2nd		-	1.18, 0.25	54	100.00
	3rd		-	1.15, 0.24	54	98.08
	1st	alike-n	29.38, 11.97	1.37, 0.33	51	100.00
	2nd		-	1.34, 0.32	60	100.00
	3rd		-	1.21, 0.28	60	96.02
	1st	alike-s	29.38, 11.97	4.66, 1.09	57	100.00
	2nd		-	4.28, 1.00	67	100.00
	3rd		-	4.03, 0.94	68	77.42
	1st	alike-t	29.38, 11.97	1.44, 0.31	57	100.00
	2nd		-	1.39, 0.29	66	100.00
	3rd		-	1.33, 0.27	67	99.38
	1st	SuperPoint	0.29, 11.97	1.38, 0.30	79	100.00
	2nd		-	1.42, 0.27	93	100.00
	3rd		-	1.33, 0.24	93	98.74

Continue on next page

stairs	1st	alike-l	26.19, 15.81	3.80, 1.03	11	100.00
	2nd		-	3.02, 0.81	12	100.00
	3rd		-	2.74, 0.82	12	97.90
	1st	alike-n	26.19, 15.81	7.19, 1.97	10	100.00
	2nd		-	6.28, 1.65	10	100.00
	3rd		-	5.96, 1.59	10	93.10
	1st	alike-s	26.19, 15.81	5.78, 1.63	70	100.00
	2nd		-	5.18, 1.49	68	100.00
	3rd		-	5.03, 1.51	68	100.00
	1st	alike-t	26.19, 15.81	5.30, 1.49	14	100.00
	2nd		-	4.38, 1.20	15	100.00
	3rd		-	4.03, 1.07	15	100.00
	1st	SuperPoint	26.19, 15.81	5.83, 1.69	27	100.00
	2nd		-	4.54, 1.21	31	100.00
	3rd		-	4.05, 1.07	31	99.90
Overall Average	1st	alike-t	26.80, 12.85	2.03, 0.60	52	99.89
	2nd		-	1.83, 0.54	57	99.24
	3rd		-	1.74, 0.50	57	99.05
	1st	alike-s	26.80, 12.85	2.31, 0.66	97	97.83
	2nd		-	2.20, 0.63	100	96.83
	3rd		-	2.11, 0.61	101	96.51
	1st	alike-n	26.80, 12.85	2.08, 0.62	47	97.18
	2nd		-	1.93, 0.55	51	96.28
	3rd		-	1.83, 0.52	51	95.89
	1st	alike-l	26.80, 12.85	1.49, 0.44	49	99.25
	2nd		-	1.37, 0.39	54	98.65
	3rd		-	1.29, 0.38	54	98.32
	1st	SuperPoint	26.80, 12.85	1.82, 0.53	73	99.81
	2nd		-	1.63, 0.44	81	99.49
	3rd		-	1.52, 0.40	81	99.33

Table 2. **6-DoF median localization errors on the 7-Scenes dataset [5]** for the various features extractors used to train FaVoR.

Scene	Iteration	Feature Extractor	Initial pose error (cm, deg)	Estimated pose error (cm, deg)	Avg. N. of inliers	Success rate (%)
Shop	1st	alike-l	136.31, 7.19	5.38, 0.27	208	100.00
	2nd		-	5.63, 0.22	231	100.00
	3rd		-	5.48, 0.25	231	100.00
	1st	alike-n	136.31, 7.19	5.27, 0.28	185	100.00
	2nd		-	5.43, 0.23	208	100.00
	3rd		-	5.09, 0.24	208	100.00
	1st	alike-s	136.31, 7.19	5.99, 0.24	225	100.00
	2nd		-	5.76, 0.25	250	100.00
	3rd		-	6.05, 0.25	250	100.00
	1st	alike-t	136.31, 7.19	5.65, 0.27	203	100.00
	2nd		-	5.89, 0.26	224	100.00
	3rd		-	5.25, 0.25	224	100.00
	1st	SuperPoint	136.31, 7.19	5.87, 0.29	204	100.00
	2nd		-	5.20, 0.26	225	100.00
	3rd		-	5.47, 0.26	224	100.00
College	1st	alike-l	289.98, 5.96	18.19, 0.25	359	100.00
	2nd		-	17.04, 0.26	373	100.00
	3rd		-	15.25, 0.23	372	100.00
	1st	alike-n	289.98, 5.96	16.82, 0.28	315	100.00
	2nd		-	17.38, 0.28	327	100.00
	3rd		-	17.61, 0.26	327	100.00
	1st	alike-s	289.98, 5.96	16.64, 0.27	327	100.00
	2nd		-	15.74, 0.26	338	100.00
	3rd		-	15.67, 0.24	338	100.00
	1st	alike-t	289.98, 5.96	17.64, 0.28	326	100.00
	2nd		-	16.33, 0.26	336	100.00
	3rd		-	16.52, 0.25	337	100.00
	1st	superpoint	289.98, 5.96	17.88, 0.27	326	100.00
	2nd		-	18.15, 0.28	336	100.00
	3rd		-	17.52, 0.27	336	100.00
Great	1st	alike-l	719.21, 9.47	32.46, 0.16	103	100.00
	2nd		-	29.48, 0.15	116	100.00
	3rd		-	27.40, 0.14	116	99.87
	1st	alike-n	719.21, 9.47	37.88, 0.21	82	100.00
	2nd		-	35.20, 0.18	91	100.00
	3rd		-	32.05, 0.18	91	99.21
	1st	alike-s	719.21, 9.47	36.18, 0.19	94	100.00
	2nd		-	34.27, 0.18	105	100.00
	3rd		-	31.78, 0.16	106	99.87
	1st	alike-t	719.21, 9.47	33.78, 0.19	101	100.00
	2nd		-	31.33, 0.15	114	100.00
	3rd		-	28.83, 0.14	114	100.00
	1st	SuperPoint	719.21, 9.47	34.69, 0.22	142	100.00
	2nd		-	30.71, 0.20	161	100.00
	3rd		-	29.09, 0.20	161	100.00

Continue on next page

Hospital	1st	alike-l	405.22, 7.58	22.18, 0.44	155	100.00
	2nd		-	21.37, 0.40	160	100.00
	3rd		-	19.37, 0.36	160	100.00
	1st	alike-n	405.22, 7.58	27.28, 0.47	128	100.00
	2nd		-	22.68, 0.44	132	100.00
	3rd		-	21.17, 0.40	131	100.00
	1st	alike-s	405.22, 7.58	25.13, 0.44	132	100.00
	2nd		-	25.71, 0.47	136	100.00
	3rd		-	20.75, 0.37	136	100.00
	1st	alike-t	405.22, 7.58	26.30, 0.51	140	100.00
	2nd		-	25.10, 0.48	145	100.00
	3rd		-	20.14, 0.41	145	100.00
	1st	SuperPoint	405.22, 7.58	31.53, 0.56	143	100.00
	2nd		-	31.05, 0.55	148	100.00
	3rd		-	27.46, 0.54	148	100.00
Church	1st	alike-l	287.61, 9.36	11.58, 0.38	201	100.00
	2nd		-	10.31, 0.31	228	100.00
	3rd		-	10.35, 0.30	228	100.00
	1st	alike-n	287.61, 9.36	12.46, 0.43	181	100.00
	2nd		-	11.53, 0.35	206	100.00
	3rd		-	10.90, 0.33	206	100.00
	1st	alike-s	287.61, 9.36	12.67, 0.42	180	100.00
	2nd		-	12.01, 0.36	203	100.00
	3rd		-	11.40, 0.35	204	99.81
	1st	alike-t	287.61, 9.36	12.18, 0.40	170	100.00
	2nd		-	11.66, 0.35	192	100.00
	3rd		-	11.21, 0.36	192	100.00
	1st	SuperPoint	287.61, 9.36	14.15, 0.49	188	100.00
	2nd		-	12.72, 0.42	220	100.00
	3rd		-	11.43, 0.38	220	99.81
Overall Average	1st	alike-t	367.67, 7.91	19.11, 0.33	188	100.00
	2nd		-	18.06, 0.30	202	100.00
	3rd		-	16.39, 0.28	202	100.00
	1st	alike-s	367.67, 7.91	19.32, 0.31	192	100.00
	2nd		-	18.70, 0.30	206	100.00
	3rd		-	17.13, 0.27	207	99.94
	1st	alike-n	367.67, 7.91	19.94, 0.34	178	100.00
	2nd		-	18.44, 0.30	193	99.92
	3rd		-	17.36, 0.28	193	99.84
	1st	alike-l	367.67, 7.91	17.96, 0.30	205	100.00
	2nd		-	16.77, 0.27	222	100.00
	3rd		-	15.57, 0.26	221	99.97
	1st	SuperPoint	367.67, 7.91	20.82, 0.37	201	100.00
	2nd		-	19.57, 0.34	218	99.96
	3rd		-	18.19, 0.33	218	99.96

Table 3. **6-DoF median localization errors on the Cambridge dataset [4]** for the various features extractors used to train FaVoR.