Supplementary Material for Efficient Consensus Maximization for Visual Localization by Globally Optimal Rotation Search

Jiawei Cai, Yanmei Jiao, Dibin Zhou, Xiumei Li, Rong Xiong, and Yue Wang

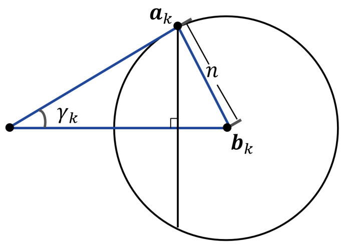
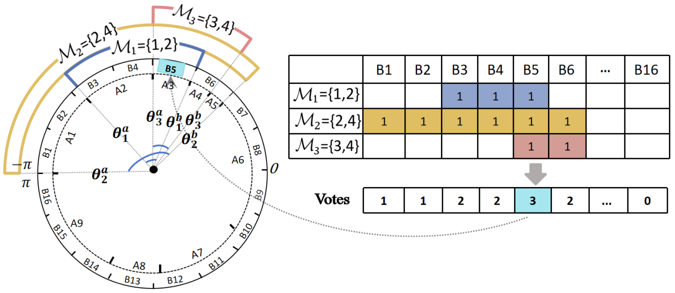


Fig. 1. Relating the Euclidean and the angular error.



	B1	B2	B3	B4	B5	B6	...	B16
$\mathcal{M}_1$={1,2}			1	1	1			
$\mathcal{M}_2$={2,4}	1	1	1	1	1	1		
$\mathcal{M}_3$={3,4}					1	1		
Votes	1	1	2	2	3	2	...	0

Fig. 2. The illustration of proposed voting and acceleration algorithm.

I. Method

A. Rotation search in SO(3)

As shown in Fig. 1, the inlier threshold γ_k (in radians) is data dependent and constraint by

$$\cos\gamma_k = \frac{\|\mathbf{a}_k\|^2 + \|\mathbf{b}_k\|^2 - n^2}{2\|\mathbf{a}_k\|\|\mathbf{b}_k\|}. \tag{1}$$

Therefore, for all inliers that satisfy $|s\tilde{\mathbf{u}}_k - \mathbf{R}\mathbf{p}_k| \leq n$, we can transform the error metric from Euclidean distance to angular distance as

$$\angle(\mathbf{b}_k, \mathbf{R}\mathbf{a}_k) \leq \gamma_k, \quad k \in \mathcal{I}, \tag{2}$$

where $\mathbf{b}_k := s\tilde{\mathbf{u}}_k$, $\mathbf{a}_k := \mathbf{p}_k$.

*1) **Dimensionality reduction of rotation**:* Inspired by GORE, we can achieve dimensionality reduction for rotation. Specifically, the process of iterating over the remaining measurements i to align with measurement k can be detailed as follows. First, the rotation that aligns measurement k according to (2) is identified as $\mathbf{R}_k$. This rotation is then decomposed into two separate rotations to explain it more clearly:

$$\mathbf{R}_k = \mathbf{A}\mathbf{B}, \tag{3}$$

where, $\mathbf{B}$ is a rotation determined by (2), $\mathbf{A}$ represents a rotation by an angle $\theta \in [-\pi, \pi]$ about the axis $\mathbf{B}\mathbf{a}_k$.

Since $\mathbf{A}$ preserves $\mathbf{B}$, the constraint on measurement k imposed by (2) is always satisfied. Then, to enable $\mathbf{R}_k$ to also transform $\mathbf{a}_i$ to $\mathbf{b}_i$, we can solve for $\mathbf{A}$. Specifically, for each measurement i, a corresponding $\mathbf{A}$ can be determined. If we denote the rotation $\mathbf{A}$ that aligns the maximum number of measurements i with measurement k as $\mathring{\mathbf{A}}$, then $\mathring{\mathbf{A}}$ must satisfy the following condition:

$$\max_{\mathbf{R}\in SO(3), \mathcal{I}_k \subseteq \mathcal{M}\setminus\{k\}} |\mathcal{I}_k| + 1$$
$$\text{subject to} \quad \angle(\mathbf{b}_i, \mathring{\mathbf{A}}\mathbf{B}\mathbf{a}_i) \leq \gamma_i, \quad i \in \mathcal{I}_k. \tag{4}$$

For the rotation $\mathbf{B}$, we define the rotation of $\hat{\mathbf{B}}$ perfectly aligned $\mathbf{a}_k$ to $\mathbf{b}_k$, denoted by

$$\mathbf{b}_k = \hat{\mathbf{B}}\mathbf{a}_k. \tag{5}$$

It is important to note that $\hat{\mathbf{B}}$ is not identical to $\mathbf{B}$. The rotation $\hat{\mathbf{B}}$ aligns $\mathbf{a}_k$ and $\mathbf{b}_k$ perfectly, whereas $\mathbf{B}$ is a rotation that satisfies (2), meaning it only aligns $\mathbf{a}_k$ with $\mathbf{b}_k$ within an angular error γ_k. This relationship is also illustrated in Fig. 1(a) in the manuscript. More formally, we can define the following region:

$$S_\gamma(\mathbf{b}) := \{\mathbf{a} \in \mathbb{R}^3 | \|\mathbf{a}\| = 1, \angle(\mathbf{a}, \mathbf{b}) \leqslant \gamma\}. \tag{6}$$

Thus, $S_{\gamma_k}(\mathbf{b}_k)$ represents the spherical region generated by rotating the measurement $\mathbf{b}_k$ within an angular distance of γ_k. Consequently, the relationship $\mathbf{B}\mathbf{a}_k \in S_{\gamma_k}(\mathbf{b}_k)$ is self-evident.

Additionally, to clearly indicate the dependency of $\mathbf{A}$ on the rotation axis $\mathbf{r}$ and angle θ, we now denote it as $\mathbf{A}_{\theta,\mathbf{r}}$:

$$\mathbf{A}_{\theta,\mathbf{r}} = \{exp(\theta[\mathbf{r}]_\times)\}. \tag{7}$$

Then we can denote the possible region obtained by applying such a rotation $\mathbf{A}$ to any unit norm point $\mathbf{p}$ as $circ(\mathbf{p}, \mathbf{r})$:

$$circ(\mathbf{p}, \mathbf{r}) := \{\mathbf{A}_{\theta,\mathbf{r}}\mathbf{p} | \theta \in [-\pi, \pi]\}. \tag{8}$$

Based on the above decomposition of $\mathbf{R}_k$, $\mathbf{B}\mathbf{a}_k$ is the rotation axis of $\mathbf{A}$, so the interior of $S_{\gamma_k}(\mathbf{b}_k)$ is also the set of possible rotation axis of $\mathbf{A}$. Then applying the same rotational procedure to $\mathbf{a}_i$ as illustrated in Fig. 1(a) in the manuscript, the possible region in which $\mathbf{R}_k\mathbf{a}_i = \mathbf{A}_{\theta,\mathbf{r}}\mathbf{B}\mathbf{a}_i$ lies can be denoted as:

$$L_k(\mathbf{a}_i) := \{circ(\mathbf{p}, \mathbf{r}) | \mathbf{p} \in S_{\gamma_k}(\hat{\mathbf{B}}\mathbf{a}_i), \mathbf{r} \in S_{\gamma_k}(\mathbf{b}_k)\}. \tag{9}$$

Therefore, when $L_k(\hat{\mathbf{B}}\mathbf{a}_i) \bigcap S_{\gamma_i}(\mathbf{b}_i) = \emptyset$, the pair $(\mathbf{a}_i, \mathbf{b}_i)$ cannot be aligned by any rotation $\mathbf{R}_k$ satisfying (2) with

$(\mathbf{a}_k, \mathbf{b}_k)$, the corresponding relationship $(\mathbf{a}_i, \mathbf{b}_i)$ can be safely removed without affecting the results. At the same time, the problem of solving for the 3D rotation $\mathbf{R}$ is thus reduced to solving for a one-dimensional angle θ.

*2) **Interval voting**:* A more illustrative voting case is shown in Fig. 2, on the left side, we visually show how $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ sequentially cover different continuous intervals from bin 1 to bin 6. After populating the binary table on the right side of Fig. 2 with these entries, we then count the votes for all bins. It becomes evident that bin 5 corresponds to the maximum intersection of $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$, receiving votes from all three. Therefore, the point pairs corresponding to bin 5 represent the union of the point pairs contained within $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$, which corresponds to the index set of point pairs $\{1, 2, 3, 4\}$.

The acceleration is primarily achieved through the use of a large binary table. However, since this table only stores binary values, it does not require significant space complexity. Additionally, the discretization resolution provides a balance between memory usage and accuracy. We have found that extremely high resolutions often decrease accuracy, as small intervals can disperse the voting results. Noise leads to uncertainty intervals that are similar but not completely identical, allowing for the merging of intervals corresponding to different sets at lower resolutions. Since this approximation only occurs near numerical values, its impact on the final average estimate of the inlier set is limited and acceptable.

*3) **Rotation search algorithm**:* The time complexity analysis of rotation search algorithm is as follows. For a given k, by traversing all $i \neq k$, we obtain K angular intervals $\{\Theta_i\}$. In response to this, we propose an improved discretized interval voting algorithm. Since this method requires frequent operations on a discretized array, we leverage the advantages of difference arrays in such scenarios to further reduce the time complexity. Overall, the time complexity of our algorithm is $O(K + L)$, where L is the length of the discretized interval. Therefore, line 9 in Alg. 2 in the manuscript takes $O(K+L)$ time. Typically, we can evenly divide the interval $[-\pi, \pi]$ using a discretization step of $0.5°$ or $1°$, resulting in L being 360 or 720, respectively. Additionally, Gore uses interval stabbing to handle these K angular intervals, with a complexity of $O(K \log K)$. In the worst case, for all k, the overall time complexity of Alg. 2 in the manuscript is $O(K(K + L))$. Compared to Gore's $O(K^2 \log K)$, the advantage of our algorithm becomes more pronounced as K increases, such as when using dense feature matching like LoFTR [1].

B. Search in SE(3)

*1) **Robustness of rotation search against translation perturbation**:* In this section, we provide the proofs related to (13) and (14) in the manuscript. As defined in the manuscript, $\mathbf{x} := \mathbf{a} - \mathbf{p}_{corg}$ and $\mathbf{x}' := \mathbf{a} - (\mathbf{p}_{corg} + \epsilon)$. Then, for the new pairs $(\mathbf{x}'_i, \mathbf{b}_i)$ and $(\mathbf{x}'_k, \mathbf{b}_k)$, we have:

$$\mathbf{x}'_i = \mathbf{x}_i - \epsilon, \tag{10}$$

$$\mathbf{x}'_k = \mathbf{x}_k - \epsilon. \tag{11}$$

Then, we first need to solve for $\hat{\mathbf{B}}$ that perfectly aligns $\mathbf{x}_k$ to $\mathbf{b}_k$, which allows us to obtain:

$$\mathbf{b}_k = \hat{\mathbf{B}}\mathbf{x}_k. \tag{12}$$

Similarly, after introducing the perturbation ϵ to $\mathbf{p}_{corg}$, it becomes:

$$\mathbf{b}_k = \hat{\mathbf{B}}'\mathbf{x}'_k. \tag{13}$$

Substituting (11) into (13) and combining with (12), we can obtain:

$$\hat{\mathbf{B}}'\epsilon = (\hat{\mathbf{B}}' - \hat{\mathbf{B}})\mathbf{x}_k. \tag{14}$$

Next, for the point pair i, combining (10) and (14), we have:

$$\begin{aligned} \hat{\mathbf{B}}'\mathbf{x}'_i &= \hat{\mathbf{B}}'(\mathbf{x}_i - \epsilon) \\ &= \hat{\mathbf{B}}'\mathbf{x}_i + \hat{\mathbf{B}}\mathbf{x}_k - \hat{\mathbf{B}}'\mathbf{x}_k. \end{aligned} \tag{15}$$

Thus, we have:

$$\begin{aligned} \hat{\mathbf{B}}'\mathbf{x}'_i - \hat{\mathbf{B}}\mathbf{x}_i &= \hat{\mathbf{B}}'\mathbf{x}_i + \hat{\mathbf{B}}\mathbf{x}_k - \hat{\mathbf{B}}'\mathbf{x}_k - \hat{\mathbf{B}}\mathbf{x}_i \\ &= \hat{\mathbf{B}}'\mathbf{x}_i + \hat{\mathbf{B}}\mathbf{x}_k - \hat{\mathbf{B}}'\mathbf{x}_k - \hat{\mathbf{B}}\mathbf{x}_i + (\hat{\mathbf{B}}\mathbf{x}_i - \hat{\mathbf{B}}\mathbf{x}_i) \\ &= \hat{\mathbf{B}}'(\mathbf{x}_i - \mathbf{x}_k) - \hat{\mathbf{B}}(\mathbf{x}_i - \mathbf{x}_k). \end{aligned} \tag{16}$$

Next, based on the triangle inequality, we can derive:

$$\begin{aligned} ||\hat{\mathbf{B}}'\mathbf{x}'_i - \hat{\mathbf{B}}\mathbf{x}_i|| &\leqslant ||(\hat{\mathbf{B}}' - \hat{\mathbf{B}})\mathbf{x}_i|| + ||(\hat{\mathbf{B}}' - \hat{\mathbf{B}})\mathbf{x}_k|| \\ &\leqslant \frac{||\mathbf{x}_i||}{||\mathbf{x}_k||}||\epsilon|| + ||\epsilon|| \\ &\leqslant (1 + \frac{||\mathbf{x}_i||}{||\mathbf{x}_k||})||\epsilon||. \end{aligned} \tag{17}$$

To maintain the consensus relationship between measurements i and k, it is necessary to ensure that

$$G_1 \subseteq G_2. \tag{18}$$

Then based on (17) and (18), we can obtain:

$$(1 + \frac{||\mathbf{x}_i||}{||\mathbf{x}_k||})||\epsilon|| \leqslant r_1 + r, \tag{19}$$

where r_1 is the axis corresponding to angle $\delta(\theta')$, while r is the axis corresponding to angle γ_i. Fig. 4 in the manuscript also provides a more intuitive depiction of this relationship.

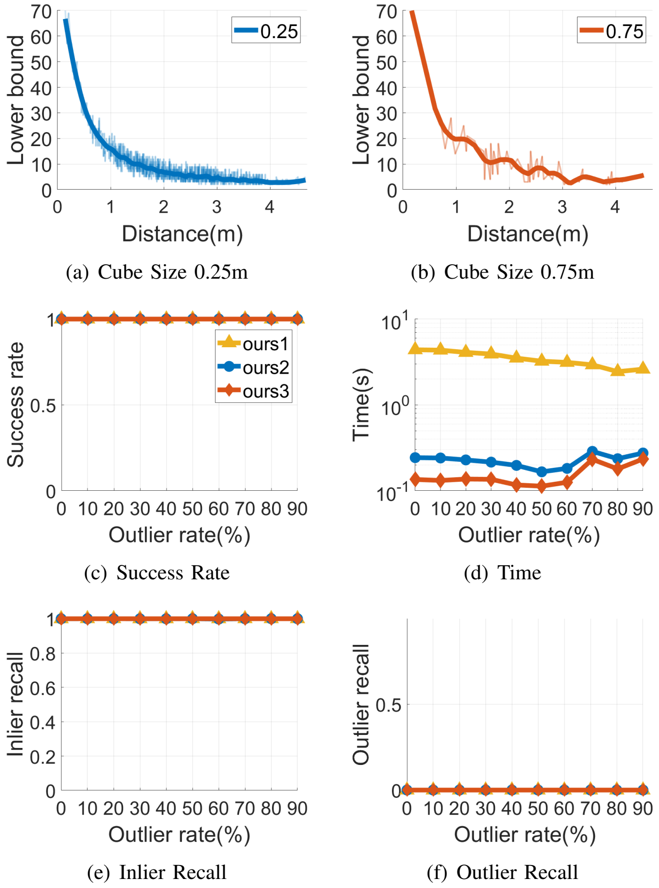
In practical applications, both γ_i and γ_k represent small angular errors, and their values are assumed to be consistent. Then for angle $\delta(\theta')$, we can further derive from $\delta(\theta) := 2|\theta| \sin\left(\frac{\gamma_k}{2}\right) + \gamma_k$ which is mentioned in Sec. II-B3 of the main text:

$$\delta(\theta') \approx (|\theta'| + 1)\gamma_k. \tag{20}$$

It is important to note that, both $\mathbf{x}_i$ and $\mathbf{x}_k$ are normalized three-dimensional vectors. Thus, utilizing the Euclidean and angular errors (1), we can obtain:

$$||\epsilon|| \le \sqrt{2(1 - \cos(|\theta'| + 2)\gamma_k)}. \tag{21}$$

Obviously, the range of the translational perturbation that rotation search can tolerate depends on both θ' and γ_k. Theoretically, the translational perturbation may be small. However, in practical applications, we can choose a fairly large size to achieve higher efficiency. This phenomenon is also shown in Sec. II-A2 of this Supp. Material.



(a) Cube Size 0.25m (b) Cube Size 0.75m

(c) Success Rate (d) Time

(e) Inlier Recall (f) Outlier Recall

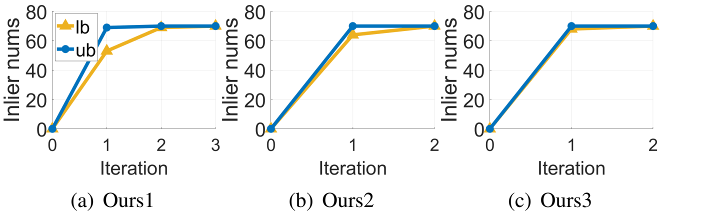
Fig. 3. Parameter selection in SE(3) for 100 points.

II. Experimental Results

A. Simulation experiments

*1) **Data generation**:* In simulation experiments, we first generate a series of 3D points within a 3D cube space $[-3,3]^3$ that simulate the map world points in the real environment. Next, we simulate the behavior of a virtual camera by projecting these 3D points into the camera's view, using random sampling within the ranges of $[-\pi,\pi]^3$ for rotation and $[-3,3]^3$ for translation. It is important to note that in the experiments concerning SO(3), no translation is introduced. Each randomly generate set of world points represents a unique simulation scene, while each randomly generated camera pose simulates the movement of the camera in that scene. In addition, we set the focal length of the virtual camera to be 800 and the resolution to be 1280×1080. Meanwhile, in order to simulate the measurement error in the actual vision system, we add random noise constrained to 5 to each 2D image point obtained from reprojection error to more closely match the actual application scene. In addition, to increase the complexity of the experiments, we also introduce outliers by projecting 3D points at random incorrect poses, and these outliers simulate erroneous or anomalous data that may occur in real applications.

*2) **Parameter selection**:* To better balance accuracy and efficiency, we select different ϵ to compare the estimation results under the guidance of real-world applications. First, we explore the choice of segment sizes, demonstrating the relationship between the lower bounds and the ground truth distances across different segment sizes. The experiments are conducted with a 30% outlier rate and a total point count of 100. The results both in Fig. 3(a)-3(b) indicate that when segmenting at sizes of 0.25m and 0.75m, the lower bounds contained within segments closer to the ground truth increased. This phenomenon forms the basis for our application of the fast search strategy in Alg. 3 in the manuscript. Also this trend is notably more pronounced with the 0.75m segmentation. This is because, with a 0.75m partition, the gradient of distance reduction toward the ground truth is steeper than that achieved with a 0.25m partition.



(a) Ours1 (b) Ours2 (c) Ours3

Fig. 4. The variations in lower and upper bounds for the three methods ours1, ours2, and ours3 are analyzed under 100 point pairs with a 70% outlier rate.

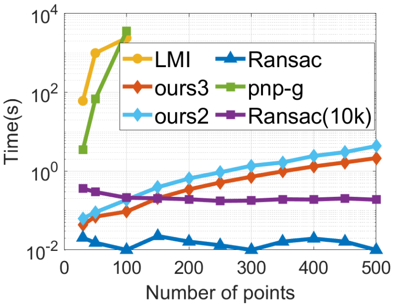
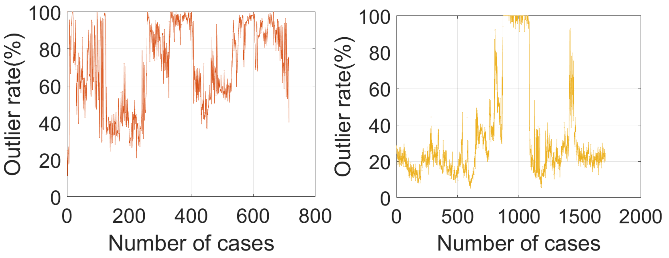


Fig. 5. Efficiency with increasing points under a 70% outlier rate.

Subsequently, based on these findings, we devise three comparative methods: ours1, ours2, and ours3, and conducted four comparative experiments evaluating their success rates, runtime, inlier recall rates, and outlier recall rates in the SE(3) space. Specifically, our1 only utilizes a 0.25m segment size for the segmentation of the 3D translation space, cutting along the x, y, and z axes simultaneously. Ours2 introduces an additional 0.75m segment size, first segmenting the 3D translation space at 0.75m to rapidly locate the block near the ground truth, then further segmenting the remaining blocks using 0.25m. Both segmentations cut across the x, y, and z axes simultaneously. For ours3, we employ the same two segment sizes as in ours2, with the distinction that during the 0.75m segmentation, we initially refrain from cutting along the z-axis. As shown in Fig. 3(c)-3(f), ours1, ours2, and ours3 exhibit minimal differences in inlier recall, outlier recall, and robustness to outliers, with all three demonstrating strong advantages. However, in comparative experiments on runtime, ours3 shows optimal performance in terms of parameter and operation selection.

Additionally, as shown in Fig. 4, we analyze the convergence rates of ours1, ours2, and ours3. It is evident that, at a 30% outlier rate, the condition for the termination of the



(a) Home5 Outlier Rate (b) Cafe1 Outlier Rate

Fig. 6. Line charts of actual outlier rates for scenes home5 and cafe1.

final iteration is characterized by $\overline{E} = E^* = 70$, this condition is also evident in line 10 of Alg. 3 in the manuscript. The iteration converges relatively slowly when ours1 is segmented using only a 0.25 division, which requiring more iterations to achieve $\overline{E} == E^*$, whereas ours3 demonstrates the fastest convergence speed.

Finally, for the efficiency comparison experiment regarding the increase in point count mentioned in the manuscript, we have also included ours2 for comparison. As shown in Fig. 5, the experimental results align with expectations, with ours3 consistently outperforming ours2 in terms of processing time.

Ultimately, considering both time and success rate, we select the best-performing method, ours3, as the segmentation approach in SE(3) space, which will serve as the basis for the subsequent synthetic and real world experiments.

B. Real world experiments

*1) **Real word data generation**:* We generate a usable set of 2D-3D point correspondences through a series of operations on the following two dataset, including map construction, image retrieval, and feature matching.

Newer College dataset: We utilize COLMAP [2] to construct a 3D map from the reference sequences. For reference image retrieval, we employ NetVLAD [3] to identify the most similar reference images to the query images. Feature extraction is carried out using SuperPoint [4], and feature matching is performed using the Nearest Neighbor method. Based on the aforementioned operations, we can obtain 2D-3D data.

Oxford dataset: Specifically, for each query image, we use NetVLAD to retrieve the top three reference images from the map session. We then apply SuperPoint and KNN matching [5] on the retrieved map images and the query image to establish 2D-2D feature correspondences. For map construction, we triangulate scene points with the observation from multiple frames and refine the reconstruction with bundle adjustment. This process ultimately yields the 2D-3D data needed for our task.

*2) **Performance analysis of home5 and cafe1 scenes**:* We show the ground truth outlier rates for both scenarios in Fig. 6. It can be observed that the proportion of cases with extreme outlier rates is significantly higher in the home5 scenario than in the cafe scenario. Specifically, in home5, 38% of cases have an outlier rate above 80%, and 17% of cases have an outlier rate exceeding 95%. For the cafe scene, cafe1, where the improvement is less significant, our analysis reveals that, in this scenario, cases with an outlier rate below 50% account for over 90%. Therefore, in such scenarios, our algorithm's advantage cannot be fully demonstrated. The strength of our method lies in its ability to guarantee deterministic optimality even under high outlier conditions.

References

[1] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8922–8931.

[2] A. Fisher, R. Cannizzaro, M. Cochrane, C. Nagahawatte, and J. L. Palmer, "Colmap: A memory-efficient occupancy grid mapping framework," *Robotics and Autonomous Systems*, vol. 142, p. 103755, 2021.

[3] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5297–5307.

[4] M. Mera-Trujillo, S. Patel, Y. Gu, and G. Doretto, "Self-supervised interest point detection and description for fisheye and perspective images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6497–6506.

[5] Q. Chen, D. Li, and C.-K. Tang, "Knn matting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 9, pp. 2175–2188, 2013.