Dominating Set Database Selection for Visual Place Recognition

Anastasiia Kornilova[1*], Ivan Moskalenko[1,2*], Timofei Pushkin[1,2], Fakhriddin Tojiboev[1],
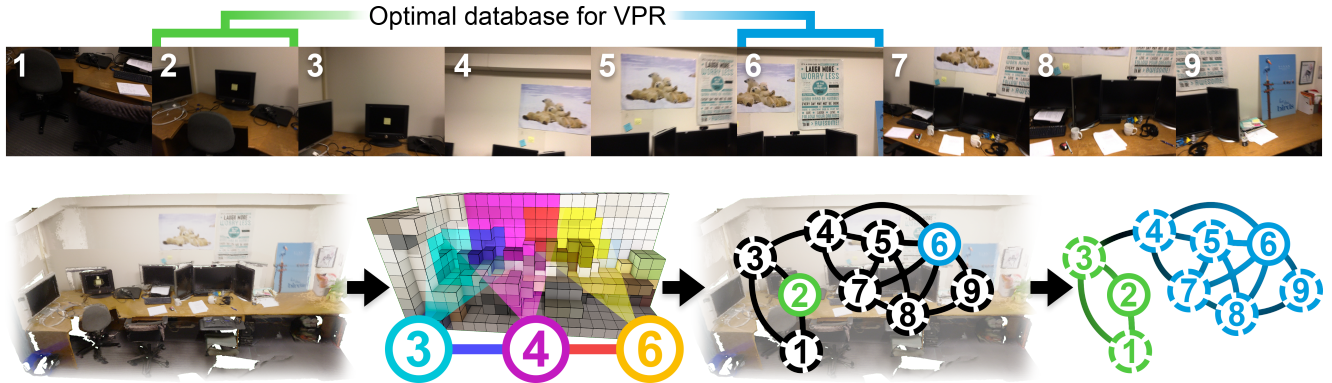Rahim Tariverdizadeh[1], and Gonzalo Ferrer[1]

Fig. 1: Overview of the proposed methodology for building an optimal database for Visual Place Recognition (VPR). *Top:* a sequence of RGBD images obtained from the environment scanning. *Bottom:* (i) a 3D environment map is built from the scanning sequence, (ii) an overlap measure is estimated for each image pair using the spatial overlap in the voxelized map, (iii) the optimal VPR database is built by solving the dominating set problem for a graph where the localized images are vertices connected based on their estimated overlap, (iv) the rest of the images are split into database classes for VPR fine-tuning on the scanned scene.

Abstract— This paper introduces a novel approach for creating a visual place recognition (VPR) database for localization in indoor environments from RGBD scanning sequences. The proposed method formulates the problem as a minimization challenge by utilizing a dominating set algorithm applied to a graph constructed from spatial information, referred to as the "DominatingSet" algorithm. Experimental results on various datasets, including 7-scenes, BundleFusion, RISEdb, and a specifically recorded sequences in a highly repetitive office setting, demonstrate that our technique significantly reduces database size while maintaining comparable VPR performance to state-of-the-art approaches in challenging environments. Additionally, our solution enables weakly-supervised labeling for all images from the sequences, facilitating the automatic fine-tuning of VPR algorithm to target environment. Additionally, this paper presents a fully automated pipeline for creating VPR databases from RGBD scanning sequences and introduces a set of metrics for evaluating the performance of VPR databases. The code and released data are available on our web-page — **https://prime-slam.github.io/place-recognition-db/**.

I. INTRODUCTION

Visual place recognition (VPR) plays a crucial role in solving the localization problem using image data alone. The applications are immense and the Robotics community as well as the Computer Vision community are actively investigating new solutions. A commonly used approach in VPR systems involves two key components: (i) a database consisting of a set of images and their corresponding 3D poses, and (ii) an algorithm that identifies the most similar image in the database to a given query image and estimates its pose relative to the database image. This paper investigates VPR from the perspective of selecting a small subset of images from a sequential stream of sensor-generated data to accurately represent the environment for VPR operation.

Scanning the environment requires a collection of visual data and its corresponding locations, that in indoors is usually solved by SLAM [1], [2] or Bundle Adjustment [3] algorithms. However, a challenge arises after the scanning process in determining which images should be included in the database. The volume of observations obtained after scanning is often massive, with thousands of observations or more, and highly redundant due to the sequential capture of the scene during movement. To optimize computational and memory resources, especially in low-computational and embedded devices, the database should be compact in size. Furthermore, it should provide sufficient data diversity and coverage of the entire scene to facilitate accurate VPR.

VPR is particularly applicable in indoor environments where the use of Global Navigation Satellite Systems (GNSS) is unavailable, and other global localization equipment such as radio or Wi-Fi beams pose challenges in terms of maintenance and cost. Over the past years, sig-

[1]The authors are with Skolkovo Institute of Science and Technology (Skoltech), Center for AI Technology (CAIT). anastasiia.kornilova, g.ferrer@skoltech.ru
[2]The authors are with Software Engineering Department, Saint Petersburg State University.
* Indicates equal contribution.

nificant efforts have been dedicated to developing robust place recognition algorithms using available benchmarks and datasets, such as Mapillary [4], Nordland [5], Pittsburgh [6], Tokyo24/7 [7], and RobotCar Seasons [8], [9]. However, these datasets primarily focus on outdoor environments and already provide pre-selected databases for VPR methods.

In this work, we propose an approach for creating VPR databases in indoor environments by utilizing RGBD information and selecting an optimal subset of images from the scanning sequence. To achieve this, we formulate a minimization problem for VPR databases and propose a solution based on the "dominating set" algorithm [10] applied to graphs and spatial information. A byproduct of this solution is the clusterization of localized images around the selected one, and its application to related tasks such as automatic creation of weakly-supervised images of the same "place" for fine-tuning neural algorithms to a specific environment. We demonstrate that our technique, which reduces database size, produces compact databases while maintaining comparable visual localization quality to state-of-the-art approaches in challenging environments.

The main contributions of the paper are as follows:

- a formal definition of an optimal VPR database in terms of size and coverage, along with an approach for its calculation;
- a fully automated pipeline for VPR database creation from an RGBD scanning sequence, which is made publicly available as a library;
- a fully automated end-to-end methodology, from sequence scanning to VPR fine-tuning for a specific environment.

II. RELATED WORK

VPR algorithms address the task of identifying the most suitable image match from a *database* that represents a given environment, based on a so-called *query* observation captured within the same environment. This problem, known as *Information Retrieval*, is common in various fields such as Natural Language Processing, Computer Vision, and Robotics.

To enable VPR, it is necessary to define an *image descriptor* and a *similarity measure* between pairs of descriptors (images). A classic approach involves calculating local image features and aggregating them into a global image descriptor using techniques like bag of words [11], VLAD [12], [13], or differentiable NetVLAD [14], [15]. Local image features can be computed using handcrafted algorithms [16]–[18] or by employing learnable methods for keypoint detection and description [19]–[21].

The calculation of the global image descriptor is an active area of research, with algorithms achieving remarkable performance. For instance, Hloc [22] learn to predict both global and local features simultaneously, while CosPlace [23] provides a learned global descriptor without intermediate local feature aggregation. In this study, we utilize state-of-the-art global image descriptors as a tool for VPR. It can be argued that only with the utilization of these advanced

VPR methods can we significantly reduce the dataset size, as proposed in this paper.

Existing VPR approaches predominantly concentrate on urban outdoor environments and datasets [4]–[9]. This emphasis arises from the availability of training data in outdoor scenarios, where reference poses can be directly obtained using GNSS technologies. Even with GNSS sensor errors of up to a meter, it is still possible to establish accurate associations for image correspondences. The creators of these datasets already provide a comprehensive database and training correspondences.

While the majority of research focuses on outdoor environments, there have been limited efforts to address indoor visual place recognition and the associated database creation. The 7-Scenes dataset [24], [25] is commonly employed for training and evaluating indoor localization methods. However, it only covers small areas like office rooms and apartment sections. Other datasets, such as TUM-RGBD [26], ScanNet [27], and BundleFusion [28], are used for RGB-D SLAM and also offer the opportunity to test visual localization methods. However, they are limited in terms of scene coverage and variations in lighting conditions. In contrast, the RISEdb dataset [29] includes large-scale scenes that encompass entire floors of various types of buildings. Additionally, this dataset provides data captured under varying lighting conditions at different times of the day.

The creation of a map database is also indirectly addressed in SLAM pipelines through a process known as *keyframe selection* and used as a basis for SLAM-graph optimization and loop closure. Various techniques have been employed in different systems to determine keyframes based on different criteria. ORB-SLAM [1] and its subsequent versions [30], [31] utilize bag-of-words representation combined with local features. In these approaches, an image is considered a keyframe if it observes a significant number of new local features compared to the existing map. BundleFusion [28] divides the data stream into chunks, each containing an equal number of RGBD images. From each chunk, a single keyframe is selected for further processing. Das et al. [32] propose two methods based on image entropy to estimate whether a keyframe will contribute to map improvement or not. These approaches evaluate the information content of the keyframe relative to the existing map. Alonso et al. [33] exploit image quality criteria such as blurriness and brightness, as well as semantic content criteria based on a CNN called MiniNet, to assess the suitability of keyframes for enhancing the map. Sheng et al. [34] introduce a joint learning approach for keyframe detection and visual odometry, where the system learns to identify keyframes that are informative for both tasks. iMap [35] and NICE-SLAM [36] adopt a depth overlap criterion with respect to the map to measure the amount of new information that a potential keyframe can contribute to the map. Mainly, the process of keyframe selection primarily targets the internal workings of the SLAM algorithm, which could potentially restrict its effectiveness for the global visual place recognition task

when compared to the technique of selecting general subsets of images that were captured in the environment.

III. METHODOLOGY

The proposed methodology for creating an optimal database for visual place recognition (VPR) is illustrated in Fig. 1. The algorithm takes color images, corresponding depth images, and their poses as input. The poses can be obtained through classical state estimation techniques like SLAM-based solutions or beacon systems. Using this information, a 3D map of the environment is generated. Next, the 3D map is divided into voxels, and the spatial overlap between pairs of images is calculated by determining the intersection between the voxel sets of the images. This overlap information enables the construction of a graph that represents the connections between the images based on their calculated overlaps. Finally, an optimal database for VPR is obtained by identifying a dominating set within this graph. Optionally, this methodology allows for the division of the remaining images from the scanning sequence into database classes, which can be utilized for training or fine-tuning VPR algorithms.

A. Problem formulation for optimal database

Consider a given set of color images that capture the environment, denoted as $C = c_1, \ldots, c_N$. For any two color images c_i and c_j, an *overlap measure* $s(c_i, c_j) \in [0,1]$ is defined to quantify the extent to which the view scopes of the images intersect. To assess the coverage of a subset of color images $\widetilde{C} \subset C$, a *coverage loss* function $f(\cdot)$ is introduced. This function quantifies the number of images covered by the subset and is defined as follows:

$$f_C(\widetilde{C}) = \sum_{c \in C} \begin{cases} 0 & \text{if } \exists \widetilde{c} \in \widetilde{C} \text{ s.t. } s(c, \widetilde{c}) > \mu \\ 1 & \text{else} \end{cases} \quad (1)$$

where μ is an overlap threshold. Our objective is to identify a subset $\widetilde{C}$ of the minimum size that ensures coverage of all frames in C:

$$C_{db} = \min_{\substack{\widetilde{C} \subseteq C \\ f_C(\widetilde{C})=0}} |\widetilde{C}|. \quad (2)$$

In this context, we define C_{db} as an **optimal image database**, which serves as a condensed representation of the color information obtained from digitizing the environment.

When the overlap measure between images satisfies symmetry, i.e., $s(c, \widetilde{c}) = s(\widetilde{c}, c)$, the problem formulation resembles the minimum dominating set problem in graph theory. In this analogy, the images are represented as vertices in a graph, where an edge exists between two vertices if their overlap measure exceeds a given threshold. The original dominating set problem aims to identify the smallest subset of vertices in the graph such that every vertex either belongs to the subset or is adjacent to a vertex within it.
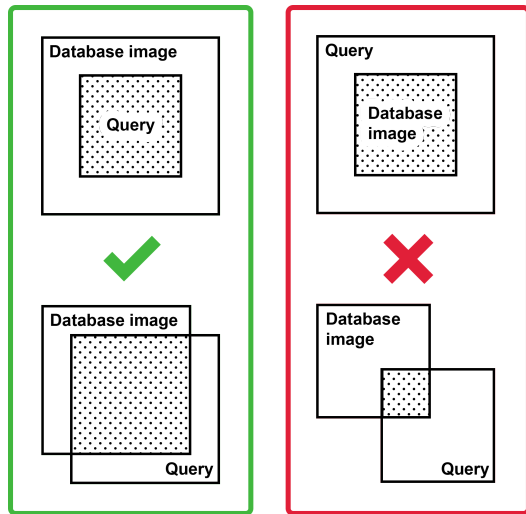

Fig. 2: Examples showcasing different overlaps between a query and a database image, interpreted as good or bad. *Left*: the database image covers a significant portion of the query image. *Right*: the query image is not well covered by the database image.

B. Spatial overlap measure

In our methodology, we propose utilizing spatial information obtained from the depth camera to estimate the overlap between two images. This approach provides an advantage over an alternative approach that relies solely on color information, such as using local features and matches between them. The use of color information alone can lead to incorrect edges being generated, particularly in situations where different locations exhibit similar textures and patterns, resulting in visual aliasing.

With access to depth information for each color frame, we can construct a 3D map of the environment using image poses. This 3D map can be represented as a set of voxels, denoted as $V = \{v_1, \ldots, v_M\}$. The sequence of color images, C, is associated with the voxels through the set $D = \{d_1, \ldots, d_N\}$, where $d_i = \{v_{i_1}, \ldots, v_{i_k}\}$ represents a subset of voxels observed in frame c_i. By considering these subsets of voxels, we can define an overlap measure, denoted as $s(d_i, d_j)$, based on the intersections of voxel sets. This overlap measure provides a more accurate estimation of the degree of overlap compared to considering only color information. In our analysis, our objective is to identify a subset $\widetilde{D} \subset D$ that satisfies the following criterion:

$$D_{db} = \min_{\substack{\widetilde{D} \subseteq D \\ f_D(\widetilde{D})=0}} |\widetilde{D}|. \quad (3)$$

C. Overlap measure

To establish an appropriate overlap measure, let us examine the examples illustrated in Fig. 2. A database image can be considered *good* in relation to a query image if it covers a significant part of the query image. Conversely, a database image can be deemed *bad* if it covers only a relatively small

part of the query image. It is important to note that this principle is not symmetric. A query image may occupy a minor portion of a database image, yet the database image may still provide sufficient coverage for the query. Formally, the overlap measure for voxel sets can be defined as follows:

$$s(d_q, d_{db}) = \frac{|d_q \cap d_{db}|}{|d_q|}, \qquad (4)$$

where $|\cdot|$ is set cardinality and d_q and d_{db} are voxel sets corresponding to the query and database images respectively.

As mentioned previously, in the case where the overlap measure exhibits symmetry, the minimization problem can be addressed using dominating set algorithms and existing solvers [37]. In our study, we propose employing the intersection over union (IoU) as the overlap measure for voxel sets:

$$s_{IoU}(d_q, d_{db}) = \frac{|d_q \cap d_{db}|}{|d_q \cup d_{db}|} \qquad (5)$$

This metric, being symmetric, imposes stricter constraints on the graphsince:

$$s_{IoU}(d_q, d_{db}) \leq s(d_q, d_{db}) \qquad (6)$$

The inclusion of stricter constraints leads to an increased number of edges in the graph, which, in turn, may result in a larger database volume when applying the dominating set solution compared to the original problem formulation.

D. Graph processing

To construct the graph, it is necessary to determine the overlap between every pair of images. However, in cases where the scanning sequence is extensive, analyzing the intersection between $\frac{N(N-1)}{2}$ pairs of voxel sets can produce computational overhead. Moreover, many pairs of voxel sets may not have any intersection. To address these challenges and optimize the process, we propose the following algorithm:

1) associate each voxel in the map with the indices of the frames that cover it;
2) generate pairs of frames for each voxel, encompassing all possible combinations;
3) for each pair, calculate the number of voxels where the same pair of frames occurs. This count reflects the size of the corresponding intersection;
4) finally, compute the Intersection over Union (IoU) for the constructed pairs using the previously calculated intersection sizes.

By employing this approach, only the voxel sets with non-empty intersections need to be considered, leading to a more efficient graph construction process.

IV. EXPERIMENTAL RESULTS

In this section, we present a comparative analysis of the quality of the VPR database constructed using our method in contrast to other existing strategies for database creation. Our evaluation focuses on two primary test cases. The first test case involves indoor localization sequences taken from the 7-scenes and BundleFusion datasets. The second test case expands upon this evaluation by conducting experiments in more challenging and large-scale environments, specifically the RISEdb dataset (with lightning changes) and our self-curated Skoltech Campus dataset. Furthermore, we showcase the performance of cutting-edge VPR algorithms that have been adapted to operate with the database generated using our methodology.

A. Datasets

To assess the quality of a database and the adapted VPR algorithm, considered datasets should resemble real-world scenarios. This entails dividing the dataset into a "scanning" sequence for database creation and VPR algorithm adaptation, and a separate "test" sequence for evaluating place recognition quality. To fulfill these requirements, our evaluation includes the two largest scenes from each considered dataset: 7-scenes dataset (RedKitchen and Office) [24], [25], BundleFusion (office0 and office1) [28], RISEdb (as3ml and auditorium) [29]. For the 7-scenes dataset, the sequences labeled as "train" in the original dataset are designated as the scanning sequence, while the remaining sequences are utilized for testing. As for the BundleFusion dataset, we manually partitioned the sequence into a dedicated scanning route and a distinct test route. In the case of the RISEdb dataset, where multiple sequences exist for each scene, we selected the largest sequence as the scanning route and the second largest sequence as the test route. As RISEdb does not provide depth images but only a 3D map, we projected the map onto the corresponding color frames in order to obtain depth information.

Furthermore, we have captured a series of sequences within the Skoltech campus using an Azure Kinect DK sensor. This particular environment presents a challenge for VPR algorithms in indoor settings due to the presence of recurring structures with similar designs, such as desks, walls, and doors, resulting in increased visual ambiguity. Illustrations depicting some instances of visual ambiguity in this scene can be seen in Fig. 3. To construct a comprehensive 3D map, we utilize the trajectory generated by the RGBD ORB-SLAM algorithm [1]. However, it should be noted that certain sensors may not provide complete depth coverage for each color frame. To address this limitation, we have implemented a mechanism to extend the depth coverage using 3D map reprojection to the precise frame.

B. Database size and reduction rate

Our database creation algorithm incorporates two hyperparameters: voxel size for map voxelization and overlap threshold for graph processing. In our experiments, we set the voxel size to 0.3 m, which provides a reasonable approximation of
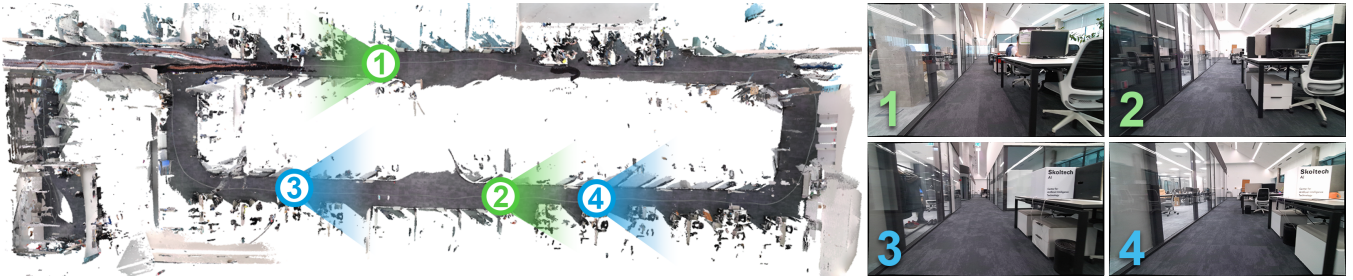
Fig. 3: Examples of different scenes captured in Skoltech campus having similar structures. *Left:* the map of the Skoltech campus sequence. *Right:* pairs of images that are very similar visually but were captured in different locations.

TABLE I: Statistics on the scanning sequence size and the size of the resulting databases

Dataset	Sequence size	Overlap threshold		
		0.1	0.3	0.5
		Database size Reduction rate Spatial coverage		
7-Scenes Office	4800	5 x960 43%	15 x320 56%	80 x60 75%
7-Scenes RedKitchen	5600	4 x1400 24%	25 x224 54%	97 x58 68%
BundleFusion office0	3547	14 x253 61%	33 x108 72%	94 x38 82%
BundleFusion office1	3622	12 x301 47%	30 x120 65%	86 x42 75%
Sk campus	6849	20 x342 50%	46 x148 65%	105 x105 72%
RISEdb as3ml	10952	102 x107 74%	363 x30 92%	845 x13 95%
RISEdb auditorium	19278	94 x205 69%	307 x63 90%	730 x26 96%

the spatial characteristics of indoor environments captured by RGBD sensors. We consider three overlap thresholds for the Intersection over Union (IoU) measure: 0.1, 0.3, and 0.5. Table 1 presents statistics on the size of the scanning sequence and the resulting size of the constructed database for different overlap threshold values. The table also includes information on the reduction rate and the percentage of spatial coverage achieved by the database. The spatial coverage is quantified as the percentage of map voxels occupied by the voxels covered by the database.

The findings of our study indicate that our proposed methodology yields a significant reduction in the size of the scanning sequence. The extent of this reduction is contingent upon the speed and comprehensiveness of the environmental data recording process. For instance, when the recording process is slow, the reduction rate is found to be the highest, as observed in the 7scenes dataset. It is important to emphasize that the spatial coverage of certain constructed

databases is relatively limited, encompassing only 25-40% of the voxels. Nevertheless, the utilization of invariance in our approach ensures that the selected frames for the database exhibit adequate overlap with other images, thereby covering the remaining voxels that have not been addressed. Additionally, it is noteworthy that the spatial overlap for the two largest sequences in the RISEdb dataset remains largely unchanged when the threshold is adjusted from 0.3 to 0.5. This observation indicates that the saturation point of information has been reached.

C. Visual place recognition

In this study, we present an evaluation of state-of-the-art visual place recognition algorithms using our methodology for database creation.

Specifically, we assess the performance of the following methods that have demonstrated superior results on popular visual place recognition (VPR) datasets: NetVLAD [15], CosPlace [23], and a combined approach using both NetVLAD and SuperGlue [38]. The combined approach utilizes the top-5 predictions from the NetVLAD algorithm and then applies the SuperGlue method to determine the prediction with the maximum number of matches. The utilization of SuperGlue for all frames is not a viable approach due to its low computational efficiency. This inefficiency stems from the fact that the entire image is processed for every request, rather than solely focusing on the descriptors of the image that can be preprocessed in advance for VPR operation.

Two types of models for NetVLAD and CosPlace are considered: the original pre-trained models provided by the authors and models that we fine-tuned specifically for each database and scene. To prepare the data, we employ the following pipeline: every fifth frame from the scanning sequence is extracted for the validation set, while the remaining images are used for training. Since the *DominatingSet* algorithm provided class labels, we trained the models on this subset of data until an "early stop" criterion was met.

All evaluations are conducted on test sequences that are not included in the database creation pipeline or VPR algorithm fine-tuning process. The evaluation metric used was Recall@1, which is a commonly used metric for VPR tasks. In order to determine whether the query frame and the database frame were correctly matched, we use metric (4)

TABLE II: Recall of Visual Place Recognition algorithms on databases generated by DominatingSet for **small-scale scenarios**. (f) stands for fine-tuned model.

Dataset	7-Scenes Office			7-Scenes RedKitchen			BundleFusion office0			BundleFusion office1		
Overlap threshold	**0.1**	**0.3**	**0.5**	**0.1**	**0.3**	**0.5**	**0.1**	**0.3**	**0.5**	**0.1**	**0.3**	**0.5**
NetVLAD	63.9	81.9	**91.3**	65.0	75.6	**88.9**	74.1	83.8	**88.1**	70.9	87.3	**96.0**
NetVLAD (f)	**99.2**	96.9	88.8	**95.3**	93.7	93.1	84.2	**90.8**	90.1	83.9	**97.1**	95.3
NetVLAD (f) + SuperGlue	–	98.6	**99.0**	–	99.1	**99.2**	91.9	96.8	**98.7**	87.4	99.2	**99.7**
CosPlace (ResNet-101, 2048)	64.7	83.8	**87.8**	68.2	63.1	**90.3**	64.7	80.7	**90.7**	62.3	79.3	**92.6**
CosPlace (f)	88.6	97.8	**98.4**	80.8	89.3	**95.8**	91.2	92.1	**95.2**	82.1	97.3	**97.4**

TABLE III: Recall of Visual Place Recognition algorithms on databases generated by DominatingSet for **large-scale scenarios**. (f) stands for fine-tuned model.

Dataset	Sk campus			RISEdb as3ml			RISEdb auditorium		
Overlap threshold	**0.1**	**0.3**	**0.5**	**0.1**	**0.3**	**0.5**	**0.1**	**0.3**	**0.5**
NetVLAD	56.9	80.3	**94.5**	25.6	63.5	**78.3**	27.7	51.0	**70.7**
NetVLAD (f)	88.0	**89.1**	83.4	31.4	59.4	**67.2**	26.3	42.5	**50.7**
NetVLAD (f) + SuperGlue	74.9	92.3	**94.1**	38.3	77.7	**82.7**	40.2	59.1	**72.9**
CosPlace (ResNet-101, 2048)	63.0	80.9	**91.9**	28.3	69.1	**83.7**	34.5	59.7	**80.5**
CosPlace (f)	84.3	93.2	**95.3**	40.4	85.0	**90.5**	68.3	84.7	**91.8**

with a threshold value equal to 0.3. The results are presented in Tab. II and Tab. III, considering small-scale and large-scale scenarios respectively.

First, it can be noticed that the overall localization quality is superior in small-scale scenarios compared to large-scale scenarios, even when employing pre-trained models. This can be attributed to the greater diversity of data captured across the scenes in small-scale scenarios, in contrast to more challenging datasets characterized by repetitive structures and variations in lighting conditions. Secondly, for the creation of databases in small-scale scenarios, an overlap threshold of 0.1 proves to be a suitable choice. This threshold enables a significant reduction in the number of frames within the database while maintaining relatively good localization quality on fine-tuned models. In the case of large-scale applications, a more reasonable choice for the overlap threshold would be 0.3 or 0.5. These values contribute to an improvement in localization quality, albeit with a marginal difference between 0.3 and 0.5. Consequently, memory consumption can be reasonably optimized by utilizing a smaller overlap threshold. Furthermore, the CosPlace method demonstrates the best overall performance. It excels in challenging environments, providing superior quality and comparable performance to other methods in small-scale environments. It is worth noting that the fine-tuning process employed in our methodology enhances the performance of the localization method by tailoring it to the specific environment. However, it should be acknowledged that when using a 0.5 overlap threshold, the fine-tuned NetVLAD exhibits lower quality compared to the original network. This can be attributed to overfitting of the network to specific frames.

In summary, our approach for database creation offers compression capabilities that yield exceptional quality in small-scale environments and comparable performance to state-of-the-art visual localization methods in challenging environments. Also, proposed approach provides a common way to formulate problem of optimal VPR database creation and, in the future, can be applied to any scanning sequences that records spatial information, for example, LiDAR data and outdoor environmetns.

V. CONCLUSION

This paper has presented a method for creating compact and descriptive VPR databases for indoor environments. To do so, our approach defines a formal definition of optimal database and provides an algorithm for its generation from sequential stream of RGBD data. In our experiments, that cover various environments with different conditions, we have demonstrated that scanning sequence can be compressed to compact database while maintaining comparable visual localization quality to state-of-the-art methods. Additionally, we have provided a methodology on top of our method to generate data for fine-tuning data for target environment, that shows improvements with respect to original models on target environments. We have made all contributions, including code and data, public.

REFERENCES

[1] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[2] S. Zhang, L. Zheng, and W. Tao, "Survey and evaluation of rgb-d slam," *IEEE Access*, vol. 9, pp. 21 367–21 387, 2021.

[3] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[4] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4990–4999.

[5] N. Sünderhauf, P. Neubert, and P. Protzel, "Are we there yet? challenging seqslam on a 3000 km journey across all four seasons," in *Proc. of workshop on long-term autonomy, IEEE international conference on robotics and automation (ICRA)*, 2013, p. 2013.

[6] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual place recognition with repetitive structures," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 883–890.

[7] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1808–1817.

[8] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The oxford robotcar dataset," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.

[9] C. Toft, W. Maddern, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, T. Pajdla, *et al.*, "Long-term visual localization revisited," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 2074–2088, 2020.

[10] M. R. Garey and D. S. Johnson, *Computers and intractability*. freeman San Francisco, 1979, vol. 174.

[11] D. Gálvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, 2012.

[12] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 3304–3311.

[13] I. Mironică, I. C. Duță, B. Ionescu, and N. Sebe, "A modified vector of locally aggregated descriptors approach for fast video classification," *Multimedia Tools and Applications*, vol. 75, pp. 9045–9072, 2016.

[14] 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.

[15] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 141–14 152.

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91–110, 2004.

[17] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Lecture notes in computer science*, vol. 3951, pp. 404–417, 2006.

[18] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*. Ieee, 2011, pp. 2564–2571.

[19] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "Lift: Learned invariant feature transform," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*. Springer, 2016, pp. 467–483.

[20] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 224–236.

[21] A. Barroso-Laguna, E. Riba, D. Ponsa, and K. Mikolajczyk, "Key. net: Keypoint detection by handcrafted and learned cnn filters," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5836–5844.

[22] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *CVPR*, 2019.

[23] G. Berton, C. Masone, and B. Caputo, "Rethinking visual geo-localization for large-scale applications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4878–4888.

[24] B. Glocker, S. Izadi, J. Shotton, and A. Criminisi, "Real-time rgb-d camera relocalization," in *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2013, pp. 173–179.

[25] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon, "Scene coordinate regression forests for camera relocalization in rgb-d images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2930–2937.

[26] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 573–580.

[27] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5828–5839.

[28] A. Dai, M. Nießner, M. Zollhöfer, S. Izadi, and C. Theobalt, "Bundle-fusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, p. 1, 2017.

[29] C. Sánchez-Belenguer, E. Wolfart, and V. Sequeira, "Rise: A novel in-door visual place recogniser," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 265–271.

[30] Y. Li, R. Yunus, N. Brasch, N. Navab, and F. Tombari, "Rgb-d slam with structural regularities," in *2021 IEEE international conference on Robotics and automation (ICRA)*. IEEE, 2021, pp. 11 581–11 587.

[31] A. Pumarola, A. Vakhitov, A. Agudo, A. Sanfeliu, and F. Moreno-Noguer, "Pl-slam: Real-time monocular visual slam with points and lines," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 4503–4508.

[32] A. Das and S. L. Waslander, "Entropy based keyframe selection for multi-camera visual slam," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 3676–3681.

[33] I. Alonso, L. Riazuelo, and A. C. Murillo, "Enhancing v-slam keyframe selection with an efficient convnet for semantic analysis," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4717–4723.

[34] L. Sheng, D. Xu, W. Ouyang, and X. Wang, "Unsupervised collaborative learning of keyframe detection and visual odometry towards monocular deep slam," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4302–4311.

[35] E. Sucar, S. Liu, J. Ortiz, and A. J. Davison, "imap: Implicit mapping and positioning in real-time," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6229–6238.

[36] Z. Zhu, S. Peng, V. Larsson, W. Xu, H. Bao, Z. Cui, M. R. Oswald, and M. Pollefeys, "Nice-slam: Neural implicit scalable encoding for slam," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 786–12 796.

[37] A. Hagberg, P. Swart, and D. S Chult, "Exploring network structure, dynamics, and function using networkx," Los Alamos National Lab.(LANL), Los Alamos, NM (United States), Tech. Rep., 2008.

[38] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4938–4947.