

Purely Geometric Scene Association and Retrieval

- A case for macro-scale 3D geometry

Rahul Sawhney^{1,2}, Fuxin Li², Henrik I. Christensen³ and Charles L. Isbell¹

Abstract—We address the problems of measuring geometric similarity between 3D scenes, represented through point clouds or range data frames, and associating them. Our approach leverages macro-scale 3D structural geometry - the relative configuration of arbitrary surfaces and relationships among structures that are potentially far apart. We express such discriminative information in a viewpoint-invariant feature space. These are subsequently encoded in a frame-level signature that can be utilized to measure geometric similarity. Such a characterization is robust to noise, incomplete and partially overlapping data besides viewpoint changes. We show how it can be employed to select a diverse set of data frames which have structurally similar content, and how to validate whether views with similar geometric content are from the same scene. The problem is formulated as one of general purpose retrieval from an unannotated, spatio-temporally unordered database. Empirical analysis indicates that the presented approach thoroughly outperforms baselines on depth / range data. Its depth-only performance is competitive with state-of-the-art approaches with RGB or RGB-D inputs, including ones based on deep learning. Experiments show retrieval performance to hold up well with much sparser databases, which is indicative of the approach’s robustness. The approach generalized well - it did not require dataset specific training, and scaled up in our experiments. Finally, we also demonstrate how geometrically diverse selection of views can result in richer 3D reconstructions.

I. INTRODUCTION

The problems of computing similarity and establishing association between range images and/or 3D point clouds of scenes (observed from a viewpoint, henceforth referred to as *scene-views*) is fundamental to robotics and computational perception in general. It plays an important role in a multitude of applications. Loop closure (identifying a place visited earlier in the trajectory) is intrinsic to metric SLAM ([9]). Localizing with respect to a previously reconstructed map or scene model, and relocalizing after a tracking failure (determining sensor pose without pose priors from trajectory) - both are essential for mapping in practice as well. Different forms of the problem are also key to many navigation scenarios, and in perception tasks such as scene-guided search / foraging or location-based context and activity recognition.

In a minimally restrictive setting, the aforementioned problems (and several others) can be formulated as a retrieval problem - to recognize / identify a scene-view by linking it to stored ones in an assorted, unorganized database. Such a setting would not require any pose priors, spatio-temporal

contiguity of collected data¹ or other additional information such as annotations or reconstructed models, and would remove the need to learn a specific pose estimator / regressor for each workspace.

While a lot of progress has been made over the years, including in the retrieval domain, competitive scene association approaches in literature have mostly been reliant on (discriminative) appearance information. Relatively few methodologies work well on noisy, imperfect 3D point clouds or depth images from the real world. Often they critically rely on additional pieces of information available in their target scenario - to prune the association hypothesis space, or obtain strong indirect priors on scene similarity, or enable construction of aggregated spatial information structures to allow its estimation (for instance, [25, 11, 7, 32, 6, 3])². Approaches also often operate under limited changes in viewpoint and / or on specific types of scene geometry (such as [42, 11, 29]) or they solve a simplified 2D problem (such as [19]). Understandably, methodologies like above are either use case limited or restrictive. Note that approaches like [10] do not ascertain association at all - these directly solve for 3D poses between pre-associated set of data frames.

The dearth of purely 3D geometric scene association approaches in the real world can be primarily attributed to the considerably more ambiguous and challenging depth / range sensing modality. In general, the modality has high local ambiguity and may not be lavish with information on the whole (in contrast to *rgb*). Data acquired from commodity 3D range / depth sensing hardware tends to be particularly noisy as well, has several imperfections. Locally smooth, isomorphic and self-similar nature of typical 3D data from indoor or structural environments makes the problem more difficult. Changes in viewpoint, occlusions and partially overlapping views / content significantly exacerbate the problem further.

We present a minimally restrictive retrieval methodology. Our approach affords means to evaluate *geometric content similarity* between 3D point sets and associate them. We show how it can be utilized to affect *geometric diversity* as well.

We generate descriptive frame-level signatures directly from range images / point clouds (any additional information or assumptions touched upon earlier are not utilized). We make use of macro scale geometry — 3D geometrical interactions (derived from relative angles and distances) over

¹ Institute of Robotics & Intelligent Machines, College of Computing, Georgia Institute of Technology, Atlanta, USA rsawhney3@gatech.edu, isbell@cc.gatech.edu ² School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, USA lif@engr.orst.edu ³ Institute of Contextual Robotics, University of California, San Diego, USA hchristensen@ucsd.edu

¹ Spatio-temporally unordered databases can store data acquired from multiple sensors, at multiple times and from disparate locations; could just constitute of snapshots covering scenes of interest.

² Quite commonly, approaches rely on spatio-temporal contiguity of frames to obtain priors or accumulate data structures to ascertain the association.

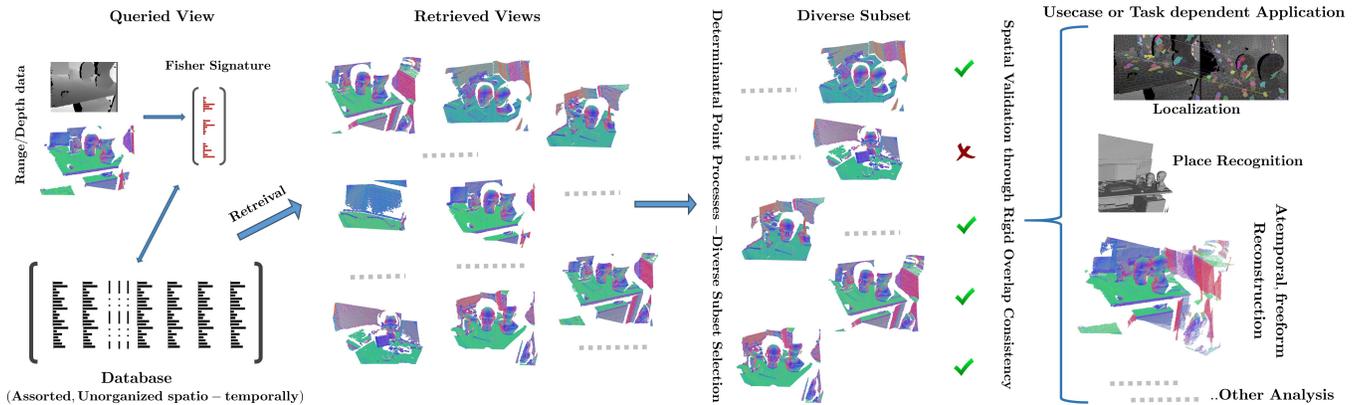


Fig. 1: Retrieval pipeline overview. The query view is indicated in the top-left. Input is a range image or a 3D point cloud. The database (bottom left) constitutes of unordered signatures from arbitrary scene-views, with no labels or ground truth pose annotations. The set of nearest-neighbor retrieved views undergo diversification and subsequent validation. The point clouds are color mapped according to the surface normals - the RGB color of a 3D point is proportional to the component values of its normal.

an arbitrarily large span, between arbitrary surfaces, primitives and structures, and their spatial arrangement (for example between walls, floor and furniture, between fixtures and equipment, or just between various parts of a given structural entity). Such interactions when considered collectively are highly discriminative. They are expressed in a learnt viewpoint invariant feature space (III-A). To characterize a scene-view, high order gradient statistics from a dense set of projected interactions are utilized (Fisher Vector, III-B). To identify a geometrically diverse subset from set of similar retrieved views (III-C), we model a Determinantal Point Process (DPP, III-D). And to establish association with some of the retrieved views, we employ a fine-grained spatial validation scheme which ascertains consistency of rigid geometry overlap (III-E).

The proposed approach not only outperformed the range / depth data baseline, but was also comparable or better than state-of-art RGB and RGB-D approaches (including ones based on CNN³) - and without relying on any additional pose annotations, apriori reconstructed 3D world models, or assumptions such as spatio-temporal contiguity of training data used by other approaches.

Experiments also indicated the learning to be general - unlike most other approaches, it did not require dataset specific training; a single learnt model performed well across the board. Experiments also indicated the performance holding up under significantly sparser databases, and under significantly increased database scale and diversity. Our empirical evaluations quantifying geometric diversity of retrievals were quite encouraging as well. They not only indicated a significant increase in viewpoint diversity of the retrieved set, but also suggested the efficacy of the proposed approach for richer reconstruction and increased workspace coverage - promising hitherto unexplored application scenarios, such as assistive structural search.

II. RELATED WORK

We refer to only more recent 3D literature among the vast and varied landscape. State-of-the-art loop closure, camera

³ Note that our approach considers surface patches as far as half a frame apart from the outset - a distinct difference from popular convolutional network based learning approaches that start by building local features.

relocalization and place recognition approaches have been primarily based on visual information ([28] presents a recent survey). Many rely on landmark-based features, such as SIFT or ORB, for instance [34, 1, 31, 26]. Approaches such as [38] have focused on the classification problem - one of categorizing similar scenes. [38] utilizes user annotated 3D data to categorize scenes with viewpoint invariance.

Recent state-of-the-art sensor relocalization approaches in real world structural settings [46, 4, 26, 47, 22] are appearance-reliant as well. They also have other critical requirements like scene-specific learning, and / or workspace models or apriori constructed feature clouds (Section IV).

As discussed earlier, high-performing scene association approaches operating solely on 3D range/depth data have been relatively scarce. A significant amount of efforts have been put on local 3D point features, such as [44, 37, 42]. There have also been work based on complete point clouds include variants of Iterative Closest Point, Normal Distributions Transform and aggregated 3D features (often position based, such as height above ground, [13]) over densely sampled keypoints. While they work well under some conditions, their performance deteriorates quickly with increasing change in viewpoint and sensor rotations - [15, 14, 39] amongst others, have noted this as well.

[39] matches surface patches between views operating on range / depth data. Our geometric property extraction is along similar lines, and our validation scheme builds upon it. In contrast to [39], which presents a localized surface patch matching algorithm based on aligning geometric sequences defined over neighborhood patches, this work focuses on capturing holistic scene level content for ascertaining geometric content similarity and retrieval.

A number of successful methods exist for shape-based retrieval. [27] presents a recent survey. Shape retrieval approaches are designed to work with CAD object models or clutter free, object-centric data, often with 3D figure-ground information (in contrast to raw, egocentric scene data from noisy sensors)⁴. There have been some successful approaches

⁴ These also involve specific assumptions - for example, watertight manifolds, surfaces with geometric texture, or disparate / distinctive topology.

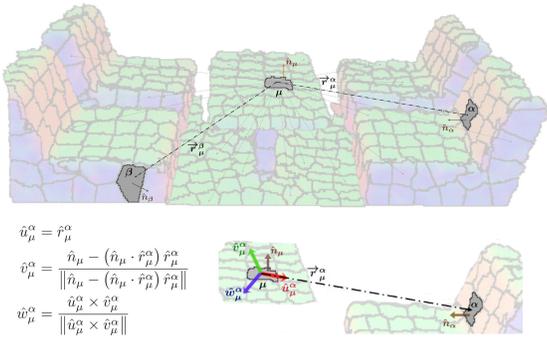


Fig. 2: For a given patch μ , relative and invariant 3D properties are extracted with respect to patches in a non-local neighborhood. To facilitate that, an orthonormal, viewpoint agnostic frame is derived using the Gram-Schmidt process.

for 3D object instance detection in clutter, by employing pre-ascertained 3D object templates, for example [21, 2]. More recently, approaches such as [36] have learnt object point clouds to identify 3D shapes with distinct topology.

III. METHODOLOGY

Given a queried scene-view, \mathcal{V}_Q , and an extant database, \mathcal{D} , of various views from various scenes, $\{\mathcal{V}_s\}_{\mathcal{D}}$ - our algorithm **a)** Retrieves a set of views which have structurally similar content as \mathcal{V}_Q , **b)** Identifies a geometrically diverse subset of views from this retrieved set, and **c)** Ascertain whether some of these views pertain to the same scene as \mathcal{V}_Q (Figure 1).

We denote $\{x_i\}_{i=1}^{c_X}$, $x_i \in X$ to indicate \mathcal{V}_X 's segmentation into smooth surface patches. $\{X_h\}_{h=1}^H$ denotes the segmentation hierarchy then. Hierarchy generation is outlined in Section III-F. To simplify notation, we only indicate the hierarchy level h when it improves clarity.

A. Geometric feature space description

Geometric property set extraction : For a given view \mathcal{V}_X , at a particular segmentation level - we first express each patch x_i through a 13- D vector set, F'_{x_i} of robust, viewpoint agnostic and macro scale 3D geometric properties. These are derived by utilizing 3D relationships relative to other patches in x_i 's neighborhood, \mathcal{N}_{x_i} (along similar lines as [39]). Note that \mathcal{N}_{x_i} is large, non-local - it could span the entire segmentation, $X - x_i$. Neighboring patch count, $|\mathcal{N}_{x_i}|$, is indicated as c_{x_i} .

For a patch $x_i \equiv \mu \in X$, we denote its mean surface normal as \hat{n}_μ and its 3D mean as l_μ . Denoting α to indicate a patch in μ 's neighborhood, with \hat{n}_α , l_α denoting its normal and mean respectively - an orthonormal basis can be derived from the spanning vectors \hat{n}_μ and $r_\mu^\alpha = l_\alpha - l_\mu$ through the Gram-Schmidt process. Figure 2 illustrates this. It also formulates the resultant orthonormal basis, $\langle \hat{u}_\mu^\alpha, \hat{v}_\mu^\alpha, \hat{w}_\mu^\alpha \rangle$, where \hat{u}_μ^α is the unit vector in the direction of r_μ^α . Note that coordinate frame spanned by this orthonormal basis is agnostic (invariant) of the sensing viewpoint, since it is a reference frame local to μ & α . Also note that this basis is seldom degenerate, as \hat{n}_μ and r_μ^α are rarely colinear, especially when data frames are captured from a projective sensing process.

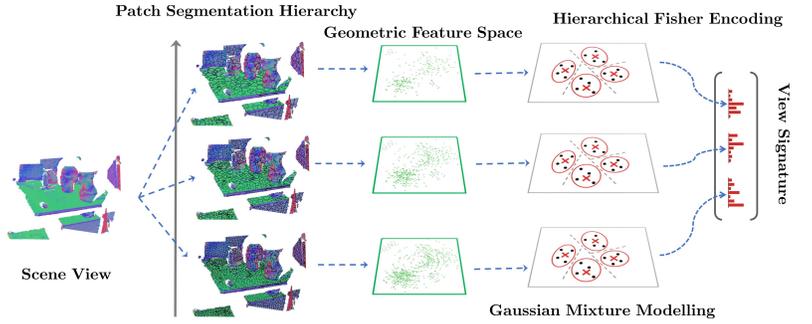


Fig. 3: Geometric properties are extracted over a hierarchy of patch segmentations. At each segmentation level, the aggregate sets of properties is first mapped to a viewpoint invariant geometric feature space. These are then jointly encoded as a view level signature using fisher vector embedding.

For each neighboring surface patch α in μ 's neighborhood, \mathcal{N}_μ , we are able to thus extract the following vector of viewpoint invariant properties, $\{f_\mu^\alpha\}_{\forall \alpha \in \mathcal{N}_\mu}$:

$$f_\mu^\alpha = [\theta_{\hat{n}_\alpha, \hat{n}_\mu}, \theta_{\hat{u}_\mu^\alpha, \hat{n}_\mu}, \theta_{\hat{u}_\mu^\alpha, \hat{n}_\alpha}, r_\mu^\alpha \cdot \hat{n}_\mu, \hat{n}_\alpha \cdot \hat{u}_\mu^\alpha, \hat{n}_\alpha \cdot \hat{v}_\mu^\alpha, \hat{n}_\alpha \cdot \hat{w}_\mu^\alpha, \dots, r_\mu^\alpha \cdot (\hat{n}_\alpha \times \hat{n}_\mu), \|r_\mu^\alpha\|, \|r_\mu^\alpha\| \cdot \text{sgn}_{e_\theta}(\hat{n}_\mu \cdot \hat{u}_\mu^\alpha), \|r_\mu^\alpha\| \cdot \text{sgn}_{e_\theta}(\hat{n}_\alpha \cdot \hat{u}_\mu^\alpha), \dots, \|r_\mu^\alpha\| \cdot \text{sgn}_{e_\theta}(\hat{n}_\alpha \cdot \hat{v}_\mu^\alpha), \|r_\mu^\alpha\| \cdot \text{sgn}_{e_\theta}(\hat{n}_\alpha \cdot \hat{w}_\mu^\alpha)]^T \quad (1)$$

The θ above refers to the angle between the indicated vectors and \times represents an outer product. $\text{sgn}_{e_\theta}(\cdot)$ is a robust signum function that clamps to zero when its parameter $\notin [\cos^{-1}(PI - e_\theta), \cos^{-1}(e_\theta)]$, with e_θ accounting for allowable tolerance to angular noise.

The feature vector f_μ^α basically represents an overcomplete characterization of relative properties between the two patches - formulated in a viewpoint agnostic fashion. The first part (first 9 features) captures angular relationships between r_μ^α , \hat{n}_α & \hat{n}_μ , characterizes \hat{n}_α in the invariant frame derived from r_μ^α and \hat{n}_μ , and characterizes r_μ^α . The second part (remaining 4) consists of robustified features - as a measure against noises arising due to estimation from real world, noisy data. Signs of projected normals' components are captured through robust signum functions and are augmented with the magnitude of relative displacement vector.

Feature space projection : A patch's property set $F'_{x_i} = \{f'_{x_i}^\alpha\}_{\alpha \in \mathcal{N}_{x_i}}$ is then projected onto a subspace learnt through nonlinear independent component analysis ([23]). The projection reduces redundancy in $f'_{x_i}^\alpha$, denoising and making the components more independent. Importantly, this fits with the component independence assumption made in Section III-B to train Gaussian mixtures with diagonal covariances.

The feature space projection results in a 12- D feature vector set $F_{x_i} = \{f_{x_i}^\alpha\}_{\alpha \in \mathcal{N}_{x_i}}$. By considering the patches in x_i 's macro scale neighborhood, \mathcal{N}_{x_i} , the feature set F_{x_i} can thus robustly express the 3D geometry in x_i 's non-local neighborhood. An aggregation of such feature sets arising from all the patches, $F^X = \{F_{x_i}\}_{i=1}^{c_X} \equiv \{f_{x_i}^\alpha | x_i \in X, \alpha \in \mathcal{N}_{x_i}\}$, can thus invariantly and richly express the geometry of the entire scene as captured by \mathcal{V}_X . Finally, the above procedure is repeated for each level in the segmentation hierarchy, to capture fine as well as coarse details. This results in a hierarchy of aggregate feature vector sets, $\{F_h^X\}_{h=1}^H$.

B. Encoding feature space statistics

To obtain a descriptive signature for a given view, \mathcal{V}_X , we encode the aggregated feature sets using Fisher vector embedding (FV, [20, 35]) - this captures the normalized gradient of the log-likelihood of the feature sets. The Fisher kernel theory, first presented in [20], introduces a similarity kernel, arising as a consequence of maximizing the log-likelihood of generatively modeled data. In this paper, Gaussian Mixture Models (GMM) were used to model the feature space distribution.

Given a learnt GMM, P_Θ , parameterized as $\Theta = \{p_g, \nu_g, \Lambda_g\}_1^G$, the FV embedding of the aggregate feature set F^X , indicated as $\phi(F^X)$, is obtained as $\phi(F^X) = L_\Theta \nabla_{\Theta} \log(P_\Theta(F^X))$. Here, L_Θ is the Cholesky decomposition factor of the inverse Fisher Information Matrix, and $\nabla_{\Theta} \log(P_\Theta(F^X))$ is the score function (log-likelihood gradient). Following similar analysis as [35], under assumptions of diagonal covariance matrices, Λ_g , and independence of the samples, $f_{x_i}^\alpha$, the embedding evaluates as $\phi(F^X) = [m_1^0, m_1^1, m_1^2, \dots, m_g^0, m_g^1, m_g^2, \dots, m_G^0, m_G^1, m_G^2]^T$ - where m_g^0, m_g^1, m_g^2 respectively capture the normalized zeroth, first and second order statistics of the sample set that falls in the g -th mixture component of the GMM. $\phi(F^X)$ has a dimensionality of $d_\phi = (2d_F + 1) \cdot G$, where G is number of mixture components, and $d_F = 12$ is the dimensionality of our geometric feature space. Below, $\mathbf{1}$ denotes an all-one vector and $\pi_{ij,g} = \frac{\exp\left[-\frac{1}{2}(f_{x_i}^{x_j} - \nu_g)^T \Lambda_g^{-1} (f_{x_i}^{x_j} - \nu_g)\right]}{\sum_{g=1}^G \exp\left[-\frac{1}{2}(f_{x_i}^{x_j} - \nu_g)^T \Lambda_g^{-1} (f_{x_i}^{x_j} - \nu_g)\right]}$.

$$m_g^0 = \frac{1}{c_X c_{x_i} \sqrt{p_g}} \sum_{i=1}^{c_X} \sum_{j=1}^{c_{x_i}} (\pi_{ij,g} - p_g) \quad (2a)$$

$$m_g^1 = \frac{1}{c_X c_{x_i} \sqrt{p_g}} \sum_{i=1}^{c_X} \sum_{j=1}^{c_{x_i}} \pi_{ij,g} \Lambda^{-1/2} (f_{x_i}^{x_j} - \nu_g) \quad (2b)$$

$$m_g^2 = \frac{1}{c_X c_{x_i} \sqrt{2p_g}} \sum_{i=1}^{c_X} \sum_{j=1}^{c_{x_i}} \pi_{ij,g} \left[\Lambda^{-1} (f_{x_i}^{x_j} - \nu_g) (f_{x_i}^{x_j} - \nu_g)^T - \mathbf{I} \right] \mathbf{1} \quad (2c)$$

$\phi(F^X)$ is then component-wise square root normalized (by replacing each component, 'a' of $\phi(F^X)$ by $|a|^{1/2} \text{sign}(a)$), and ℓ_2 normalized. The square root normalization serves to alleviate the dominant effect of relatively indiscriminate samples occurring with high frequency (for example, arising from patches on a wall or ceiling) and the ℓ_2 normalization helps generalization across different scenes by normalizing the energy content. The desired view signature vector for \mathcal{V}_X , denoted as $\psi(X)$, is obtained by evaluating the embedding at each level in hierarchy, and concatenating them — $\psi(X) = [\phi(F_1^X)^T, \dots, \phi(F_h^X)^T, \dots, \phi(F_H^X)^T]^T$

C. Similarity and Retrieval

The thus obtained view signature, $\psi(X)$, captures discriminative 3D geometrical properties, and is robust to viewpoint changes, sensor noise, occlusions and other data imperfections by design. As experiments indicate, a metric based on such view signatures is a reliable measure of 3D geometric similarity. We tried ℓ_1 & ℓ_2 distance metrics, and used ℓ_1 for all experiments in the paper as it performed better. Thus the

similarity between two given views \mathcal{V}_X & \mathcal{V}_Y can be denoted as, $s(X, Y) = -(\sum_1^{25GH} |\psi(X) - \psi(Y)|_1)$.

Given a queried view, \mathcal{V}_Q , and a database \mathcal{D} of view signatures, one can thus retrieve a set of putative view associations in the geometric sense through nearest neighbor queries. We indicate this retrieved set of putatively associated views as $\mathcal{R} = \{\mathcal{V}_X\}_{X=1}^{c_{\mathcal{R}}}$.

D. Diversity Sampling with Determinantal Point Processes

Depending on the distribution of scenes' views in the database, $\mathcal{R} = \{\mathcal{V}_X\}_{X=1}^{c_{\mathcal{R}}}$ could be overwhelmed with views which are *near duplicates* (all being very similar to each other, hence almost equally similar to the queried view). This may not be desirable since the subset of top retrievals could just be flooded with near duplicates of false putative associations, resulting in complete failure. By filtering out near duplicates, a diversity based subset selection procedure may still be able to salvage correct, albeit lower ranked, putative associations present in \mathcal{R} with further post-processing validation.

A diverse set of retrievals is generally desirable. It would provide assorted and possibly complementary information, which could be made use of thereon. For instance, it could be potentially beneficial in reconstruction or coverage tasks, where diverse viewpoints observing the environment with only partially overlapping content are more desirable than having redundant views from nearly the same perspective. A querying human user could also be better assisted by being provided with a diverse set of the retrievals to choose from.

Determinantal point processes ([24]) are employed to select a diverse subset of candidate views, \mathcal{C} , from \mathcal{R} . A point process \mathcal{P}_L is called an L - ensemble k -*determinantal point process* if for every random subset, \mathcal{C} , of \mathcal{R} , such that $|\mathcal{C}| = k$, drawn according to \mathcal{P}_L , we have $\mathcal{P}_L(\mathcal{C}; \mathcal{R}) = \frac{\det(L_{\mathcal{C}})}{\sum_{\mathcal{V}_A \in \mathcal{R}, |\mathcal{A}|=k} \det(L_{\mathcal{A}})}$. L here is a symmetric positive semi-definite similarity matrix indexed by the elements of \mathcal{R} . $L_{\mathcal{C}}$ is the principal minor (submatrix) with rows and columns from L indexed by the elements in subset \mathcal{C} . Thus the probability of selecting a subset \mathcal{C} , ($|\mathcal{C}| = k = c_{\mathcal{C}}$) elements is directly proportional to the determinant of the submatrix indexed by it. Note that higher diagonal values would proportionately encourage their inclusion in a selected subset \mathcal{C} as they lead to higher determinants. Similarly, the off-diagonal values determine correlation between different elements, and a high value decreases the determinant overall. Thus two elements with a high similarity value tend not to co-occur in \mathcal{C} . DPP sample sets are therefore able to balance the net significance of their constituent elements with their diversity. We modeled L accordingly as follows

$$\{L\}_{X,Y} = \rho_X \rho_Y \kappa e^{-\frac{s(X,Y)}{\sigma}}, \quad 1 \leq X, Y \leq c_{\mathcal{R}} \quad (3)$$

where $\rho_X = e^{\frac{1}{2} \frac{s(X,Q)}{\omega}}$, $\exists X \in \mathcal{R}$ models the similarity of a retrieved view \mathcal{V}_X to the queried view \mathcal{V}_Q . The similarity between two given views \mathcal{V}_X and \mathcal{V}_Y is captured by the rightmost term. Positive valued parameters σ , ω and κ can be tuned to balance the need for both diversity and similarity to \mathcal{V}_Q . A lower sigma would induce a higher resolution in

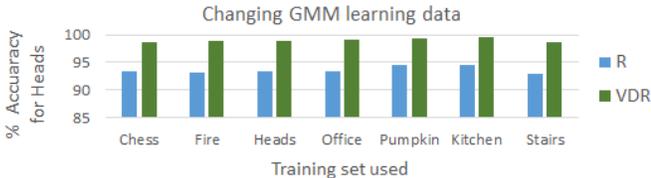


Fig. 4: Consistency in GMM learning. Similar retrieval accuracies were achieved with GMMs learnt from each of the 7 training sets.

similarity scores between retrieved views, and hence would result in a more diverse subset selection.

While the MAP inference on \mathcal{P}_L to determine the most probable subset is NP-hard, efficient sampling algorithms exist which provide good approximate solutions in practice. For our purposes, a greedy procedure based on [24] which results in $\mathcal{O}(k \log k)$ -approximation worked well.

E. Validating candidate views for association

We employ a finer grained spatial validation step before finally associating the queried view with some of the views in the candidate set, $\mathcal{C} = \{\mathcal{V}_X\}_{X=1}^c$. This is done by directly leveraging the rigid 3D spatial arrangement of surface patches to ascertain surface alignment. We make use of the patch matching scheme presented in our prior work [39]. It utilizes a sequence alignment scheme over similarly motivated patch properties to find standalone correspondences based on 3D neighborhood similarity. A semi-dense set of correspondences can be ascertained. Rigid transform between two views of a given scene can then be robustly, accurately computed through consensus of patch associations.

When views from scenes with different geometrical content are matched through [39], the matches would likely be inconsistent with respect to the computed transform. We exploit this understanding to validate associations with candidate views. For each candidate view, $\mathcal{V}_X \in \mathcal{C}$, and the queried view, \mathcal{V}_Q , we utilize randomly sampled patches to estimate rigid transformations both ways, that is, $T_Q^X \equiv (R_Q^X, t_Q^X)$ and $T_X^Q \equiv (R_X^Q, t_X^Q)$ and check whether they are consistent with each other. We ascertain a candidate $\mathcal{V}_X \in \mathcal{C}$ as associated with \mathcal{V}_Q when $\|\log(R_Q^X R_X^Q)\|_2 \leq \epsilon_{val}^r$ & $\|t_Q^X + t_X^Q\|_2 \leq \epsilon_{val}^t$ - we are basically ensuring that the magnitude of the rotation and translation components in the residual transform, $T_Q^X T_X^Q$, are below certain thresholds $\{\epsilon_{val}^r, \epsilon_{val}^t\}$.

F. Further details and discussion

The approach is amenable to any boundary-preserving patch segmentation scheme, as long as it results in superpixels / patches that are geometrically regularized for smoothness and compactness. For example [33], which segments volumetrically, could be used while working with point clouds; and surface segmentation schemes such as one presented in [39] could be employed when working with depth / range images. Both [39] and [33] performed well in our experiments. Starting with the base level segmentation, a segmentation hierarchy can be built in either fine to coarse (agglomerative), or coarse to fine (divisive) fashion. Each successive level has patches reduced (increased, in case of divisive) by a constant factor - this can be approximately

ensured by employing K-Means in 3D with near uniform surface component seeding (any resultant patches below a certain size / surface area are merged back). We used four levels of segmentation hierarchy ($H = 4$). The number of mixture components were also kept fixed, $G = 1250$. The GMMs were learnt through an expectation maximization scheme, and the mixture components were initialized from the result an iteration of K-Means++ procedure. Our empirical analysis indicated the learnt feature space distribution to be general for similar sensor types⁵. Figure 4 suggests that as well. In fact, a single set of Gaussian mixture (and ICA) models were utilized for all the experiments shown in the article (except Figure 4).

In practice, for efficiency, while encoding feature space statistics (Section III-B), it suffices to approximately ascertain F_h^X by sampling patches from X_h , and subsequently sampling the neighborhoods of the sampled patches. This also partly corroborates our assertion that the methodology is robust to occlusions. Databases were indexed as KD-trees. Our current straight up implementation is not optimized for efficiency (on a 4.2 GHz, 4 core setup, III-A - III-D takes $\sim .3$ ms, 1000 superpixels), though the methodology is GPU parallelizable. Most of the procedures outlined in Sections III-A, III-B, III-C, III-D and III-E can be GPU parallelized in a straightforward fashion. The computational bottleneck arises during validation, which is quadratic in number of superpixels (~ 1 s for segmentation with 1000 superpixels at finest level, but again naturally parallelizable). Note that it suffices to validate at a coarse hierarchical level (~ 250 superpixels) — the result, T_Q^X , can then be used as reliable initialization and be quickly refined iteratively as per task.

IV. EXPERIMENTS

In all experiments, the method indicated 'R' refers to our retrieval approach (till Section III-B), without the diverse subset selection and validation steps. 'DR' refers to our approach till Section III-D, with diversification but without the validation step. 'VDR' would then refer to the complete approach, resulting in the set \mathcal{C}_{val} - diverse retrievals which have been validated through rigid overlap consistency. The retrievals in both the sets \mathcal{C} and \mathcal{C}_{val} follow the same order (by $s(X, Y)$) as they appear in the initial retrieval set \mathcal{R} . All analysis is done on the top few results from these sets.

The retrieval and association problems can be subjective - two views with only partially overlapping geometric content can be evaluated differently by users. We employed an objective measure - evaluating our retrieval approach on a sensor relocalization task. We utilized the 7-scenes datasets from [12, 40], the standard benchmark for indoor RGB/RGB-D relocalization. The objective is to localize the sensor (ascertain pose) with respect to the workspace within the maximal allowable translation and orientation errors (5 cm and 5 deg respectively). The datasets are collected from different workspaces (although some scenes in *Redkitchen* and *Pumpkin*

⁵ Sufficient number of GMM components should be utilized to span the extent of the geometric feature space. This is a function of maximum scene scale captured, and thus sensor range.

Data	Appearance Reliant (RGB or RGB-D)							Depth – Only		
Approach	Reconstruction Truth Needed for Relocalization							Retrieval		
Method	<i>Spr</i> [40]	[5] <i>C</i>	<i>DSc</i> [4]	[40]	[16]	[46]	[5]	<i>D</i> [40]	<i>R</i>	<i>VDR</i>
<i>Chess</i>	70.7	94.9	97.4	92.6	96	99.4	99.6	82.7	97.3	99.5
<i>Fire</i>	49.9	73.5	74.3	82.9	90	94.6	94.0	44.7	92.3	97.8
<i>Heads</i>	67.6	48.1	71.7	49.4	56	95.9	89.3	27.0	93.5	98.9
<i>Office</i>	36.6	53.2	71.2	74.9	92	97.0	93.4	65.5	89.7	98.4
<i>Pumpkin</i>	21.3	54.5	53.6	73.7	80	85.1	77.6	15.1	78.3	82.8
<i>Kitchen</i>	29.8	42.2	51.2	71.8	86	89.3	91.1	61.3	87.9	93.7
<i>Stairs</i>	9.2	20.1	4.5	27.8	55	63.4	71.7	13.6	54.8	61.0
<i>Average</i>	40.7	55.2	60.1	67.6	79.3	89.2	88.1	44.3	84.8	90.3
<i>Combine</i>	38.6	55.2	62.5	-	-	-	-	-	84.8	90.4

TABLE I: The presented approaches (*R*, *VDR*) are compared with baselines through localization accuracies on the standard 7-scenes datasets from [12, 40]. All methods utilize RGB-D data during training, except [40] *D*, and our *R* and *VDR*, which are based on range / depth data. During test time, the three leftmost approaches only take RGB images as input, while the three rightmost approaches only take range / depth images - the rest operate on RGB-D. *Average* indicates the average among the 7 datasets. *Combine* indicates performance when jointly considering all 7 scenes as a single database. *VDR* outperforms all the RGB-D approaches while using depth information *only*. *R* performs very well as well, outperforming all but two RGB-D approaches.

are quite similar). Standard train - test splits are provided, with the viewpoints in the test set differing significantly from the training set. This makes it most appropriate for use in the evaluation⁶. 7-scenes also provide additional training information - global sensor pose annotations, as well as reconstructed volumetric workspace models.

In our approach, *R*, *DR* and *VDR*, depth images for training were simply encoded as an unordered view-signature database. A given query image from the test split was localized by computing the relative transform with respect to the top retrieval (in the sets \mathcal{R} , \mathcal{C} and \mathcal{C}_{vld} respectively), and the localization accuracy was computed by evaluating the disparity between the estimated and ground truth relative poses. Same as in the baselines, 5 cm and 5 deg are the allowable error. Our approach did not require additional information accompanying the datasets to operate (pose truth annotations and workspace reconstructs). Importantly, it also did not require specific training for each dataset. This differs from most of our baselines which required some additional information or dataset-specific training.

Baselines: We compare our approach against many baselines. Approaches like [41, 45, 16, 22, 4, 47, 30] require additional information and dataset specific training. They rely on annotations, workspace models, and involve regression against absolute sensor poses or 3D coordinates of pixels. Deep-CNN based regressors have been proposed as well, such as [22, 4, 47, 30]. Such approaches can overfit on the training data, and are difficult to generalize to scenes that are not similar to the training. Some baseline results were not shown in Table I — [12], which presents a random ferns based retrieval method over RGB-D, report accuracies differently; but they indicate the achieved results to be weaker than some of the baselines considered in Table I. Methods like [22, 47, 30] report localization accuracies as median errors - since the lowest reported median errors, that we are aware of, are greater than 10 cm (translation, implicitly includes

⁶ As opposed to mapping, visual odometry or semantic scene datasets such as [43, 17, 48, 8, 18]. These either do not have enough loop closures and/or are synthetic, or lack ground truth for quantitative evaluation or standard train-test splits for loop closure.

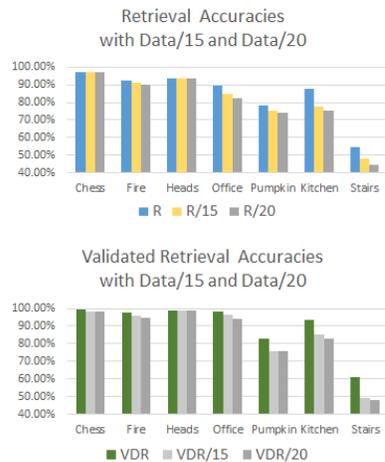


Fig. 5: Accuracies with significantly sparser acquisition. Database sizes were reduced to 1/15 and 1/20.

orientation errors as well), these methods are also not as accurate as some of the baselines in Table I. Approaches [26] and *Sparse* [40] employ frame to model matching for relocalization. They match local features from the query frame to a global feature cloud accumulated and reconstructed *a priori* from the training data and the pose ground-truth annotations. [26] shows nice results, though we were unable to obtain exact numbers from the authors. However, *VDR* in Table I does seem to perform better than [26] in 4 out of 7 datasets in comparison. *VDR* also seems to outperform [26] in at least 6 out of 7 datasets when only 1/15 of the training data is used (Figure 5). All the aforementioned approaches are appearance-reliant as well (except [40] which additionally present a depth only variant). We also tried a retrieval methodology similar to ours with local 3D geometric point-features (such as [44]), but their performance was worse than those shown in Table I.

As Table I shows, *VDR* achieved state-of-the-art results through pure geometry alone - without needing any additional annotations, assumptions or appearance features. Equally promising were the results from *R* which were obtained by simply using the first retrieval in \mathcal{R} (no diversification or validation), which were better than all baselines but two. *DR* gives the same results as *R* in the relocalization experiments and is hence not shown. This is because the accuracies were evaluated with respect to only the top retrieval - this is the same for *R* and *DR* since the greedy algorithm we used for *k*-DPP automatically selects the top-scoring retrieval as the first one. These results support our hypothesis that macro-level 3D geometry holds immense discriminative information.

In the last row of Table I, we combined all training data from the 7 datasets into one single database, and evaluated accuracies of the combined test splits. As can be seen, the results held up quite well in the combined experiment, when the database size and complexity (variety, aliasing) was drastically increased.

We also tabulated the affect of significantly reducing the database sizes - by re-evaluating results with databases built from only 1/15th and 1/20th of the available train-

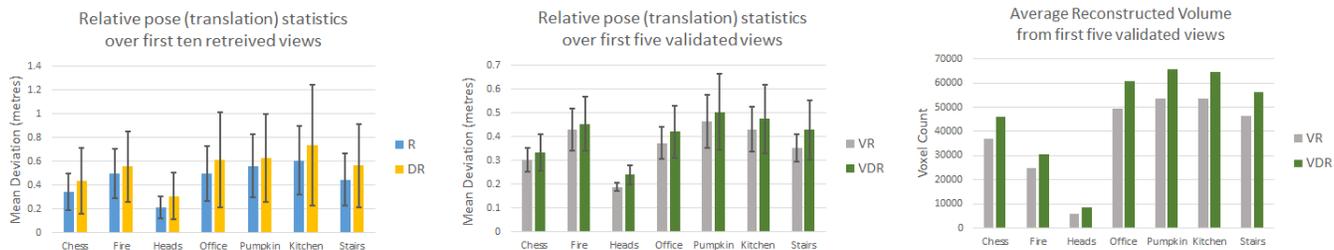


Fig. 6: Quantifying diversity. Left, Middle: The average relative translation of the retrieved views with respect to the queried view. One can see DR improves diversity over R, and VDR improves over VR. Right: Efficacy of diverse viewpoints for reconstruction task. The average number of voxels (in a 8 cm^3 occupancy grid) occupied by ground truth reconstructs from the first five validated retrievals from VR and VDR are plotted. From the same number of initial views, VDR results in richer reconstructs that capture significantly more voxels in the scene.

splits for each dataset. With a much sparser coverage of the environment, both retrieval and subsequent validation and localization becomes much more difficult. The frames were sampled at uniform intervals, thus may have steep viewpoint changes, much reduced content overlap and significantly increased occlusions. As Figure 5 indicates, the accuracies of both R and VDR held up quite well. This is indicative of the approach’s robustness to these practical challenges.

The approach generalizes well. Our experiments do not suggest a need for scene specific training - a single set of learnt gaussian mixtures and ICA projection matrices were utilized in all our experiments (except Figure 4). The training data was taken from the train split of Redkitchen in [40], and from datasets in [48, 43], a reasonably rich and diverse set of samples. Figure 4 shows the robustness of the GMM parameters with respect to the dataset used to train it. As can be seen the results stay consistent.

Finally, we conducted experiments to quantify the effect of our diversification approach, and its role in generating significantly richer reconstructions. As Figure 6(left, middle) show, the diversity of retrieved viewpoints is greatly improved due to our DPP-based diversification. Note that DR and VDR select views which are not only further off than the queries (higher relative mean), but result in view sets which have significantly more viewpoint variance amongst themselves as well (significantly higher standard deviations). And as Figure 6(right) shows, the reconstruction volume improves significantly when a diversified set of views is utilized. Figure 7 shows a qualitative example. One can see that the diversified retrievals are significantly more diverse, from varied viewpoints, and are resulting in an appreciably richer reconstruction. In general, retrieval as well as geometric diversity is often desirable - apart from reconstruction, it would prove useful in other tasks such as structure and semantic analysis, and ‘human in loop’ selection tasks.

V. CONCLUSION

We presented a robust solution to the problems of measuring 3D geometric similarity between 3D range images or point clouds, and determining whether they come from the same scene. A general-purpose retrieval approach was proposed, based on encoding (FV) of viewpoint-invariant features that are hand-crafted to capture 3D geometry at macro scales. The approach performed well in real world settings - including ones that involved sharp viewpoint changes, partially overlapping and occluded content. It scaled well,

and did not require scene-specific training - making it useful in a variety of scenarios. As experiments established, the approach is powerful and did better than specifically fitted solutions such as CNNs trained on RGB or RGB-D data. Furthermore, we introduced a way to obtain *geometrically diverse* retrievals (DPP), and showed how such retrievals can help generate richer reconstructions. Interestingly, in contrast to CNN approaches which begin with a local neighborhood, our approach utilized macro scale features from start. The combination of both paradigms would be explored in future work, for this and other tasks involving 3D recognition.

Acknowledgements

RS and FL are supported in part by NSF grant IIS-1464371.

REFERENCES

- [1] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. Seitz, and R. Szeliski. Building Rome in a day. *Communications of the ACM*, 2011.
- [2] L. Bo, X. Ren, and D. Fox. Depth kernel descriptors for object recognition. In *Intelligent Robots and Systems (IROS)*. IEEE, 2011.
- [3] M. Bosse and R. Zlot. Place recognition using keypoint voting in large 3D lidar datasets. In *Robotics and Automation (ICRA)*. IEEE, 2013.
- [4] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother. DSAC - differentiable RANSAC for camera localization. *CoRR*, abs/1611.05705, 2016.
- [5] E. Brachmann, F. Michel, A. Krull, M. Y. Yang, S. Gumhold, and C. Rother. Uncertainty-driven 6D pose estimation of objects and scenes from a single RGB image. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [6] S. Choi, Q.-Y. Zhou, and V. Koltun. Robust reconstruction of indoor scenes. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [7] R. Cupec, E. K. Nyarko, D. Filko, A. Kitanov, and I. Petrović. Place recognition based on matching of planar surfaces and line segments. *The International Journal of Robotics Research*, 2015.
- [8] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3D reconstructions of indoor scenes. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [9] H. Durrant-Whyte and T. Bailey. Simultaneous localization and mapping: part I. *IEEE robotics & automation magazine*, 2006.
- [10] G. D. Evangelidis, D. Kounades-Bastian, R. Horaud, and E. Z. Psarakis. A generative model for the joint registration of multiple point sets. In *European Conference on Computer Vision*, 109–122. Springer, 2014.
- [11] E. Fernandez-Moral, W. Mayol-Cuevas, V. Arevalo, and J. González-Jiménez. Fast place recognition with plane-based maps. In *Robotics and Automation (ICRA)*. IEEE, 2013.
- [12] B. Glocker, S. Izadi, J. Shotton, and A. Criminisi. Real-time RGB-D camera relocalization. In *Mixed and Augmented Reality (ISMAR)*, 2013.
- [13] K. Granström, T. B. Schön, J. I. Nieto, and F. T. Ramos. Learning to close loops from range data. *The international journal of robotics research*, 30(14):1728–1754, 2011.
- [14] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, and J. Wan. 3D object recognition in cluttered scenes with local surface features: A survey. *Pattern Analysis and Machine Intelligence*, 2014.
- [15] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, J. Wan, and N. M. Kwok. A comprehensive performance evaluation of 3D local feature descriptors. *International Journal of Computer Vision*, 1–24, 2015.

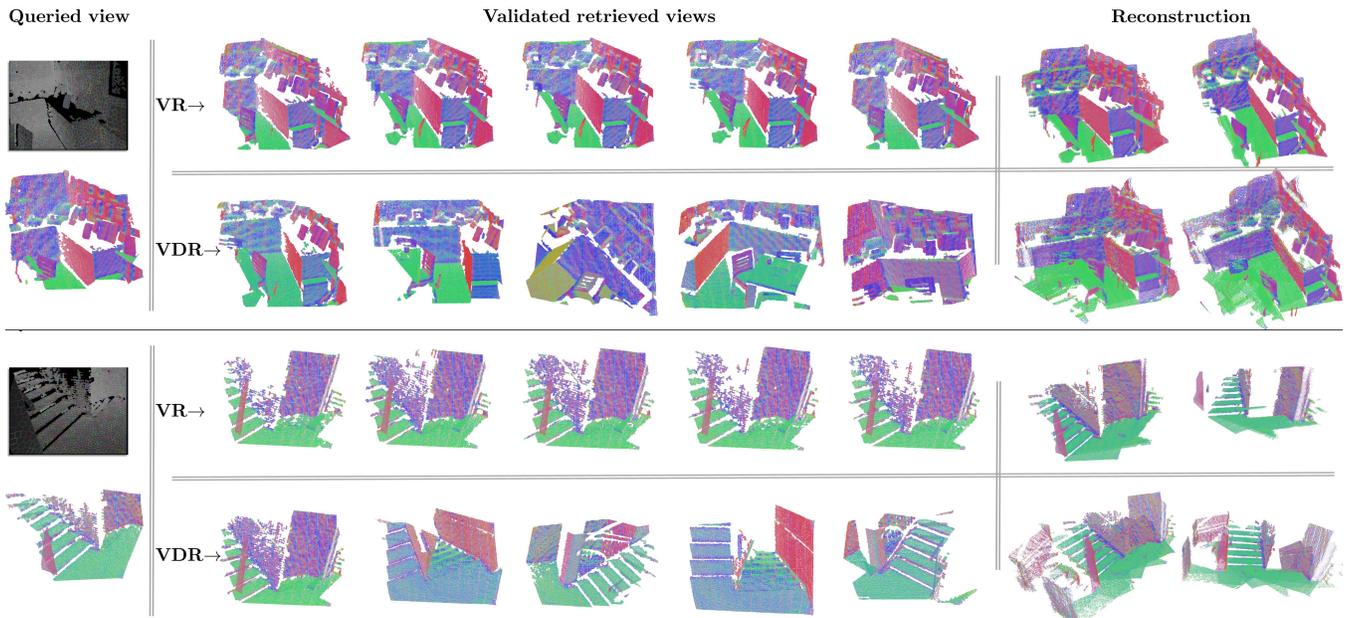


Fig. 7: Retrievals and reconstructions are shown for two example scenes (top and bottom). For each scene, queried view is shown on the left as a depth image with overlaid patch boundaries. Views on top row are the top-five retrieved and validated views without using DPP. Views on the bottom row are the top-five validated views with DPP. Reconstructed scene models from the respective sets are shown on the right from two perspectives. Note that viewpoints vary significantly in the diversified retrievals, and results in a much larger reconstructed volume (over 1.5x).

- [16] A. Guzman-Rivera, P. Kohli, B. Glocker, J. Shotton, T. Sharp, A. Fitzgibbon, and S. Izadi. Multi-output learning for camera relocalization. In *Computer Vision and Pattern Recognition*, 2014.
- [17] A. Handa, T. Whelan, J. McDonald, and A. J. Davison. A benchmark for RGB-D visual odometry, 3D reconstruction and slam. In *Robotics and automation (ICRA)*. IEEE, 2014.
- [18] A. Handa, V. Patraucean, V. Badrinarayanan, S. Stent, and R. Cipolla. Scenenet: Understanding real world indoor scenes with synthetic data. *arXiv preprint arXiv:1511.07041*, 2015.
- [19] M. Himstedt and E. Maehle. Geometry matters: Place recognition in 2D range scans using geometrical surface relations. In *Mobile Robots (ECMR)*, 1–6. IEEE, 2015.
- [20] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. *Advances in neural information processing systems*, 487–493, 1999.
- [21] A. E. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3D scenes. *Pattern Analysis and Machine Intelligence*, 1999.
- [22] A. Kendall and R. Cipolla. Geometric loss functions for camera pose regression with deep learning. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [23] Z. Koldovskiy, P. Tichavsky, and E. Oja. Efficient variant of algorithm FastICA for independent component analysis attaining the Cramér-Rao lower bound. *IEEE Transactions on Neural Networks*, 2006.
- [24] A. Kulesza and B. Taskar. Determinantal point processes for machine learning. *arXiv preprint arXiv:1207.6083*, 2012.
- [25] M. Labbe and F. Michaud. Online global loop closure detection for large-scale multi-session graph-based SLAM. In *Intelligent Robots and Systems (IROS)*. IEEE, 2014.
- [26] S. Li and A. Calway. RGBD relocalisation using pairwise geometry and concise key point sets. In *Robotics and Automation (ICRA)*, 2015.
- [27] B. Li, Y. Lu, C. Li, A. Godil, T. Schreck, M. Aono, M. Burtcher, Q. Chen, N. K. Chowdhury, B. Fang, et al. A comparison of 3D shape retrieval methods based on a large-scale benchmark supporting multimodal queries. *Computer Vision and Image Understanding*, 2015.
- [28] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford. Visual place recognition: A survey. *IEEE Transactions on Robotics*, 32(1):1–19, 2016.
- [29] M. Magnusson, H. Andreasson, A. Nüchter, and A. J. Lilienthal. Automatic appearance-based loop detection from three-dimensional laser data using the normal distributions transform. *Journal of Field Robotics*, 26(11-12):892–914, 2009.
- [30] I. Melekhov, J. Ylioinas, J. Kannala, and E. Rahtu. Image-based localization using hourglass networks. *CoRR*, abs/1703.07971, 2017.
- [31] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 2015.
- [32] E. Olson. Recognizing places using spectrally clustered local matches. *Robotics and Autonomous Systems*, 57(12):1157–1172, 2009.
- [33] J. Papon, A. Abramov, M. Schoeler, and F. Worgotter. Voxel cloud connectivity segmentation - supervoxels for point clouds. In *Computer Vision and Pattern Recognition (CVPR)*, 2013.
- [34] R. Paul and P. Newman. FAB-MAP 3D: Topological mapping with spatial and visual appearance. In *Robotics and Automation (ICRA)*, 2010.
- [35] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *Computer Vision (ECCV)*. 2010.
- [36] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, 5105–5114, 2017.
- [37] R. B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms for 3D registration. In *Robotics and Automation (ICRA)*, 2009.
- [38] S. Satkin and M. Hebert. 3DNN: Viewpoint invariant 3D geometry matching for scene understanding. In *Computer Vision (ICCV)*, 2013.
- [39] R. Sawhney, F. Li, and H. Christensen. GASP: Geometric association with surface patches. In *3D Vision (3DV)*, 2014.
- [40] 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 *Computer Vision and Pattern Recognition*, 2013.
- [41] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 2013.
- [42] B. Steder, M. Ruhnke, S. Grzonka, and W. Burgard. Place recognition in 3D scans using a combination of bag of words and point feature based relative pose estimation. In *Intelligent Robots and Systems (IROS)*, 2011.
- [43] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *Intelligent Robots and Systems (IROS)*, 2012.
- [44] F. Tombari, S. Salti, and L. Di Stefano. Unique signatures of histograms for local surface description. In *Computer Vision (ECCV)*, 2010.
- [45] J. P. Valentin, S. Sengupta, J. Warrell, A. Shahrokni, and P. H. Torr. Mesh based semantic modelling for indoor and outdoor scenes. In *Computer Vision and Pattern Recognition (CVPR)*, 2013.
- [46] J. Valentin, M. Nießner, J. Shotton, A. Fitzgibbon, S. Izadi, and P. Torr. Exploiting uncertainty in regression forests for accurate camera relocalization. In *Computer Vision and Pattern Recognition*, 2015.
- [47] F. Walch, C. Hazirbas, L. Leal-Taixe, T. Sattler, S. Hilsenbeck, and D. Cremers. Image-based localization using LSTMs for structured feature correlation. In *Computer Vision (ICCV)*, 2017.
- [48] J. Xiao, A. Owens, and A. Torralba. SUN3D: A database of big spaces reconstructed using SfM and object labels. In *Computer Vision*, 2013.