Relocalization with Submaps: Multi-session Mapping for Planetary Rovers Equipped with Stereo Cameras

Riccardo Giubilato[1,2], Mallikarjuna Vayugundla[2], Martin J. Schuster[2], Wolfgang Stürzl[2], Armin Wedler[2], Rudolph Triebel[2] and Stefano Debei[1]

***Abstract*—To enable long term exploration of extreme environments such as planetary surfaces, heterogeneous robotic teams need the ability to localize themselves on previously built maps. While the Localization and Mapping problem for single sessions can be efficiently solved with many state of the art solutions, place recognition in natural environments still poses great challenges for the perception system of a robotic agent. In this paper we propose a relocalization pipeline which exploits both 3D and visual information from stereo cameras to detect matches across local point clouds of multiple SLAM sessions. Our solution is based on a Bag of Binary Words scheme where binarized SHOT descriptors are enriched with visual cues to recall in a fast and efficient way previously visited places. The proposed relocalization scheme is validated on challenging datasets captured using a planetary rover prototype on Mount Etna, designated as a Moon analogue environment.**

***Index Terms*—Localization; Space Robotics and Automation; Mapping**

I. INTRODUCTION

THE capability of autonomous robots to recognize previously visited places is crucial for the success of long-term missions. In case of GPS denied environments, such as for planetary scenarios, exploration robots can map and localize themselves by performing SLAM (Simultaneous Localization and Mapping). Place recognition enables multiple agents to join their maps under a common reference frame or to merge subsequent mapping sessions of a single robot. Images are widely used for place recognition and loop closure detection in Visual SLAM [1]. However, in the presence of varying visual appearance such as in case of strong illumination and viewpoint changes, traditional approaches for matching visual information are severely compromised. LiDAR (Light Detection and Ranging) sensors allow instead to exploit the 3D information of an environment for SLAM purposes and increasing attention is directed towards place recognition and localization with 3D point clouds [2]. However, stereo cameras are preferable as space qualifiable instruments thanks to their mechanical simplicity.

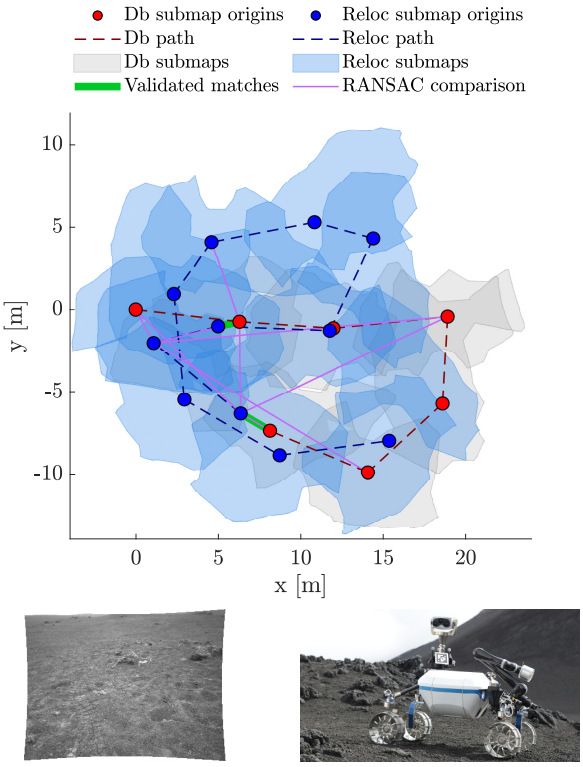


Fig. 1. Top: Relocalization on Mt. Etna (top view, session *Etna_easy*). Submaps are represented as transparent patches, shaded in blue for the active relocalization session ("Reloc") and in grey for the existing map in the data base ("Db"), and their origins by circular markers. Green lines connect the origins of matching submaps validated by our relocalization pipeline, which, compared to the standard RANSAC approach (magenta lines) are outlier-free. Left: Example view (rectified) of the Etna environment. Right: LRU (*Lightweight Rover Unit*) [3] on Mount Etna, Italy.

This work builds upon a 6D Stereo SLAM framework [4] which combines the benefits of using cameras as the main source of perception and exploiting the invariant nature of point clouds. The environment is discretized in *submaps* whose associated local reference frames are connected in a graph by visual-inertial odometry constraints. Submap growth

Manuscript received: September 9, 2019; Revised: November 26, 2019; Accepted: December 19, 2019.

This paper was recommended for publication by Editor Sven Behnke upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by the Helmholtz Association, project alliance ROBEX (contract number HA-304) and project ARCHES (contract number ZT-0033).

[1]R. Giubilato and S. Debei are with CISAS "Giuseppe Colombo", University of Padova, via Venezia 15, 35131, Padova, Italy {riccardo.giubilato@phd.unipd.it} {stefano.debei@unipd.it}

[2]R. Giubilato, M. Vayugundla, M. J. Schuster, W. Stürzl, A. Wedler and R. Triebel are with German Aerospace Center (DLR), Institute of Robotics and Mechatronics, Münchener Str. 20, 82234, Wessling, Germany {firstname.lastname@dlr.de}

Digital Object Identifier (DOI): see top of this page.

is bounded on pose uncertainty and loop closures are found relying on pose priors.

The main contribution presented in this paper is a pipeline for localizing a robot without priors in previous maps, targeted at planetary environments. Binarized 3D SHOT descriptors [5], [6] are enriched with texture information and used to recall similar places in a Bag of Binary Words approach. As the first mapping session is completed, binary 3D descriptors are clustered in a k-d tree to build a vocabulary of words which will serve in the following sessions to recall similar places. However, obtaining a general purpose vocabulary is not feasible since 3D features are generally not scale invariant and very specific to the observed shapes. We address the vocabulary incompleteness by modifying the well known DBoW2 library [7] with a re-weighting scheme.

As stereo depth uncertainty affects the uniqueness and descriptiveness of 3D descriptors [8], we enhance binary SHOT with visual cues by appending a limited size component inspired by Local Binary Patterns [9] and directly computed over the monochrome intensity values associated to the points in the cloud. We will refer to this descriptor as B-Tex-SHOT in the course of this paper. The uniqueness associated to image intensities especially around obstacles should balance the effect of 3D noise without overpowering, which could be dangerous in presence of shadows or changing lighting conditions. Submap correspondences, selected from BoW (Bag of Words) vector similarity, are validated by matching the original SHOT descriptors and by using a voting scheme for the suggested transformation between the origins of each mapping session. To summarize, this paper presents the following contributions:

- we propose a relocalization pipeline for stereo vision systems based on binary 3D descriptors and Bag of Words.
- we exploit the monochrome intensity of 3D points to build a short binary descriptor which we demonstrate to improve recall precision of similar submaps.
- a novel transformation voting scheme is introduced for removing outliers amongst 3D descriptor matches. We prove the superiority of the proposed validation scheme over traditional RANSAC (RANdom SAmple Consensus) approaches.
- we show the effectiveness of the proposed pipeline in several experiments including a challenging outdoor planetary analogue environment. In this experiments, the proposed system is able to correctly localize the rover in previous maps with 100% precision after validation.

This paper is organized as follows: in Section II, we give a brief overview of existing work on place recognition with a focus on point cloud based methods. In Section III, the relocalization pipeline is explained in detail and, in Section IV, we validate the approach with indoor and outdoor experiments.

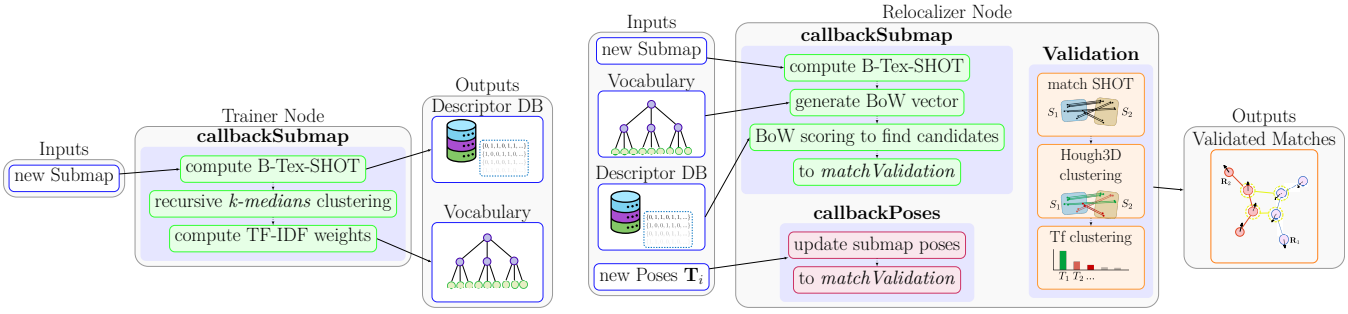
II. RELATED WORK

Traditional approaches for re-localization using cameras involve matching sparse visual features with a database [10], mostly by means of aggregation techniques such as FAB-MAP2 [11], BoW or VLAD [12]. Many Visual SLAM systems, such as ORB-SLAM2 [13] or LDSO [14], rely on bag of binary words [7] in conjunction with the ORB descriptor. While enabling fast and accurate re-localization, visual similarity based on descriptor matching decreases in presence of viewpoint differences and changing environment appearances. To lower the dependency on viewpoint, map densification through local meshes is used in [15] to increase the number of candidate keypoint matches selected using the DBoW2 library and BRISK descriptors. A wider set of candidates and a careful geometric verification increases the chance of successful relocalization. Invariance to changing appearance can be obtained by relocalizing on multiple overlapping maps built in different weather [16] or lighting conditions [17]. As mentioned also in [18], this requires to initialize the observer position.

Exploiting geometry for place recognition helps to overcome the limitations of pure visual localization [19]. A BoW scheme is used in [20] to match range images with a database, which could be applied also to stereo vision systems. However, while relying on geometry and not visual appearance, dependency on the viewpoint would remain unsolved. Local 3D feature descriptors are evaluated in [2] to merge LiDAR maps from a ground vehicle to dense maps from monocular SLAM captured by an aerial vehicle acknowledging the robustness of SHOT [5]. Learned 3D descriptors and matching metrics [21]–[23] show promising results although require powerful hardware, which limits the implementation on resource constrained vehicles. SHOT enriched with LiDAR intensity data is used in [24] along with a probabilistic selection scheme for multi-session localization. Instead of local features, the SegMatch algorithm [25] extracts segments from dense 3D LiDAR clouds and matches them with a database using an ensemble of handcrafted global features or, as published later in [26], from learned descriptors. However, it is not clear how to segment stereo point clouds from natural environments in a repeatable way. Other means for multi-agent relocalization involve matching planar segments from point clouds [27], which is only feasible in mostly artificial environments, or from Monte Carlo Localization as in [28] where an UAV augments the map of an UGV equipped with an RGB-D camera. This, however, is not feasible for large scale environments due to computational limitations.

III. RELOCALIZATION ON POINT CLOUDS

Fig. 2 gives an overview of the proposed relocalization pipeline. During the first mapping session, each submap pushed to the SLAM system is processed by binarizing already computed SHOT descriptors and appending texture information (more details in Section III-A). At the end of this session, descriptors are stored in a database and a vocabulary is generated by clustering descriptors in a k-d tree using the DBoW2 [2] library. During a second mapping session, descriptors computed over each submap are converted into BoW vectors accounting for the fact that a vocabulary built over a single session is very likely to be incomplete (Section III-B). Candidate submap matches between the two sessions selected from BoW similarity are then validated by matching the original SHOT descriptors and clustering the suggested



(a) Trainer node (b) Relocalizer node

Fig. 2. Relocalization pipeline architecture. (a) In a first mapping session B-Tex-Shot descriptors are created, i.e. SHOT features are extracted from submaps and binarized, and compact (binary) texture descriptors appended. At the session end, a vocabulary of B-Tex-SHOT is generated and the binary descriptors are stored as part of a database. (b) During a second session, each submap is processed as in (a) by computing B-Tex-SHOT descriptors. Using the vocabulary, binary descriptors in each submap are transformed into BoW vectors which are compared with the database to find the most similar ones. A candidate selection strategy (Section III-B) identifies a minimal set of submap matches from BoW similarity. Candidate submap pairs are validated by matching the original SHOT descriptors and grouping them using a Hough3D approach. Each Hough3D group votes for a transformation between the origins of the two mapping sessions and relocalization is triggered if the two highest voted transformations satisfy (6).

transformations in order to determine the most likely alignment between the maps (Section III-C).

A. Binary Descriptors

As the full SLAM pipeline running onboard the LRU (Lightweight Rover Unit) [3] depends on SHOT descriptors, we leverage the binarization scheme described in [6] to obtain a set of lightweight and computationally efficient version of the SHOT descriptors referred to as B-SHOT. The size of this new descriptor is 352 as the original SHOT but can be matched using the Hamming distance with much less effort than the original float vector. To increase the descriptive power, we design, inspired by [29], a short additional descriptor of the intensity information associated with the point clouds. With a size of just 44, this binary descriptor acts as a visual cue, increasing the precision of the full descriptor without risking to undermine the invariance given by 3D information. We first recall the BOARD frames [30] used to compute the original SHOT descriptors. Points which lay inside the support region are projected onto the x-y plane defined by the reference frame axes (where z is the normal direction). A grid of 9×9 cells is centered on the local origin as shown in Fig. 3, and the intensities of points which lay in the same cells are averaged.

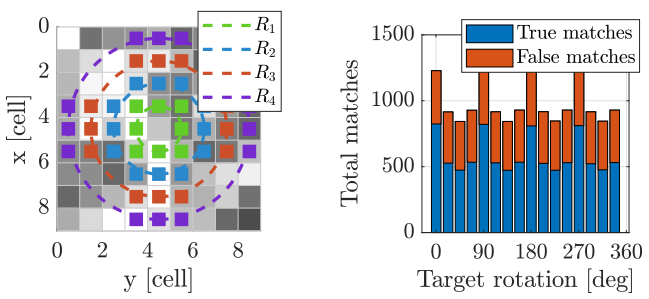


Fig. 3. Left: Pattern for the binary texture descriptor superimposed over the cell grid and example texture patch in grayscale. Cell size depends on the support region radius: $l_{\text{cell}} = \sqrt{2}R_{\text{sr}}/9$. Right: $d_{\text{B-Tex}}$-only match test on a submap from the RMC laboratory applying various degrees of rotation and a Gaussian noise to the intensity of $\sigma = 0.01 I_{\max}$

A reference value I_{ref} is defined as the average intensity of the central 3×3 cells and binary values of the descriptor are defined by comparing the intensities $I_{\text{c}}(i)$ of all designated cells (44 out of $9 \times 9 = 81$) with I_{ref}:

$$d_{\text{B-Tex}}(i) = \begin{cases} 1 & \text{if } I_{\text{c}}(i) > I_{\text{ref}} + t \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where t is a small threshold to account for noise (set to 1 in our experiments) and $i = 1, 2, \ldots, 44$. $d_{\text{B-Tex}}$ tolerates small orientation errors around the normal axis as it considers average intensities in the cells and tolerates displacements of keypoints as it uses the average intensity of the central pattern as reference. Fig. 3 shows the pattern used to generate $d_{\text{B-Tex}}$ as well as the result of a matching test performed on an LRU submap, demonstrating good accuracy even in absence of 3D information. For the latter, $d_{\text{B-Tex}}$ descriptors were generated with Gaussian noise added to the point intensities. The final descriptor, of size 396, is obtained as $d_{\text{B-Tex-SHOT}} = d_{\text{B-SHOT}} \cup d_{\text{B-Tex}}$.

B. Candidate Selection from Bags of Binary Words

The set of binary descriptors belonging to each submap is converted into BoW vectors by traversing the vocabulary provided from the first mapping session. However, as the vocabulary does not provide a full representation of the features that can be observed in the scene, many new descriptors can provide wrong contributions to the BoW vector because the Hamming distance between each of them and the closest vocabulary node at leaf level might be high. To overcome this, we establish a re-weighting scheme which dampens the contribution to the *idf* (inverse document frequency) in case of low similarity between descriptors and closest leaves: let $w_i^* \in [0, 1]$ be a coefficient which depends on the Hamming distance H between the vocabulary leaf c_i (cluster centers at the lowest level) and binary descriptor d_i. We design a simple function which does not influence close descriptor-leaf pairs (i.e. H lower than a value H_t) and penalizes higher distances (i.e. $H > H_t$). Amongst all possible solutions to this

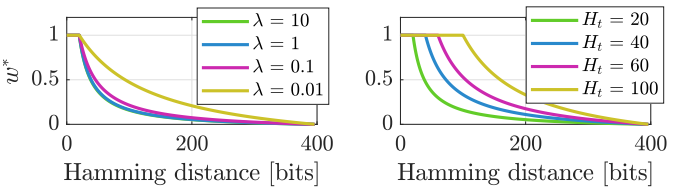


Fig. 4. Visualization of the re-weighting factor from (2), varying the free parameters λ (left, with $H_t = 20$) and H_t (right, with $\lambda = 1$).

problem, we choose a formulation which empirically delivers satisfactory results:

$$w_i^* = \begin{cases} \frac{\alpha}{1+\lambda H(c_i,d_i)} + \beta & \text{if } H > H_t \\ 1 & \text{otherwise.} \end{cases} \quad (2)$$

Fig. 4 visualizes w_i^* as a function of λ and H_t. In our experiments λ is set to 1. This coefficient w_i^* is multiplied with the traditional (as default in DBoW2) inverse document frequency (*idf*) and term frequency (*tf*) associated with d_i to obtain the final contribution $\bar{w}_i$ to the BoW vector $\mathbf{v} = \{\langle \mathrm{id}_0, \bar{w}_0 \rangle, ..., \langle \mathrm{id}_i, \bar{w}_i \rangle\}$. id_i is an index of leaf c_i and $\bar{w}_i$ is defined as

$$\bar{w}_i = w_i^* \cdot idf \cdot tf. \quad (3)$$

α and β from (2) are determined from the constraints $w^*(396) = 0$ and $w^*(H_t) = 1$. Bag of words vectors from the second sessions $\mathbf{v}_i^q$ are compared with the database ones $\mathbf{v}_j^{db}$ using the L1 score $s(\mathbf{v}_i^q, \mathbf{v}_j^{db})$ proposed in [7]. The absolute value of this score is, however, dependent on how the vocabulary that generated $\mathbf{v}$ has been built. It is not possible then to compare s with any fixed threshold to determine if two submaps are similar or not. Furthermore, we expect that in each mapping session subsequent submaps do not overlap often. For this reason, approaches typical of visual SLAM [13] where new images are compared with a window of old keyframes to search for consistent temporal similarity are not applicable in our case. We observe instead the highest and lowest values of the scores computed so far and wait for a minimum number $n_{\min}$ of received submap pairs before any decision on candidate selections is made. Then, a relative threshold t_{rel} is applied to discriminate tentative matches. Submaps i and j are considered as matching if

$$s(\mathbf{v}_i^q, \mathbf{v}_j^{db}) > t_{\mathrm{rel}} \cdot (s_{\max} - s_{\min}) + s_{\min} \quad (4)$$

where $s_{\max}$ and $s_{\min}$ are updated each time a new submap is used to evaluate the BoW score against the database.

C. Match Validation from Tf Clustering

Submap matches proposed by the previous step are validated to eliminate false positives, which is done by evaluating multiple transformation hypotheses to find a safe consensus. The point clouds aggregated by the stereo camera mounted on the LRU pan-tilt head suffer from noise proportionally to the square of depth. When matching SHOT descriptors across two submaps, the number of false matches can easily surpass the number of correct matches. As we show later in Fig. 7, traditional feature matching followed by RANSAC outlier rejection is likely to fail in such challenging conditions.

We first group all the pairwise SHOT correspondences using the Hough3D voting technique described in [31] and adapted from the PCL implementation to return 4D transformations instead of full 6D. As submaps are in fact always guaranteed to be gravity aligned [4], the transformation between submaps can be described by 3D translation vector (x, y, z) and yaw rotation angle ϕ. This step returns coherent groups of keypoint matches resembling *model* to *model* (such as individual rocks) correspondences that are robust to clutter and occlusions, and discards uncorrelated wrong keypoint matches. This behavior is important for point clouds built from aggregated stereo observations as partial views and submap boundaries easily result in holes and truncated models. Each matching group of keypoints after Hough3D suggests a transformation between submap i of the query session and submap j of the database in the form of $\{x, y, z, \phi\}$ which, as each submap is rigid, should, if correct, be equal to $\mathbf{T}_j^i$ (see Fig. 5). As many groups after Hough3D might be returned even in absence of correct keypoint matches, RANSAC methods would likely select the resulting correspondences as inliers for a wrong rototranslation model. For this reason we choose to cluster the 4D transformations determined by the Hough3D groups to find a wide consensus across multiple submaps.

As illustrated in Fig. 5, the pose of submap i from the query session is $\mathbf{T}_i^q$ relative to the reference frame $\mathbf{O}_q$, while the pose of database submap j is $\mathbf{T}_j^{db}$ with respect to cordinate system $\mathbf{O}_{map}$ of the global map. By suggesting a transformation $\mathbf{T}_j^i(k)$ to align the two submaps, Hough3D keypoint cluster k defines also a transformation between the reference frame of the query session and the global map:

$$\mathbf{T}_q^{db}(k) = \mathbf{T}_j^{db} \cdot \mathbf{T}_i^j(k) \cdot (\mathbf{T}_i^q)^{-1} \quad (5)$$

with $\mathbf{T}_i^j(k) = (\mathbf{T}_j^i(k))^{-1}$. All transformations $\mathbf{T}_q^{db}(k)$ are clustered using an incremental k-means scheme (see Alg. 1). A transformation belongs to a cluster only if the 4 coordinates are closer than a pre-defined threshold. After assigning a transformation to its cluster, the center is updated as the mean of all contained elements. Under the hypothesis that wrong matches vote for random transformations, if any correct match is present, the votes for that particular transformation should be recognizable. Let n_l be the total number of keypoint matches

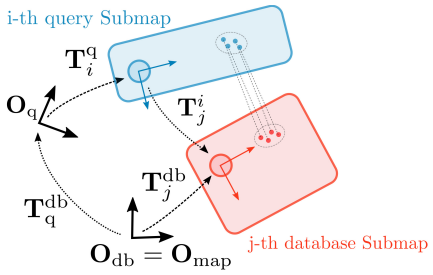
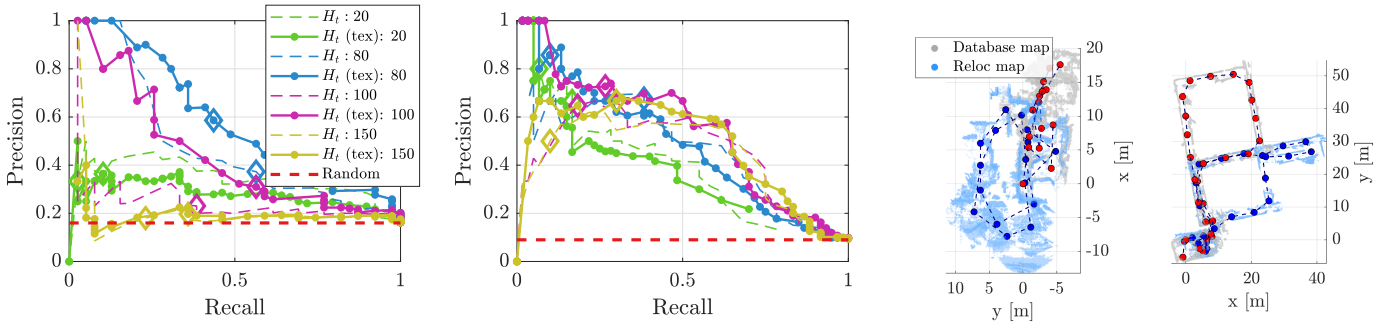


Fig. 5. Highlighted keypoints belong to a Hough3D cluster defining the transformation between the submaps of different sessions ($\mathbf{T}_j^i$) in a local coordinate frame. Via the transformations or poses of the submaps in the respective session reference frame, they also provide an estimate of the transformation between the different session origins ($\mathbf{O}_{db}$ and $\mathbf{O}_q$).



(a) PR curve in *RMC_in_out* (b) PR curve in *RMC_corridor* (c) Map of in *RMC_in_out* (d) Map of *RMC_corridor*

Fig. 6. Precision-recall (PR) curves and maps for the two indoor experiments. (a,b) PR curves comparing the effect of embedding texture information with plain 3D binarized SHOT (B-Tex-SHOT versus B-SHOT). The curves are generated by varying t_{rel} of (4) from 0 to 1. Markers denote the precision-recall values at $t_{\text{rel}} = 0.7$. A red dashed line denotes the precision-recall curve of a random classifier, whose precision is computed as the fraction of overlapping submaps over the number of all possible pairs. (c,d) Top views of the maps. The invidual maps of the two sessions are highlighted in gray and blue. Filled circle indicate positions of submap origins.

Algorithm 1: Incremental Tf Clustering

Input :
- $\{\mathbf{T}_q^{db}(k)\}$: 4D transform hypotheses for all clusters
- n_k: number of keypoints voting for $\{\mathbf{T}_q^{db}(k)\}$
- t_{xyz}, t_ϕ, r_t: spatial, yaw and ratio thresholds

Output:
- $\mathbf{T}_{\text{max}}$: consensus transformation for relocalization

$C = \{\}$: init clusters
foreach *submap pair* $\in$ *candidate pairs* **do**
  foreach *cluster* $k \in$ *submap pair* **do**
    search C for cluster l that is closest to cluster k;
    if $|\mathbf{T}_q^{db}(k) - \mathbf{T}_q^{db}(l)| < t_{xyz}$ & t_ϕ **then**
      add $\mathbf{T}_q^{db}(k)$ to cluster l;
      add n_k to n_l;
      update $\mathbf{T}_q^{db}(l)$;
    else
      add new cluster with $\mathbf{T}_q^{db}(k)$
rank clusters for number of votes n_l;
if $1 - n_{2\text{nd}}/n_{\text{max}} > r_t$ **then**
  return $\mathbf{T}_{\text{max}}$

voting for $\mathbf{T}_q^{db}(l)$. Being n_{max} and $n_{2\text{nd}}$ the number of votes for highest and second highest voted clusters, the transformation corresponding to n_{max} is selected as winner if

$$1 - n_{2\text{nd}}/n_{\text{max}} > r_t \tag{6}$$

where the ratio r_t is defined as 0.5 in our experiments.

IV. EXPERIMENTS

We tested our relocalization system in both laboratory and outdoor scenarios. First, experiments are performed in the robotics laboratory and hallways of the DLR Robotics and Mechatronics Center to characterize the performances of the relocalizer in selecting valid candidate matches. The full pipeline is then evaluated both in the laboratory dataset and on outdoor sequences captured on Mount Etna, a designated planetary analogue environment as part of the ROBEX mission [32].

A. Candidate Selection Accuracy

A laboratory dataset is used to evaluate the performances of the submap candidate selection step (Section III-B) and consists of two sessions denominated *RMC_corridor* and *RMC_in_out*. The rover explores a laboratory environment, where several big rocks are placed to imitate the appearance of natural scenes, then travels in the hallways or outside the lab, where 3D features are lacking, as it is mostly flat. Both sessions contain a significant number of overlapping submaps which should be detected by the relocalization pipeline. For these datasets, 3D keypoints are extracted over a voxel grid of 5 cm size by segmenting obstacle point clouds from depth images [4], [33].

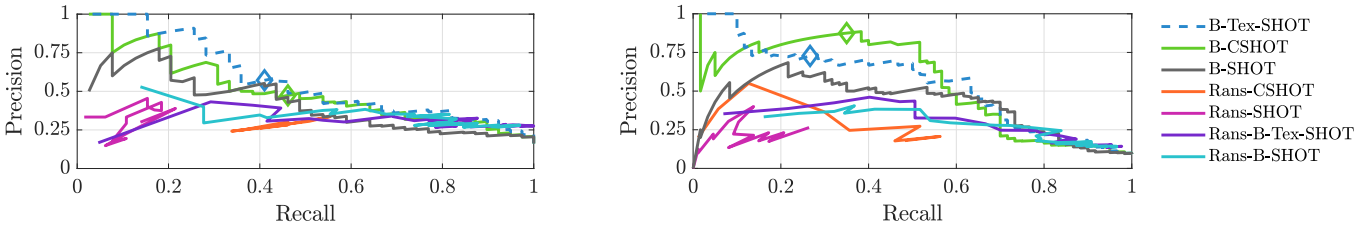
As a vicon ground truth reference is only partially available, for each dataset the two sessions are manually aligned in order to average the effects of minor drifts in the pose estimates. To generate ground truth submap correspondences, we perform a nearest-neighbor search within the point clouds of all possible submap pairs. We define the grade of overlap between submaps s_i and s_j as

$$o(s_i, s_j) = 2n_{\text{pairs}}/(n_i + n_j) \tag{7}$$

where n_{pairs} is the number of unique point pairs with distance below 0.5 meter and n_i and n_j are the total number of points contained in s_i and s_j. We consider ground truth submap matches those for which the grade of overlap is higher than 0.1. We measure the quality of candidate selection using the *precision-recall* metric:

$$\text{Precision} = T_p/(T_p + F_p), \quad \text{Recall} = T_p/(T_p + F_n) \tag{8}$$

where T_p and F_p are the number of true and false matches and F_n is the number of missed matches, i.e. the number of overlapping submaps which were not detected by the pipeline. Fig. 6(a) and 6(b) show the precision-recall curves for the two indoor data sets for multiple values of H_t (see section III-B). Each curve is generated by varying the relative threshold t_{rel} from 0 to 1 and computing precision and recall. The curves highlight the effect of appending texture information



(a) PR curves for *RMC_in_out*

(b) PR curves for *RMC_corridor*

Fig. 7. Performance Comparison with Baseline. B-CSHOT refers to the candidate selection scheme using C-SHOT descriptors binarized following the approach in [6] adapted to a longer descriptor. The B-Tex-SHOT, B-CSHOT and B-SHOT curves are selected as the highest for various values of H_t (see Fig. 6). "Rans" refers to brute-force matching of descriptors rejecting outliers in a RANSAC scheme. RANSAC curves are generated by varying the SHOT and C-SHOT matching threshold from 0 to 1 (min and max L_2 distance between descriptors) and a threshold on the Hamming distance for B-Tex-SHOT from 0 to 396. In the RANSAC case, the results are definitive as matches are already validated by selecting a consensus model.

to the original B-SHOT descriptor, showing a significant performance improvement for $H_t = 80$ and $H_t = 100$ especially at high t_{rel} levels. The fact that many curves in Fig. 6 start at zero precision, such as for $H_t = 20$ and $H_t = 150$ suggest that the proposed re-weighting scheme is necessary to build appropriate BoW vectors. Associating distant descriptor-leaves can lead to high BoW similarity for non-overlapping submap pairs, hence the zero precision at highest t_{rel} levels. While in the *RMC_in_out* dataset all submaps tend to include significant 3D information, in the *RMC_corridor* dataset many submaps (as visible in Fig. 6(c)) contain only points from partial views of straight walls, which do not deliver useful information for relocalization reducing the recall score. In Fig. 7 we compare the performances of our system with alternative solutions. First we binarize C-SHOT [34] descriptors adapting the approach in [6] (resulting in a 1344 bits descriptor) replacing our B-Tex-SHOT for the candidate selection stage. As the curves highlight, none of the two consistently outperform the other, therefore our descriptor in conjunction with binary SHOT allows to save computational time without penalizing recall performances. Secondly, we match SHOT, C-SHOT and B-Tex-SHOT in a brute force way across all submap pairs and reject bad correspondences using a RANSAC scheme. This also replaces the match validation stage by selecting the appropriate transformation between submaps, therefore the outcomes are to be considered as definitive. Positives are submap pairs for which RANSAC returns a model (implemented using PCL class `pcl::CorrespondenceRejectorSampleConsensus`, P=0.99). Our pipeline outperforms the baseline RANSAC approach even before match validation. As it is also found in [8], noise in point clouds from dense stereo compromises the effectiveness of descriptor matching. Our approach, however, provides a robust solution to this problem by focusing the attention on a selected set of submaps and searching for a wide consensus of transformation hypotheses.

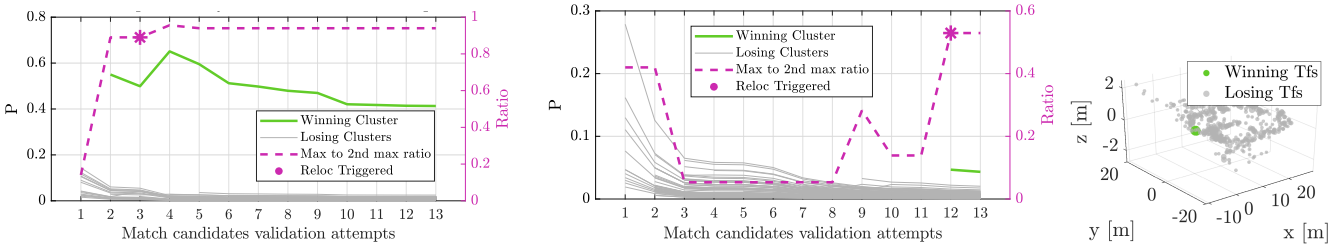
B. Relocalization Experiments

Here we evaluate the capability of our algorithm to relocalize the LRU rover over subsequent mapping sessions on the previously introduced laboratory dataset and on a dataset captured in a planetary analogue environment (Table I), where the repetitiveness of visual features (see Fig. 1) and the lack of conspicuous 3D objects pose a big challenge for place recognition. In the first sequence of the Etna datasets, *Etna_easy*, the LRU rover drives autonomously along given waypoints exploring an area which is partly covered in small sized rocks. The maps of two short sessions, covered without a pause in between, overlap in a few areas containing rocks. As the LRU frequently uses the pan-tilt camera head for obstacle avoidance, the resulting submap point clouds cover a relatively wide area, maximizing the chance of gathering useful information. In the second sequence, *Etna_hard*, the LRU rover performs autonomous exploration experiments [35], using frequently the pan-tilt head to estimate traversability. Two subsequent sessions, performed with a two hours pause in between, partially overlap in terms of submap positions and origins. However, useful 3D information is present in only a very limited part of the second map. The parts where submaps overlap most are mainly observing flat areas lacking unique features, both 3D and visual. In fact, for comparing the relocalization performances of our algorithm, we processed all 4 sequences involved in the two Etna datasets with the visual SLAM system ORB-SLAM2 [13]. However, the visual front-end failed to track frequently.

As in this scenario the distinction between obstacles and travellable ground is more subtle, we extract keypoints from high curvature regions, where curvature is computed from the eigenvalues of the scatter matrix as $\sigma = \lambda_{\text{min}} / \sum_{i=1}^{3} \lambda_i$. Keypoints are obtained at positions where $\sigma > 0.02$. For recalling candidate submaps, in all datasets a Hamming threshold H_t of 100 was used and the relative BoW score threshold was set to $t_{\text{rel}} = 0.7$ as suggested by the precision-recall tests in Fig. 6. The full submap point clouds were downsampled to 3 cm voxels before processing in order to reduce the computational effort. SHOT descriptors are matched during validation employing k-d trees. Fig. 1 shows the two sessions after alignment in the *Etna_easy* sequence, highlighting the

TABLE I
TEST DATASETS

	Etna_easy		*Etna_hard*		*RMC_in_out*		*RMC_corridor*	
	Db	Q	Db	Q	Db	Q	Db	Q
submaps	6	10	12	28	14	15	27	22
area [m^2]	285	440	272	330	318	263	482	413
time [min]	7.5	17.4	21.6	38.5	8.9	8.9	12	12.2



(a) Normalized Tf cluster score and ratio metric (*Etna_easy*)

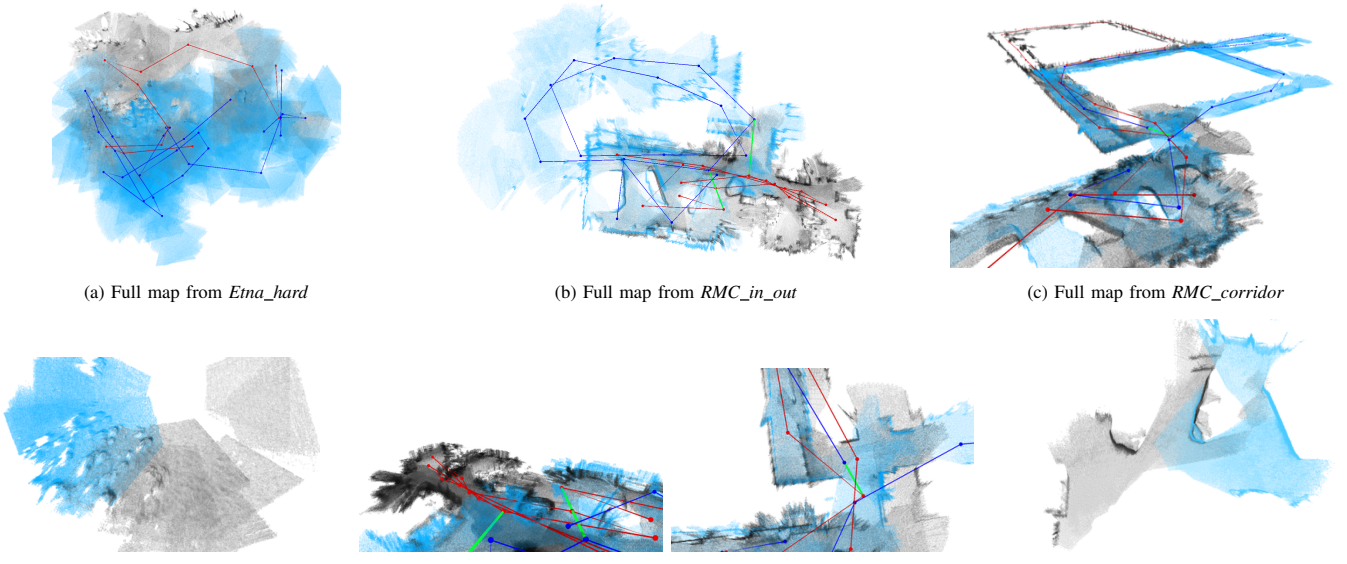
(b) Normalized Tf cluster score and ratio metric (*Etna_hard*)

(c) Tf clusters in *Etna_hard*

Fig. 8. Results of the Tf voting scheme for the *Etna_easy* and *Etna_hard* mapping sessions. (a,b) Normalized Tf cluster score $P_l = n_l(\sum_i n_i)^{-1}$ and ratio $r = 1 - n_{2\text{nd}}/n_{\max}$ during subsequent validations. A marker (*) denotes the moment at which relocalization is triggered. (c) 3D visualization of the Tf cluster centers (only $\{x, y, z\}$). Marker size is proportional to the number of votes.

submap discretization, and compares the outcomes of our pipeline (green lines) to CSHOT+RANSAC (magenta lines). While our validation stage selects only true matches, RANSAC results contain outliers. Fig. 8(a) shows the normalized Tf cluster scores and ratios in the same session over subsequent evaluations of candidate submap matches. The winning cluster is detected in the second candidate pair and maintains a high ratio score throughout the entire run, triggering relocalization at $n_{\text{min}} = 3$ pairs received. Fig. 9 shows the aligned maps from the *Etna_hard* dataset as well as the matching submaps (Fig. 9(d) and 9(f)) The visible overlap between point clouds demonstrates the relocalization accuracy, which as Fig. 8(b) shows, was triggered during the second to last evaluated match, where the winning Tf cluster appeared. The same figure shows also that in the first candidate submap match, wrong clusters can be selected as winners using the ratio metric because not enough voting keypoint matches are present. For this reason, relocalization is triggered only after evaluating at least three candidate matches ($n_{\text{min}} = 3$).

The pipeline is run also on the *RMC_in_out* and *RMC_corridor* sequences using the same parameters as for the Etna dataset triggering correctly the relocalization respectively with ratio scores r_t 0.81 and 0.52. The aligned maps are visible in Fig. 9(b) and 9(c). Fig. 9(d) to 9(f) show details of the matched submaps aligned using the winning transformations, visually testifying the relocalization quality. To provide a coarse value for the alignment precision, we compute point-to-point distances between matched submaps for all test sequences, which span from 5 cm to 7 cm which is comparable with the cloud resolution. Timings for each part of the pipeline are reported in Table II listing the average time required for computing keypoints, SHOT, B-Tex-SHOT, then generating and scoring BoW vectors and finally performing match validation. Timings in Table II were measured on a desktop Intel Xeon E5-1620 v3 @3.5GHz whose single-thread performances are quite similar to the Intel i7-3740QM running at 2.7 GHz onboard the LRU. Being n the number of keypoints, computing descriptors and generating BoW vectors (n independent searches in the k-d tree) are operations of $\mathcal{O}(n)$ time complexity. The high standard deviations in Table II are due to the varying number of keypoints detected in each submap. The time required for keypoint selection depends on the submap size and the time required for computing BoW scores depends on the number of database submaps.



(a) Full map from *Etna_hard*

(b) Full map from *RMC_in_out*

(c) Full map from *RMC_corridor*

(d) Submap match in *Etna_hard*

(e) Details from *RMC_in_out* and *RMC_corridor*

(f) Submap match in *RMC_corridor*

Fig. 9. (a-c) Aligned maps after relocalization. (d-f) Detail views of matching submaps and aligned maps testifying relocalization accuracy from visual inspection. Green lines connect the origins of matching submaps.

TABLE II
TIMINGS PER SUBMAP (MEAN μ AND STD σ) ON *Etna_hard*

	Key-Points	B-Tex-SHOT		BoW gen.	BoW scores	Match Valid.
		SHOT	*B-Tex*			
μ [s]	0.09	1.01	0.49	0.04	0.01	9.08
σ [s]	0.02	0.59	2.26	0.21	0.01	10.40

Both computation of SHOT and the full binary descriptor takes significantly longer. We reckon that in particular the latter, where most time is spent at generating the texture part, can be accelerated by better optimization at code-level. The bottleneck of the system can be found in the match validation stage, where matching full SHOT descriptors is particularly expensive. As this task is performed on a dedicated thread and called each time a new submap is published by the system (roughly every 10-20 seconds depending on the robot motion), the computational overhead for the validation stage is not compromising the efficiency of the full SLAM system.

V. CONCLUSION AND FUTURE WORK

We presented a 3D submap based relocalization pipeline for mobile robots using binary descriptors in a modified bags of binary words scheme. The pipeline copes with extreme outlier ratios thanks to a match validation scheme based on transformation clustering. Laboratory and outdoor tests on a designated planetary analogous environment demonstrate the effectiveness of our approach. Future developments will involve capturing new datasets on Etna during a planned test campaign for the ARCHES program [36] in different times of the day with changing lighting conditions. Furthermore, we will investigate how to reduce the current bottlenecks using robust binary descriptors for match validation and more efficient descriptor clustering strategies such as HBST [37].

REFERENCES

[1] C. Cadena, *et al.*, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1309–1332, 2016.

[2] A. Gawel, *et al.*, "3d registration of aerial and ground robots for disaster response: An evaluation of features, descriptors, and transformation estimation," in *SSRR*, 2017, pp. 27–34.

[3] M. J. Schuster, *et al.*, "Towards autonomous planetary exploration," *Journal of Intelligent & Robotic Systems*, vol. 93, no. 3-4, pp. 461–494, 2019.

[4] ——, "Distributed stereo vision-based 6D localization and mapping for multi-robot teams," *Journal of Field Robotics*, vol. 36, no. 2, pp. 305–332, 2019.

[5] F. Tombari, *et al.*, "Unique signatures of histograms for local surface description," in *ECCV*, 2010, pp. 356–369.

[6] S. M. Prakhya, *et al.*, "B-SHOT: a binary 3D feature descriptor for fast keypoint matching on 3D point clouds," *Autonomous Robots*, vol. 41, no. 7, pp. 1501–1520, 2017.

[7] D. Galvez-Lopez 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.

[8] Y. Guo, *et al.*, "A comprehensive performance evaluation of 3D local feature descriptors," *International Journal of Computer Vision*, vol. 116, no. 1, pp. 66–89, 2016.

[9] T. Ahonen, *et al.*, "Face description with local binary patterns: Application to face recognition," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 12, pp. 2037–2041, 2006.

[10] S. Lowry, *et al.*, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2015.

[11] M. Cummins and P. Newman, "Appearance-only slam at large scale with fab-map 2.0," *The International Journal of Robotics Research*, vol. 30, no. 9, pp. 1100–1123, 2011.

[12] R. Arandjelovic and A. Zisserman, "All about VLAD," in *CVPR*, 2013, pp. 1578–1585.

[13] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[14] X. Gao, *et al.*, "LDSO: Direct sparse odometry with loop closure," in *IROS*, 2018, pp. 2198–2204.

[15] F. Maffra, *et al.*, "Real-time wide-baseline place recognition using depth completion," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1525–1532, 2019.

[16] M. Bürki, *et al.*, "Vizard: Reliable visual localization for autonomous vehicles in urban outdoor environments," *arXiv preprint arXiv:1902.04343*, 2019.

[17] S. Chiodini, *et al.*, "Robust visual localization for hopping rovers on small bodies," in *ICRA*, 2018, pp. 897–903.

[18] T. Schneider, *et al.*, "maplab: An open framework for research in visual-inertial mapping and localization," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1418–1425, 2018.

[19] N. Piasco, *et al.*, "A survey on visual-based localization: On the benefit of heterogeneous data," *Pattern Recognition*, vol. 74, pp. 90–109, 2018.

[20] B. Steder, *et al.*, "Place recognition in 3D scans using a combination of bag of words and point feature based relative pose estimation," in *IROS*, 2011, pp. 1249–1255.

[21] A. Dewan, *et al.*, "Learning a local feature descriptor for 3d lidar scans," in *IROS*, 2018, pp. 4774–4780.

[22] Z. J. Yew and G. H. Lee, "3DFeat-Net: Weakly supervised local 3d features for point cloud registration," in *ECCV*, 2018, pp. 630–646.

[23] Z. Gojcic, *et al.*, "The perfect match: 3D point cloud matching with smoothed densities," *arXiv preprint arXiv:1811.06879*, 2018.

[24] J. Guo, *et al.*, "Local descriptor for robust place recognition using LiDAR intensity," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1470–1477, 2019.

[25] R. Dubé, *et al.*, "Incremental-segment-based localization in 3-d point clouds," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1832–1839, 2018.

[26] ——, "SegMap: 3D segment mapping using data-driven descriptors," in *RSS*, 2018.

[27] H. Surmann, *et al.*, "3D mapping for multi hybrid robot cooperation," in *IROS*, 2017, pp. 626–633.

[28] C. Forster, *et al.*, "Air-ground localization and map augmentation using monocular dense reconstruction," in *IROS*, 2013, pp. 3971–3978.

[29] D. Schlegel and G. Grisetti, "Adding cues to binary feature descriptors for visual place recognition," in *ICRA*, 2019, pp. 5488–5494.

[30] A. Petrelli and L. Di Stefano, "On the repeatability of the local reference frame for partial shape matching," in *ICIV*, 2011, pp. 2244–2251.

[31] F. Tombari and L. Di Stefano, "Object recognition in 3d scenes with occlusions and clutter by hough voting," in *Pacific-Rim Symposium on Image and Video Technology*, 2010, pp. 349–355.

[32] A. Wedler, *et al.*, "First results of the ROBEX analogue mission campaign: Robotic deployment of seismic networks for future lunar missions," in *68th International Astronautical Congress (IAC)*, 2017.

[33] C. Brand, *et al.*, "Stereo-vision based obstacle mapping for indoor/outdoor SLAM," in *IROS*, 2014, pp. 1846–1853.

[34] F. Tombari, *et al.*, "A combined texture-shape descriptor for enhanced 3d feature matching," in *ICIP*, 2011, pp. 809–812.

[35] H. Lehner, *et al.*, "Exploration with active loop closing: A trade-off between exploration efficiency and map quality," in *IROS*, 2017, pp. 6191–6198.

[36] A. Wedler, *et al.*, "From single autonomous robots toward cooperative robotic interactions for future planetary exploration missions," in *69th International Astronautical Congress (IAC)*, 2018.

[37] D. Schlegel and G. Grisetti, "HBST: A Hamming Distance Embedding Binary Search Tree for Feature-Based Visual Place Recognition," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3741–3748, 2018.