IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 17, 2020, accepted February 29, 2020, date of publication March 4, 2020, date of current version March 13, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2978406

Deep Learning for Underwater Visual Odometry Estimation

BERNARDO TEIXEIRA[1], HUGO SILVA[1], ANIBAL MATOS[1,2], AND EDUARDO SILVA[1,3]

[1] INESC TEC - Institute for Systems and Computer Engineering, Technology and Science, 4200-465 Porto, Portugal
[2] FEUP - Faculty of Engineering, University of Porto, 4200-465 Porto, Portugal
[3] ISEP - School of Engineering, Porto Polytechnic Institute, 4200-072 Porto, Portugal

Corresponding author: Bernardo Teixeira (bernardo.g.teixeira@inesctec.pt)

This work was supported in part by the National Funds through the Portuguese funding agency, Fundacao para a Ciencia e Tecnologia (FCT), under Project UID/EEA/50014/2019, and in part by the European Commission under the H2020 EU Framework Programme for Research and Innovation (H2020-WIDESPREAD-2018-03, 857339).

ABSTRACT This paper addresses Visual Odometry (VO) estimation in challenging underwater scenarios. Robot visual-based navigation faces several additional difficulties in the underwater context, which severely hinder both its robustness and the possibility for persistent autonomy in underwater mobile robots using visual perception capabilities. In this work, some of the most renown VO and Visual Simultaneous Localization and Mapping (v-SLAM) frameworks are tested on underwater complex environments, assessing the extent to which they are able to perform accurately and reliably on robotic operational mission scenarios. The fundamental issue of precision, reliability and robustness to multiple different operational scenarios, coupled with the rise in predominance of Deep Learning architectures in several Computer Vision application domains, has prompted a great a volume of recent research concerning Deep Learning architectures tailored for visual odometry estimation. In this work, the performance and accuracy of Deep Learning methods on the underwater context is also benchmarked and compared to classical methods. Additionally, an extension of current work is proposed, in the form of a visual-inertial sensor fusion network aimed at correcting visual odometry estimate drift. Anchored on a inertial supervision learning scheme, our network managed to improve upon trajectory estimates, producing both metrically better estimates as well as more visually consistent trajectory shape mimicking.

INDEX TERMS Artificial intelligence, computer vision, deep learning, visual odometry, robot navigation, visual SLAM.

I. INTRODUCTION

Achieving persistent and reliable autonomy for underwater mobile robots in challenging field mission scenarios is a long time quest for the Robotics research community, to which a great amount of research has been devoted to. In underwater scenarios, since GPS is not available and Inertial Measurement Units (IMU) are prone to failures, complementary sensorization has to be added to allow for a reliable robot localization and navigation.

Visual odometry estimation from outdoors imagery is always challenging due to multiple factors that generate blur, shadows and other illumination artifacts which lead to low signal-to-noise ratios in images. In spite of that, technological advancements in computer vision perception and processing has allowed the development of ever more robust and reliable algorithms capable of deriving camera self-motion from a sequence of images with significant accuracy, posing as a viable solution to help tackle the robot navigation problem.

Recently, deep learning has been garnering a lot of attention in the field of Computer Vision, even managing to become the "go to solution" for most visual based object detection and classification tasks [1]–[4]. In recent years, there were multiple relevant tasks in the scope of Robotics where there have been surfacing deep learning approaches, including but not limited to: depth estimation [5], semantic mapping [6], sensor fusion [7] and place recognition [8].

Bearing in mind the significant improvements on accuracy and performance that deep learning applications were able to achieve, camera pose and motion estimation frameworks

The associate editor coordinating the review of this manuscript and approving it for publication was Junchi Yan.

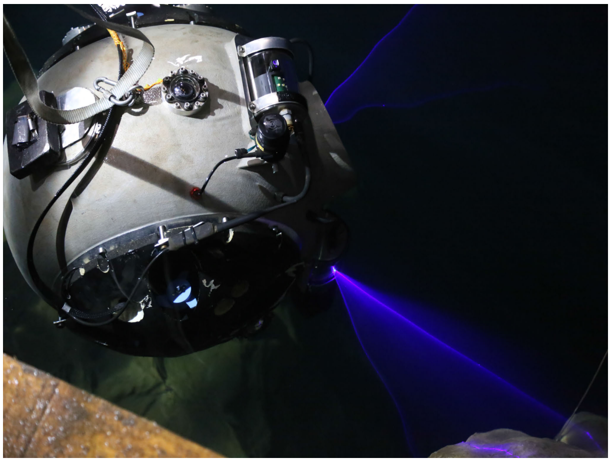


FIGURE 1. UX-1 robot photo at an operational mission scenario, courtesy of the UNEXTMIN[1] project.

also started to be designed to take advantage of the potential increased robustness of data-centric approaches, helping to tackle the robot navigation problem and thus serving as a precursor for sustained autonomy for robots and other autonomous systems.

However, the general design focus of Visual Odometry systems is historically not placed on underwater scenarios, with much more emphasis on designing and tailoring visual odometry systems to urban dataset scenarios [9]–[11]. The underwater context poses some additional challenges to visual-based navigation applications in comparison to the urban scenarios they were designed to tackle, especially under harsh operational mission scenarios. Underwater deep sea and/or inland flooded mines present uniquely challenging operational conditions, as repetitive image patterns and low texture, coupled with unfavourable lighting conditions, push the limits of visual odometry methods and shed light on most of their degenerate failure conditions.

In this work, we are expanding the formulation presented in [12], presenting additional benchmark information through a more thorough comparison against classical VO and v-SLAM methods, using datasets gathered using the UX-1 robot, see Fig.1. It is worth noting that we are interested in assessing the performance and accuracy of visual-based navigation methods with respect to the underwater scenarios we have been facing in the field, by performing a comparison between VO estimation methods, encompassing both feature-based, direct and especially learning-based algorithms (i.e. Deep Learning methods).

This article outline is as follows: Section II contains a review of motion estimation literature, starting from classical VO and v-SLAM techniques and ultimately culminating in novel deep learning approaches. In section III, we describe the different underwater dataset scenarios acquired and developed in the scope of this work. Section IV discusses the multiple renown VO, v-SLAM and deep learning frameworks that we evaluate on our underwater dataset sequences. In section V, an outline of the design of a novel Visual-Inertial Fusion Network approach is presented and in section VI comparison is drawn with the previously referenced visual odometry estimation methods. Finally, in section VII, some conclusions are drawn from the obtained results and future research directions in the scope of this work are layed out.

II. RELATED WORK

VO is the process of estimating solely the self-motion of an agent over consecutive camera image frames, and can serve as a prerequisite input for v-SLAM, which is a process by which a robot is required to simultaneously localize itself and build a map of the environment without any prior knowledge. This means VO mainly focuses on achieving higher local consistency in its estimates through incrementally estimating trajectories pose after pose and possibly performing local optimization operations. v-SLAM algorithms, on the other hand, aim to obtain a globally consistent map and therefore estimate pose mostly through reducing odometry drift in loop closure steps (i.e. understanding and recognizing when trajectories re-visit the same places in the map).

Recently, novel VO and v-SLAM learning-based data-centric approaches to the motion estimation problem, especially deep learning architectures, have been the focus of research in the motion estimation field, as advances in GPU technology and availability means that implementing large convolutional network architectures is no longer an insurmountable computational problem.

Through leveraging powerful high-level feature representations and exploring parallel convolutional implementations to the fullest, deep learning camera pose and motion estimation algorithms can replace the classical VO estimation pipeline and theoretically allow for the design of end-to-end systems that can be offer more robustness to different application scenarios and camera calibration parameterization.

In the remainder of this section, we will review previous work with regards to motion estimation, respecting this same taxonomy, separately reviewing Visual Odometry, Visual SLAM and Learning-based approaches.

A. VISUAL ODOMETRY

The need for the development of VO applications stems from the increased proliferation of the use of mobile robotic systems in a wide range of modern world tasks, across very different application scenarios. As such, mobile robots are required to extend their perception and cognition capabilities, so as to be able to navigate effectively in complex unstructured environments, where they may be deprived of the most commonly used IMU/GPS data for navigation purposes.

VO is performed through determining instantaneous camera displacement over consecutive frames, concatenating the obtained relative translational and rotational deltas onto a trajectory on a global reference frame. Robotic systems determine self-motion by measuring their displacement relative to

[1] https://www.unexmin.eu/

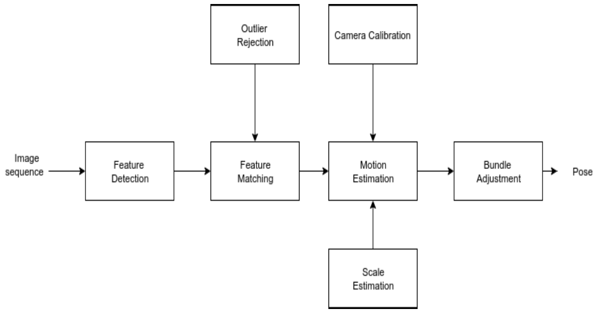


FIGURE 2. Feature-based VO typical processing pipeline: Features are detected and tracked in an image sequence, allowing for motion estimation. Additional modules such as camera intrinsic calibration, outlier rejection and bundle adjustment are utilized so as increase accuracy and performance of VO methods.

static key points in the environment and may or may not have algorithmic components that introduce robustness to ''hard to handle'' situations such as occlusions, non-Lambertian surfaces and dynamic elements in the scene.

VO algorithms take advantage of either monocular, stereo or RGB-D camera setups but the processing pipeline is roughly similar regardless of the sensor configuration (see Fig.2). The literature taxonomy divides VO algorithms in two different categories: feature-based (e.g., Howard *et al.* [13]) and dense (e.g., Engel *et al.* [14]) approaches. Nonetheless, there are some instances of hybrid approaches [15] that leverage both feature correspondence and dense methods for different estimation tasks.

1) FEATURE-BASED APPROACHES

Visual Odometry research can be traced back to the 1960's when a lunar rover started being constructed by Stanford University. Following this, Moravec [16] provided a description of the first motion estimation pipeline (whose main definitions remain basically intact today) as well of his renowned corner detector. This work formed a basis for further research in motion estimation using stereo vision systems, extended by means of correcting absolute orientation [17]. When a robust outlier rejection scheme denoted as RANSAC [18] was developed, Cheng [19] was able to introduce a a least-squares motion estimation step that was robust enough for the definitive VO implementation on board the Mars planetary Rover.

Lots of different VO formulations have been presented in the literate since then, amongst which Nister *et al.* [20] seminal work that provided the first step-by-step motion estimation framework for both the monocular and stereo cases and coined the popular term ''visual odometry''. Later, Engels *et al.* [21] estimated pose using this five-point algorithm followed by a global refinement strategy, denoted as bundle adjustment.

Feature-based approaches rely heavily on keypoint detection, description and matching of corresponding image points. The traditional edge and corner detection strategies employed by Moravec [16] and Harris [22] have fallen off slightly in favour of incremental improvements around scale and transformation invariant feature extraction methodologies such as SIFT [23], SURF [24], ORB [25] or BRIEF [26].

Extensive work followed on different Visual Odometry estimation techniques, including with different camera topologies [27], [28], camera calibration estimation and pose refinement bundle adjustment strategies [29]. In 2011, Scaramuzza *et al.* [30] proposed a 1-point algorithm for motion estimation, leveraging physical constraints to help reduce model complexity.

Other influential work by Kitt *et al.* [31], was made available as an opensource visual odometry estimation library named LIBVISO that is able to estimate 6-DoF pose from both monocular and stereo camera setups.

Combining visual input with other complementary sensorization [32], such as GPS or IMU information, it is also sometimes a worthwhile technique, as pose estimate computation becomes generally more robust. Another interesting approach was proposed by Kasik [33], emulating a stereo camera setup using non-overlapping fields-of-view cameras. The principle is to estimate monocular VO from each of the cameras and later impose the stereo constraints thanks to the known baseline distance.

2) DENSE APPROACHES

Algorithms for aligning images and estimating motion in video sequences are widely used in computer vision since the early days of digital cameras and primitive image stabilization features [34].

Dense methods are methods by which the intensity information of all image pixels or subregions of it is computed in order to perceive motion in sequential images. Motion estimation can then be achieved through optimizing photometric error metrics.

Optical flow is one of the fundamental principles that define egomotion of an observer as per Gibson ecological optics [35]. Optical flow methods, which involves minimizing the brightness or color difference between corresponding pixels over a time frame, are often divided in three main categories [36], [37]:

- **Differential methods:** The motion is computed from spatio-temporal derivatives or filtered versions of the image [38], [39]. [40]
- **Frequency methods:** These methods work by applying spatio-temporal filters in the frequency domain [41], [42].
- **Correlation methods:** Correlation based methods attempt to find matching image regions by maximizing some similarity measure between them, all under the assumption that the image has not been overly distorted over a local region for a short period of time [43]–[45].

More recently, Engel *et al.* [46] proposed a direct sparse odometry scheme, which optimizes the photometric error in a framework similar to sparse bundle adjustment. It avoids the use of geometric priors commonly encoded in feature-based approaches at the same time it uses all image points to estimate egomotion even in lower texture environments.

3) HYBRID APPROACHES

Feature-based methods are known to provide reliable data at the cost of disregarding a lot of information from the image. Dense methods on the other hand, take the entirety of the image data with the downside of usually introducing reconstruction errors on some patches of the image. Some authors propose to combine both approaches in a way to further increase algorithmic robustness to complex unstructured environments.

Scaramuzza *et al.* [15] used dense methods for rotation estimation whilst translation was handled by feature extraction on the ground plane. Forster *et al.* proposed SVO [47], where pose estimates are refined through minimizing the reprojection error introduced in the feature-alignment step. Silva *et al.* proposes a dense egomotion estimation technique that works in tandem with a feature-based translation scale factor estimation, later refined through a Kalman filter [48]. This work was later extended through employing a fully dense probabilistic stereo egomotion framework, that reports increased robustness against more challenging image scenarios [49].

B. VISUAL SLAM

Probabilistic robotics is a growing area in robotics, concerned with perception and cognition in the face of uncertainty [50]. One such algorithm in the scope of probabilistic robotics is Simultaneous Localization and Mapping (SLAM).

As described in this chapter introduction, SLAM is a method whereby a robot has to localize itself in an unknown environment, all while it is constructing a map of the surrounding environment.

Visual SLAM has merited a strong degree of interest mostly because low-cost cameras can provide rich visual information about the environment. The tradeoff is the higher computational burden when compared with laser scanner based SLAM, but due to recent CPU and GPU advances in computational power, this is becoming less of a factor.

There are three main building blocks in a SLAM system:

1) **Initialization**: Define initial coordinate system and reconstruct an initial map of the environment
2) **Tracking**: The reconstructed map is used to track the current position with respect to the map. To do so, it is necessary to solve the 2D-3D correspondences between the current image and map.
3) **Mapping**: The map is extended everytime the camera observes a region previously unseen, by computing the 3D structure of the environment

Additional modules became increasingly more relevant for SLAM systems, mostly for application purposes.

Relocalization is required to account for fast camera motions that may cause the system to lose track of its position. This problem is known in the literature as ‘‘kidnapped robot’’ and can be tackled through recomputing camera pose with respect to the map, thus making the system more robust and versatile.

Loop-Closure is also a commonplace technique aimed at global map optimization. The reconstructed map in SLAM systems generally has a tendency to accumulate estimation errors proportional to camera movement. To mitigate these errors, modern algorithms detect when the acquired images match an already visited part of the map, enforcing geometric consistency to the estimated trajectory and thus suppressing most of the accumulated odometry drifts. Bundle adjustment [21] (minimizing the reprojection error of the map by jointly optimizing map and pose) or pose-graph optimization (representing pose and map relationships as a computational graph to be optimized) strategies can also be employed to the purpose of obtaining a more globally concsist map and pose estimate.

One of the first proposed v-SLAM frameworks is Mono-SLAM [51], [52]. They employed a monocular camera setup using feature-based Shi-Tomasi [53] operators and matched features to previously observed ones though SSD correlation.

In order to solve the problem of the high computational cost, PTAM [54] enforced computation parallelism, splitting the tracking and mapping into separate CPU threads. This became the standard for later SLAM implementations. PTAM was also more effective as its implementation allows for real-time handling of larger numbers of points than Mono-SLAM EKF-based approach.

LSD-SLAM [55] is an extension of previous semi-dense VO, where camera motion is estimated by view synthesis generation from the reconstructed map and coupled with the introduction of loop closure and 7 DoF pose-graph optimization for obtaining globally consistent maps.

LDSO [56] is an extension of DSO [46]. also introducing loop-closure and pose-graph optimization to DSO local geometric consistency, thus transforming it in a SLAM system.

ORB-SLAM [57], [58] is one of the most renown SLAM systems, mostly to its versatility and accuracy whether for monocular, stereo or RGB-D camera setups. It includes an automatic initialization that is able to work in both planar and non-planar scenes thanks to a cleverly designed heuristic. It is also able to close large loops in estimated trajectories and perform real-time relocalization in some situations it loses tracking of the environment.

RGB-D SLAM techniques [59], [60] leverage the capabilities of emergent structured light-based RGB-D cameras, namely the ability to automatically acquire absolute scale of the environment. This methods however, tend to work only indoors, as it somewhat difficult to detect the emitted IR in outdoors environments as well as the range limitations of RGB-D camera sensors. SLAM++ [60] has the particularity of registering an object database prior, with estimated map refinement around detected database objects.

SLAM systems can also benefit from incorporating deep learning components in one or more of its component tasks. As an example, CNN-SLAM [61] leverages convolutional neural networks to perform depth estimation, recovering pose

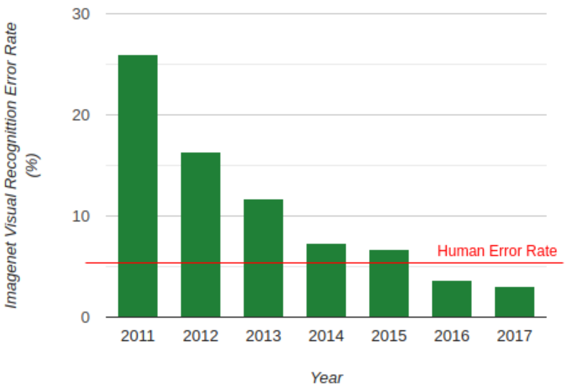


FIGURE 3. Imagenet Error statistics: As of the advent of Deep Learning, the performance on visual classification task sharply rose, inclusively surpassing human ability. Graphic adapted from [62].

and performing pose graph optimization using conventional feature-based SLAM.

C. LEARNING-BASED APPROACHES

Even tough the onset of deep learning methods is still fairly recent, it is already established that deep learning architectures can outperform most other methods for some computer vision tasks. As an example, concerning visual classification, deep learning methods managed to the take the Imagenet competition by storm, massively outperforming classical methods and even Humans at classifying objects in a picture (see Fig.3).

Building upon the success of deep learning frameworks for some visual tasks, a lot of research has been devoted to taking advantage of deep learning potential for a wider range of tasks. In this light, the egomotion between consecutive image frames can also be estimated with the use of deep neural architectures inspired by geometric models. The key principle is that for the egomotion estimation task we are interested in capturing the motion undergone by the camera system between consecutive images using an end-to-end deep neural architecture, bypassing almost all the modules in the classical VO pipeline (see Fig. 2) with a learned motion estimation scheme.

FlowNet [63] and its successive iterations garnered immense attention as a reliable deep learning framework for learning optical flow and paved the way for early egomotion estimators. Wang *et al.* proposed a monocular visual odometry system called DeepVO [64], which trains a RCNN to estimate camera motion in an end-to-end fashion, inferring pose directly from a sequence of raw RGB images in a video clip while bypassing all usual modules in the conventional VO pipeline. The advantage of such approach is to simultaneously factor in both feature extraction and sequential modelling through combining CNN's and RNN's.

As labeling data in large scale significantly hinders the application of supervised learning methods to robotic applications, Li *et al.* proposed UnDeepVO [65], a monocular system that uses stereo image pairs in the training phase for scale recovery. After training with unlabeled stereo images, UnDeepVO can simultaneously perform visual odometry and depth estimation with monocular images.

SfMLearner [66] is a solution that established an influential framework for Deep Learning for Visual Odometry research. It uses a monocular image sequence in order to estimate depth and pose simultaneously in an end-to-end unsupervised manner, through enforcing geometric constraints between image pairs in the view synthesis process. SfMlearner++ [67] improved upon the results in both depth and pose estimation by using the Essential matrix [68], obtained using the Five Point Algorithm [69], to enforce epipolar constraints on the loss function, effectively discounting ambiguous pixels.

GeoNet [70] is a similar approach, a jointly unsupervised learning framework for monocular depth, optical flow and egomotion estimation that decouples rigid scene reconstruction and dynamic object motion, making use of this knowledge to further tailor the geometric constraints to the model. Vijayanarasimhan *et al.* [71] presented SfM-Net, innovating through adding motion masks to photometric losses to jointly estimate optimal flow, depth maps and egomotion.

Luo *et al.* [72] also presented a three-prone architecture, composed of networks to predict the camera motion, dense depth map, and per-pixel optical flow between two frames, binding all together through adaptive consistency checks. Additionally, an holistic 3D motion parser is introduced, so as to distinguish between the camera system motion and dynamic objects in the scene, introducing extra robustness to occlusions.

Valada *et al.* [73] proposed a novel architecture that encompasses both global pose localization and a relative pose estimation, jointly regressing global pose and odometry and learning inter-task correlations and shared features through parameter sharing. This method is denoted as Deep Auxiliary Learning. This work was later extended to also perform pixel-wise semantic segmentation, still exploring the same inter-task parameter sharing strategy [74].

Learning pose corrections after estimation [75] can also be a very interesting approach. Visual Odometry methods are particularly sensitive to rotation errors, as small early drifts can have a large influence on final trajectory pose estimates. Peretroukhin [76] proposed HydraNet, a deep learning structure aimed at improving attitude estimates, able to be fused with classical visual methods. Through regressing unit quaternions, modeling rotation uncertainty and producing 3D covariances, HydraNet manages to improve visual algorithms at predicting 6-DoF pose estimates.

Another application Deep learning architectures are currently being tested on is sensor fusion. VINet [77] is a proposed framework that fuses pose estimates from DeepVO [64] with inertial data, showing comparable performance to traditional fusion systems. The same method was also adopted to fuse other kinds of information such as magnetic sensors, GPS, INS or wheel odometry [7], [78]. Sensor fusion can be easily incorporated into deep learning architectures and jointly trained end-to-end with pose regression, thus making a potentially interesting solution for Visual Odometry applications as it can be used for a wide variety of purposes (e.g. recovering absolute scale on monocular camera systems or correcting visual odometry drift). Deep-VIO [79] also employs IMU status update upon trajectory estimates, improving the overall accuracy of the method by taking advantage of the sensor fusion architecture.

Zhu *et al.* [80] proposed to combine unsupervised stereo optical flow estimation and monocular disparity with classical model-based optimization for camera pose estimation. Through using RANSAC [18] for relative pose estimation and outlier rejection, it reports to provide extra guarantees about instantaneous camera pose estimation robustness.

Generative Adversarial Networks (GAN's) have recently surfaced as a new type of network architecture to be exploited for the purpose of estimation egomotion [81], [82]. The operating principle is to have two different networks: a generator and a discriminator. The generator is trained to perform view synthesis reconstruction and pose estimation, while the discriminator is fed both the reconstructed image and the real one, learning to detect reconstruction errors and thus helping the generator also becoming more accurate at view reconstruction and pose estimation. A fully trained generator is then able to produce accurate pose estimates from visual input alone.

In this work, we employ a sensor fusion Recurrent Neural Network architecture, using inertial data as training set supervision to correct visual input predictions and limit the effects of accumulating VO drifts.

III. UNDERWATER VISUAL APPLICATION SCENARIOS

In the underwater context, there are not many publicly available large datasets. In addition, the ones that exist are mostly focused on vehicles navigating near the seabed [83], with close to no data pertaining to inland flooded mines. This work serves the dual purpose of wanting to critically assess visual odometry method performance for the robot navigation task in complex underwater scenarios, as well as exploring promising novel deep learning approaches to the problem. In particular, we are interested to work with the UNEXMIN UX-1 [84] robot data, as it poses several interesting challenges to visual-based navigation, given that operational mission scenarios in flooded deep mines is not a commonplace scenario in the visual odometry literature.

Data acquisition and dataset construction was therefore a key step towards the goal of benchmarking VO methods in our target scenarios. During this work, we were able to gather huge amounts of visual data automatically labeled by the UX-1 navigation software module, allowing for the construction of a dataset with significant localization accuracy without the need for any kind of manual labelling.

Deep learning methods in particular usually require vast amounts of data in order to properly train its neural architectures. This is particularly true in robotic applications, since autonomous systems can operate in very complex environments, often under extreme conditions. As so, the availability of large scale datasets with ever increasing variability spanning different scenarios and situational motions, is crucial for further development of deep learning algorithms and for improving upon generalization ability, as that leads to improved robustness when being deployed in large complex environments.

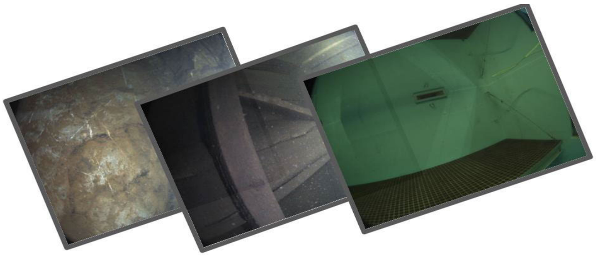


FIGURE 4. Underwater Visual Dataset image example: indoor pool and inland flooded mine imagery.

With this in mind, we can assert that the data used in this work represents a novel and varied underwater focused dataset collected with the UX-1, tailored for visual odometry method implementation and evaluation, with which we pretend to assess performance of state-of-the-art methods for VO estimation and Visual SLAM in different underwater scenarios. In Fig.4, we can observe example images of our dataset sequences, that showcase the different environments included in our dataset.

In this section, we are discussing in detail the data acquisition process, specifically describing the UNEXMIN UX-1 robot and all the technology contained within it, while providing related remarks about the image acquisition methodology, specifically the camera setup, the reasoning and assumptions of the process.

A. DATA ACQUISITION METHODOLOGY

As previously mentioned, the dataset was constructed using data acquired with the UX-1 robot. This robot was built for exploring and mapping decommissioned flooded mines so as to assess its geological potential, and is therefore equipped with a plethora of different sensors, including 5 cameras. In this work, mostly due to UX-1 design constraints which prevent the use of visual stereo methods, we are focusing on monocular camera setups, and as so, we chose to analyze the left camera system, with the goal of estimating robot pose in the central reference frame (i.e. pose estimates in the camera system reference frame has to be later transformed to the robot body reference frame). Groundtruth data is generated by the navigation module of the UX-1 software, a filtered calibration of sensor fusion from multiple local sensor sources (IMU, Doppler Velocity Logger, Structured Laser System, etc), progressively refined through multiple operation missions in complex settings and extremely challenging operational conditions.

In the scope of this work, we are working with the underlying assumption that this navigation data corresponds exactly to the real robot pose, which is not easily verifiable in

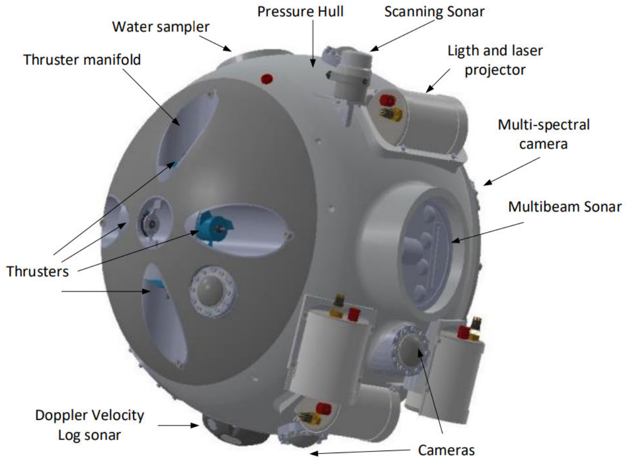


FIGURE 5. UNEXMIN UX-1 robot description.

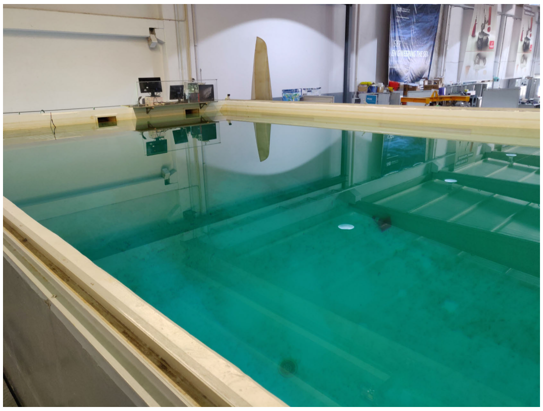


FIGURE 6. CRAS indoor pool.

operational mission scenarios. However, it can be asserted, with relative confidence, that this data represents a close approximation of the real robot position and can, therefore, be used as groundtruth for our use case. The groundtruth data file consists of a.txt file where each line contains 8 scalars, representing a timestamp and 6-DoF poses with a 3D translation vector and an orientation quaternion.

B. APPLICATION SCENARIOS

For the purpose of constructing a complete and thorough dataset, we utilize two different application scenarios, which pose different types of problems to visual-based methods:

1) **The CRAS pool** sequence depicts a fully known environment, ideal for calibrating some aspects of visual-based navigation, since all navigation information is fully verifiable. However, it is a rather non feature rich environment with lack of appropriate illumination conditions, which complicates visual-based navigation.
2) **The Urgeirica uranium mine** is a decommissioned flooded mine in Viseu, Portugal. It is mostly composed of vertical shafts that lead to 15-30m wide galleries. It is a real operational mission scenario for the UX-1, which was tasked with exploring and mapping the mine.

FIGURE 7. Urgeirica mine entrance.

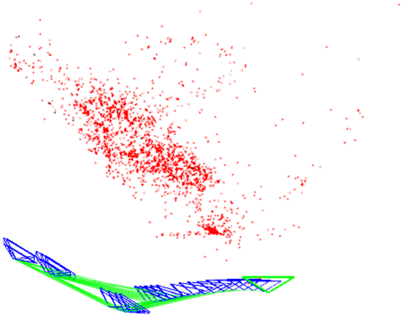


FIGURE 8. ORBSLAM-2 reconstructed map on the CRAS pool sequence.

IV. CAMERA POSE AND MOTION ESTIMATION

As mentioned before, our target scenarios for visual odometry estimation is our own underwater dataset. The visual perception and estimation of motion in this type of imagery still poses a big research challenge, as real-time persistent navigation accuracy with significant robustness to real world conditions is still not feasible.

So as to assess the extent to which it is possible to rely on visual input for navigation purposes on our challenging operational mission scenario, we decided to evaluate some of the most renown visual-based frameworks, namely ORBSLAM 2 [58], LDSO [56] and LIBVISO 2 [85]. In addition, following the significant attention deep learning methods have managed to garner in recent years, we also evaluate SfMLearner [66] and Geonet [70].

ORBSLAM 2 [58] is a complete feature-based SLAM systems that works in many different camera configurations and scenarios. In our work, we are sticking to the monocular use case and our dataset indoor pool and flooded mine scenario. It was designed to take advantage of the same feature set for all the modules: tracking, mapping, relocalization, and loop closure. Feature choice was ORB [25], mostly because of its extremely fast computation and matching, coupled with invariance to multiple viewpoints.

Initialization was designed to be robust to both planar or non-planar scenes, utilizing a heuristic to decide between the

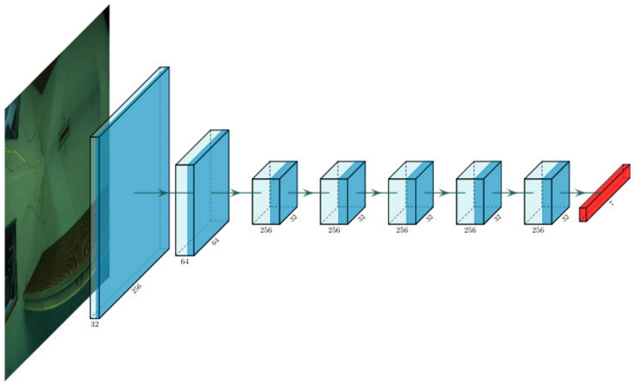


FIGURE 9. Representation of the SfMlearner PoseNet, the framework component responsible for regressing 6-DoF pose estimates. It consists of 7 blocks of convolutional layers followed by ReLU activations, outputting a 6-dimensional vector that comprises a 3D translation vector and Euler angles orientation representation.

parallelly computed homography assuming a planar scene and a fundamental matrix assuming a non-planar scene. In addition, the system has embedded a Bag-of-Words (BoW) [86] place recognition module that is used for performing both loop-closing and relocalization, built with the open-source library DBoW2 [87], taking advantage of an offline computation of the discretization of the descriptor space, which is known as the visual vocabulary.

LDSO [56] is an extension of the influential DSO [46], a direct probabilistic model that jointly considers minimizing a photometric error with consistently introducing the geometric notions of camera motion and pixel inverse depth on the reference frame. The introduction of loop closure detection, also through a Bag-of-Words approach, and also pose-graph optimization, allow the already robust tracking to be further improved through correction rotation, translation and scale drifts.

LIBVISO 2 [85] is another influential work egomotion estimation, based on computing the camera motion by minimizing the sum of reprojection errors and refining the obtained velocity estimates by means of a Kalman filter. It can work either with rectified stereo image sequences or for the monocular use case and produces an output 6D vector with estimated linear and angular velocities.

SfMLearner [66] is an unsupervised learning pipeline for depth and egomotion estimation. The unsupervised objective is fulfilled based on the following intuition: given knowledge of camera self-motion within a sequence of images and the depth of every pixel in those images, we can gain an unsupervised target by performing view synthesis.

As mentioned above, we are interested in evaluating Zhou's PoseNet, the SfMLearner framework component responsible for regressing 6-DoF pose estimates. The PoseNet architecture is essentially a temporal convolutional network which processes a sequence of n images by predicting relative transformation from the center image of the sequence (the image at the central position of the snippet, as shown in Fig. 10) to the other images in the sequence, outputting a *n-1* transformation vector composed of a 3D translation vector and a Euler angle orientation vector for each transformation.

FIGURE 10. CRAS pool 5-sequence length snippet.

The network itself is a convolutional regressor model with seven convolutional layers with stride-2 followed by ReLU activations, leading to a final linear convolution that outputs the aforementioned *6 x (n-1)* - dimensional channels. On top on this network, an "explainability" mask is used to down-weight the loss on image patches undergoing motion external to the camera motion (e.g. a car or pedestrian moving in the frame).

GeoNet [70] is a jointly trained end-to-end unsupervised learning framework for monocular depth, optical flow and egomotion estimation. Specifically, this framework focuses on extracting geometric relationships in the input data by separately considering static and dynamic elements in the scene. Significant performance gains have been reported, mostly due to increased robustness towards texture ambiguity and occlusions in the scene.

The framework is composed of two stages: the Rigid Structure Reconstructor and the Non-rigid Motion Localizer. The first stage is tasked with understanding the scene layout and structure and it consists of two sub-networks, i.e. the DepthNet and the PoseNet. The second stage concerns itself with dynamic objects in the scene and it utilized for the purpose of refining imperfect results from the first stage due to motion external to the camera motion, as well as help deal with high pixel saturation and extreme lighting conditions.

Similarly to SfMlearner, view synthesis at different stages works as a synthetic supervision for the unsupervised learning architecture, with image appearance similarity enforcing geometric and photometric consistency within the loss function.

The most relevant part of the framework in the scope of our work is the Pose Net, which consists of 7 convolutional layers followed by batch normalization and Relu activation (see Fig. 11). The prediction layers outputs the 6-DoF camera poses, i.e. translational vectors and orientation Euler angles.

V. VISUAL-INERTIAL FUSION NETWORK

Regardless of the algorithm, traditional monocular VO solutions are unable to observe the scale of the scene and are subject to scale drift and scale ambiguity. This is not different for deep neural architectures, as reported in the previously studied frameworks. The most common approach for pose optimization in the literature is to fuse visual and inertial data as a way to enforce global consistency with respect to the groundtruth data and therefore it would make sense to investigate analogous deep learning approaches to perform this task.

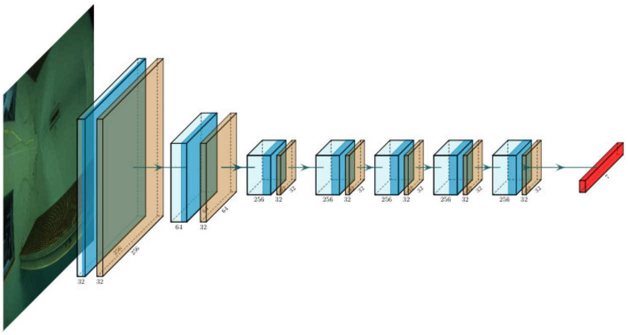


FIGURE 11. Representation of the GeoNet PoseNet, the framework component responsible for regressing 6-DoF pose estimates. It consists of 7 blocks of convolutional layers followed by ReLU activations and additional batch normalization layers, outputting a 6-dimensional vector that comprises a 3D translation vector and euler angles orientation representation.

In this work, we propose a Recurrent Neural Network architecture anchored in a supervised learning scheme whereby we use filtered IMU readings as a supervision for 6-DoF pose estimate optimization.

The **input space** of this network are the concatenated egomotion predictions of both SfmLearner and GeoNet, i.e. global trajectory estimates in the robot central body frame. For this purpose, and due to deep learning architectures requiring large amounts of data to converge to a robust model, we had to run multiple predictions from both frameworks so as to synthetize a dataframe dataset.

The **network** itself consists of stacked LSTM units working with progressively smaller time step lags leading to a multilayer perceptron that regresses the optimized trajectory estimate. The goal is to process the data as a sequence-to-sequence problem, optimizing the input trajectory estimates to a more globally consistent trajectory.

At training time, the network is fed both the visual odometry pose estimates and the IMU readings. At inference time, only the visual odometry poses estimates are fed to the network, as it is able to generalize the pose corrections that would be introduced by the inertial readings.

The **fundamental assumption** driving this architecture is that the output space of the optimized trajectory estimate lie in a manifold much smaller than 6-DoF space. Implicitly constraining the output prediction space to a minimization of the mean square error between visual and inertial data helps to avoid the curse of dimensionality.

For **loss function** design, the intuition was that we needed to make use of the quaternion parametrization to penalize rotation errors in a meaningful way. In this light, we decoupled the translation and rotation components and formulated a loss function that takes the mean squared error for translation and the quaternion distance between estimate and groundtruth in the SO(3) group.

$$loss = \sqrt{\sum(E_x^2 + E_y^2 + E_z^2) + \sum |q_e - q|} \qquad (1)$$

where $E_{x\ldots z}$ represents the computation of distance between estimate and groundtruth position. Quaternion distance is computed as the norm of the difference between estimate and groundtruth quaternions. In addition, we constrained the equation to take into account the fact that q and -q encode the same rotation, only considering the smaller of the two possible distances in the loss function calculation. This loss function design allowed for a more meaningful handling of rotations, especially around capturing the most significant motion direction.

VI. EXPERIMENTAL RESULTS

A. IMPLEMENTATION DETAILS AND TRAINING PROCEDURES

In this section, we focus on the experimental results we were able to obtain for the previously mentioned egomotion estimation frameworks. In addition, we will show the impact of the Visual-Inertial Fusion Network so as to optimize the trajectory estimate and correct inherent VO drift on the trajectory data generated by SfMLearner.

For testing SLAM systems, we are aware that the underwater environment is very different from the environments they were designed to tackle, with several different challenges they were not tailored to. As an attempt to further calibrate the environment, a bag-of-words visual dictionary was compiled taking into consideration the underwater scenarios we want to work on. For this purpose, we utilized the DBOW2 library [87] and constructed a representation of the descriptor space spanning both of our underwater environments with sufficient data for providing the SLAM systems with an additional tool for place recognition and pose correction.

SfMLearner and GeoNet share the data preprocessing step whereby the input image sequence is split into 5 sequence length snippets (see Fig. 10). In conjunction with camera intrinsic calibration and image timestamps, the 416 × 128 snippets were fed to the frameworks and the neural networks were trained using tensorflow [88] running on a CUDA enabled Nvidia GTX 1080. It is also worth noting that a post-processing step was implemented in order to recover full concatenated trajectory from the 5-snippet length predicts, so as to analyze also the global trajectory errors. Some context finetuning was performed, empirically adapting the network to penalize heavier errors in rotation as large global trajectory errors were being introduced due to early rotation errors unaligning the pose estimates with the groundtruth, thus accumulating significant drift.

For the Visual-Inertial Fusion Network on the other hand, and given that there was no prior knowledge about how to tune a pose optimization network, we adopted a grid-search learning scheme to sweep multiple combinations of hyperparameters and return the one that converges to smaller loss values. This is only feasible in a short timeframe because we are working with low dimensional data (i.e. dataframes instead of high resolution imagery) but for this application, it is perfectly suited for finding an optimal solution for hyperparameter tuning.

TABLE 1. Absolute Trajectory Error (ATE) evaluation.

Absolute Trajectory Error (ATE)

	CRAS Pool	Urgeirica Mine
ORBSLAM 2	0.3491	n/a
LDSO	0.1664	n/a
LIBVISO2-m	0.3091	0.5618
SfMLearner	0.1699	0.1502
GeoNet	0.1175	0.1382
ours	**0.0714**	**0.1112**

B. RESULTS

In order to analyze and benchmark the obtained trajectory estimate results and method performance, we are following quantitative VO literature [89]. This envolves processing the pose estimates with scale correction optimization and alignment with groundtruth data, so as to resolve scale ambiguity and minimize the impact of early drift accumulation errors.

In the remaining of this section, we will present and discuss the results considering the full concatenated trajectory, for all considered methods. Particularly, when it comes to deep learning methods, we are escaping the snippet representation and recomputing errors with respect to groundtruth pose for the entire test sequence trajectories under analysis. Mean Absolute Trajectory Error (ATE) is an single value error metric for position, rotation and velocity estimation, which makes it easy for comparisons.

Table 1 shows the best results we were able to obtain for our two test sequences. Results were only considered when the algorithms manage to produce an estimate of at least two thirds of the trajectory sequence. SLAM systems struggle a lot in the Urgeirica mine scenario, both in achieving a robust initialization and also keeping track of the environment for long periods of time. It was noted that after losing tracking, most of the times the systems could not recover and reinitialize. Monocular LIBVISO2 manages to get an somewhat consistent estimate, as shown in fig. 12.

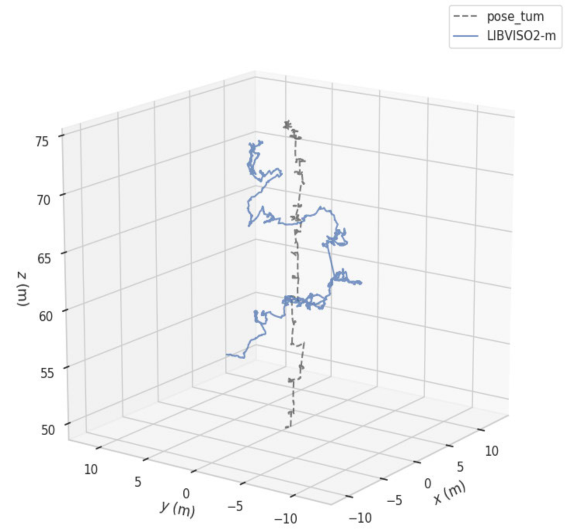


FIGURE 12. Monocular LIBVISO2 estimation for the Urgeirica mine test sequence: trajectory estimate against groundtruth data.

The designed Visual-Inertial Fusion Network makes use of IMU data in the training step of the algorithm, but at inference time it uses only visual images as input, thus working “in essence” as a monocular vision algorithm. Nonetheless, it is sensible to draw a comparison with VIO algorithms, since the input data is readily available and the addition of inertial data could be acting as the difference maker for overall accuracy improvements. To this effect, two renown opensource VIO frameworks, namely VINS-mono [90] and R-VIO [91] were also tested on our underwater scenarios. Despite tight camera intrinsic and camera to IMU extrinsic calibration, VINS-mono [90] was not able to generate a sequence estimate for any of our dataset test sequences, especially because the

TABLE 2. Result compilation for Absolute Pose Error w.r.t. translation.

		Absolute Pose Error (APE)					
		"raw" comparison		scale-corrected		SIM(3) Umeyama aligment	
		Avg.Error(m)	RMSE (m)	Avg.Error(m)	RMSE (m)	Avg.Error(m)	RMSE (m)
CRAS POOL	SfMlearner	3.301±2.049	3.996	2.755±1.573	3.049	0.731±0.440	0.905
	GeoNet	28.739±14.613	29.912	20.846±6.687	20.087	5.345±1.112	5.475
	ours	**2.329±1.781**	**2.877**	**1.380±1.259**	**1.380**	**0.570±1.005**	**0.637**
Urgeirica Mine	SfMlearner	52.709±1.199	52.461	20.354±3.366	19.129	0.7208±0.584	1.158
	GeoNet	55.392±2.728	56.096	22.043±1.041	22.475	0.839±0.543	1.077
	ours	**46.269±2.928**	**47.973**	**4.177±0.219**	**4.227**	**0.168±0.106**	**0.212**

TABLE 3. Average position and orientation errors for RVIO [91].

	CRAS pool	Urgeirica mine
Absolute Position Error (m)	1.318±1.691	5.096 ±3.954
RMSE	1.720	6.450
Absolute Orientation Error (○)	2.43 ±0.21	22.190 ±5.269
RMSE	2.440	22.532

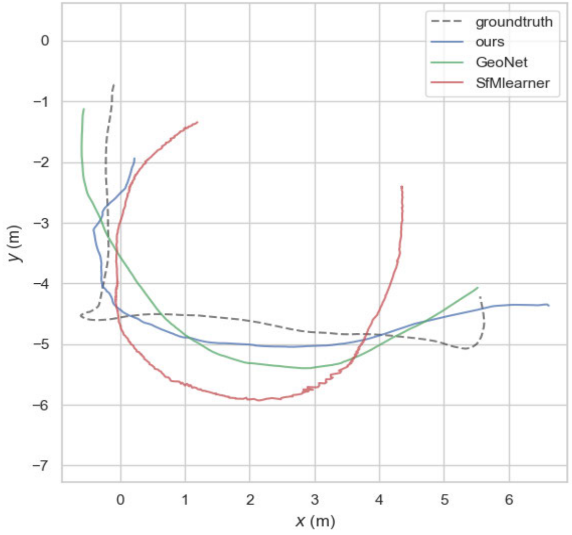


FIGURE 13. Results for the CRAS pool sequence: trajectory estimates against groundtruth data.

visual component prevented robust initialization due to low parallax and rather low feature environments. R-VIO [91], on the other hand, was able to generate trajectory estimates, as reported in table 3.

For our underwater visual dataset test sequences, most of the tested visual only or visual-inertial odometry estimation algorithms struggle with providing a consistent estimate for our trajectory test sequence. We attribute this to domain-specific characteristics that shed light to method degenerate conditions in repetitive low feature environments. In that regard, the performance of Deep Learning methods showed comparatively great potential, as their data-centric nature appears to have been able to obtain a better representation of the feature space within the images, thus helping improve the overall accuracy of the methods.

Zooming in on deep learning methods, as they have shown to be more robust to the environment characteristics, we want to quantify the trajectory error with respect to its translational component. For the sake of coherent representation, and abiding to quantitative VO metrics, we will present the trajectories before and after SIM(3) alignment, which consists of the application of a post-processing step denoted as Umeyama alignment [92]. This algorithm performs a least-squares estimation of transformation parameters translation, rotation and scale between estimates and groundtruth pose data, so that the RMSE between groundtruth and aligned estimates can then be computed. Table 2 shows that comparison, reporting on metric average errors with standard deviation and additionally the RMSE.

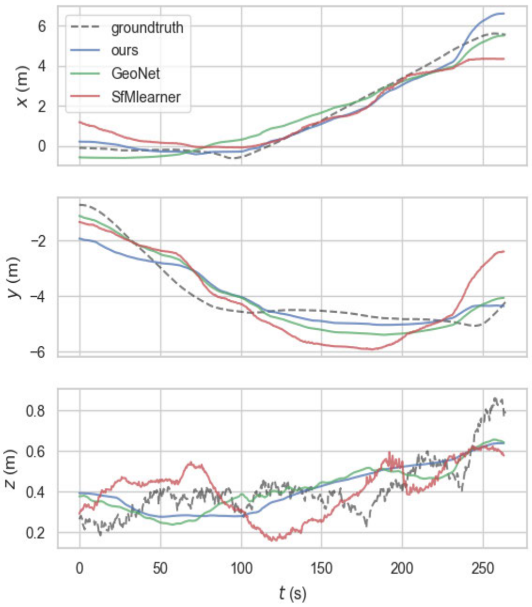


FIGURE 14. Results for the CRAS pool sequence: trajectory estimates against groundtruth data decoupled by translational component.

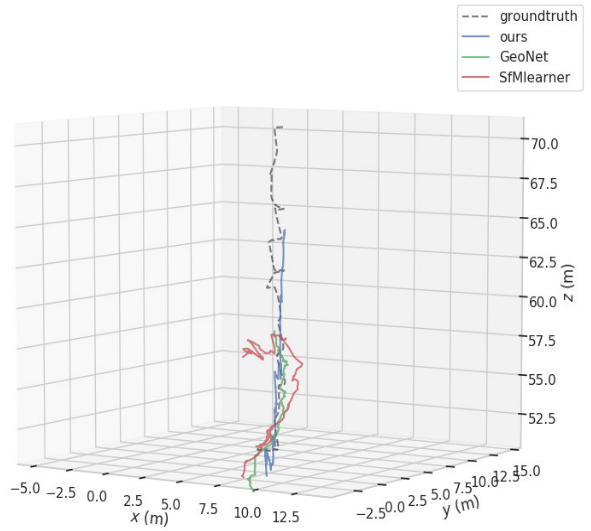


FIGURE 15. Results for Urgeirica Mine sequence: computed trajectory estimates against groundtruth data.

As it can be observed in Fig.13 and Fig.14, our Visual-Inertial Fusion Network component is outperforming the other visual methods with respect to mimicking groundtruth trajectory shape. In a somewhat textureless environment, the introduction of inertial information pose correction is allowing for capturing the robot self-motion in a more adequate fashion, especially around handling rotations.

In Fig. 15, that is even more evident, as the Urgeirica mine sequence is significantly more challenging. In this case, it is also possible to see that our method is able to capture more adequately the most significant motion direction, not overly considering lateral motions to it. Again, it is shown to be the closest approximation to groundtruth pose information in this scenario, as numerically reported in table 1 and table 2.

VII. CONCLUSION

A. SUMMARY

In this paper, the focus is placed on analysing visual-based robot navigation, exploring the different ways in which robot egomotion can be estimated. A special emphasis is placed upon data-driven learning-based approaches and, also, the underwater application scenario.

As detailed in section III, a comprehensive underwater visual dataset was constructed, which encompasses different texture environments and provides different types of challenges for visual-based pose and/or motion estimation. This was achieved through the use of data acquired with the UX-1 robot, a robot tailored for exploration and mapping of indoor decommissioned flooded mines. The dataset presents both a fully known controlled environment in our CRAS indoor pool, as well as an operational mission scenario in the Urgeirica uranium mine in Viseu, Portugal.

Robot localization and motion estimation literature was thoroughly reviewed and some of most renown opensource software was tested, benchmarked and evaluated on our underwater visual dataset. There was a particular interest to see how deep learning algorithms would compare to classical VO and v-SLAM approaches in the context of underwater visual environments, as there was close to no information about that in the literature.

We soon realized that SLAM systems were not suited for handling our use case challenging underwater scenarios. Initialization proved to be a difficult problem since sometimes there is no reliable texture and/or volume of features to build an initial map of environment. Tracking also proved to lose itself multiple times over the test sequences, with relocalization again presenting a difficult issue. Overall, it results in systems that can only estimate subsets of the target trajectory, despite the high accuracy on those trajectory patches.

Classical VO, specifically monocular LIBVISO-2, performed much better in that regard, estimating much more successfully the test sequence trajectories. However, it was also not able to extract the translation and rotation for every pair of consecutive frames, probably due to low amounts of features and small inter-frame displacements. The same can be said for the tested VIO systems, as low parallax and small amount of features detected significantly hinder their performance on our dataset test sequences.

Deep learning architectures, on the other hand, present a significant advantage in that regard, as its data-centric mathematical formulation allows for constant pose estimation for all frames, even in challenging operational mission scenarios. The intrinsic high-level representation of the feature space enforces consistency along the image sequence and additional robustness to the environment challenges. However, some challenging issues could be clearly identified, as poor handling of rotations and drift accumulation compromised the accuracy and reliability of pose estimates, as required by real robotic systems.

A Visual-Inertial Fusion Network was presented in section V, with the purpose of addressing the aforementioned egomotion performance problems. The proposed solution consists of a Visual-Inertial Fusion Network, aimed at improving global pose estimates through an inertial supervision learning scheme. This supervised architecture proved to significantly improve results on global pose estimation, with around 40% better error rates.

In this work, real-time implementation of deep learning algorithms was not addressed, mainly because the UX-1 does not possess any type of GPU hardware, therefore rendering any conclusion from on board implementations non-viable. In addition, and although the robot possesses multiple cameras, visual stereo implementations are significantly hard to design for this particular robot, due to non-overlapping camera fields-of-view.

B. FUTURE WORK

There are several open opportunities and challenges for underwater data-centric motion estimation. In particular, and following the groundwork layed on in this article, the following research action is suggested:

- Sensitivity analysis of image preprocessing steps (i.e. haze removal and vignetting mitigation).
- Integration of visual-inertial fusion within end-to-end deep learning for underwater robot navigation pipelines. Further study and calibration of inertial measurement integration for underwater mobile robots.
- Assessment and testing of visual stereo implementations on top of deep learning architectures for the underwater context. This work focused on monocular camera setups mostly due to the UX-1 design constraints, yet it would be interesting to investigate the performance of deep learning architectures also for the stereo use case.
- Closing the loop also poses as an interesting challenge for VO estimation deep learning methods, as VO can greatly benefit from global optimization steps to correct VO drift accumulation. Place recognition or higher level semantics are worth exploring as a solution for loop-closure in DL algorithms.
- Real-time implementation and testing of deep learning architectures for both relocalization and egomotion tasks for the underwater context.

ACKNOWLEDGMENT

The authors would like to thank the UNEXMIN project for providing the operational mission scenario data utilized in the scope of this work.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, vol. 1. 2012, pp. 1097–1105. [Online]. Available: http://dl.acm.org/citation.cfm?id=2999134.2999257

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2015, *arXiv:1506.02640*. [Online]. Available: http://arxiv.org/abs/1506.02640

[3] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," 2015, *arXiv:1506.01497*. [Online]. Available: http://arxiv.org/abs/1506.01497

[5] R. Garg, V. K. Bg, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 740–756.

[6] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[7] S. Pillai and J. J. Leonard, "Towards visual ego-motion learning in robots," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 5533–5540.

[8] A. Kendall, M. Grimes, and R. Cipolla, "PoseNet: A convolutional network for real-time 6-DOF camera relocalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2938–2946.

[9] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, Aug. 2013.

[10] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[11] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The Oxford RobotCar dataset," *Int. J. Robot. Res.*, vol. 36, no. 1, pp. 3–15, Nov. 2016.

[12] B. Teixeira, H. Silva, A. Matos, and E. Silva, "Deep learning approaches assessment for underwater scene understanding and egomotion estimation," in *Proc. OCEANS MTS/IEEE SEATTLE*, Oct. 2019, pp. 1–9.

[13] A. Howard, "Real-time stereo visual odometry for autonomous ground vehicles," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2008, pp. 3946–3952.

[14] J. Engel, J. Sturm, and D. Cremers, "Semi-dense visual odometry for a monocular camera," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1449–1456.

[15] D. Scaramuzza, F. Fraundorfer, M. Pollefeys, and R. Siegwart, "Closing the loop in appearance-guided structure-from-motion for omnidirectional cameras," in *Proc. 8th Workshop Omnidirectional Vis., Camera Networks Non-Classical Cameras (OMNIVIS)*, Marseille, France, R. Swaminathan, V. Caglioti, and A. Argyros, Eds., Oct. 2008.

[16] H. P. Moravec, "Obstacle avoidance and navigation in the real world by a seeing robot rover," Dept. Comput. Sci., Stanford Univ. Stanford, CA, USA, Tech. Rep. STAN-CS-80-813, 1980.

[17] C. F. Olson, L. H. Matthies, M. Schoppers, and M. W. Maimone, "Robust stereo ego-motion for long distance navigation," in *Proc. CVPR*, Jun. 2000, pp. 453–458.

[18] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[19] Y. Cheng, M. Maimone, and L. Matthies, "Visual Odometry on the mars exploration rovers," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, vol. 1, Oct. 2005, pp. 903–910. [Online]. Available: http://ieeexplore.ieee.org/document/1571261/

[20] D. Nister, O. Naroditsky, and J. Bergen, "Visual odometry," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jul. 2004, p. 1. [Online]. Available: https://www.engineeringvillage.com/share/document.url?mid=inspec_480457%fff4e3c658M603419255120119&database=ins

[21] C. Engels, H. Stewénius, and D. Nistér, "Bundle adjustment rules," *Photogramm. Comput. Vis.*, vol. 2, p. 1–6, Sep. 2006.

[22] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. Alvey Vis. Conf.*, 1988, vol. 15, no. 50, p. 5244.

[23] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[24] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2006, pp. 404–417.

[25] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2564–2571.

[26] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: Binary robust independent elementary features," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 778–792.

[27] P. Chang and M. Hebert, "Omni-directional structure from motion," in *Proc. IEEE Workshop Omnidirectional Vis.*, Jun. 2000, pp. 127–133.

[28] D. Scaramuzza, A. Martinelli, and R. Siegwart, "A toolbox for easily calibrating omnidirectional cameras," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2006, pp. 5695–5701.

[29] C. Engels, H. Stewénius, D. Nistér, B. Triggs, P. F. McLauchlan, R. I. Hartley, A. W. Fitzgibbon, M. Quigley, K. Conley, B. P. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, "Bundle adjustment–a modern synthesis," in *Proc. ICRA Workshop Open Source Softw.*, vol. 2. Berlin, Germany: Springer, 2009, pp. 298–372.

[30] D. Scaramuzza, "1-Point-RANSAC structure from motion for vehicle-mounted cameras by exploiting non-holonomic constraints," *Int. J. Comput. Vis.*, vol. 95, no. 1, pp. 74–85, Apr. 2011.

[31] B. Kitt, A. Geiger, and H. Lategahn, "Visual odometry based on stereo image sequences with RANSAC-based outlier rejection scheme," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2010, pp. 486–492.

[32] L. Kneip, M. Chli, and R. Siegwart, "Robust real-time visual odometry with a single camera and an IMU," in *Proc. Brit. Mach. Vis. Conf.*, 2011, pp. 1–12.

[33] T. Kazik, L. Kneip, J. Nikolic, M. Pollefeys, and R. Siegwart, "Real-time 6D stereo visual odometry with non-overlapping fields of view," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1529–1536.

[34] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," M.S. thesis, Dept. Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA, 1981.

[35] E. J. Gibson, "Where is the information for Affordances?" *Ecol. Psychol.*, vol. 12, no. 1, pp. 53–56, Feb. 2000.

[36] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques," *Int. J. Comput. Vis.*, vol. 12, no. 1, pp. 43–77, Feb. 1994.

[37] J. J. K. Lim, "Egomotion estimation with large field-of-view vision," Australian Nat. Univ., Canberra, ACT, Australia, Tech. Rep., 2010.

[38] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *Int. J. Comput. Vis.*, vol. 56, no. 3, pp. 221–255, Feb. 2004.

[39] J. Weber and J. Malik, "Robust computation of optical flow in a multi-scale differential framework," *Int. J. Comput. Vis.*, vol. 14, no. 1, pp. 67–81, Jan. 1995.

[40] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, nos. 1–3, pp. 185–203, Aug. 1981.

[41] D. J. Heeger, "Optical flow using spatiotemporal filters," *Int. J. Comput. Vis.*, vol. 1, no. 4, pp. 279–302, Jan. 1988.

[42] D. J. Fleet and A. D. Jepson, "Computation of component image velocity from local phase information," *Int. J. Comput. Vis.*, vol. 5, no. 1, pp. 77–104, Aug. 1990.

[43] M. Ogata and T. Sato, "Motion-detection model with two stages: Spatiotemporal filtering and feature matching," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 9, no. 3, pp. 377–387, Mar. 1992.

[44] A. Goshtasby, S. H. Gage, and J. F. Bartholic, "A two-stage cross correlation approach to template matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vols. PAMI–6, no. 3, pp. 374–378, May 1984.

[45] T. A. Camus, "Method for real time correlation of stereo images," U.S. Patent 6 516 087, 2003.

[46] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 611–625, Mar. 2018.

[47] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast semi-direct monocular visual odometry," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2014, pp. 15–22.

[48] H. Silva, A. Bernardino, and E. Silva, "Probabilistic egomotion for stereo visual odometry," *J. Intell. Robotic Syst.*, vol. 77, no. 2, pp. 265–280, Apr. 2014.

[49] H. Silva, A. Bernardino, and E. Silva, "A voting method for stereo egomotion estimation," *Int. J. Adv. Robotic Syst.*, vol. 14, no. 3, Jun. 2017, Art. no. 172988141771079.

[50] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotic*. Cambridge, MA, USA: MIT Press, 2005.

[51] A. J. Davison, "Real-time simultaneous localisation and mapping with a single camera," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, 2003, pp. 1403–1410.

[52] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "MonoSLAM: Real-time single camera SLAM," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1052–1067, Jun. 2007.

[53] J. Shi and Tomasi, "Good features to track," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 1994, pp. 593–600.

[54] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *Proc. 6th IEEE ACM Int. Symp. Mixed Augmented Reality*, Nov. 2007, pp. 1–10.

[55] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 834–849.

[56] X. Gao, R. Wang, N. Demmel, and D. Cremers, "LDSO: Direct sparse odometry with loop closure," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 2198–2204.

[57] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.

[58] R. Mur-Artal and J. D. Tardos, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, Oct. 2017.

[59] R. A. Newcombe, A. Fitzgibbon, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, and S. Hodges, "KinectFusion: Real-time dense surface mapping and tracking," in *Proc. 10th IEEE Int. Symp. Mixed Augmented Reality*, Oct. 2011, pp. 127–136.

[60] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. J. Kelly, and A. J. Davison, "SLAM++: Simultaneous localisation and mapping at the level of objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1352–1359.

[61] K. Tateno, F. Tombari, I. Laina, and N. Navab, "CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jul. 2017, pp. 6243–6252.

[62] C. P. Langlotz, B. Allen, B. J. Erickson, J. Kalpathy-Cramer, K. Bigelow, T. S. Cook, A. E. Flanders, M. P. Lungren, D. S. Mendelson, J. D. Rudie, G. Wang, and K. Kandarpa, "A roadmap for foundational research on artificial intelligence in medical imaging: From the 2018 NIH/RSNA/ACR/The academy workshop," *Radiology*, vol. 291, no. 3, pp. 781–791, Jun. 2019.

[63] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. V. D. Smagt, D. Cremers, and T. Brox, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2758–2766.

[64] S. Wang, R. Clark, H. Wen, and N. Trigoni, "DeepVO: Towards end-to-end visual odometry with deep recurrent convolutional neural networks," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 2043–2050.

[65] R. Li, S. Wang, Z. Long, and D. Gu, "UnDeepVO: Monocular visual odometry through unsupervised deep learning," 2017, *arXiv:1709.06841*. [Online]. Available: http://arxiv.org/abs/1709.06841

[66] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017. vol. 2, no. 6, p. 7.

[67] V. Prasad and B. Bhowmick, "SfMLearner++: Learning monocular depth & ego-motion using meaningful geometric constraints," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 2087–2096.

[68] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. New York, NY, USA: Cambridge Univ. Press, 2003.

[69] D. Nister, "An efficient solution to the five-point relative pose problem," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 6, pp. 756–770, Jun. 2004.

[70] Z. Yin and J. Shi, "GeoNet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1983–1992.

[71] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, and K. Fragkiadaki, "SfM-net: Learning of structure and motion from video," 2017, *arXiv:1704.07804*. [Online]. Available: http://arxiv.org/abs/1704.07804

[72] C. Luo, Z. Yang, P. Wang, Y. Wang, W. Xu, R. Nevatia, and A. Yuille, "Every pixel counts ++: Joint learning of geometry and motion with 3D holistic understanding," 2018, *arXiv:1810.06125*. [Online]. Available: http://arxiv.org/abs/1810.06125

[73] A. Valada, N. Radwan, and W. Burgard, "Deep auxiliary learning for visual localization and odometry," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 6939–6946.

[74] N. Radwan, A. Valada, and W. Burgard, "VLocNet++: Deep multitask learning for semantic visual localization and odometry," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 4407–4414, Oct. 2018.

[75] V. Peretroukhin and J. Kelly, "DPC-net: Deep pose correction for visual localization," *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 2424–2431, Jul. 2018.

[76] V. Peretroukhin, B. Wagstaff, M. Giamou, and J. Kelly, "Probabilistic regression of rotations using quaternion averaging and a deep multi-headed network," 2019, *arXiv:1904.03182*. [Online]. Available: http://arxiv.org/abs/1904.03182

[77] R. Clark, S. Wang, H. Wen, A. Markham, and N. Trigoni, "VINet: Visual-inertial Odometry as a sequence-to-sequence learning problem," in *Proc. 31st AAAI Conf. Artif. Intell.* Dec. 2017, pp. 3995–4001.

[78] M. Turan, J. Shabbir, H. Araujo, E. Konukoglu, and M. Sitti, "A deep learning based fusion of RGB camera information and magnetic localization information for endoscopic capsule robots," *Int. J. Intell. Robot. Appl.*, vol. 1, no. 4, pp. 442–450, Nov. 2017.

[79] L. Han, Y. Lin, G. Du, and S. Lian, "DeepVIO: Self-supervised deep learning of monocular visual inertial odometry using 3D geometric constraints," 2019, *arXiv:1906.11435*. [Online]. Available: http://arxiv.org/abs/1906.11435

[80] A. Zihao Zhu, W. Liu, Z. Wang, V. Kumar, and K. Daniilidis, "Robustness meets deep learning: An End-to-End hybrid pipeline for unsupervised learning of egomotion," 2018, *arXiv:1812.08351*. [Online]. Available: http://arxiv.org/abs/1812.08351

[81] Y. Almalioglu, M. R. U. Saputra, P. P. B. D. Gusmao, A. Markham, and N. Trigoni, "GANVO: Unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 5474–5480.

[82] T. Feng and D. Gu, "SGANVO: Unsupervised deep visual odometry and depth estimation with stacked generative adversarial networks," 2019, *arXiv:1906.08889*. [Online]. Available: http://arxiv.org/abs/1906.08889

[83] M. Ferrera, V. Creuze, J. Moras, and P. Trouvé-Peloux, "AQUALOC: An underwater dataset for visual–inertial–pressure localization," *Int. J. Robot. Res.*, vol. 38, no. 14, pp. 1549–1559, Oct. 2019.

[84] S. Domínguez, C. Rossi, A. Martins, J. Almeida, C. Almeida, A. Dias, N. Dias, J. Aaltonen, A. Heininen, K. Koskinen, C. Vörös, S. Henley, M. McLoughlin, H. van Moerkerk, J. Tweedie, B. Bodo, N. Zajzon, and E. Silva, "UX 1 system design—A robotic system for underwater mining exploration," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 1494–1500.

[85] A. Geiger, J. Ziegler, and C. Stiller, "StereoScan: Dense 3D reconstruction in real-time," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2011, pp. 963–968.

[86] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," *Int. J. Robot. Res.*, vol. 27, no. 6, pp. 647–665, Jun. 2008.

[87] D. Galvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Trans. Robot.*, vol. 28, no. 5, pp. 1188–1197, Oct. 2012.

[88] M. Abadi. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. [Online]. Available: https://www.tensorflow.org/

[89] Z. Zhang and D. Scaramuzza, "A tutorial on quantitative trajectory evaluation for Visual(-Inertial) odometry," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 7244–7251.

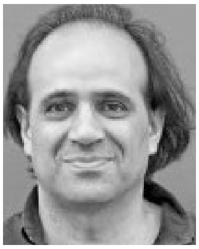
[90] T. Qin, P. Li, and S. Shen, "VINS-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 1004–1020, Aug. 2018.

[91] Z. Huai and G. Huang, "Robocentric visual-inertial odometry," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 6319–6326.

[92] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 4, pp. 376–380, Apr. 1991.

BERNARDO TEIXEIRA received the M.Sc. degree in electrical and computer engineering from the Faculdade de Engenharia, Universidade do Porto (FEUP), in 2019. He is currently a Researcher with INESC TEC, Centre for Robotics and Autonomous Systems (CRAS). His research activities pertain mostly to deep learning applications in the scope of robotic systems, with a focus on robotic relocalization and visual odometry estimation tasks.

HUGO SILVA was born in Porto, Portugal, in 1979. He received the licentiate degree in electrical and electronic engineering from the ISEP Porto Polytechnic School, in 2004, and the master's degree in electronics and computers engineering from the IST University of Lisbon, in 2008. In 2009, he obtained a PhD Scholarship from the Portuguese Science Foundation (FCT), and graduated (Ph.D.) in electronics and computers engineering from the IST University of Lisbon, in 2014. He currently works in INESC TEC as a Senior Researcher, where he is currently a project member in several international FP7 and H2020 projects. He is the main author of several research publications in the domains of computer vision and mobile robotics applications.



ANIBAL MATOS received the Ph.D. degree in electrical and computer engineering from Porto University, in 2001. He is currently a Research Coordinator of the Centre for Robotics and Autonomous Systems with INESC TEC and also an Assistant Professor with the Faculty of Engineering, Porto University. His main research interests are related to perception, sensing, navigation, and control of autonomous marine robots, being the author or coauthor of more than 80 publications in international journals and conferences. He has participated and lead several research projects on marine robotics and its application to monitoring, inspection, search and rescue, and defense.

EDUARDO SILVA received the Ph.D. degree in electrical and computer engineering from Porto Polytechnic Institute. He is currently an Assistant Professor with the Porto Polytechnic Institute. He is currently a Robotics and Autonomous Systems Research Coordinator at INESC TEC. His main research areas are control architectures and navigation for autonomous mobiles robots. He has participated in over 25 research projects and has contributed with more than 50 publications in the area of field robotics.