Deep Learning Approaches Assessment for Underwater Scene Understanding and Egomotion Estimation

Bernardo Teixeira*‡, Hugo Silva*, Anibal Matos*‡, Eduardo Silva*†

* INESCTEC - Institute for Systems and Computer Engineering, Technology and Science
† ISEP - School of Engineering, Porto Polytechnic Institute, Portugal
‡ FEUP - Faculty of Engineering, University of Porto, Portugal
Email: bernardo.g.teixeira, hugo.m.silva, anibal.matos, eduardo.silva@inesctec.pt

***Abstract*—This paper address the use of deep learning approaches for visual based navigation in confined underwater environments. State-of-the-art algorithms have shown the tremendous potential deep learning architectures can have for visual navigation implementations, though they are still mostly outperformed by classical feature-based techniques.**

In this work, we apply current state-of-the-art deep learning methods for visual-based robot navigation to the more challenging underwater environment, providing both an underwater visual dataset acquired in real operational mission scenarios and an assessment of state-of-the-art algorithms on the underwater context. We extend current work by proposing a novel pose optimization architecture for the purpose of correcting visual odometry estimate drift using a Visual-Inertial fusion network, consisted of a neural network architecture anchored on an Inertial supervision learning scheme.

Our Visual-Inertial Fusion Network was shown to improve results an average of 50% for trajectory estimates, also producing more visually consistent trajectory estimates for both our underwater application scenarios.

***Index Terms*—Artificial intelligence, Computer vision, Deep learning, Visual Odometry, Robot navigation**

I. Introduction

Deep Learning has become the ”Holy Grail” approach for visual based classification tasks. Most of the renown novel methods for computer vision classification tasks [1][2][3][4] are based on deep learning architectures (e.g. neural network approaches) and deep learning is becoming ubiquitous in most of today's Computer Vision applications.

Based on the significant improvements on accuracy and performance obtained in visual detection and classification tasks, novel deep learning approaches for other applications such as camera pose estimation and visual motion estimation (i.e. Visual Odometry (VO)), started to surface, laying the groundwork for the acquisition of robust reliable data that can be used to feed visual SLAM systems. Motion estimation for visual based navigation applications is one of the key challenges in Computer Vision that is undergoing extensive research in the field of Robotics research, but also in the context of autonomous driving applications. This is partly due to the development and proliferation of more powerful and cheaper GPU alternatives, which has prompted the surge of data-driven methods such as deep learning to also tackle VO tasks. In recent years, learning based VO has been drawing significant attention, as it can combine powerful feature representation capability with increased robustness especially in more complex scenarios.

Fig. 1. UX-1 Robot photo at Kaatiala Mine courtesy of UNEXTMIN project

One of most complex scenarios is the underwater environment, where visual based navigation methods tend to fail or have lackluster performance due to the lack of appropriate lighting conditions, water turbidity, backscattering effect, lack of image texture and vignetting effect. In this work, we benchmarked deep learning egomotion estimation methods performance in underwater motion estimation scenarios using indoor pool image sequences but also real operation mission scenarios from the UNEXTMIN UX-1[1] [5] robot. The dataset was acquired while the robot was in operation exploring and mapping flooded caved mines.

Our contribution in this paper is twofold: (i) assessment and evaluation of deep learning motion estimation frameworks

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

on the underwater context; (ii) a novel Visual-Inertial Fusion Network that manages to improve around **50%** upon global trajectory estimate errors

The remaining of the paper has the following organization: Section II contains a review of relevant work in the deep learning for computer vision research field, with the focus placed on VO tasks. In section III, we describe the different dataset scenarios. In section V we describe the design of the novel Visual-Inertial Fusion Network approach. Experimental Results and training methodology are presented and discussed in section VI. Lastly, 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

Usually, VO taxonomy divides geometric based Visual Odometry methods into feature-based or direct/dnese methods for VO estimation. Both approaches have made great success in the past decade [6]. However, they still face many challenging issues, in particular when being deployed in large scale robotic applications and facing complex environment application scenarios.

In recent years, deep learning methods have risen to predominance by showing good capability for cognitive and perceptual tasks in computer vision applications, whether at analyzing unknown features, capturing image depth or even perceiving egomotion between image frames. Thus, the development of learning-based applications aimed at improving visual-based robotic navigation has had a significant surge as of late.

We start our related work with a brief review of Visual Odometry implementations running on top of deep learning architectures by analyzing key contributions and nuances between different deep VO estimation methods, further dividing the state-of-the-art into depth, global pose and egomotion estimation.

A. Depth Estimation

Depth estimation methods take advantage of camera displacement or difference in the apparent position of an object viewed along two different lines of sight to estimate depth.

Early work by Eigen *et al* [7] proposed a supervised method for depth estimation with a ground-truth depth map and a scale-invariant error as a cost function for training. The work was further extended by further integrating convolution neural networks improving accuracy and efficiency on both segmentation tasks and depth estimation. CNN-SLAM [8] is a proposed monocular SLAM system that relies on convolutional neural networks solely to estimate depth, recovering pose and graph optimization from conventional feature-based SLAM. This approach demonstrated that deep learning architectures can also work hand-in-hand with vision-based systems, improving upon overall robustness and accuracy of said algorithms.

Unsupervised schemes have recently emerged, also posing as viable alternatives. Garg's idea [9] was to use CNN's to predict the depth map for the left input image, reconstructing the left image from the right image and using the photometric reconstruction error (eq. 1) between the original left image *I* and the new synthetized left image *I'* in the training phase of the algorithm.

$$E = \sum \|I - I'\|^2 \tag{1}$$

SfMLearner [10] 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++ [11] improved upon the results in both depth and pose estimation by using the Essential matrix, obtained using Nisters Five Point Algorithm [12], to enforce epipolar constraints on the loss function, effectively discounting ambiguous pixels.

GeoNet [13] 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*.[14] presented SfM-Net, innovating through adding motion masks to photometric losses to jointly estimate optimal flow, depth maps and egomotion.

B. Global Pose Estimation

Localization is a crucial component for autonomous systems development, since it enables a robot to determine where it is on an environment, which serves as a precursor to any type of action execution or planning.

The main purpose of data-driven pose estimation is to estimate pose without explicitly modeling the camera motion. PoseNet [15] was the first instance of CNN usage for pose estimation, starting from a supervised scheme with a 6-DoF pose groundtruth. Making use of geometry to design meaningful constraints to the loss function [16] proved to yield significant improvements to method performance and accuracy. This method showed very robust performances in relocalization tasks and was further extended to support both color and depth inputs, improving upon its accuracy in challenging environments, such as night-time.

The application of deep RCNN's architectures to Visual Odometry task have been gaining favor in the past years, as them allow for bypassing the need for almost all blocks in the conventional VO pipeline, allowing for end-to-end pose inference. The Recurrent Neural Network component can be used for exploring either temporal dynamics or spatial analysis of image sequences, thereby reducing the uncertainty of pose estimation and generally improving upon method performance. The introduction of LSTM units to neural network design as showcased in [17] proved to improve results in localization tasks making use of structured correlation in feature space using LSTM units.

TABLE I
RELATED WORK IN DEEP LEARNING FOR COMPUTER VISION APPLICATIONS

Algorithm	Year	Supervised	Depth	Global Pose	Egomotion
Eigen et al.[7]	2014	✓	✓		
PoseNet[15]	2015	✓		✓	
DeepVO [18]	2017	✓			✓
SfMLearner [10]	2017	no	✓		✓
PoseLSTM [17]	2017	✓		✓	
UnDeepVO [19]	2017	no			✓
CNN-SLAM [8]	2017	✓	✓		
VINET [20]	2017	no			✓
VLocNet [21]	2018	no		✓	✓
GeoNet [13]	2018	no	✓		✓

C. Egomotion Estimation

Building upon the success of absolute pose estimation, 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 rather than just determining the position and attitude of the observer. FlowNet [22] 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 [18], 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 [19], 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.

Valada *et al* [21] 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.

Visual Odometry methods are particularly sensitive to rotation errors, as small early drifts can have a large influence on final trajectory pose estimates. Peretroukhin [23] 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 [20] is a proposed framework that fuses pose estimates from DeepVO [18] 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 [24] [25]. 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).

In table I, a brief comparison of state-of-the-art methods is presented, detailing the tasks they perform.

III. UNDERWATER VISUAL DATASET

Deep learning methods 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 is crucial for further development of deep learning algorithms and its respective generalization ability, therefore improving upon its robustness when being deployed in full-scale large complex environments.

In the underwater context, there are not many publicly available large datasets and there is none widely regarded as a comprehensive benchmark for method evaluation. In the scope of this work, we also wanted to assess method performance using one of CRAS robotic solutions, namely the UNEXMIN UX-1 robot. With this in mind, we developed a deep visual underwater dataset, an 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 deep learning architectures for VO estimation in different underwater scenarios. In Fig. 2, we can observe example images of our dataset sequences, that showcase the different environments included in our dataset.

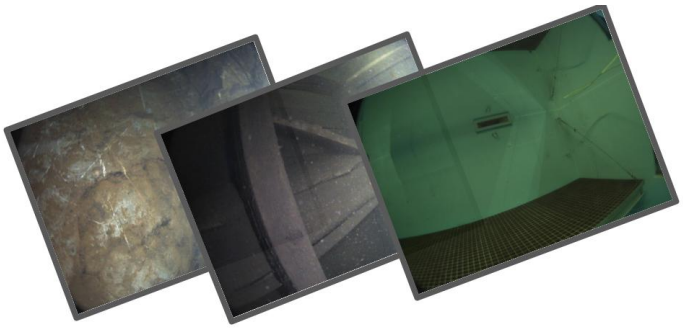

Fig. 2. Dataset image examples

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 is equipped with a plethora of different sensors, including 5 cameras. In this work, and especially since the UX-1 does not have a great overlay of camera fields-of-view, we are focusing on monocular visual methods, and as so, we choose 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 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.

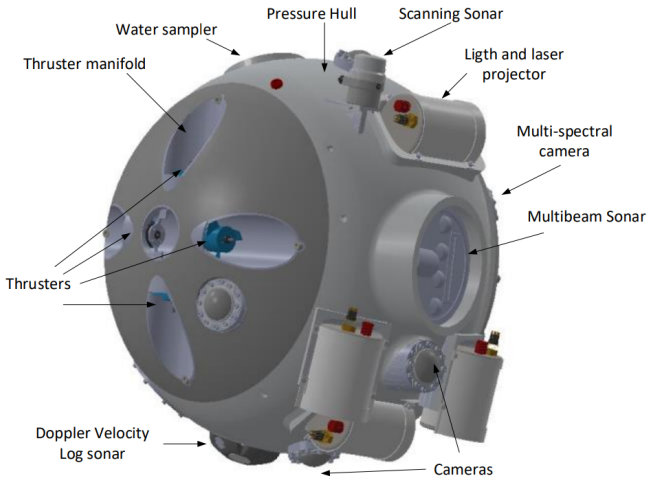

Fig. 3. UNEXMIN UX-1 robot description

B. Dataset 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 ilummination conditions, which complicates visual-based navigation.

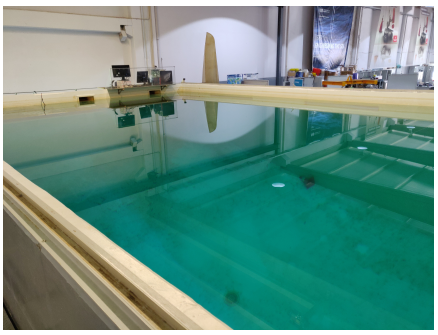

Fig. 4. CRAS indoor pool

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.

Fig. 5. Urgeirica mine entrance

IV. Deep Learning Approaches for Visual-based Robot Navigation

A. Egomotion Estimation

In the scope of underwater robotics research, and specifically in the context of our work, the most interesting application we are interested in exploring are unsupervised deep learning frameworks for egomotion estimation.

For the purpose of estimating motion dynamics, we are turning our attention to two similar state-of-the-art deep convolutional visual frameworks: SfmLearner [10] and GeoNet [13]. Though both frameworks also estimate monocular depth (and optical flow in the case of GeoNet), we are only focusing on camera motion estimation CNN's.

SfMLearner[10] 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

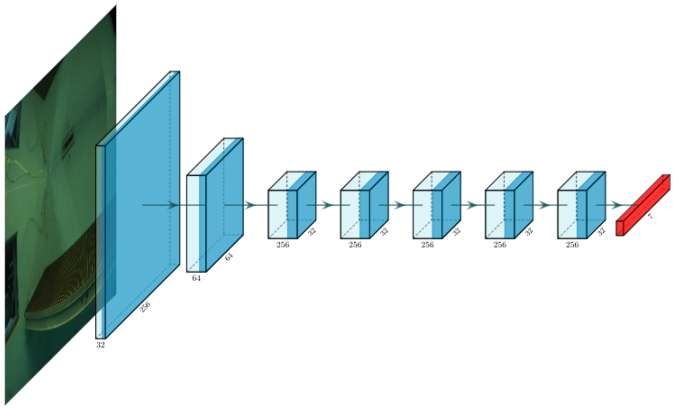

Fig. 6. Representation of the SfMlearner PoseNet, the framework component responsible for regressing 6-DoF pose estimates. It consists of 7 blocks of convolutional layers followed by ReLU activations, outputting a 6-dimensional vector that comprises a 3D translation vector and euler angles orientation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

by predicting relative transformation from the center image of the sequence (the image at the central position of the snippet, as shown in Fig. 7) 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.

Fig. 7. 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 downweight the loss on image patches undergoing motion external to the cameras motion (e.g. a car or pedestrian moving in the frame).

GeoNet[13] 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. 8). The prediction layers outputs the 6-DoF camera poses, i.e. translational vectors and orientation Euler angles.

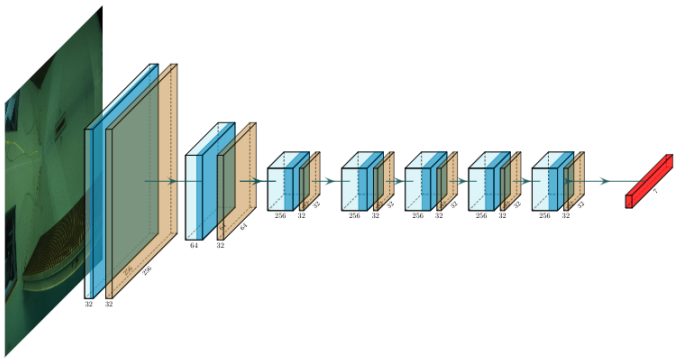

Fig. 8. Representation of the GeoNet PoseNet, the framework component responsible for regressing 6-DoF pose estimates. It consists of 7 blocks of convolutional layers followed by ReLU activations and additional batch normalization layers, outputting a 6-dimensional vector that comprises a 3D translation vector and euler angles orientation representation.

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 scale consistency with respect to the groundtruth data and therefore it would make sense to investigate analogous deep learning approaches to perform this task.

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 optmization.

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.

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.

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

where $E_{x...z}$ 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.

VI. Experimental results

A. Training Procedure and Hyperparameter Details

In this section, we focus on the experimental results for 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 data generated by the previously studied egomotion estimation frameworks. SfMLearner and Geonet share the data preprocessing step whereby the input image sequence is split into 5 sequence length snippets (see Fig. 7). In conjunction with camera intrinsic calibration and image timestamps, the 416x128 snippets were fed to the frameworks and the neural networks were trained using tensorflow[26] 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.

B. Results

Results on our dataset are presented in two different forms. First, we evaluate in a similar fashion to how both authors presented them, by computing estimate errors within the previously mentioned 5-frame sequences, with scale correction optimization and alignment with groundtruth data, so as to resolve scale ambiguity and minimize the impact of early drift accumulation errors.

TABLE II
Absolute Trajectory Error (ATE) evaluation

	KITTI seq 09	KITTI seq 10	CRAS Pool	Urgeirica Mine
SfMLearner	0.016 ± 0.009	0.013 ± 0.009	0.016 ± 0.006	0.028 ± 0.086
GeoNet	0.012 ± 0.007	0.012 ± 0.009	0.012 ± 0.006	0.026 ± 0.081

In the remaining of this section, we will present and discuss the results considering the full concatenated trajectory, thereby escaping the snippet representation and recomputing errors with respect to translation for all sequence trajectories under analysis. For the sake of coherent representation we will present the trajectories after the application of a post-processing step denoted as Umeyama alignment [27], commonly used in VO quantitative trajectory error metrics. It consists of a least-squares estimation of transformation parameters translation, rotation and scale between estimates and groundtruth pose data.

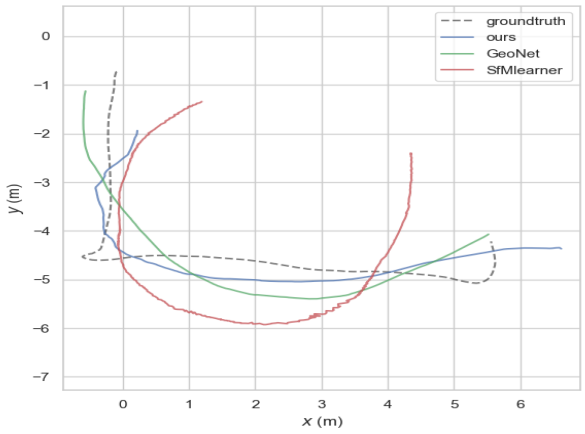

Fig. 9. Results for the CRAS pool sequence: trajectory estimates against groundtruth data

As it can be observed in table III, our Visual-Inertial Fusion Network was able to synthesize the best results for global trajectory estimation with or without any type of preprocessing step. It performs on average around **40%** better for the CRAS pool sequence while showing an average improvement of around **55%** in the urgeiria mine sequence. It is important to note, however, that both SfMlearner and GeoNet are unsupervised frameworks, and the devised solution leverages a supervised learning scheme.

TABLE III
RESULT COMPILATION FOR ABSOLUTE POSE ERROR W.R.T. TRANSLATION

		Absolute Pose Error (APE)					
		"raw" comparison		scale-corrected		SIM(3) Umeyama aligment	
		Avg.Error	RMSE (m)	Avg.Error	RMSE (m)	Avg.Error	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**

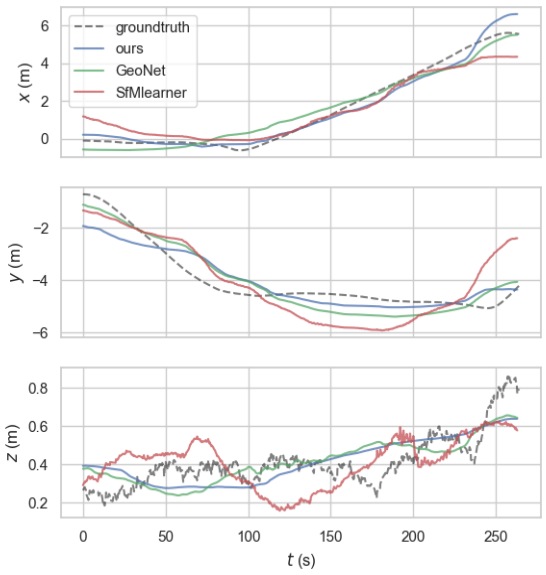

Fig. 10. Results for the CRAS pool sequence: trajectory estimates against groundtruth data decoupled by translational component

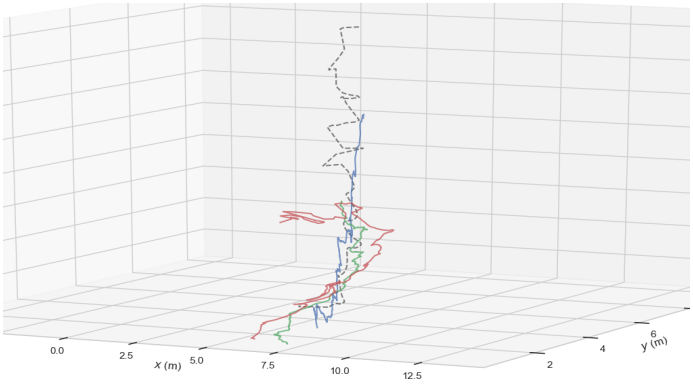

Fig. 11. Results for Urgeirica Mine sequence: computed trajectory estimates against groundtruth data

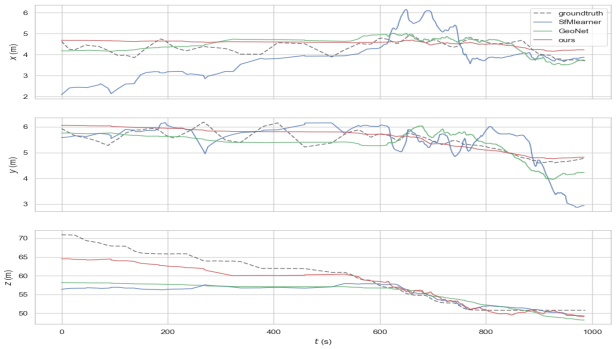

Fig. 12. Results for Urgeirica Mine sequence: trajectory estimates against groundtruth data decoupled by translational component

C. Discussion

The performance of state-of-the-art deep learning methods for egomotion estimation can be analyzed through different perspectives, leading to the following conclusions:

- First of all, as it can be observed in table II, we were able to produce similar results to those presented in the literature only for our CRAS pool sequence. It is still a good indication that it was possible to achieve such results in the underwater context, however, it is important to note that it was only true for our fully known structured environment. Real mission operational scenarios like the urgeiria mine sequence pose greater challenges to visual-based motion estimation algorithms and that is reflected on higher magnitude error rates.
- Secondly, it is possible to observe that both networks performs fairly better at regressing translational displacement than rotational movement. Rotation, and in particular pure rotations, are not handled well in any of the studied methods.
- In accordance to the expectations, and in agreement with both authors result presentation, pose estimates only present persuasive results with a post-processing step. The need for scale correction is a consequence of the use of monocular camera setups, but some type of groundtruth alignment algorithm is also required.

- Though relative motion estimates seem at first glance to show potential due to small average error rates, their concatenation onto the full trajectory reveals that the drift accumulation results in poor trajectory shape mimicking. In conclusion, there is still room for improvement when it concerns to global pose estimation derived from unsupervised egomotion estimation frameworks.
- We introduced a visual-inertial fusion network, anchored on a recurrent neural network architecture with an inertial supervision learning scheme. It was shown to improve results an average of **50%** for trajectory estimates, also producing more visually consistent trajectory estimates for both our application scenarios. This approach can later be integrated with egomotion estimation frameworks in an end-to-end fashion, leading to more accurate and reliable robot trajectory estimates.

VII. Conclusion

A. Summary

In this paper, the focus was placed on deep learning approaches for visual-based robot navigation, with particular interest on evaluating the potential for learning-based visual method application on complex underwater operational mission scenarios.

Firstly, a review of state-of-the-art deep learning approaches for Visual Odometry applications was conducted, detailing the progress in performance and accuracy deep learning methods have managed to achieve in recent years, as well as its shortcomings. It was concluded that there was close to no information about the performance of deep learning methods for VO estimation in underwater context scenarios, and would therefore be particularly interesting and relevant to assess the performance of some of the most renown state-of-the-art algorithms in operational mission underwater scenarios.

The next step was to construct a comprehensive dataset encompassing different texture environments and providing different types of challenges to visual-based pose and/or motion estimation. As reported in III, this was achieved through the use of data acquired with the UX-1 robot, and presents three novel image sequences that all pose different challenges to visual-based VO estimation.

In order to access the performance of learning-based visual methods on our dataset image sequence, we focused on two different tasks: absolute relocalization and egomotion estimation. We came to the conclusion that relocalization algorithms have an overall good performance across different scenarios, but lack generalization ability when exposed to more than one different mapping during training. It is reasonable to assume that we could achieve good performance from the application of this methods in real robotic solutions, though real time testing was not performed and thus validation is still required.

As for egomotion estimation, the results were not as accurate and reliable as expected. Relative motion estimates of state-of-the-art algorithms show small errors in translation yet rotations still pose some challenges these methods are not able to overcome. Analyzing concatenated trajectories, we can easily observe that pure rotations and accumulated drifts lead to failures in pose estimation, thus making the algorithm unable to provide consistent and reliable estimates, as required by real robotic systems.

In section V, we again address the issue of the aforementioned poor performance of egomotion estimation methods, presenting a possible solutions for obtaining the global pose optimization objective. 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 **50%** 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

The following future work in this research scope is suggested:

- Integration of visual-inertial fusion within end-to-end deep learning for robot navigation pipelines. Further study of inertial integration without losing the unsupervised learning objective.
- 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.
- Real-time implementation and testing of deep learning architectures for both relocalization and egomotion tasks for the underwater context. The low budget recommended option would be using a Nvidia Jetson Nano and TensorRT for fast inference implementation.

Acknowledgments

We want to thank the UNEXMIN project for providing the data utilized in the scope of this work.

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. NIPS’12. USA: Curran Associates Inc., 2012, pp. 1097–1105. URL: http://dl.acm.org/citation.cfm?id=2999134.2999257.

[2] Joseph Redmon et al. “You Only Look Once: Unified, Real-Time Object Detection”. In: *CoRR* abs/1506.0 (2015). URL: http://arxiv.org/abs/1506.02640.

[3] Ross Girshick. “Fast r-cnn”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448.

[4] Shaoqing Ren et al. “Faster {R-CNN:} Towards Real-Time Object Detection with Region Proposal Networks”. In: *CoRR* abs/1506.0 (2015). URL: http://arxiv.org/abs/1506.01497.

[5] Sergio Domínguez et al. “UX 1 system design -A robotic system for underwater mining exploration”. In: 2018.

[6] Davide Scaramuzza and Friedrich Fraundorfer. “Tutorial: Visual odometry”. In: *IEEE Robotics and Automation Magazine* 18.4 (2011), pp. 80–92. ISSN: 10709932. DOI: 10.1109/MRA.2011.943233.

[7] David Eigen, Christian Puhrsch, and Rob Fergus. “Depth map prediction from a single image using a multi-scale deep network”. In: *Advances in neural information processing systems*. 2014, pp. 2366–2374.

[8] Keisuke Tateno et al. “CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2. 2017.

[9] Ravi Garg et al. “Unsupervised cnn for single view depth estimation: Geometry to the rescue”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 740–756.

[10] Tinghui Zhou et al. “Unsupervised learning of depth and ego-motion from video”. In: *CVPR*. Vol. 2. 6. 2017, p. 7.

[11] Vignesh Prasad and Brojeshwar Bhowmick. “SfMLearner++: Learning Monocular Depth & Ego-Motion using Meaningful Geometric Constraints”. In: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2019, pp. 2087–2096.

[12] David Nistér. “An efficient solution to the five-point relative pose problem”. In: *IEEE transactions on pattern analysis and machine intelligence* 26.6 (2004), pp. 756–770.

[13] Zhichao Yin and Jianping Shi. “Geonet: Unsupervised learning of dense depth, optical flow and camera pose”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 1983–1992.

[14] Sudheendra Vijayanarasimhan et al. “Sfm-net: Learning of structure and motion from video”. In: *arXiv preprint arXiv:1704.07804* (2017).

[15] Alex Kendall, Matthew Grimes, and Roberto Cipolla. “Posenet: A convolutional network for real-time 6-dof camera relocalization”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 2938–2946.

[16] Alex Kendall and Roberto Cipolla. “Geometric loss functions for camera pose regression with deep learning”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 5974–5983.

[17] Florian Walch et al. “Image-based localization using LSTMs for structured feature correlation”. In: *ICCV*. 2017. URL: https : / / github . com / NavVisResearch / NavVis-Indoor-Dataset.

[18] Sen Wang et al. “DeepVO : Towards End-to-End Visual Odometry with Deep Recurrent Convolutional Neural Networks”. In: (2017), pp. 2043–2050.

[19] Ruihao Li et al. “UnDeepVO: Monocular Visual Odometry through Unsupervised Deep Learning”. In: (2017). DOI: 10.1109/ICRA.2018.8461251. URL: http://arxiv.org/abs/1709.06841.

[20] Ronald Clark et al. “VINet: Visual-Inertial Odometry as a Sequence-to-Sequence Learning Problem”. In: (2017). ISSN: 1098-0121. DOI: 10.1109/ISMAR.2016.19.

[21] Abhinav Valada, Noha Radwan, and Wolfram Burgard. “Deep auxiliary learning for visual localization and odometry”. In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2018, pp. 6939–6946.

[22] Alexey Dosovitskiy et al. “Flownet: Learning optical flow with convolutional networks”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 2758–2766.

[23] Valentin Peretroukhin et al. “Probabilistic Regression of Rotations using Quaternion Averaging and a Deep Multi-Headed Network”. In: *arXiv preprint arXiv:1904.03182* (2019).

[24] Sudeep Pillai and John J Leonard. “Towards visual ego-motion learning in robots”. In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2017, pp. 5533–5540.

[25] Mehmet Turan et al. “A deep learning based fusion of RGB camera information and magnetic localization information for endoscopic capsule robots”. In: *International journal of intelligent robotics and applications* 1.4 (2017), pp. 442–450.

[26] Mart\'\in˜Abadi et al. {*TensorFlow*}*: Large-Scale Machine Learning on Heterogeneous Systems*. 2015. URL: https://www.tensorflow.org/.

[27] Shinji Umeyama. “Least-squares estimation of transformation parameters between two point patterns”. In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 4 (1991), pp. 376–380.