DEEP LEARNING TO SUPPORT 3D MAPPING CAPABILITIES OF A PORTABLE VSLAM-BASED SYSTEM

N. Padkan [1,2], R. Battisti [1], F. Menna [1], F. Remondino [1]

[1] 3D Optical Metrology (3DOM) unit, Bruno Kessler Foundation (FBK), Trento, Italy
Web: http://3dom.fbk.eu – Email: <npadkan><rbattisti><fmenna><remondino>@fbk.eu
[2] Dept. Mathematics, Computer Science and Physics, University of Udine, Italy

Commission II

KEY WORDS: vSLAM, Real-Time, Low-Cost, Mobile Mapping, Deep Learning, Image Segmentation, Moocular Depth Estimation

ABSTRACT:
The use of vision-based localization and mapping techniques, such as visual odometry and SLAM, has become increasingly prevalent in the field of Geomatics, particularly in mobile mapping systems. These methods provide real-time estimation of the 3D scene as well as sensor's position and orientation using images or LiDAR sensors mounted on a moving platform. While visual odometry primarily focuses on the camera's position, SLAM also creates a 3D reconstruction of the environment. Conventional (geometric) and learning-based approaches are used in visual SLAM, with deep learning networks being integrated to perform semantic segmentation, object detection and depth prediction. The goal of this work is to report ongoing developments to extend the GuPho stereo-vision SLAM-based system with deep learning networks for tasks such as crack detection, obstacle detection and depth estimation. Our findings show how a neural network can be coupled to SLAM sequences in order to support 3D mapping application with semantic information.

1. INTRODUCTION

Vision-based localization techniques, such as visual odometry (VO) and Simultaneous Localization And Mapping (SLAM), are getting more and more common in Geomatics and a key component in many mobile mapping systems, especially portable ones (Torresani et al., 2021a; Otero et al., 2020; Nocerino et al., 2019a; Blaser et al., 2018; Schöps et al., 2017; Nüchter et al., 2015). VO and SLAM provide real-time estimation of the position and orientation of the sensor moving in an environment based solely on a sequence of images or LiDAR profiles captured by one or more sensors rigidly mounted on a platform. They are often combined with other positioning systems such as GNSS and IMU to provide a seamless and more robust navigation and mapping solution. While VO primarily focuses on the camera's position, reconstructing sensor trajectories, SLAM also creates a 3D sparse, semi-dense or dense reconstruction of the environment (Yang et al. 2022; Taketomi et al., 2017; Scaramuzza and Fraundorfer, 2011).
SLAM-based 3D surveyinng is nowadays used in multiple applications and field: underwater mapping (Nocerino et al., 2018), rail tunnel inspection (Panella et al., 2020), exploration (Steenbeek and Nex, 2022), autonomous driving (Singandhupe and La, 2019), Augmented Reality (Torresani et al., 2021b), etc.
The aim of the work is to introduce the on-going developments to extend our stereo-vision, SLAM-based, lightweight and modular system, called GuPho (Menna et al., 2022; Torresani et al., 2021) with deep learning neural networks in order to perform:

- Semantic segmentation, e.g., for crack detection: the system is used in monitoring or inspect tasks and it identifies in real-time cracks in structures; leveraging on the stereo-vision, metric information can be retrieved;
- Object detection, such as rocks: when GuPho is used to automatically guide a moving robot, the detection of obstacle is a fundamental task for avoidance and re-routing;
- Monocular Depth Estimation (MDE): depth prediction is useful to improve scene understanding, support autonomous navigation and complement conventional MVS methods in textureless areas.

The paper is organized as follows: Section 2 briefly recall the low-cost, lightweight and portable modular prototype system, GuPho. Section 3 reports single, stereo or multi-sensor SLAM solutions for 3D mapping purposes. Deep learning solutions are mentioned in Section 4. Data preparation is discussed in Section 5 whereas experiments, evaluations and results are presented in Section 6. Finally Section 7 concludes the paper.

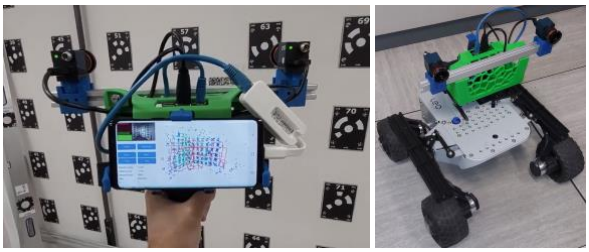

Figure 1: The GuPho stereo-vision system for real-time 3D mapping in its handheld (a) and robotic (b) version.

2. THE GUPHO SYSTEM

GuPho (Guided Photogrammetry system) is a low-cost, lightweight and portable modular prototype system based on stereo vision and vSLAM method (Menna et al., 2022; Torresani et al., 2021a; Di Stefano et al, 2021). GuPho is equipped with a Raspberry Pi 4 model B, with a roadcom BCM2711, Quad core Cortex-A72 (ARM v8) 64-bit SoC @ 1.8GHz and 8gb RAM. It was developed to provide real-time guidance to the surveyor during the image capturing phase, ensuring a more reliable and effective photogrammetric data acquisition and processing. GuPho can use rectilinear or fisheye lenses to survey indoor or outdoor scenarios, including underwater environments. Real-time 3D mapping capabilities are provided through OpenVSLAM (Sumikura et al., 2019) which builds upon ORB-SLAM2 (Mur-Artal and Tardós, 2017). Real-time computation and visualisation capabilities are used to introduce visual feedbacks to users, including camera-to-object distance warnings to guarantee the expected ground sample distance (GSD) or speed warnings to avoid motion blur. Also, it uses a novel automatic exposure algorithm that exploits 3D information of the observed scene. Figure 1 shows the realized GuPho system,

either in its handheld version or coupled to a ground robot (Leo Rover[1]) for autonomous navigation and 3D mapping.

3. SLAM SOLUTIONS

The literature is populated by single, stereo or multi-camera photogrammetric systems designed for portable mobile mapping applications and SLAM processing (Perfetti and Fassi, 2022; Torresani et al., 2021; Ortiz-Coder and Sánchez-Ríos, 2020; Meyer et al., 2020; Menna et al., 2019; Nocerino et al., 2019b; Koehl et al., 2016; Teo et al., 2015; Shortis et al., 2007). For mobile mapping applications, real-time processing is mandatory and performed with SLAM approaches (Lai, 2022). In particular, vSLAM can be divided into two categories: traditional and learning-based vSLAM (Chen et al., 2022). Traditional vSLAM uses geometric features, such as points and lines extracted from the images, or the pixel intensity values to understand and map the environment. Learning-based vSLAM methods rely on deep learning-based feature descriptors (Bruno and Colombini, 2021), hybrid methods (Tang et al., 2019) or complete end-to-end approaches (Wang et al., 2017). Convolutional Neural Networks (CNN) have been also integrated into SLAM pipeline (Tateno et al., 2017): the estimation of camera pose is performed by minimizing photometric error whereas learning is used to compute depth information. Steenbeek and Nex (2022) proposed a similar concept applied to UAV video sequences.
Novel approaches are also integrating Neural Radiance Fields (NeRF - Mildenhall et al. 2021) into SLAM pipeline in order to offer novel geometric and photometric 3D mapping solutions for accurate and real-time scene reconstruction from monocular images (Rosinol et al., 2022). Sucar et al. (2021) introduced iMAP, the first real-time NeRF-based dense online SLAM model that optimizes camera pose and the implicit scene representation in a hand-held RGB-D camera system. The iMAP system employs an iterative two-step approach of tracking and mapping and utilizes keyframe selection. Zhu et al. (2022) introduced NICE-SLAM, a dense RGB-D SLAM system that uses a hierarchical scene representation incorporating information at multiple levels and pre-trained geometric priors, resulting in detailed reconstructions of large indoor scenes that are more scalable, efficient, and robust than other recent SLAM systems using neural networks. The successive NICER-SLAM (Zhu et al., 2023) is a dense RGB SLAM system that optimizes for camera poses and a hierarchical neural implicit map representation, which allows for high-quality novel view synthesis. The system incorporates additional supervision signals, including monocular geometric cues and optical flow, and a simple warping loss to enforce geometry consistency.
SLAM algorithms have been also coupled to neural networks to enhance recognition capability in images or classification algorithms in 3D space (Pillai and Leonard, 2015; Zhag et al., 2018; Duan et al., 2019).

4. DEEP LEARNING SOLUTIONS

In recent years, machine learning techniques have been applied to images or point clouds with promising results. Convolutional neural networks (CNNs) and other deep learning models can provide high accuracy, recall and prediction speed, allowing for real-time application in SLAM-based applications.
Our developments focused on coupling deep learning methods to image sequences acquired by the GuPho system for semantic segmentation and objected detection as well as monocular depth estimation (Section 4.3).

- Deep Learning for semantic segmentation and object detection: we rely on Yolov8 (Ultralytics, 2023), designed to detect and localize objects within images or video frames. It can be re-trained to detect a wide range of (new) objects, ensuring real-time performances at 30 fps or higher on medium GPU. Yolov8 is based on multiple layers of convolutional and pooling operations, followed by several fully connected layers. The network takes an input image and processes it through the layers, gradually learning to recognize and locate objects within the image. We have retrained and generalized the method to our scenarios.
- Deep Learning for MDE: we build upon MiDaS (Ranftl et al., 2022) which demonstrated to clearly outperform competing methods across diverse datasets. It includes a flexible loss function and a robust training objective invariant to changes in depth range and scale, advocating the use of principled multi-objective learning to combine data from different sources.

4.1. Instance segmentation and object detection

The primary objective of object detection is to identify the (precise) location of various objects present in a given scene and assign relevant labels to the bounding boxes of these objects. On the other hand, instance segmentation is a technique that identifies and labels individual objects in an image and their components at pixel level. This allows for a more precise understanding of objects and their relationships. The state-of-the-art neural network for object detection in images is YOLO. The YOLO (You Only Look Once) algorithm (Redmon et al., 2016) was a cutting-edge object detection method that could achieve both high precision and speed. YOLO differs from traditional classifiers as it examines the image just once and can identify objects within it. YOLO gained rapid popularity due to its high speed and accuracy in object detection and image segmentation. As a one-stage object detectors, YOLO directly predicts the bounding boxes and class probabilities of objects in a single pass through the network. These models are known for their speed and efficiency, making them well-suited for real-time applications. Different variants of YOLO (Redmon and Farhadi, 2017; Bochkovskiy at al., 2020; Wang et al., 2022; Ultralytics, 2023) were released throughout the years (Figure 2), with successive improvements in terms of speed, accuracy, efficiency and generalization. Tiny implementation of YOLO on single-board devices (e.g. Raspberry Pi, Jetson, etc.) were also proposed (Ayoub and Schneider-Kamp, 2021; Chan et al., 2022).
With respect to the latest version, YOLOv8 (Ultralytics, 2023), there are five models (YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x) for detection, segmentation and classification. YOLOv8n is the fastest and smallest, while YOLOv8x is the most accurate yet the slowest among them.

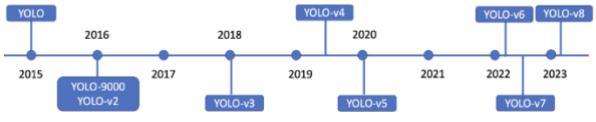

Figure 2: Timeline of You Only Look Once (YOLO) variants (Zhang, 2023).

Beside identification and tracking of people or animals (Kajabad and Ivanov, 2019; Tang et al., 2023), the YOLO network has been used to detect pavement or side-walk cracks (Yang et al.,

[1] https://www.leorover.tech/

2019; Liu et al., 2020; Qui and Lau, 2023; Wang et al, 2023), rocks in mining areas (Loncomilla et al., 2022) and concrete bridge defects (Zhang et al. 2019). It has been shown that YOLO is faster than other object detector methods that use a two-stage deep learning approach like Faster R-CNN (Yin et al., 2019).

4.2. Monocular depth estimation

Although humans find it easy to estimate the depth of a scene from a single image, it is a challenging task for computational models due to the ill-posed problem and high resource requirements. Monocular Depth Estimation (MDE) refers to the process of estimating depth from a single RGB image (Ming et al., 2021). Being able to estimate depth from a single image has several benefits, such as aiding in scene comprehension, 3D modelling, robotics, autonomous driving, etc. The recovery of depth information is particularly important in these applications when other information such as stereo images, optical flow or point clouds are not available. Real-time depth estimation has traditionally been performed using stereo images or video sequences, as evidenced by the research in (Ha et al., 2016; Kong and Black, 2015; Cheng and Huang, 2015; Karsch. et al., 2015). However, these methods are resource-intensive and require more data compared to monocular depth estimation. Therefore, MDE has become increasingly popular, leading to the development of several deep learning methods. These methods do not rely on hand-crafted features and utilize deep convolutional neural networks. Among different tested networks, we have chosen Zero-shot Transfer by Combining Relative and Metric Depth (ZoeDepth) framework (Bhat et al., 2023): it combines both monocular depth estimation (MDE) and relative depth estimation (RDE) approaches in a two-stage framework (Figure 3). In the first stage, an encoder-decoder structure is trained to estimate relative depths from the input image. This model is trained on a large variety of datasets, which improves its generalization to different scenes and environments. It builds upon the MiDaS (Ranftl et al., 2020) training strategy for relative depth prediction which uses a loss that is invariant to scale and shift. In the second stage, components responsible for estimating metric depth are added as an additional head. This stage helps to refine the depth estimates by incorporating metric depth information, which is the absolute distance between objects in the scene.

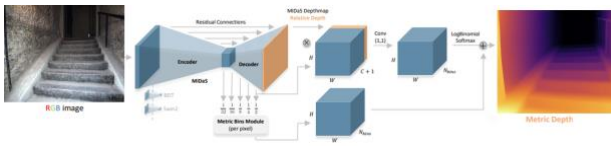

Figure 3: The ZoeDepth architecture. An RGB image is fed into the MiDaS depth estimation framework to predict a depth (after Bhat et al., 2023).

5. DATA PREPARATION FOR OBJECT DETECTION

The datasets utilized to evaluate neural networks were acquired via GuPho using rectilinear or fisheye lenses. The image sequences have a resolution of 1280 x1024 pixels and feature cracks in asphalt or cement surfaces or sidewalks or building walls, off-road paths with rocks, tunnels with fall obstacles, etc.
Given our objects of interest, a manual process of image annotation to improve detection performances was necessary. Stones were annotated using bounding boxes (Figure 4a-b) whereas cracks were annotated using polygons (Figure 4c-d). This latter type of annotation is useful for detecting irregularly shaped objects and provides more precise information about the object's shape. In order to boost the model's performance, some augmentation techniques were applied, such as rotation (90 deg clockwise and counter-clockwise), brightness (+/-15%), blurring (up to 2.5px), shearing (±5 deg horizontal, ±5 deg vertical), cropping (0% minimum zoom, 25% maximum zoom).

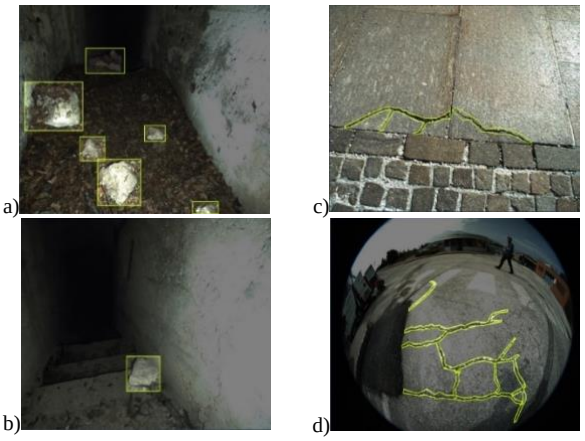

Figure 4: Annotated images: stones in a tunnel (a-b), cracks in tiles (c) or on asphalt (captured by fisheye lens).

6. EXPERIMENTAL EVALUATION AND RESULTS

YOLOv8 was chose due to its accuracy and speed in comparison with other versions. For computational limitations, the learning-based functionalities are applied to monocular images of GuPho. The extracted semantic information, coupled to the stereo-vision capabilities of GuPho, allows to retrieve metric information and deliver added-value 3D mapping results.
For these initial tests, the processing and analyses were performed "offline", using an 12th Gen Intel® Core™ i9-12950HX 2.30 GHz with 32 GB RAM and NVIDIA RTX A3000 12GB GPU.
To evaluate detection results, metrics like Recall R and mean Average Precision are used:

$$Recall = \frac{tp}{tp+ fp},$$

$$Precision = \frac{tp}{tp+fp},$$

where tp is the true positive, fn is the false negative and fp is the false positive;

$$AP = \sum_n (recall_n - recall_n) precision_n.$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i,$$

where N is the number of classes and i is the corresponding class. Some of the prediction/detection results on test images are shown in Figure 5. Each predicted stone is recognised by a bounding box and a confidence score which shows how likely the box contains an object of interest and how confident the classifier is about it. Predicted cracks are shown with a bounding box, a polygon mask and also confidence score. The confidence threshold was set to 0.25, i.e., the minimum score for which the model considers the prediction to be a true prediction.
For stone detection, various iterations of the YOLOv8 model were tested (Table 1), leading to the conclusion that YOLOv8s performed optimally for detecting the stones. Specifically, the highest level of detection accuracy was achieved after conducting 150 epochs, with a Recall of 0.70 and mAP of 0.64. Besides, the processing and inference time are important and Yolov8s

performed better (0.3 ms for processing and 8 ms for inference time). Results indicated that there was no significant improvement in detection accuracy beyond 150 epochs.

Model	Epochs	Processing Time (ms)	Inference Time(ms)	R	mAP
YOLOV8S	70	4.3	3.6	0.65	0.64
YOLOV8S	100	4.2	18.2	0.69	0.65
YOLOV8S	150	0.3	8.0	0.70	0.64
YOLOV8m	100	4.6	18.4	0.69	0.65

Table 1: Object detection evaluation of YOLOv8.

For crack detection, YOLOv8x model was found to be the fastest (0.2 ms for processing) and the most accurate one (R of 0.73) after 100 epochs (Table 2). Figure 5 reports some detection results on GuPho frames extracted from sequences in the field.

Model	Epochs	Processing Time (ms)	Inference Time (ms)	R	mAP
YOLOV8m	50	0.2	13	0.66	0.61
YOLOV8m	100	3.2	35.7	0.62	0.59
YOLOV8S	150	13.2	0.62	0.56	0.59
YOLOV8X	100	0.2	14.4	0.73	0.69
YOLOV8S	100	4.2	18.7	0.65	0.65
YOLOV8n	200	0.0	7.1	0.16	0.17

Table 2: Instance segmentation evaluation of YOLOv8.

For monocular depth estimation, a single camera sequence from GuPho was considered. We chose the two-stage framework that combined MDE and RDE, named ZoeDepth. As shown in Figure 6, the learning-based approach have a good performance on our dataset (for rectilinear or fisheye lenses) and could be coupled to conventional photogrammetric approaches for depth estimation.

7. CONCLUSIONS

The paper introduced an extension of stereo-vision, SLAM-based, lightweight and modular GuPho system with deep learning neural networks in order to perform semantic segmentation, object detection and depth estimation. We focused on rock detection (to aid autonomous navigation and obstacle avoidance in robotics applications), crack detection (to support structural monitoring and inspection) and depth prediction (to complement conventional stereo-vision methods in areas with non-collaborative surfaces). Our findings show how a neural network can be couple to SLAM sequences in order to support 3D mapping application with semantic information.

In order to achieve on-board real-time processing of both SLAM and deep learning tasks, we plan to extend GuPho with a more powerful board (e.g., NVIDIA Jetson Nano, equipped with an NVIDIA GPU with 128 CUDA cores) to allow computationally intensive tasks on the GPU.

The final aim in the long run is to transform GuPho into an intelligent system that can automatically and swiftly identify objects and obstacles in real-time. GuPho can be operated manually to identify damages on man-made structures or can navigate a robotic platform in challenging environments, like forests or tunnels. Incorporating deep learning methods, GuPho will obtain a profound and intelligent understanding of its surroundings for application and deployment in various fields.

Figure 5: Results of rocks (above) and cracks (below) detection in some images of a GuPho sequence (b).

ACKNOWLEDGEMENTS

This study was carried out within the Interconnected Nord-Est Innovation Ecosystem (iNEST) and received funding from the European Union Next-GenerationEU (Piano Nazionale Di Ripresa e Resilienza (PNRR) – Missione 4, Componente 2, Investimento 1.5 – D.D. 1058 23/06/2022, ECS00000043). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

REFERENCES

Ayoub, N., Schneider-Kamp, P., 2021. Real-Time On-Board Deep Learning Fault Detection for Autonomous UAV Inspections. Electronics, Vol. 10(9):1091.

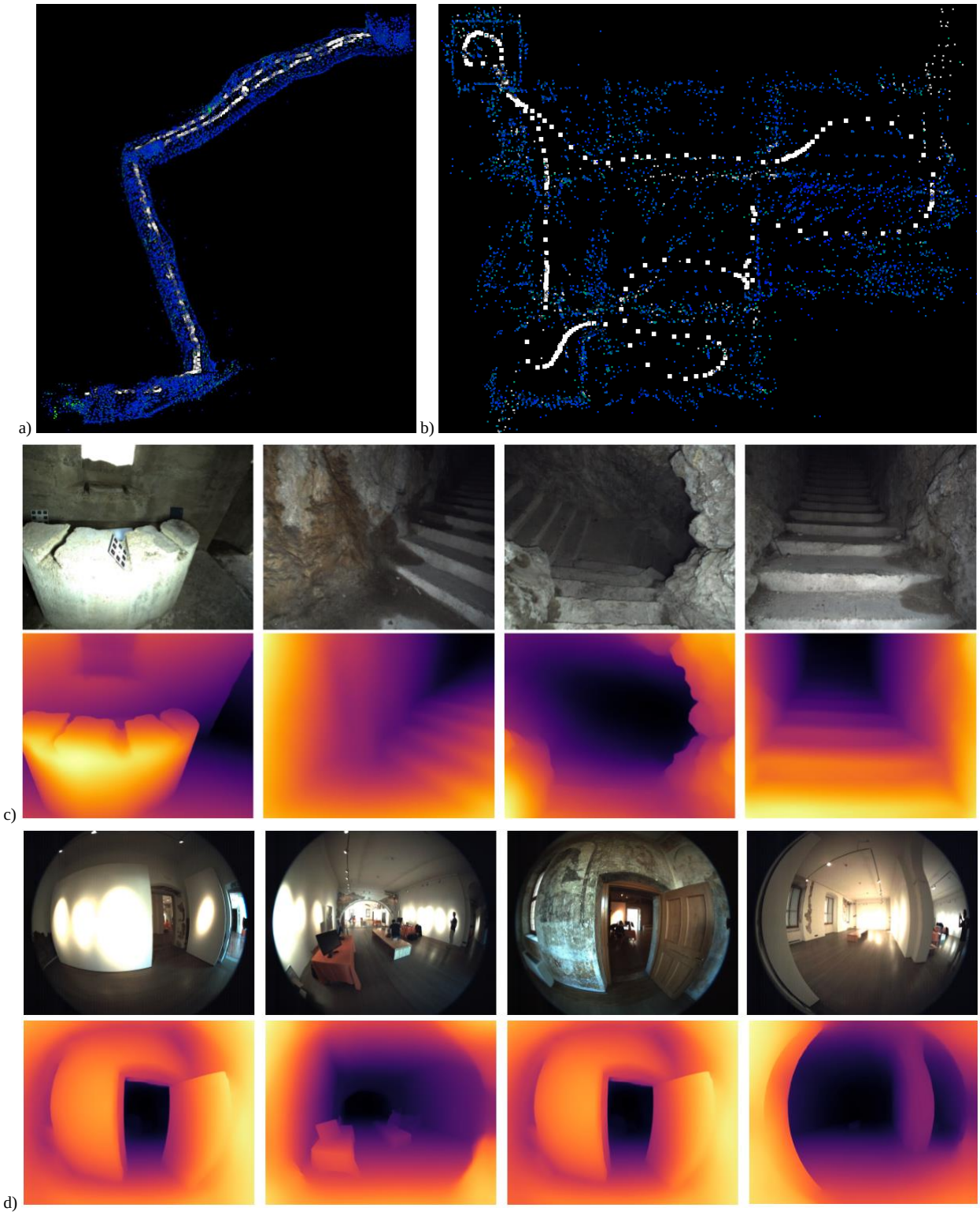

Figure 6: Results of two GuPho's SLAM processes in a WWI tunnel (a) and indoor building (b). Monocular depth estimations for some images of the sequences: tunnel (c) and indoor building (d).

Bhat, S.F., Birkl, R., Wofk, D., Wonka, P. and Müller, M., 2023. Zoedepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288.

Bhat, S.F., Birkl, R., Wofk, D., Wonka, P. and Müller, M., 2023. Zoedepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288.

Blaser, S., Cavegn, S., Nebiker, S., 2018. Development of a portable high performance mobile mapping system using the

robot operating system. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. IV-1, pp. 13-20.

Bochkovskiy, A., Wang, C.Y. and Liao, H.Y.M., 2020. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.

Bruno, H.M.S. and Colombini, E.L., 2021. LIFT-SLAM: a deep-learning feature-based monocular visual SLAM method. arXiv:2104.00099v2

Cha, Y.J., Choi, W. and Büyüköztürk, O., 2017. Deep learning-based crack damage detection using convolutional neural networks. Computer-Aided Civil and Infrastructure Engineering, 32(5), pp. 361-378.

Chan, C.J., Reyes, E.J., Linsangan, N. Juanatas, R., 2022. Real-time Detection of Aquarium Fish Species Using YOLOv4-tiny on Raspberry Pi 4. IEEE Proc. IICAIET.

Chen, W., Shang, G., Ji, A., Zhou, C., Wang, X., Xu, C., Li, Z., Hu, K., 2022. An overview on visual SLAM: From tradition to semantic. Remote Sensing, 14(13): 3010.

Chen, T.Y.H., Ravindranath, L., Deng, S., Bahl, P. and Balakrishnan, H., 2015. Glimpse: Continuous, real-time object recognition on mobile devices. Proc. 13th ACM Conference on Embedded Networked Sensor Systems, pp. 155-168.

Cheng, F.H., Huang, K.Y., 2015. Real-time stereo matching for depth estimation using GPU. Proc. IEEE UMEDIA.

Di Stefano, F., Torresani, A., Farella, E.M., Pierdicca, R., Menna, F., Remondino, F., 2021. 3D Surveying of Underground Built Heritage: Opportunities and Challenges of Mobile Technologies. Sustainability, Vol.13, 13289.

Duan, C., Junginger, S., Huang, J., Jin, K., Thurow, K., 2019. Deep Learning for Visual SLAM in Transportation Robotics: A review. Transp. Saf. Environ., 1, 177–184.

Ha, H., Im, S., Park, J., Jeon, H.-G, Kweon, I.S., 2016. High Quality Depth from Uncalibrated Small Motion Clip. Proc. IEEE CVPR.

Jocher, G., Chaurasia, A., Qiu, J., 2023. YOLO by Ultralytics." https://github.com/ultralytics/ ultralytics, 2023. Accessed: April 30, 2023.

Kajabad, E.N., Ivanov, S., 2019. People Detection and Finding Attractive Areas by the use of Movement Detection Analysis and Deep Learning Approach. Procedia Computer Vision, Vol. 156, pp. 327-337.

Karsch, K., Liu, C., Bing Kang, S., 2012. Depth transfer: Depth extraction from videos using non-parametric sampling. Proc. ECCV.

Koehl, M., Delacourt, T., Boutry, C., 2016. Image capture with synchronized multiple-cameras for extraction of accurate geometries. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., Vol. XLIB1, pp. 653-660.

Kong, N. and Black, M.J., 2015. Intrinsic depth: Improving depth transfer with intrinsic images. Proc. ICCV.

Lai, T., 2022. A Review on Visual-SLAM: Advancements from Geometric Modelling to Learning-based Semantic Scene Understanding. arXiv:2209.05222.

Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., Ke, Z., Li, Q., Cheng, M., Nie, W., et al., 2022. YOLOv6: A single-stage object detection framework for industrial applications. arXiv preprint arXiv:2209.02976.

Liu, J., Yang, X., Lau, S., Wang, X., Luo, S., Lee, V.C.-S., Ding, L., 2020. Automated pavement crack detection and segmentation based on two-step convolutional neural network. Comput. Aided Civil Infrastruct. Eng. 2020, 35, pp. 1291-1305.

Loncomilla, P., Samtani, P., Ruiz-del-solar, J., 2022. Detecting rocks in challenging mining environments using convolutional neural networks and ellipses as an alternative to bounding boxes. Expert Systems with Applications, Vol.194, 116537.

Menna, F., Nocerino, E., Nawaf, M.M., Seinturier, J., Torresani, A., Drap, P., Remondino, F. and Chemisky, B., 2019. Towards real-time underwater photogrammetry for subsea metrology applications. Proc. IEEE OCEANS, pp. 1-10.

Menna, F., Torresani, A., Battisti, R., Nocerino, E., Remondino, F., 2022. A modular and low-cost portable VSLAM system for real-time 3D mapping: from indoor and outdoor spaces to underwater environments. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLVIII-2/W1-2022, pp. 153-162

Meyer, D. E., Lo, E., Klingspon, J., Netchaev, A., Ellison, C., Kuester, F., 2020. TunnelCAM- a HDR spherical camera array for structural integrity assessments of dam interiors. Electronic Imaging, Vol. 2020(7).

Mildenhall, B., Srinivasan, P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R., 2021. Nerf: Representing scenes as neural radiance fields for view synthesis." Communications of the ACM 65, no. 1, pp. 99-106.

Ming, Y., Meng, X., Fan, C., Yu, H., 2021. Deep learning for monocular depth estimation: A review. Neurocomputing, Vol. 438: 14-33.

Mur-Artal, R. and Tardós, J.D., 2017. ORB-SLAM2: An open-source slam system for monocular, stereo, and RGB-D cameras. IEEE Transactions on Robotics, 33(5), pp. 1255-1262.

Nocerino, E., Nawaf, M. M., Saccone, M., Ellefi, M. B., Pasquet, J., Royer, J.-P., Drap, P., 2018. Multi-camera system calibration of a low-cost remotely operated vehicle for underwater cave exploration. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., Vol. XLII-1, pp. 329-337.

Nocerino, E., Rodríguez-Gonzálvez, P. and Menna, F., 2019a. Introduction to mobile mapping with portable systems. In Laser Scanning (pp. 37-52). CRC Press.

Nocerino, E., Menna, F., Farella, E., Remondino, F., 2019b. 3D Virtualization of an underground semi-submerged cave system. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., Vol. XLII-2/W15, pp. 857-864.

Nüchter, A., Borrmann, D., Koch, P., Kühn, M., May, S., 2015. A man-portable, IMU-free mobile mapping system. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. II-3/W5, pp. 17-23.

Ortiz-Coder, P. and Sánchez-Ríos, A., 2020. An integrated solution for 3D heritage modeling based on videogrammetry and V-SLAM technology. Remote Sensing, 12(9), p.1529.

Otero, R., Lagüela, S., Garrido, I. and Arias, P., 2020. Mobile indoor mapping technologies: A review. Automation in Construction, 120, p.103399.

Panella, F., Roecklinger, N., Vojnovic, L., Loo, Y., Boehm, J., 2020. Cost-benefit analysis of rail tunnel inspection for photogrammetry and laser scanning. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., Vol. XLIII-B2-2020.

Perfetti, L. and Fassi, F., 2022. Handheld fisheye multicamera system: surveying meandering architectonic spaces in open-loop mode – accuracy assessment,.Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLVI-2/W1-2022, 4pp. 35-442.

Pillai, S., Leonard, J., 2015. Monocular SLAM supported object recognition. arXiv 2015, arXiv:1506.01732.

Qin, Z., Li, Z., Zhang, Z., Bao, Y., Yu, G., Peng, Y.,Sun, J., 2019. ThunderNet: Towards real-time generic object detection on mobile devices." Proc. ICCV, pp. 6718-6727.

Qiu, Q. and Lau, D., 2023. Real-time detection of cracks in tiled sidewalks using YOLO-based method applied to unmanned aerial vehicle (UAV) images. Automation in Construction, Vol. 147: 104745.

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K. and Koltun, V., 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE TPAMI, 44(3), pp. 1623-1637.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. Proc. CVPR, pp. 779-788.

Redmon, J. and Farhadi, A., 2017. YOLO9000: better, faster, stronger. Proc. CVPR, pp. 7263-7271.

Rosinol, A., Leonard, J.J., Carlone, L., 2022. NeRF-SLAM: Real-Time Dense Monocular SLAM with Neural Radiance Fields. arXiv:2210.13641.

Scaramuzza, D. and Fraundorfer, F., 2011. Visual odometry [tutorial]. IEEE robotics & automation magazine, 18(4), pp.80-92.

Schöps, T., Sattler, T., Häne, C., Pollefeys, M., 2017. Large-scale outdoor 3d reconstruction on a mobile device. Computer Vision and Image Understanding, Vol. 157, pp. 151-166.

Singandhupe, A., La, H.M., 2019. A Review of SLAM Techniques and Security in Autonomous Driving. Proc IEEE IRC, pp. 602-607.

Steenbeek, A., Nex, F., 2022. CNN-Based Dense Monocular Visual SLAM for Real-Time UAV Exploration in Emergency Conditions. Drones, vol. 6(3):79.

Sucar, E., Liu, S., Ortiz, J., Davison, A.J., 2021. iMAP: Implicit mapping and positioning in real-time." Proc. CVPR, pp. 6229-6238.

Sumikura, S., Shibuya, M. and Sakurada, K., 2019. OpenVSLAM: A versatile visual SLAM framework. In Proc. 27th ACM International Conference on Multimedia, pp. 2292-2295.

Taketomi, T., Uchiyama, H. and Ikeda, S., 2017. Visual SLAM algorithms: A survey from 2010 to 2016. IPSJ Transactions on Computer Vision and Applications, 9(1), pp.1-11.

Tang, J., Ericson, L., Folkesson, J., Jensfelt, P., 2019. GCNv2: Efficient correspondence prediction for real-time SLAM. IEEE Robotics and Automation Letters.

Tang, F., Yang, F., Tian, X., 2023. Long-Distance Person Detection Based on YOLOv7. Electronics, Vol. 12, 1502.

Tateno, K., Tombari, F., Laina, I., Navab, N., 2017. CNN-SLAM: Real-Time Dense Monocular SLAM with Learned Depth Prediction. Proc. IEEE CVPR.

Teo, T., 2015. Video-based point cloud generation using multiple action cameras. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., Vol. XL-4/W5, pp. 55-60.

Torresani, A., Menna, F., Battisti, R., Remondino, F., 2021a. A V-SLAM guided and portable system for photogrammetric applications. Remote Sensing, Vol.13(12), 2351.

Torresani, A., Rigon, S., Farella, E.M., Menna, F., Remondino, F., 2021b. Unveiling large-scale historical contents with V-SLAM and markerless mobile AR solutions. ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLVI-M-1-2021, 761-768.

Ultralytics, 2023: YOLOv8 in PyTorch. https://github.com/ultralytics/ultralytics. Accessed: April 30, 2023.

Yang, F., Zhang, L., Yu, S., Prokhorov, D., Mei, X., Ling, H., 2019. Feature pyramid and hierarchical boosting network for pavement crack detection. IEEE Transactions on Intelligent Transportation Systems 21, no. 4: 1525-1535.

Yang, M., Sun, X., Jia, F., Rushworth, A., Dong, X., Zhang, S., Fang, Z., Yang, G. and Liu, B., 2022. Sensors and sensor fusion methodologies for indoor odometry: A review. Polymers, 14(10), p.2019.

Yin, X., Chen, Y., Bouferguene, A., Zaman, H., Kurach, L., 2019. A deep learning-based framework for an automated defect detection system for sewer pipes. Autom. Constr, 109, 102967.

Wang, S., Clark, R., Wen, H., Trigoni, N., 2017. DeepVO: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. Proc. IEEE ICRA, pp. 2043-2050.

Wang, R., Li, A., Ling, C., 2018. Pelee: A real-time object detection system on mobile devices. Advances in neural information processing systems, 31.

Wang, C.-Y., Bochkovskiy, A., Liao, H.Y.M., 2022. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv:2207.02696

Wang, S., Chen, X., Dong, Q., 2023. Detection of Asphalt Pavement Cracks Based on Vision Transformer Improved YOLO

V5. Journal of Transportation Engineering, Part B: Pavements 149, no. 2: 04023004.

Zhang, L., Wei, L., Shen, P., Wei, W., Zhu, G., Song, J., 2018. Semantic SLAM Based on Object Detection and Improved Octomap. IEEE Access, 6, 75545-75559.

Zhang, C., Chang, C.C., Jamshidi, M., 2019. Concrete bridge surface damage detection using a single-stage detector. Comput. Aided Civil Infrastruct. Eng, Vol. 35, 389-409.

Zhang, Y., 2023. Stall Number Detection of Cow Teats Key Frames. arXiv preprint arXiv:2303.10444.

Zhu, Z., Peng, S., Larsson, V., Xu, W., Bao, H., Cui, Z., Oswald, M.R., Pollefeys, M., 2022. Nice-SLAM: Neural implicit scalable encoding for SLAM. Proc. IEEE CVPR, pp. 12786-12796.