Learning-based Underwater Depth Maps Generation

Inès Larroche [1] Clémentin Boittiaux [1,2,3] Maxime Ferrera [1] Ricard Marxer [2]

[1]Ifremer, Centre de Méditerranée, France
[2]Université de Toulon, Aix Marseille Univ, CNRS, LIS, Toulon, France
[3]Université de Toulon, COSMER, Toulon, France

Motivation

- **Images** acquired with **Remotely Operated Vehicules** (ROV) provide insightful data. Yet, 2D images only are hardly sufficient to get a general idea of the **scene 3D structure**.
- **3D reconstruction** from 2D images provide very useful additional data.
- **Depths Maps** are essential in the 3D reconstruction process.
- **Deep-Learning** approaches could provide more accurate depth maps than those obtained through hand-crafted geometric or photometric methods.

Goal

- Investigate the **potential of Deep Learning** based methods for **Depth Map estimations**.
- **Handle the lack of ground-truth labels** due to the underwater context.
- Evaluate the **balance** between **real-time** and **accuracy** focused inference.

Approach

Road Map

- Apply **supervised learning** methods through the **generation of pseudo ground-truth** Depth Maps from Structure-from-Motion (SfM) models.
- Bring **prior 3D information from SfM** to help in the training process.
- Investigate **self-supervised learning** methods to **get rid of the ground-truth requirement**.

The supervised learning approach

Our neural network follows a U-net architecture. U-Nets are neural networks composed of an encoder and a decoder, as well as connections between those two parts.

- **Encoder:** In the encoder, convolutional blocks learn interesting features from the image and encode information it contains to a tensor of smaller dimension.
- **Decoder:** In the decoder, convolutional blocks decode information and generate a depth map from the encoder output.
- **Output and training:** The reconstructed depth maps are compared to ground truth depth maps obtained thanks to a ray-casting process.

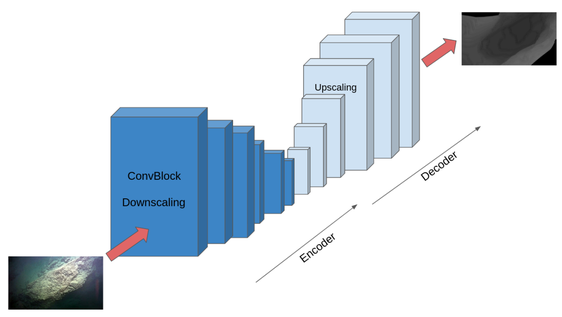


Figure 1. The supervised learning process

The different architectures used

- **MiDaS architecture [3]** U-Net structured. Pretrained on mixed datasets extracted from different environnements (outdoor and images, 3D movie frames...) (1.9 M images). This pre-built network is used to establish a proof of concept.
- **Basic U-Net architecture [2]** Encoder/ Decoder structure. A ResNet50 architecture is used for the encoder part. ResNets are neural networks composed of Residual Blocks, in which there are connections between input and outputs of different layers.

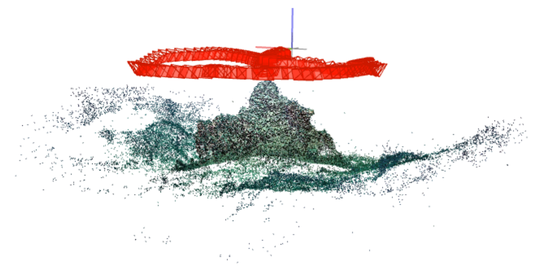


Figure 2. Sparse 3D Point Cloud Generated through Colmap, a general-purpose Structure-from-Motion (SfM) and Multi-View Stereo (MVS) pipeline

Qualitative Results

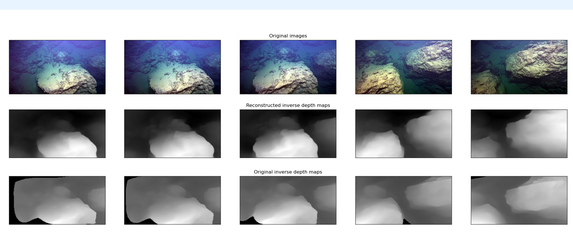


Figure 3. Reconstructed Inverse Depth Maps (MiDas network)

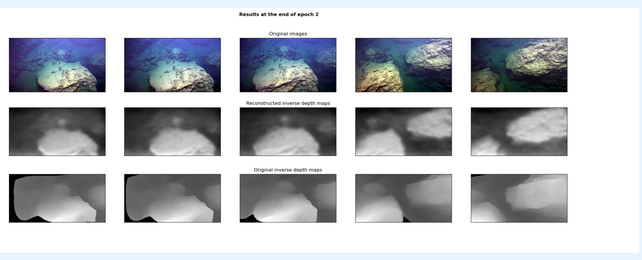


Figure 4. Reconstructed Inverse Depth Maps (ResNet50 AutoEncoder)

Guided image completion

Sparse Depth Maps: During SfM based 3D reconstruction, sparse depth maps are generated. Those are reliable and can be used as inputs to our network to increase its performances.

Guided Convolutional Network v1: The same auto-encoder architecture is used, now taking in input a RGB image, and its associated sparse depth map, in order to have prior structure information.

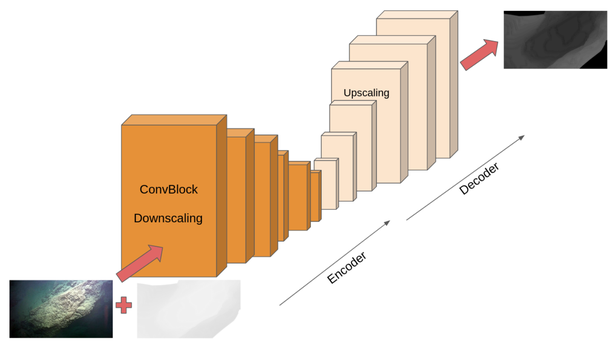


Figure 5. Sparse auto-encoder architecture

Guided Convolutional Network v2: Building upon Tang al. [4], we use two parallel U-Net structures to reconstruct a depth map. One takes in input a RGB image and the other its associated sparse depth map. The first networks guides the second during the learning process.

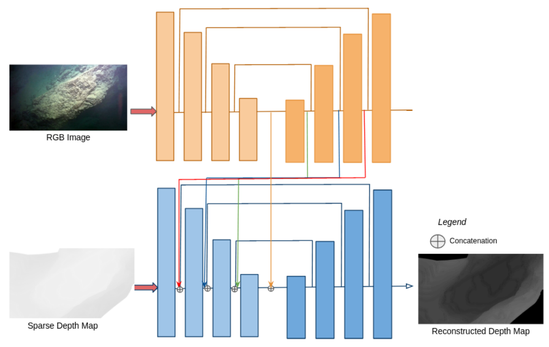


Figure 6. Our guided network architecture

Quantitative results

Metrics for evaluation : the root mean square error (RMSE), log root mean square error (RMSLE), the absolute relative error (Abs. Rel.). Those shown are obtained after 50 epochs of training.

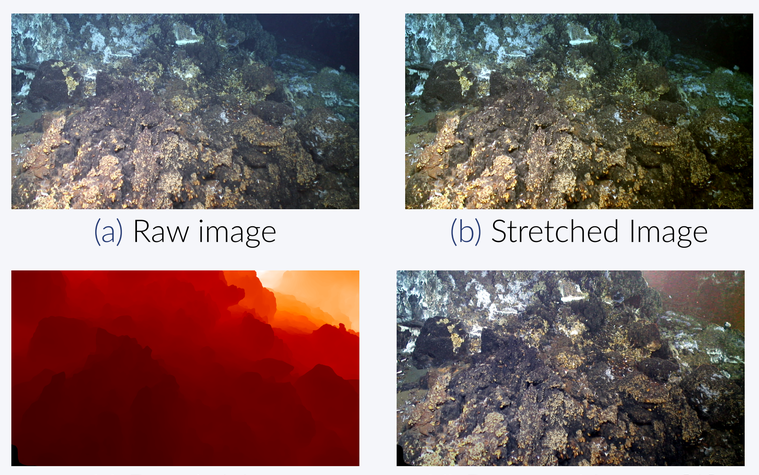
Network name	RMSE	RMSLE	Abs. rel.
MiDaS Small Net	3.01	0.236	0.659
ResNet50 based AutoEncoder	5.0988	1.1518	0.9207
ResNet50 Sparse AutoEncoder	2.7348	0.2238	0.2432

Underwater image color restoration using learned depth maps

Because of light propagation in underwater medium, underwater images exhibit particular characteristics. Two major characteristics of underwater images include color attenuation and backscatter. Both are strongly correlated with the distance between the scene and the camera. Akkaynak *et al.* introduced Sea-Thru [1], a method to restore colors by estimating an underwater image formation model. This model relies on the distance z between the scene and the camera:

$$I_c = J_c\, e^{-\beta_c(z)z} + B_c\,(1 - e^{-\gamma_c z})$$

where c is the color channel, I_c is the measured pixel intensity, J_c is the restored pixel intensity, β_c is the color attenuation coefficient, B_c is the backscatter color and γ_c is the backscatter coefficient. Both color attenuation and backscatter coefficient depend on z, which could be retrieved from the depth maps estimated with the tested approaches.



(a) Raw image (b) Stretched Image (c) Depth Map (d) Restored image [1]

Figure 7. Color restoration using histogram stretching (b) and a Depth-based method [1] (d).

Discussion & Future Work

- Sparse depth maps addition $\longrightarrow$ better reconstruction performance
- Use more than one image as input of the networks to provide multiview information
- Bypass the ground truth constraint by exploring self-supervised methods

References

[1] Derya Akkaynak and Tali Treibitz. Sea-thru: A method for removing water from underwater images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[2] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[3] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. In *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[4] Jie Tang, Fei-Peng Tian, Wei Feng, Jian Li, and Ping Tan. Learning guided convolutional network for depth completion. In *IEEE Transactions on Image Processing*, 2020.

Clementin.Boittiaux@ifremer.fr

Ines.Larroche@ifremer.fr