2019 International Conference on Robotics and Automation (ICRA)
Palais des congres de Montreal, Montreal, Canada, May 20-24, 2019

UWStereoNet: Unsupervised Learning for Depth Estimation and Color Correction of Underwater Stereo Imagery

Katherine A. Skinner[1], Junming Zhang[2], Elizabeth A. Olson[1] and Matthew Johnson-Roberson[3]

***Abstract*— Stereo cameras are widely used for sensing and navigation of underwater robotic systems. They can provide high resolution color views of a scene; the constrained camera geometry enables metrically accurate depth estimation; they are also relatively cost-effective. Traditional stereo vision algorithms rely on feature detection and matching to enable triangulation of points for estimating disparity. However, for underwater applications, the effects of underwater light propagation lead to image degradation, reducing image quality and contrast. This makes it especially challenging to detect and match features, especially from varying viewpoints. Recently, deep learning has shown success in end-to-end learning of dense disparity maps from stereo images. Still, many state-of-the-art methods are supervised and require ground truth depth or disparity, which is challenging to gather in subsea environments. Simultaneously, deep learning has also been applied to the problem of underwater image restoration. Again, it is difficult or impossible to gather real ground truth data for this problem. In this work, we present an unsupervised deep neural network (DNN) that takes input raw color underwater stereo imagery and outputs dense depth maps and color corrected imagery of underwater scenes. We leverage a model of the process of underwater image formation, image processing techniques, as well as the geometric constraints inherent to the stereo vision problem to develop a modular network that outperforms existing methods.**

I. INTRODUCTION

Underwater stereo vision is a critical perception component for many marine robotic systems. Relative to other sensors, stereo cameras are compact, inexpensive, and capable of providing high resolution color imagery and depth of subsea scenes. Still, there are many challenges to deploying stereo camera systems in underwater environments. Underwater image formation is affected by water column effects, such as attenuation of light and backscattering, which lead to degradation of underwater images. Many of these effects are range-dependent and wavelength-dependent, so image degradation increases with range from the camera, and it alters the ratio of color channels in resulting images. This can have severe consequences for traditional stereo vision algorithms, which rely on image contrast and brightness or photometric consistency across views to detect and match image features [1] [2] [3]. In order to develop a robust underwater stereo vision system, we must account for these water column effects to restore photometric consistency, or we must develop methods that are invariant to the inconsistencies seen across varying viewpoints of underwater scenes.

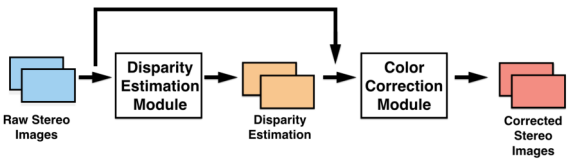


Fig. 1: Overview of proposed network structure for simultaneous depth estimation and color correction from raw underwater stereo imagery.

In recent years, deep learning has led to many advances in robotic perception. Applications of deep learning to the underwater domain have shown great potential to solve many open problems in the field [4] [5]. However, once again, there are unique challenges to applying these methods in underwater environments. First, it is particularly challenging to gather large training datasets with ground truth structure and color of underwater scenes. This motivates the development of unsupervised or self-supervised learning approaches. Second, data gathered in underwater environments is highly degraded. Due to underwater lighting effects, images taken underwater can also vary widely depending on factors such as time-of-day and water properties. This makes it difficult to gather representative training data and to develop generalizable networks that work across a range of subsea scenes. One thing we can leverage is prior knowledge from traditional computer vision, image processing and underwater imaging to provide constraints for approaching this problem.

In this paper, we develop a novel unsupervised network that addresses two challenging problems in underwater stereo vision: dense depth estimation and color correction of raw color underwater images. Our method exploits the process of underwater image formation, insights from image processing, as well as the geometry of stereo camera systems. To our knowledge, this is the first approach to develop a deep learning approach that estimates both dense depth and water column color correction directly from underwater stereo imagery. We present experiments on real underwater data collected at different field sites, with ground truth structure and color to provide both quantitative and qualitative evaluation of results. We show that our method improves upon traditional and state-of-the-art approaches. We also provide a discussion of insights gained on the application of deep learning to underwater stereo vision of robotic systems.

[1]K. Skinner and E. Olson are with the Robotics Institute, University of Michigan, Ann Arbor, MI 48109 USA. {kskin,lizolson}@umich.edu

[2]J. Zhang is with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 USA. junming@umich.edu

[3]M. Johnson-Roberson is with the Department of Naval Architecture and Marine Engineering, University of Michigan, Ann Arbor, MI 48109 USA. mattjr@umich.edu

The remainder of this paper is organized as follows: Section II includes relevant prior work, Section III details our technical approach, Section IV presents experiments and results, and, lastly, Section V summarizes conclusions and suggestions for future work.

II. RELATED WORK

A. Learning from Stereo Imagery

Recent work for terrestrial applications has focused on learning dense depth maps directly from rectified stereo images [6] [7] [8] [9] [10]. These methods require dense ground truth depth maps, which are difficult to gather underwater. We instead focus on developing an unsupervised network. Prior work on unsupervised disparity estimation has leveraged the constraints of stereo geometry with an image warping loss function [11] [12] [13]. For our disparity estimation network, we leverage a state-of-the-art architecture, DispSegNet [14]. DispSegNet uses a five-dimensional cost volume to estimate a coarse initial disparity. This initial disparity is then refined further using smoothness and semantic segmentation labels. We modify this network so that semantic segmentation is not required.

B. Underwater Image Restoration

One approach to restore underwater images is to use image processing techniques, such as histogram equalization, to stretch the effective contrast of the image. This can improve feature detection and matching and often results in visually appealing images. However, image processing techniques have no knowledge of the physical process of underwater image formation; thus they do not account for range-dependent effects to restore photometric consistency across different viewpoints. This makes these techniques unreliable for consistent color correction across changing viewpoints.

Our work proposes a network for underwater image restoration that incorporates the physical model of underwater light propagation. Prior work has incorporated knowledge of range-dependent water column effects to perform image restoration of monocular underwater images [15]. Other work has incorporated this model into simultaneous localization and mapping (SLAM) and 3D reconstruction frameworks [16] [17] [18] [19]. These approaches leverage the dense depth output from 3D reconstruction methods as input to the range-dependent model. Bryson et al. also incorporates information such as vehicle lighting configuration [17]. Our approach does not require pre-processing or full 3D reconstruction. We input only unlabeled raw stereo imagery and output dense disparity maps and restored images directly.

Recently, work has been done to provide further insight and experimental validation of traditional models of underwater light propagation that are commonly used for underwater image restoration [20] [21]. This has led to development of an updated model for underwater light propagation, which inspires the structure of our proposed network, discussed in more detail in Section III.

C. Learning for Underwater Vision

State-of-the-art methods use deep learning architectures for monocular underwater image restoration. Many methods rely on synthetic underwater datasets that are augmented from in-air datasets. These training sets are either constructed manually using a known physical model for underwater light propagation [22], or images are augmented within a learned pipeline [4] [5]. We train our network on real underwater images. To our knowledge, our work is the first to develop a learning-based framework for both dense depth estimation and color correction from raw underwater stereo imagery.

III. TECHNICAL APPROACH

Figure 1 shows an overview of our proposed network structure. The network is a modular, two-stage network. The first stage performs disparity estimation based on [14]. This stage takes input rectified raw color stereo images, I_L and I_R, and estimates disparity maps for each image, D_L and D_R. The second stage is a color correction module. This module takes in D_L and D_R and converts disparities to metric depth values using the stereo camera calibration. The depth is then concatenated with the raw stereo images and input to the color correction module. The color correction network outputs restored underwater stereo images, C_L and C_R. The following subsections describe further details of each network module.

A. Disparity Estimation

In this work we employ a Siamese network architecture based on DispSegNet [14] to learn dense disparity maps from input left and right images. This network structure was initially developed to refine disparity estimates based on ground truth semantic segmentation [14]. Here we modify it by removing the Pyramid Scene Parsing (PSP) module [23] in order to only perform unsupervised disparity estimation without reliance on ground truth semantic segmentation. The network structure contains two branches, one for the left image and another for the right image. Each branch has a ResNet [24] structure and weights are shared across both branches. Feature extraction is first performed on the input stereo images, to output a feature map 1/4 of the size of input image. The left and right output feature maps are concatenated to form a five-dimensional cost volume, one for each view. Three-dimensional convolution can be applied to these cost volumes to output a coarse disparity map as the initial estimate. This coarse estimate is then input into a refinement network to achieve more finegrained disparity estimation.

We use different losses for the initial and refinement stages of the disparity estimation process. The loss for initial disparity estimate (L_{disp_init}) forces the model to estimate a coarse disparity map weighting all pixels equally, while the refinement loss (L_{disp_ref}) forces the model to focus on regions of high difficulty (low texture and high noise). The losses are defined as following:

$$Loss = \alpha_1 L_{disp\_init} + \alpha_2 L_{disp\_ref} \quad (1)$$

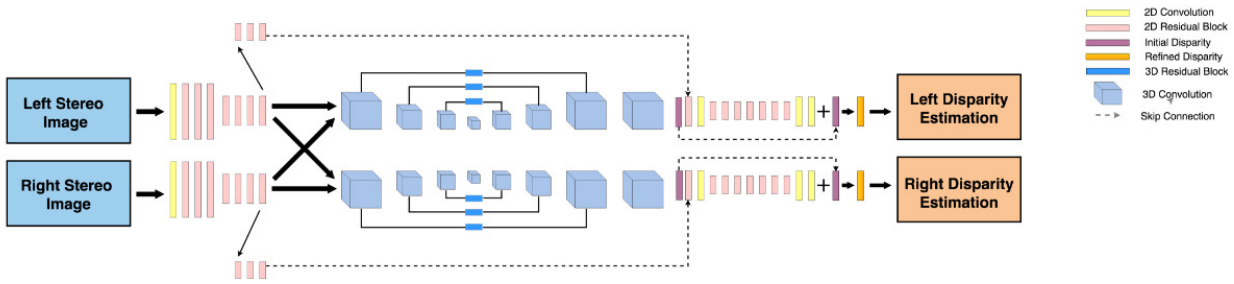


Fig. 2: Network structure for disparity estimation module [14].

$$L_{disp\_init} = \beta_1 L_{disp\_warp} + \beta_2 L_{consist} + \beta_3 L_{reg} \quad (2)$$

$$L_{disp\_ref} = \gamma_1 L_{disp\_warp} + \gamma_2 L_{consist} + \gamma_3 L_{smooth} \quad (3)$$

where α_1 and α_2 are scalars that trade off between initial and refinement loss and β_1, β_2 and β_3 trade off between photometric loss (L_{disp_warp}), consistency loss ($L_{consist}$) and regularization loss (L_{reg}) respectively, each of which are described below. The γ scalars serve the same purpose for each respective loss in the L_{disp_ref} refinement loss.

The key component of the total loss during unsupervised disparity estimation is the photometric warping loss. This loss leverages geometric constraints inherent to the stereo vision problem and is enabled by a differentiable bilinear sampler [11] [12]. To construct this loss, we use the estimated disparity map to warp the left image to the right image, and vice versa. After this warping, we use photometric reconstruction error that measures the visual difference between the warped image and the real image. Photometric loss is computed for both sides of the stereo pair. For the left side, it is defined as following:

$$L_{disp\_warp} = \lambda_1 S(I_L, I_L^{'}) + \lambda_2 |I_L - I_L^{'}| + \lambda_3 |\nabla I_L - \nabla I_L^{'}| \quad (4)$$

where I_L is left input image, $I_L^{'}$ is the reconstructed left image from the right input image, ∇ is the first derivative and $S()$ is structural similarity which is used to increase robustness against errors in ill-posed regions of the image. The λ scalars were set experimentally. The photometric loss for the right camera is symmetric.

With strict photometric loss during training, the predicted disparity map typically contains a high amount of noise. Regularization loss L_{reg} is used to enforce locally smooth results. Consistency loss $L_{consist}$ enforces that given the reconstructed left image $I_L^{'}$, we can also warp it back to the right view $I_R^{''}$ using the right camera disparity map D_R. This additional loss forces the left and right disparity branches of the network to be coupled. This loss minimizes the visual difference between $I_R^{''}$ and I_R. After convergence the initial disparity still contains error in challenging regions. We locate these regions by finding the inconsistencies between the initial disparities of the left and right disparity maps [25]. This informs a smoothness loss that minimizes differences in these regions. This loss is only applied during refinement because it requires initialized disparity estimates. Note that we do not use the supervised segmentation loss. For additional details, please refer to the original paper [14].

B. Color Correction

Figure 3 shows a detailed diagram of the color correction network module. The color correction network structure is motivated by a physics-based model of underwater image formation [26] [27]:

$$I(x) = J(x)e^{-\beta d(x)} + B(1 - e^{-\beta d(x)}) \quad (5)$$

where d is the range between the camera and the scene, β is an attenuation coefficient, x are spatial coordinates in the image, B is the veiling or environmental light, J is the image before attenuation, and I is the resulting underwater image. The first component, $J(x)e^{-\beta d(x)}$, is the direct transmission subject to attenuation; the additive component contributes to backscattering, which appears as haze in the resulting image.

Here we focus on modeling the direct transmission component to correct for the effects of attenuation. Future work will address backscattering for hazy scenes. Our approach requires estimation of range, or depth maps, d of the scene, as well as the attenuation coefficient β. In classical models, β is considered to be wavelength-dependent, such that different wavelengths are attenuated at different rates, leading to the characteristic color of underwater imagery. This would mean estimating a scalar β_c per color channel. Recently, Akkaynak et al. presented a novel model of attenuation that suggests that estimating β in this way induces error in color correction algorithms [21]. Instead, [21] proposes, and validates through experimentation, that β is dependent on many factors including sensor characteristics, water properties, and range to the scene. These factors are challenging to measure or estimate for all scenes. Instead, we propose a learning module to estimate and account for attenuation in order to correct color of underwater imagery.

Our color correction network (Fig. 3) is based on [28], which was initially developed for terrestrial image restoration. We make several modifications for our underwater image restoration pipeline. Our network features two branches with shared weights. Each branch takes a raw underwater image and its respective disparity map output from the disparity estimation module. Disparity maps are converted to depth maps and scaled to improve numerical stability. The paired depth map and image are concatenated and input into a fully convolutional network. The output of our network

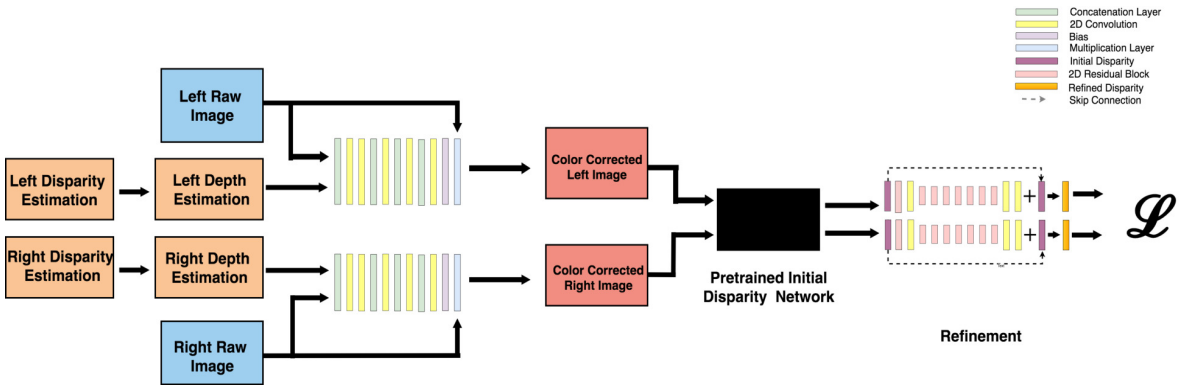


Fig. 3: Network structure for color correction and disparity refinement module.

is then multiplied by the raw underwater image to output a color corrected image. Relating back to the model of underwater image restoration, our network estimates $e^{\beta d}$, the inverse of the transmission.

We use a two-stage loss for our network. The initial and refinement color correction losses are given by: $L_{color\_init} = \theta_1 L_{color\_warp} + \theta_2 L_{color\_cyc} + \theta_3 L_{gray} + \theta_4 L_{IQ}$ and $L_{color\_ref} = w_{init} L_{color\_init} + w_{ref} L_{disp\_ref}$, where θ and w are hyperparameters for relative weighting of loss components determined experimentally.

Photometric Warping Loss (L_{color_warp}): To ensure that both the left and right images appear with consistent colors after image restoration, we use a photometric warping error similar to that in the disparity estimation network, where the estimated disparity maps are used to warp one corrected image to the other view. For example, the left corrected image C_L is warped to the right view $C_R^{'}$.

$$L_{color\_warp} = \|C_L^{'} - C_L\|^2 + \|C_R^{'} - C_R\|^2 \tag{6}$$

Cyclic Reconstruction Loss (L_{cyc}): Inspired by recent advances in cyclic networks [29], we employ a cyclic reconstruction loss. This helps to ensure that the learned color correction parameters do not collapse to trivial results. Inverse color correction is performed for each corrected image, producing $\widehat{I_L}$ and $\widehat{I_R}$. That is, each corrected image is divided by the network output to estimate the original raw input images. Reconstruction loss is used to ensure that the inverted images align with the original raw inputs:

$$L_{cyc} = \|I_L - \widehat{I_L}\|^2 + \|I_R - \widehat{I_R}\|^2 \tag{7}$$

Image Quality Losses (L_{IQ}, L_{gray}): We also leverage knowledge from image processing to force our output to have high image quality using the gray world assumption, acutance, and contrast metrics. The gray world loss L_{gray} is based on prior knowledge of natural image statistics and assumes that the average color in natural images is gray. This constraint minimizes the distance between the average color of each color channel and gray. The contrast constraint q_c maximizes gain in contrast of the corrected image compared to the raw underwater image, which typically suffers from reduced contrast [30]. Lastly, the acutance constraint q_a indicates sharpness in an image and can be computed by the gradient strength of the image [30]. We use an image quality loss given by $L_{IQ} = 1.0 - (w_a q_a + w_c q_c)$, where w_a and w_c are scalars.

Disparity Loss on Corrected Color (L_{disp_ref}): Lastly, we leverage the interdependency between structure and color to provide further constraints on our output color imagery. Our corrected color images are input into the pre-trained coarse disparity network. The coarse disparity output is then refined to give a smooth disparity map. The weights of the residual refinement network are trained. We compute the loss L_{disp_ref} as described in the above section. This loss on the corrected color image ensures that resulting image quality and features are sufficient for accurate disparity estimation.

Note that this approach is completely self-supervised, where the target output for each stage is input or generated by a previous stage. The network is designed to be modular, so that each learned component can be used on its own or substituted for another network or approach. See `https://github.com/kskin/UWStereo` for further implementation details.

IV. EXPERIMENTS & RESULTS

Fig. 4: (left) Stereo camera configured on bottom of the bottom of the BlueROV and (right) artificial rock platform with attached color board for ground truth structure and color

A. Data collection

To validate our method, we deploy a BlueRobotics BlueROV2 remotely operated vehicle equipped with a custom downward facing stereo camera system to gather underwater stereo imagery (Fig. 4). A color board and artificial rock platform are submerged for ground truth (Fig. 4). We collected two datasets near the Hawaii Institute of Marine Biology (HIMB). Each site contained different features and

had different water column properties. The first site, HIMB #1, was in an open bay containing coral. There are 1371 training images and 10 test images. The second site, HIMB #2, features rocks and manmade objects such as cement blocks. This site was in a sheltered canal. There are 2676 training images and 5 test images. Note that all images containing the ground truth color board and artificial rock platform were removed from the training set.

B. Training Details

We pre-train the disparity module using an in-air dataset [31] for 100k iterations, which takes approximately 1 day on a Titan V GPU. We then finetune the disparity network on both HIMB datasets for 10k iterations. For training the color correction network, we train only on an individual HIMB dataset for 4k iterations, and show testing on the corresponding test set. Training of the color correction module takes under 1 hour, which we believe is reasonable for adapting to different environments.

C. Qualitative Results

Figure 5 shows qualitative results of our color correction network. We compare our method to more traditional image processing approaches, including histogram equalization and gray world image correction [32], as these are popular methods for preprocessing of underwater data for high level vision tasks. We also compare to UGAN, a state-of-the-art deep learning approach for underwater image restoration, using the pretrained model [5]. Qualitatively, histogram equalization gives the sharpest resulting image with enhanced contrast. UGAN results in a haloing effect in areas that were not fully corrected. Compared to other methods, our method results in a consistent color across different viewpoints.

Figure 6 shows qualitative results of disparity estimation. The traditional Semi-Global Matching (SGM) approach [33] results in the sparsest disparity map, especially around edges and occluded regions. The proposed disparity network provides much denser results with notable improvement between pretraining on in-air data and finetuning on underwater data.

D. Quantitative Evaluation of Disparity Estimation

To measure the accuracy of each DNN's disparity output quantitatively, we create ground truth by scanning the rock platform (Fig. 4 right) with an ASUS Xtion Pro, and inputting the scan into ElasticFusion [2]. The resulting cloud was next filtered and transformed into a mesh, from which one million points were sampled to obtain a dense, evenly distributed set of reference points. To enable comparison, output disparity from each method is cropped to mask only the rock platform region. This section is then converted to a point cloud, which is filtered for invalid disparity ranges. The point cloud is hand-registered to the reference ground truth point cloud in Cloud Compare [34]. This coarse registration is improved through the Iterative Closest Point algorithm [35]. Finally, a modified Hausdorff distance between the two clouds is calculated. Table I shows that SGM outperforms the proposed network when it is only pre-trained in air, but once finetuned on underwater data, our approach achieves the highest performance of any technique compared. Through experimentation, it was also noted that the disparity estimation network trained directly on underwater imagery is robust to varying water conditions. However, to achieve this it was is important to have representative features in the data. We noted that models trained only on the coral site performed poorly where models trained across both sites were more successful on the quantitative rock evaluations.

TABLE I: The mean and standard deviation of modified Hausdorff distance across point clouds generated from test images for each trained DNN and traditional Semi-Global Matching when compared against the ground truth point cloud for the artificial rock structure.

Dataset	**HIMB #1 - Coral**		
Method	**Mean**	**STD**	**# Pts.**
SGM	0.0862	0.0250	8936
Pretrained	0.1444	0.0936	17611
Finetuned	0.0709	0.0341	14658

Dataset	**HIMB #2 - Rocky**		
Method	**Mean**	**STD**	**# Pts.**
SGM	0.0738	0.0129	15837
Pretrained	0.2136	0.1264	41491
Finetuned	0.0632	0.0346	44636

E. Quantitative Evaluation of Color Correction

Table II shows quantitative evaluation of color correction. For each test image, we took the average value for each color on the color board. We then computed RMSE in RGB-space compared to the ground truth color board imaged in air. Here we show the mean and standard deviation of RMSE across the test set for each site. Our method outperforms the other traditional and state-of-the-art methods.

TABLE II: The mean and standard deviation of RMSE (m) from ground truth color board across color corrected test images.

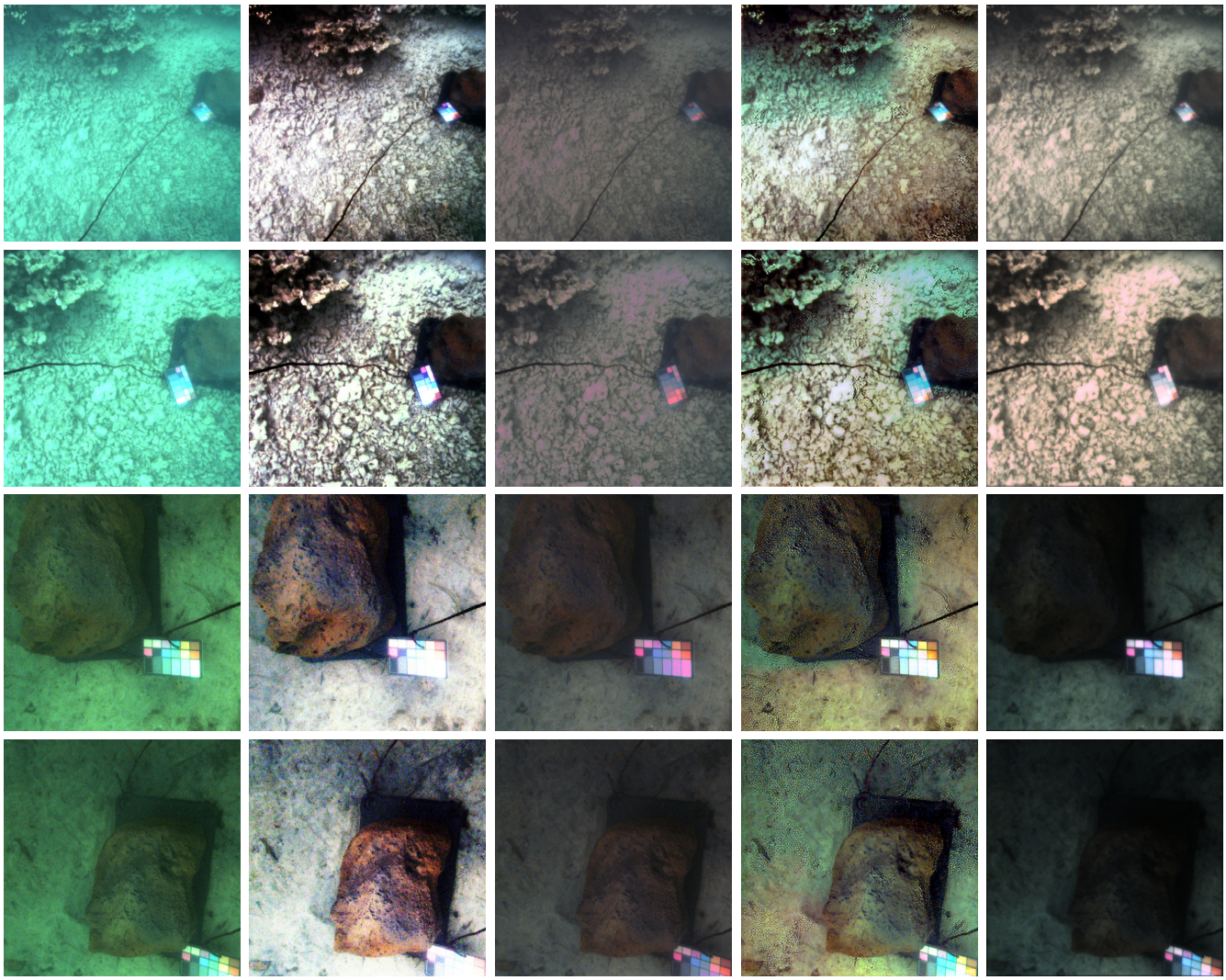
Dataset	**HIMB #1 - Coral**		**HIMB #2 - Rocky**	
Method	**Mean RMSE**	**STD**	**Mean RMSE**	**STD**
Raw	0.2203	0.0439	0.2219	0.0554
Hist. Eq.	0.1301	0.0232	0.1132	0.0108
Gray World	0.1579	0.0365	0.1204	0.0310
UGAN [5]	0.1461	0.0245	0.1555	0.0414
Our Method	**0.1065**	0.0187	**0.1037**	0.0159

V. CONCLUSIONS

Our proposed method is a novel, modular learning pipeline for dense depth estimation and color correction of raw underwater stereo imagery. By leveraging knowledge from image processing, computer vision, and underwater light propagation, we are able to perform these tasks without supervision. Our experiments validate our method quantitatively and qualitatively to show that we outperform existing methods on both tasks. Future work will focus on integrating this learned vision system onto an underwater robotic platform in the field.

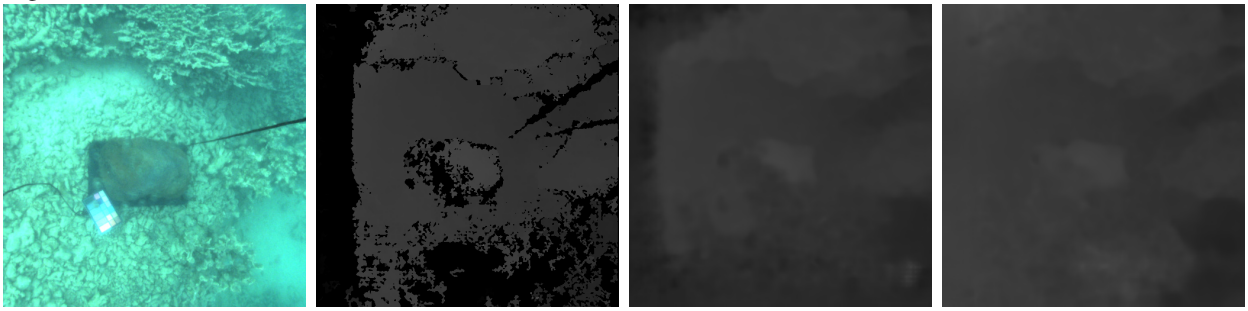
ACKNOWLEDGMENT

This work was supported by the National Science Foundation under Award 1452793, and by the Department of Energy under Award DE-EM0004383 (subcontract through Carnegie Mellon University).



Raw Stereo Image | Histogram Equalization | Gray World | UGAN | UWStereoNet

Fig. 5: Color correction: We compare the results of each method on four test image sets. The first column displays a raw stereo image, followed by histogram equalization, which shows the sharpest image but becomes oversaturated. Gray World results in an over-amplified red channel on the color board. The fourth column contains the output of UGAN, which has unnatural coloring on the ocean floor. Finally, UWStereoNet's output is provided in the last column, with photorealistic coloring for the natural terrain and color board.



Raw Stereo Image | SGM | Pretrained - in-air | Finetuned - underwater

Fig. 6: Disparity estimation: The leftmost image is a raw stereo image, followed by the results of SGM when implemented on the respective stereo pair. We then provide the results of UWStereoNet, first from pretraining on in-air images, then of finetuning with underwater imagery. This practice yields a denser map of the scene, particularly on the side of the rock, as well as better texturing of the coral.

References

[1] P. Horn K. Berthold and B. G. Schunck. "Determining optical flow". In: *Artificial Intelligence* 17.1-3 (1980), pp. 185–203.

[2] Thomas Whelan, Stefan Leutenegger, Renato F Salas-Moreno, Ben Glocker, and Andrew J Davison. "ElasticFusion: Dense SLAM without a pose graph". In: *Proceedings of Robotics: Science and Systems (RSS)*. 2015.

[3] David G. Lowe. "Distinctive image features from scale-invariant keypoints". In: *International Journal of Computer Vision* 60.2 (Nov. 2004), pp. 91–110.

[4] J. Li, K. A. Skinner, R. M. Eustice, and M. Johnson-Roberson. "WaterGAN: Unsupervised generative network to enable real-time color correction of monocular underwater images". In: *IEEE Robotics and Automation Letters* 3.1 (Jan. 2018), pp. 387–394.

[5] Cameron Fabbri, Md Jahidul Islam, and Junaed Sattar. "Enhancing underwater imagery using generative adversarial networks". In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. Brisbane, Queensland, Australia, May 2018, pp. 7159–7165.

[6] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 4040–4048.

[7] Wenjie Luo, Alexander G Schwing, and Raquel Urtasun. "Efficient deep learning for stereo matching". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 5695–5703.

[8] Yiliu Feng, Zhengfa Liang, and Hengzhu Liu. "Efficient deep learning for stereo matching with larger image patches". In: *10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE. 2017, pp. 1–5.

[9] 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.

[10] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. "DeepMVS: Learning multi-view stereopsis". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 2821–2830.

[11] 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, pp. 270–279.

[12] Yiran Zhong, Yuchao Dai, and Hongdong Li. "Self-supervised learning for stereo matching with self-improving ability". In: *arXiv preprint arXiv:1709.00930* (2017).

[13] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. "Unsupervised learning of depth and ego-motion from video". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6612–6619.

[14] Junming Zhang, Katherine A. Skinner, Ram Vasudevan, and Matthew Johnson-Roberson. "DispSegNet: Leveraging semantics for end-to-end learning of disparity estimation From stereo imagery". In: *IEEE Robotics and Automation Letters* 4.2 (Apr. 2019), pp. 1162–1169.

[15] Nicholas Carlevaris-Bianco, Anush Mohan, and Ryan M. Eustice. "Initial results in underwater single image dehazing". In: *Proceedings of IEEE/MTS OCEANS Conference*. Seattle, WA, USA, Sept. 2010, pp. 1–8.

[16] Mitch Bryson, Matthew Johnson-Roberson, Oscar Pizarro, and Stefan B. Williams. "Colour-consistent structure-from-motion models using underwater imagery". In: *Robotics: Science and Systems VIII* (2013), pp. 1–8.

[17] Mitch Bryson, Matthew Johnson-Roberson, Oscar Pizarro, and Stefan B Williams. "True color correction of autonomous underwater vehicle imagery". In: *Journal of Field Robotics* 33.6 (2016), pp. 853–874.

[18] Anne Jordt. "Underwater 3D Reconstruction Based on Physical Models for Refraction and Underwater Light Propagation". PhD thesis. Kiel University, 2013.

[19] Katherine A Skinner, Eduardo Iscar, and Matthew Johnson-Roberson. "Automatic color correction for 3D reconstruction of underwater scenes". In: *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2017, pp. 5140–5147.

[20] Derya Akkaynak and Tali Treibitz. "A revised underwater image formation model". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 6723–6732.

[21] Derya Akkaynak, Tali Treibitz, Tom Shlesinger, Yossi Loya, Raz Tamir, and David Iluz. "What is the space of attenuation coefficients in underwater computer vision?" In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2017, pp. 568–577.

[22] Young-Sik Shin, Younggun Cho, Gaurav Pandey, and Ayoung Kim. "Estimation of ambient light and transmission map with common convolutional architecture". In: *OCEANS 2016 MTS/IEEE Monterey*. 2016, pp. 1–7.

[23] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. "Pyramid scene parsing network". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 2881–2890.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 770–778.

[25] Jure Zbontar and Yann LeCun. “Stereo matching by training a convolutional neural network to compare image patches”. In: *Journal of Machine Learning Research* 17.1-32 (2016), pp. 2287–2318.

[26] J. S. Jaffe. “Computer modeling and the design of optimal underwater imaging systems”. In: *IEEE J. Oceanic Engin.* 15.2 (1990), pp. 101–111.

[27] B. L. McGlamery. *Computer analysis and simulation of underwater camera system performance*. Tech. rep. UC San Diego, 1975.

[28] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. “AOD-Net: All-in-one dehazing network”. In: *IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017, pp. 4780–4788.

[29] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. “Unpaired image-to-image translation using cycle-consistent adversarial networks”. In: *IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 2242–2251.

[30] Walysson V. Barbosa, Henrique G.B. Amaral, Thiago L. Rocha, and Erickson R. Nascimento. “Visual-quality-driven learning for underwater vision enhancement”. In: *IEEE International Conference on Image Processing (ICIP)* (2018), pp. 3933–3937.

[31] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. “The cityscapes dataset for semantic urban scene understanding”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 3213–3223.

[32] M. Johnson-Roberson, Mitch Bryson, Ariell Friedman, Oscar Pizarro, Giancarlo Troni, Paul Ozog, and Jon C. Henderson. “High-resolution Underwater Robotic Vision-based Mapping and 3D Reconstruction for Archaeology”. In: *J. Field Robotics* (2016), pp. 625–643.

[33] Heiko Hirschmuller. “Accurate and efficient stereo processing by semi-global matching and mutual information”. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. 2. IEEE. 2005, pp. 807–814.

[34] Daniel Girardeau-Montaut. “Cloud compare—3d point cloud and mesh processing software”. In: *Open Source Project* (2015).

[35] Paul J. Besl and Neil D. McKay. “A Method for Registration of 3-D Shapes”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 14.2 (Feb. 1992), pp. 239–256.