Self-Supervised Underwater Caustics Removal and Descattering via Deep Monocular SLAM

Jonathan Sauder and Devis Tuia

Environmental Computational Science and Earth Observeration Lab (ECEO)
EPFL, Switzerland



(a) (b) (c)

Fig. 1: Given an input underwater image (a), CausticsNet removes the caustics effects (b) and BackscatterNet removes detrimental scattering from the water column (c).

Abstract. Underwater scenes are challenging for computer vision methods due to color degradation caused by the water column and detrimental lighting effects such as caustic caused by sunlight refracting on a wavy surface. These challenges impede widespread use of computer vision tools that could aid in ecological surveying of underwater environments or in industrial applications. Existing algorithms for alleviating caustics and descattering the image to recover colors are often impractical to implement due to the need for ground-truth training data, the necessity for successful alignment of an image within a 3D scene, or other assumptions that are infeasible in practice. In this paper, we propose a solution to tackle those problems in underwater computer vision: our method is based on two neural networks: CausticsNet, for single-image caustics removal, and BackscatterNet, for backscatter removal. Both neural networks are trained using an objective formulated with the aid of self-supervised monocular SLAM on a collection of underwater videos. Thus, our method does not requires any ground-truth color images or caustics labels, and corrects images in real-time. We experimentally demonstrate the fidelity of our caustics removal method, performing similarly to state-of-the-art supervised methods, and show that the color restoration and caustics removal lead to better downstream performance in Structure-from-Motion image keypoint matching than a wide range of methods.

1 Introduction

Underwater scenes often pose significant challenges to computer vision methods due to detrimental effects caused by the water column. Underwater, light is scattered and attenuated following an intricate non-linear relationship with light wavelength as well as with the distance between the object and the camera, the depth below the surface, and the optical properties of the water at hand [4]. The latter can vary extremely depending on the body of water, which makes it extremely challenging to devise methods to invert the detrimental effects of underwater scenes across water bodies without a-priori calibration [5]. Furthermore, lighting effects near the surface can add to the difficulty: as sunlight hits a wavy water surface, which acts like a moving lens, the diffracted rays bundle and form rapidly changing noisy *caustic* patterns on the seafloor [72].

Due to these effects, computer vision methods often struggle in underwater environments. Structure from Motion (SfM) photogrammetry in underwater scenes only works reliably if the underlying images are collected in favorable conditions: the effect of the water should be minimally apparent, with photos being taken close to the substrate and excluding images which contain a background water column [9]. Caustics, in particular, are extremely detrimental for image keypoint matching [25], which forms the basic building block in almost all Simultaneous Localization and Mapping (SLAM) and SfM pipelines [47, 8, 64]. The effect of the water column and caustics is also detrimental to other tasks such as semantic segmentation and human pose estimation [65]. Counteracting such detrimental effects in underwater scenes could improve a large range of underwater computer vision applications, for example ecosystem monitoring in coral reefs [23, 14, 83, 62], seagrass beds [48], antarctic benthic habitats [59], surveying submerged archaeological sites [31], and industrial applications [41, 51].

Meanwhile, models that describe the physics of underwater image formation have been proposed [39, 53, 4], which in turn have sparked several approaches aiming at inverting these models to achieve color restoration [13, 5, 54, 42, 10]. However, underwater color restoration from only a single image is an ill-posed problem. Therefore, these restoration methods often have impractical requirements: prior successful SfM alignment with a collection of reference images or a well calibrated stereo camera setup is necessary to reconstruct the geometry of the scene, and the acquired images must be in raw format. Similarly, caustic removal methods have been proposed [2, 3, 67], but often also make a set of impractical requirements on the image being successfully aligned within a 3D scene, or require ground truth annotations of the caustics to be removed.

Recently, SLAM methods based on deep learning have been introduced, promising to learn features that are robust to the environments they are trained in. In particular, self-supervised deep SLAM objectives [86, 27, 11] enable fast estimation of unscaled depth on single images and camera movement on collections of image sequences in absence of ground-truth. In fact, even in challenging reef scenes marked by caustics and other artifacts from the water column, self-supervised SLAM systems work reliably [62]. However, the total accuracy of depths and poses still lags behind conventional SfM systems [64, 27]: processing

underwater images to lead to improved feature-based image matching remains an important frontier of computer vision.

In this paper, we propose neural network-based solutions for underwater image restoration that can be readily applied to RGB images and are consistent within larger scenes and videos. To do so, we propose CausticsNet for unsupervised caustics removal, and BackscatterNet for self-supervised backscatter removal. Our approach works on single images, but also produces video-consistent results. Despite being unsupervised, CausticsNet rivals supervised methods in caustics segmentation performance in a zero-shot out-of-domain task. Our results show that BackscatterNet and CausticsNet lead to significantly and consistently improved keypoint matching. To summarize, our main contributions are:

- We propose a method for single-image caustics and backscatter removal in underwater scenes which requires no ground-truth color images or caustics labels. The method is based on a self-supervised SLAM formulation to learn from unlabeled underwater videos.
- Contrarily to most state of the art, our method works in real time and operates on RGB images as opposed to requiring images in raw format.
- Experimentally, our method is effective in caustics segmentation and leads to significantly improved downstream image keypoint matching.

2 Background & Related Work

2.1 Underwater Image Formation

While many underwater image formation models have been proposed [39, 53, 55, 63, 12], the most comprehensive model to date is the Akkaynak-Treibitz image formation model [4]. These underwater image formation models model the signal I recorded by the camera as a sum of a direct signal D, a backscatter B from light reflected from particles suspended in the water column, and a forward-scattering term that can often be ignored due to its small contribution [63]. In the simplified form (assuming an image taken horizontally at a constant depth), for each color channel (red, green, blue), this is expressed as:

$$I = D + B = D + B^{\infty}(1 - e^{-\beta_B \cdot Z}) \tag{1}$$

Where Z refers to the depth (distance to camera), and β_B is a coefficient describing how quickly the image signal reaches the final backscatter intensity B^{∞} as the distance increases. Note that the underwater imaging model is in general only true for raw images, and does not hold for RGB images that have already undergone the processing, which commonly happens in digital cameras in a black-box fashion (e.g. histogram stretching, white-balancing). Furthermore, all mentioned underwater imaging models assume a caustics-free scene.

2.2 Underwater Color Restoration and Enhancement

Since the development of underwater imaging models, there has been a strong interest in inverting these models, and restoring the colors of the imaged scenes

as if the water were not present. In Bryson et al. [13] and Nakath et al. [54], AUV imaging models assuming onboard lighting as the only illumination source are inverted. In Berman et al. [10], haze-lines are used for underwater image restoration. In Sea-Thru [5], the Akkaynak-Treibitz imaging model is approximately inverted, with a per-image optimization scheme. In a similar way, Sea-Thru-NeRF [42] inverts the Akkaynak-Treibitz model in Neural Radiance Fields [52] of underwater scenes using the obtained depth maps & integrating the model in the differentiable rendering formulation. All of these methods have in common that they require the scene to have ground-truth depth maps, or at least be aligned within a 3D scene, and that the images are available in raw format.

A popular method to mitigate the effect of scattering media (in particular atmospheric haze) from single images is the Dark Channel Prior (DCP) [30]. DCP estimates scattering medium & relative depth from patches of dark pixels in images, and has been extended to work better under water by modeling the wavelength dependency of scattering in water [20, 46, 80, 66]. A modified DCP is used as a component in Sea-Thru [5]. However, a major drawback of DCP methods is the required presence of dark patches in an image.

It is important to disambiguate underwater color restoration, which aims at recovering the true color of the scene by inverting imaging models, to underwater image enhancement (UIE), which aims to improve the perceived visual quality of underwater imagery. Some UIE methods use components of physical imaging models [18, 56, 84, 78, 43, 1, 58, 66], while others use mostly heuristic approaches [57, 45, 15, 16, 33, 36, 22, 34]. Even though UIE datasets have been established [44, 40, 32], UIE is notoriously challenging to evaluate objectively, as visual quality of images remains subjective and depends on the downstream application. Moreover, UIE benchmarks often do not measure the consistency of a method within a scene, which is necessary for SfM/SLAM applications.

2.3 Caustics

Caustics are formed by a wavy water surface acting as a rapidly locally changing lens to the light above the surface. Caustics lead to rapidly changing noisy patterns on the substrate, and have been demonstrated to be highly detrimental to image matching [25, 67, 7]. In Forbes et al. [25], synthetic caustics are generated on images, and subsequently used to train neural networks to identify and remove the observed caustics, followed by a qualitative evaluation on real-world data. In Agrafiotis et al. [2], the evaluation is extended to image matching, showing a higher number of valid keypoints after applying the method from Forbes et al. [25]. In order to allow a quantitative evaluation, the R-CAUSTICS dataset [3] was introduced, containing images with ground-truth caustics. In Gracias et al. [28], caustics are removed from videos by registering all frames with SfM, and then using the color given by the temporal median of the scene viewed from the registered video frames. In Swirski and Schechner [67], a stereo photogrammetry rig is used in a similar fashion to remove caustics and simultaneously estimate underwater imaging model parameters in a controlled setting in a pool.

All of these methods either require ground-truth annotated caustics, rely on synthetic data (which may not generalize to other forms of caustics), or rely on previous alignment with SfM, which may in fact be prevented by the existence of strong caustics themselves. There is no single-image caustics removal that works in absence of an in-domain dataset of ground-truth annotations. In this work, we propose a method to fill this gap.

2.4 Self-Supervised Deep Monocular SLAM

Deep learning-based SLAM methods promise to learn features that are robust to the scenarios in which they are trained. They can be trained in a supervised manner to predict depth [21, 49], optical flow [19, 35], camera movement [71, 85], or a combination of these tasks that enables SLAM [85, 70, 68, 74]. However, in some scenarios, for example under water, ground-truth depth and odometry is extremely challenging to obtain. In these cases, self-supervised formulations of rigid transformation can be used to learn SLAM from image sequences [26, 86].

The underlying concept is that for two overlapping images I_1 and I_2, a neural network f^{SfM} with learnable parameters θ predicts estimated depths $\hat{Z}_1, \hat{Z}_2$ (pixel-wise distances to camera) and the 6D camera pose transform $\hat{T}_{1\to 2}$ between the two images:

$$\hat{Z}_1, \hat{Z}_2, \hat{T}_{1\to 2} = f_\theta^{\text{SfM}}(I_1, I_2) \tag{2}$$

By choosing an appropriate camera model, an image I_1 and its depth estimate $\hat{Z}_1$ can be reprojected to the camera position of I_2, obtaining a reprojected image $\hat{I}_{1\to 2}$ via differentiable operations [38]:

$$\hat{I}_{1\to 2} = \text{Reproject}(I_1, \hat{Z}_1, \hat{T}_{1\to 2}) \tag{3}$$

A photometric error, combining a pixel-wise ℓ_1-norm and structural similarity (SSIM) [76] between the original image and the reprojection forms the base of a loss objective for self-supervised monocular SLAM:

$$\mathcal{L}_P(I_1, \hat{I}_{2\to 1}) = \lambda \|I_1 - \hat{I}_{2\to 1}\|_1 + (1-\lambda)\big(1 - \text{SSIM}(I_1, \hat{I}_{2\to 1})\big)/2 \tag{4}$$

The seminal SfMLearner [86] adds a regularization term $\mathcal{L}_S(I, Z)$, enforcing smooth depth predictions in regions where the image has low texture. In the Scale-Consistent (SC)-SfMLearner [11], the geometric scale is stabilized by adding a geometric consistency loss $\mathcal{L}_G(\hat{Z}_1, \hat{Z}_2)$ on reprojected predicted depth maps. The loss terms are combined into an objective to train the neural network over a dataset of overlapping image pairs (e.g. video frames), with the photometric loss weighted by geometrically consistency:

$$\min_\theta \mathbb{E}_{I_1, I_2}\Big[\mathcal{L}_P(I_1, \hat{I}_{2\to 1}) \odot \big(1 - \mathcal{L}_G(\hat{Z}_1, \hat{Z}_2)\big) + \lambda_S \mathcal{L}_S(I_1, \hat{Z}_1) + \lambda_G \mathcal{L}_G(\hat{Z}_1, \hat{Z}_2)\Big] \tag{5}$$

In practice, the loss maps are averaged after filtering out stationary points , and the losses are calculated in both directions, by using $\hat{Z}_2$ and the inverse camera

pose transform $\hat{T}_{2\to1} = \hat{T}_{1\to2}^{-1}$. Variants of the SC-SfMLearner have been trained on underwater scenarios [6, 62, 81, 73].

Numerous extensions of self-supervised deep SLAM systems have been proposed, including architectural changes [82, 27], optical flow prediction [50, 19, 35, 60, 82, 69], the usage of cost volumes [85, 69, 70], multiple frames for training or inference [29, 77], or non-linear camera models [22]. In this paper, we base the underlying SLAM system on the SC-SfMLearner, due to its relative simplicity and the ease of obtaining single-image depth predictions and image-pair losses, but our method is applicable to any of the more complex approaches.

3 Method

Our method leverages the Self-Supervised Monocular SLAM setup to define a learning objective which then allows scene-consistent underwater image restoration. In this Section we describe how a SC-SfMLearner trained on underwater imagery is used to formulate an objective loss for CausticsNet (to remove caustics effects) and BackscatterNet (to alleviate the backscatter).

3.1 CausticsNet

CausticsNet takes an image and learns to predict the effect of the caustics by predicting the expected residual reprojection error, allowing for caustics identification & removal. Using the re-projections in the dense monocular SLAM formulation in Eq (3) allows to formulate a residual for a pair of images:

$$R_1 := I_1 - \hat{I}_{2\to1} = I_1 - \text{Reproject}(I_2, \hat{Z}_2, \hat{T}_{2\to1}) \tag{6}$$

When assuming an ideal caustics-free rigid scene with constant ambient lighting, the residual of a trained monocular SLAM system on the overlapping area of the image pair stems from the actual rigid re-projection error from the misestimation of depths and poses. However, in a scene where caustics are present, the residual is mostly dominated by the caustics of I_1 and the (negative) caustics of I_2. CausticsNet takes an image as input, and is trained to predict the expected residual with a randomly sampled overlapping frame, as shown in Figure 2.

As CausticsNet only sees one image at a time and has no a-priori information about the second image used to compute the residual, it can only learn to predict the part of the residual related to the caustics visible on I_1. Letting $\hat{C} = f_\mu^{\text{Caustics}}(I)$ denote the caustics prediction by a neural network f^{Caustics} with learnable parameters μ, we can write the objective function for CausticsNet as:

$$\min_\mu \mathbb{E}_{I_1,I_2}\Big[\|\hat{C}_1 - R_1\|_2^2 + \lambda_{\text{TV}}\|\hat{C}_1\|_{\text{TV}} + \lambda_{\ell_2}\|\hat{C}_1\|_2^2 + \lambda_{[0,1]}\mathcal{L}_{[0,1]}(\hat{C}_1)\Big] \tag{7}$$

In this loss, we regularize the predicted caustics $\hat{C}_1$ using a Total-Variation (2-)norm to promote smoothness and an ℓ_2 norm to enforce small magnitude of

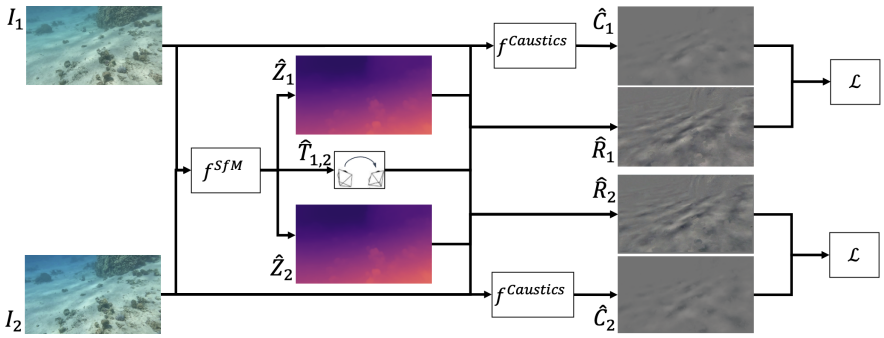


Fig. 2: A schematic overview of the training of f^{Caustics}. The predictions of f^{SfM} are used to compute the reprojection residual between a pair of images, which is dominated by the caustics effects and becomes the main learning target for f^{Caustics}.

caustics. We also incorporate a clip loss, which adds a pixel-wise penalty if the cleaned image is outside of the $[0, 1]$ range:

$$\mathcal{L}_{[0,1]}(I) = \| \min(\max(I, 0), 1) - I) \|_2^2 \tag{8}$$

For practical consideration, the image pairs should have a sufficient temporal distance for the caustics to change in between - an expected temporal difference of a fraction of a second is sufficient in our experimentation.

3.2 BackscatterNet

BackscatterNet takes as input an RGB image and estimates the six parameters defining the backscatter, as described in Eq. (1) (B^∞ and β_B for each of the three color channels). Using the estimated depth $\hat{Z}$ from a monocular depth estimation network, the backscatter is computed. BackscatterNet is a neural network $f^{\text{Backscatter}}$ with learnable parameters γ. For splitting the image into its descattered direct signal and the backscatter, we use the shorthand $\hat{D} = f_\gamma^{\text{Backscatter}}(I)$ summarizing both the estimation of the coefficients and removal of backscatter via Eq. (1) with $\hat{D} = I - \hat{B}$. We use the following loss function:

$$\min_\gamma \mathbb{E}_{I_1,I_2}\Big[\mathcal{L}_P(\hat{D}_1, \hat{D}_2) + \lambda_{\text{I}}\mathcal{L}_{\text{I}}(I_1) + \lambda_{\text{B}}\mathcal{L}_{\text{B}}(I_1, \hat{Z}_1) + \lambda_{[0,1]}\mathcal{L}_{[0,1]}(\hat{D}_1)\Big] \tag{9}$$

where $\mathcal{L}_{[0,1]}(I)$ is the clip loss also used in CausticsNet (Eq. (8)). The other components of this loss encode desirable properties for backscatter removal:

1. Under different water conditions, the backscatter removal should yield the same result. In other words, the current color & turbidity of the water column should minimally influence the descattered image. To enforce this property, we introduce a style independence loss. As a proxy to varying water conditions, we simply use color jitter augmentations. We minimize the photometric error between a descattered image and a jittered version of itself:

$$\mathcal{L}_{\text{I}}(I) = \mathcal{L}_{\text{P}}\Big(f_\gamma^{\text{Backscatter}}(I), f_\gamma^{\text{Backscatter}}(\text{Jitter}(I))\Big) \tag{10}$$

2. In parts of the image which show only the water column, the estimated backscatter should have the same color as the input image. To enforce this property, we calculate the mean color of the 5% of pixels that have are the furthest away (have the highest z-value). On pixels further away than this value (we denote this masking operation by $Q_{95}^{\hat{Z}}(\cdot)$), the estimated backscatter should have the color of the input image, weighted by the depth of the image:
$$\mathcal{L}_{\mathrm{B}}(I, \hat{Z}) = \hat{Z} \odot \|Q_{95}^{\hat{Z}}(\hat{B}) - \mathrm{Mean}\big(Q_{95}^{\hat{Z}}(I)\big)\|_2^2 \tag{11}$$
3. The color of the substrate should not vary with distance, as it does in water (objects further away appear in blue/green and reveal their color as the distance decreases). For this third and arguably most important goal, we leverage the self-supervised SLAM system with the photometric loss $\mathcal{L}_P(I_1, \hat{I}_{2\to1})$ defined in Eq (4): this loss is calculated between a pair of images where both images have been descattered by BackscatterNet.

A schematic of the components of the BackscatterNet objective is shown in Figure 3. For practical consideration, for the photometric loss objective to be effective, the change in distance of the viewed scene (and the induced color change) should be rather large. In the case of ego-motion divers videos considered in this paper, given that the camera mostly moves forward or backward at a consistent swimming speed, an expected temporal distance of one second is an effective choice, but this depends on the dataset at hand.

We train BackscatterNet on the unmodified RGB video frames of the dataset, but find that it works reliably even on images from which CausticsNet has removed caustics. For visualization purposes, to maintain the perceived brightness of images, the descattered images can be min-max stretched or the color-channel mean of the subtracted backscatter can be added. BackscatterNet is trained in conjunction with a depth prediction network, requiring no metric depth, only consistent relative depths. While the parameters B^∞, β_B are resolution-independent, the de-facto backscatter resolution depends on the resolution of the estimated depth map.

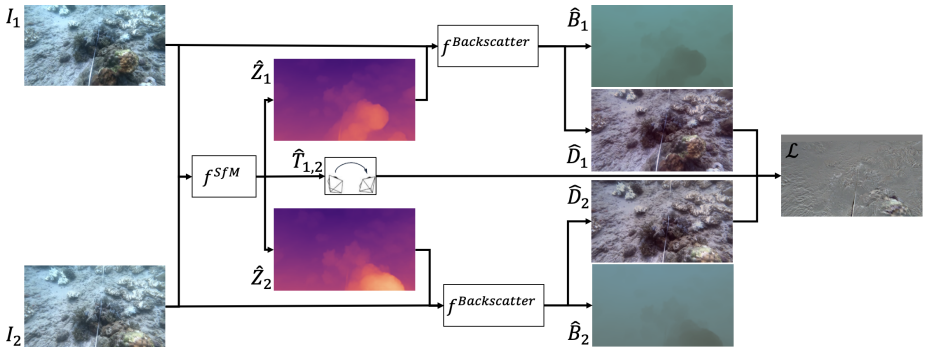


Fig. 3: Schematic overview of training of $f^{\mathrm{Backscatter}}$. The image is split into the estimated direct signal and the backscatter, minimizing a regularized photometric loss.

4 Experiments & Results

4.1 Deep Monocular SLAM on Underwater Video Dataset

We train the SC-SfMLearner on a large dataset of videos collected by divers in coral reefs, containing 145 hours of video from 35 dive sites, all collected with the same camera model [1] in Israel, Jordan, and Djibouti. In spite of the videos being marked by challenging conditions, including strong caustics and sometimes poor visibility, the trained monocular SLAM system is able to predict reasonably good depth and pose estimates for the videos in the dataset. Implementation details can be found in Appendix A.1, samples showcasing the diversity of the dataset as well as the challenging conditions are found in B.3.

4.2 Caustics Removal on R-CAUSTICS

We quantitatively evaluate CausticsNet in a caustics segmentation scenario on the R-Caustics dataset [3]. The R-Caustics dataset contains 1465 images at 4000×3000px resolution taken from stationary cameras in shallow waters, displaying strong rippling caustics on the seafloor. R-Caustics spans 7 scenes, with some scenes containing image sets from multiple views (totaling 20 views). Ground-truth was obtained by covering the entire scene with a sun-blocking cover, eliminating caustics in the scene, and computing a chromatic aberration between the covered scene and the scene with sunlight-caused caustics. This is used to compute ground-truth caustics presence maps via converting from RGB to LAB color space and thresholding in the L channel, which is done per-split. In particular, R-Caustics only contains labels for *positive* caustics (bright ripples) from focused light, and not dark patches where the waves lens the water away.

[1] Videos were collected during expeditions of the Transnational Red Sea Center hosted at EPFL (`http://trsc.org`)

	CausticsNet (global threshold)		CausticsNet (per-split threshold)		Supervised [3]		
Test Splits	F1	Accuracy	F1	Accuracy	Train Splits	F1	Accuracy
1,2,3,4,5,6,7	68.61	92.28	**72.37**	**93.61**	-	-	-
1,2,3	71.82	92.97	77.71	**94.56**	4,5,6,7	**78.62**	92.96
1,2	76.73	95.43	77.18	**95.60**	3,4,5,6,7	81.25	93.82
5,7	67.18	91.80	67.20	92.13	1,2,3,4,6	79.72	95.16
4,6	62.41	91.89	62.58	92.63	1,2,3,5,7	77.97	94.36
2,4,5,7	67.80	92.78	68.25	**93.06**	1	**70.91**	92.30
2,4,5,7	67.74	92.78	**68.25**	**93.06**	3	65.42	89.74
2,4,5,7	67.74	92.78	68.25	93.06	6	**69.84**	**93.60**

Table 1: Quantitative results in terms of F1-score and accuracy (%) on R-Caustics.

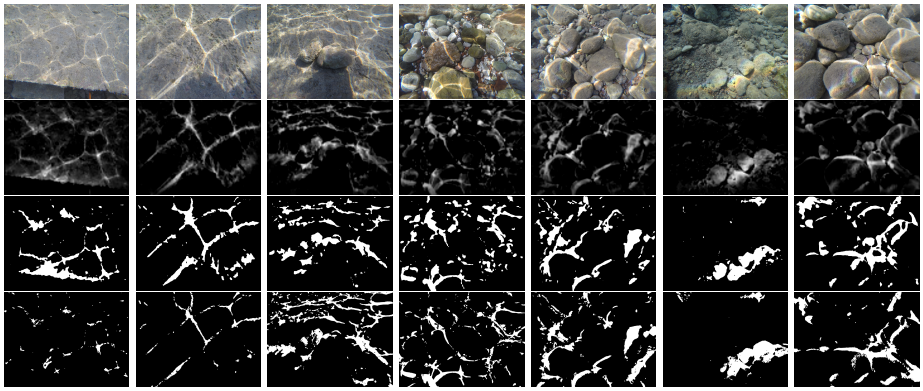


Fig. 4: Samples of zero-shot caustics prediction on the R-Caustics dataset. Input RGB images (top), predicted caustics (second row), thresholded predicted caustics (third row), and ground-truth caustics annotations (bottom row).

The images are resized to 352×576px resolution and passed to CausticsNet, obtaining the predicted caustics in a single forward pass with no training on the R-Caustics dataset in a zero-shot domain-transfer setting. The predicted caustics are resized back to the full resolution with bilinear interpolation and the difference between the input image and the image without the caustics is thresholded along the L channel after converting to LAB color space to obtain segmentation masks. These are compared against the ground-truth masks. As R-Caustics uses significantly varying thresholds between splits to obtain the ground-truth masks, we show results for one global (best) threshold, as well as for best thresholds chosen individually for each test split.

The method from Agrafiotis et al. [3] serves as a supervised baseline: the scenes of R-Caustics are split into training and testing splits, and a DeepLabV3+ network with a ResNet50 backbone is trained on patches of size 128×128 on the training set to predict the ground-truth caustics segmentation masks. Predictions on the test set are obtained moving this patch-based prediction with a stride of 32, thus requiring 10890 forward passes to get the caustics segmentation.

The results, comparing the F1-score and the total accuracy are displayed in Table 1. Despite being trained in a self-supervised manner, in a zero-shot setting, and in a domain transfer setting from coral reef scenes to rocky substrate images, CausticsNet is able to match the accuracy of the supervised network on R-Caustics. CausticsNet does so at a speedup of three orders of magnitude compared to the supervised method.

Samples from each of the scenes from the R-Caustic dataset and CausticsNet's prediction are displayed in Figure 4. Qualitatively, the predicted caustics match the true caustics closely. The masks also generally match, but in some cases (two leftmost columns) the ground-truth seems to be thresholded at a very conservative value. It is also visible that the ground-truth masks are computed with severely differing thresholds: for example, the first two columns use a very conservative threshold, whereas columns 3-7 are more lenient.

4.3 Downstream Image Matching

We evaluate the effectiveness of our method on image matching. For this, we use 6 held-out test videos from challenging and diverse reef scenes containing caustics and/or strong color effects from the water column. We extract 1000 frames from each video with 6 FPS at 352 × 608px and use COLMAP [64] for image feature extraction and (sequential) matching to then compare the number of geometrically verified matched keypoints. This evaluates whether a given underwater image restoration/enhancement method improves the quality with regards to the input in a scene-consistent manner.

We compare the unmodified RGB input to the images processed by CausticsNet and BackscatterNet, and compare against a suite of underwater image enhancement and restoration methods [75] that do not require raw input format. The number of geometrically verified matched keypoints, with the median of five independent COLMAP runs (whose behavior is randomized due to use of RANSAC [24]), as well as the average improvement over the unmodified RGB baseline (per scene) is shown in Table 2.

In 5 out of 6 scenes, our method leads to the highest number of keypoints matched. While all the competing methods lead to a worse performance than the unmodified inputs on many scenes, our method consistently improves the number of keypoints matched (with only one exception where three keypoints

Input↓ Scene →	1	2	3	4	5	6	Total	Avg. Improvement
RGB (unmodified)	114	371	744	596	78	2608	4512	1.0
CausticsNet	**347**	793	**979**	1011	**206**	2749	6085	1.98
BackscatterNet	111	462	748	644	87	2675	4572	1.07
CausticsNet + BackscatterNet	275	**821**	795	**1120**	194	2734	5696	1.85
DCP [30]	164	349	453	634	71	2101	3773	0.96
UDCP [20]	70	316	500	527	33	2180	3627	0.71
LowComplexityDCP [80]	179	355	773	528	98	**3020**	4954	1.14
NewOpticalModel [78]	113	406	464	484	46	2680	4194	0.86
GBdehazingRCorrection [43]	100	284	338	551	34	2030	3339	0.71
RayleighDistribution [1]	65	503	373	589	27	2061	3619	0.76
IBLA [58]	153	393	415	548	68	2301	3880	0.94
ULAP [66]	129	543	721	657	73	2458	4581	1.09
MIP [15]	139	478	600	584	79	2710	4591	1.06
RoWS [16]	74	399	468	422	66	2831	4260	0.83
Gamma Correction	138	482	526	642	113	2258	4160	1.1
RGHS [33]	121	442	604	592	102	2524	4386	1.05
CLAHE [61]	80	569	608	563	76	2643	4540	1.0
ICM [36]	139	456	532	631	98	2608	4465	1.08
UCM [37]	120	534	629	542	101	2607	4534	1.09
HE [34]	97	535	530	711	27	2355	4257	0.91

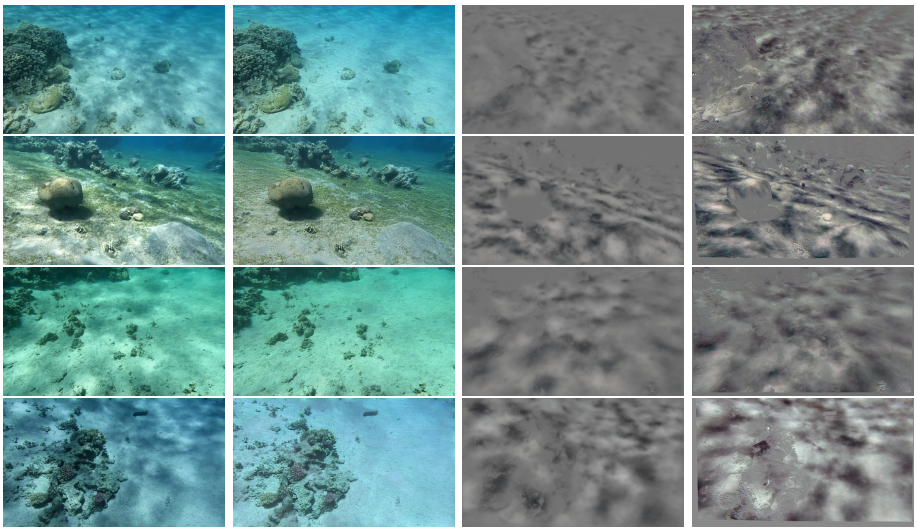
Table 2: Number of geometrically verified matched keypoints using COLMAP using images processed with various underwater image restoration/enhancement methods. Methods with less matched keypoints than the unmodified RGB input are shown in red. Best method is displayed in **bold**.

less were matched). In particular, CausticsNet leads to the highest number of total keypoints and the largest average improvement, followed by the combination of CausticsNet and BackscatterNet. Even though our method performs best, the total number of matched keypoints remains insufficient to create a full SfM reconstruction of all 1000 frames for any of the scenes: underwater scenes remains a challenging frontier of photogrammetry due to repetitive features and low-texture regions. A qualitative comparison of the different methods in this experiment is provided in the Appendix (Figure 3).

4.4 Qualitative Analysis

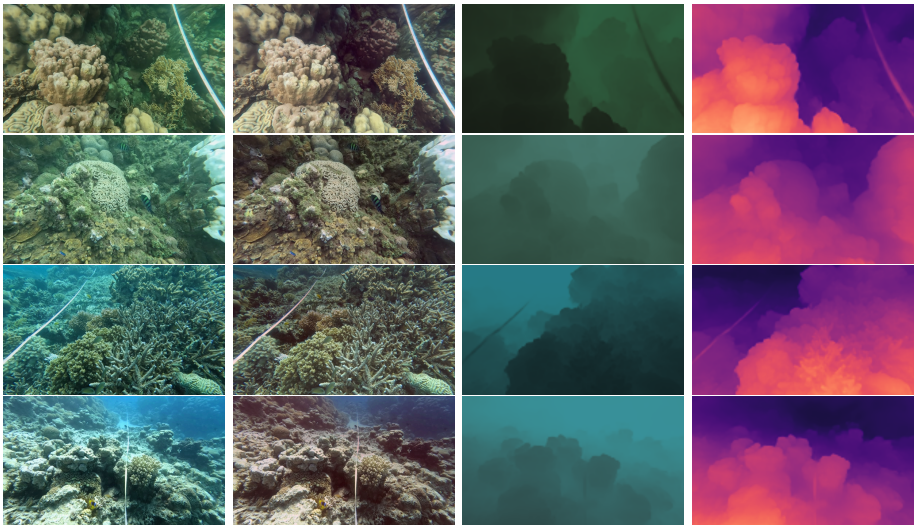
To evaluate the quality of BackscatterNet and CausticsNet, we show samples of held out validation videos. Figure 5 shows random samples from videos that contain caustics, as well as the cleaned versions and the target reprojection error with a randomly sampled overlapping frame. The caustics removal substantially alleviates the detrimental caustics and leads to a much better visual quality. CausticsNet is the first method which can alleviate caustics from a previously unseen image in a single forward pass.

Qualitative samples from held-out videos of BackscatterNet are shown in Figure 6, along with the predicted backscatter and the estimated depth from the monocular SLAM system. The visual quality is clearly improved by counteracting the green and blue tint from the water column. As the subtracted backscatter becomes more intense with increasing depth, the improved colors are particularly



(a) RGB Input **(b)** Remove Caustics **(c)** Caustics **(d)** Target

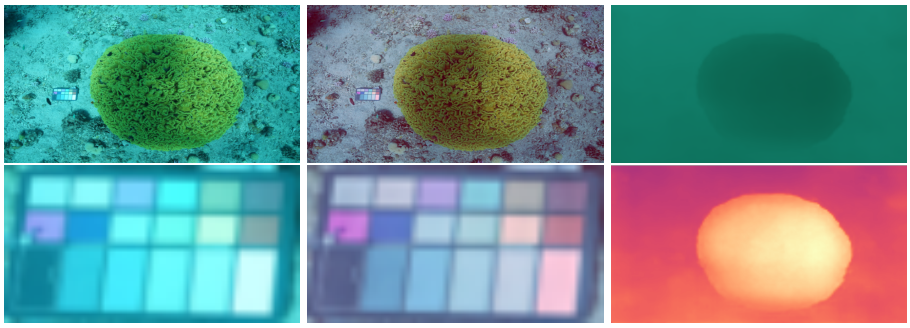
Fig. 5: Example input RGB images (a), the cleaned versions by CausticsNet (b), the predicted caustics (c), and the residual with a randomly selected overlapping frame (d), serving as the learning target.



(a) RGB Input **(b)** Descattered **(c)** Backscatter **(d)** Depth

Fig. 6: Input RGB images (a), color-restored images (b), estimated backscatter (c), estimated depth by learned monocular SLAM (d).

noticeable in further away regions. In none of the images, BackscatterNet leads to unrealistic colors or overblown contrasts, as common with most underwater image enhancement and some underwater image restoration methods. This is highlighted in a qualitative comparison of CausticsNet and BackscatterNet with all 16 methods compared in the image matching task in Appendix (Fig. 3). Almost all other methods lead to the caustics being even more present and exaggerated than in the input image. Furthermore, we observe that CausticsNet and BackscatterNet are highly video-consistent (see videos in supplementary material), even though they operate frame by frame.



(a) RGB Input **(b)** Descattered **(c)** Backscatter, Depth

Fig. 7: BackscatterNet in practice: an input image from Akkaynak and Treibitz [5] (left), the output of BackscatterNet (center), as well as the predicted backscatter and depth (right). The colors of the hard substrate and in the color chart are improved.

Finally, we apply BackscatterNet on an image from a coral reef scene containing a color chart from Akkaynak and Treibitz [5] in Figure 7. The depth estimation and resulting backscatter estimation work reliably, and the color chart is enhanced substantially, in particular in the red, yellow, green and purple tones. The green-blue tint of the gray tones is improved, but not completely alleviated, and the white patch has a slight red tint instead of a blue tint.

5 Discussion & Future Work

We present an unsupervised neural-network method for single-image underwater restoration which consists of CausticsNet and BackscatterNet, which effectively remove caustics and backscatter from single images in a single forward pass. Our method does not require raw image input format and is competitive against supervised caustics segmentation methods in a zero-shot domain-transfer setting. Compared against 16 underwater image restoration/enhancement methods, our method is the most effective method in a downstream image matching task, and the only one yielding consistent improvements across scenes.

We use a deep monocular SLAM system to formulate the objective losses for our method. In this paper, we constrain ourselves to first training the SLAM system and then fixing it for our method. The interactions of jointly training the SLAM system with BackscatterNet and CausticsNet are left as future work. Furthermore, the output resolution of BackscatterNet and CausticsNet is bounded by the resolution of the depth maps produced by the monocular SLAM system. We expect future developments in deep monocular SLAM to produce more accurate poses and depths at higher input image resolutions, and to better handle the diffraction in underwater scenarios. This should translate to clear improvements in our general method.

BackscatterNet achieves excellent and consistent color restoration by estimating the backscatter via parameters of the underwater image formation model, but this is only an approximation to the physical model, as it operates on RGB images pre-processed by the camera instead of raw images. Future work could extend BackscatterNet to a more flexible backscatter estimation which could follow the physics even in spite of pre-processing via histogram stretching and white balancing. Furthermore, we only model the additive backscatter signal, whereas in practice (in particular in deep waters), the attenuation may have a dominant effect on color loss [4, 5]. A model that includes learned attenuation using a SLAM objective could further improve results.

BackscatterNet and CausticsNet are remarkably video-consistent even though they are applied frame by frame. Nonetheless, the method could be extended to take multiple frames into consideration for increased scene consistency. Furthermore, the predictions of CausticsNet could re-used in a generative setting, allowing scene-conditioned caustics generation, which has many applications in computer graphics and the animation industry.

Bibliography

[1] Ahmad Shahrizan Abdul Ghani and Nor Ashidi Mat Isa. Underwater image quality enhancement through composition of dual-intensity images and rayleigh-stretching. *SpringerPlus*, 3(1):1–14, 2014.

[2] Panagiotis Agrafiotis, Dimitrios Skarlatos, Timothy Forbes, Charalambos Poullis, Margarita Skamantzari, and Andreas Georgopoulos. Underwater photogrammetry in very shallow waters: main challenges and caustics effect removal. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42:15–22, 2018.

[3] Panagiotis Agrafiotis, Konstantinos Karantzalos, and Andreas Georgopoulos. Seafloor-invariant caustics removal from underwater imagery. *IEEE Journal of Oceanic Engineering*, 2023.

[4] Derya Akkaynak and Tali Treibitz. A revised underwater image formation model. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6723–6732, 2018.

[5] 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*, pages 1682–1691, 2019.

[6] Shlomi Amitai, Itzik Klein, and Tali Treibitz. Self-supervised monocular depth underwater. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1098–1104. IEEE, 2023.

[7] Josep Aulinas, Marc Carreras, Xavier Llado, Joaquim Salvi, Rafael Garcia, Ricard Prados, and Yvan R Petillot. Feature extraction for underwater visual slam. In *OCEANS 2011 IEEE-Spain*, pages 1–7. IEEE, 2011.

[8] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.

[9] Daniel TI Bayley and Andrew OM Mogg. A protocol for the large-scale analysis of reefs using structure from motion photogrammetry. *Methods in Ecology and Evolution*, 11(11):1410–1420, 2020.

[10] Dana Berman, Deborah Levy, Shai Avidan, and Tali Treibitz. Underwater single image color restoration using haze-lines and a new quantitative dataset. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2822–2837, 2020.

[11] Jia-Wang Bian, Huangying Zhan, Naiyan Wang, Zhichao Li, Le Zhang, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth learning from video. *International Journal of Computer Vision*, 129(9):2548–2564, 2021.

[12] Henryk Blasinski and Joyce Farrell. A three parameter underwater image formation model. *Electronic Imaging*, 2016(18):1–8, 2016.

[13] Mitch Bryson, Matthew Johnson-Roberson, Oscar Pizarro, and Stefan B Williams. True color correction of autonomous underwater vehicle imagery. *Journal of Field Robotics*, 33(6):853–874, 2016.

[14] JHR Burns, D Delparte, RD Gates, and M Takabayashi. Integrating structure-from-motion photogrammetry with geospatial software as a novel technique for quantifying 3d ecological characteristics of coral reefs. *PeerJ*, 3:e1077, 2015.

[15] Nicholas Carlevaris-Bianco, Anush Mohan, and Ryan M Eustice. Initial results in underwater single image dehazing. In *Oceans 2010 Mts/IEEE Seattle*, pages 1–8. IEEE, 2010.

[16] Liu Chao and Meng Wang. Removal of water scattering. In *2010 2nd international conference on computer engineering and technology*, volume 2, pages V2–35. IEEE, 2010.

[17] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[18] John Y Chiang and Ying-Ching Chen. Underwater image enhancement by wavelength compensation and dehazing. *IEEE transactions on image processing*, 21(4):1756–1769, 2011.

[19] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[20] Paul Drews, Erickson Nascimento, Filipe Moraes, Silvia Botelho, and Mario Campos. Transmission estimation in underwater single images. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 825–830, 2013.

[21] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

[22] Jiading Fang, Igor Vasiljevic, Vitor Guizilini, Rares Ambrus, Greg Shakhnarovich, Adrien Gaidon, and Matthew R Walter. Self-supervised camera self-calibration from video. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 8468–8475. IEEE, 2022.

[23] Renata Ferrari, Liam Lachs, Daniel R Pygas, Adriana Humanes, Brigitte Sommer, Will F Figueira, Alasdair J Edwards, John C Bythell, and James R Guest. Photogrammetry as a tool to improve ecosystem restoration. *Trends in Ecology & Evolution*, 36(12):1093–1101, 2021.

[24] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[25] Timothy Forbes, Mark Goldsmith, Sudhir Mudur, and Charalambos Poullis. Deepcaustics: Classification and removal of caustics from underwater imagery. *IEEE Journal of Oceanic Engineering*, 44(3):728–738, 2018.

[26] 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*, pages 270–279, 2017.

[27] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019.

[28] Nuno Gracias, Shahriar Negahdaripour, Laszlo Neumann, Ricard Prados, and Rafael Garcia. A motion compensated filtering approach to remove sunlight flicker in shallow water images. In *OCEANS 2008*, pages 1–7. IEEE, 2008.

[29] Vitor Guizilini, Rareș Ambruș, Dian Chen, Sergey Zakharov, and Adrien Gaidon. Multi-frame self-supervised depth with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 160–170, 2022.

[30] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010.

[31] Jon Henderson, Oscar Pizarro, Matthew Johnson-Roberson, and Ian Mahon. Mapping submerged archaeological sites using stereo-vision photogrammetry. *International Journal of Nautical Archaeology*, 42(2):243–256, 2013.

[32] Guojia Hou, Yuxuan Li, Huan Yang, Kunqian Li, and Zhenkuan Pan. Uid2021: an underwater image dataset for evaluation of no-reference quality assessment metrics. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(4):1–24, 2023.

[33] Dongmei Huang, Yan Wang, Wei Song, Jean Sequeira, and Sébastien Mavromatis. Shallow-water image enhancement using relative global histogram stretching based on adaptive parameter acquisition. In *MultiMedia Modeling: 24th International Conference, MMM 2018, Bangkok, Thailand, February 5-7, 2018, Proceedings, Part I 24*, pages 453–465. Springer, 2018.

[34] Robert Hummel. Image enhancement by histogram transformation. *Unknown*, 1975.

[35] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.

[36] Kashif Iqbal, Rosalina Abdul Salam, Azam Osman, and Abdullah Zawawi Talib. Underwater image enhancement using an integrated colour model. *IAENG International Journal of computer science*, 34(2), 2007.

[37] Kashif Iqbal, Michael Odetayo, Anne James, Rosalina Abdul Salam, and Abdullah Zawawi Hj Talib. Enhancing the low quality images using unsupervised colour correction method. In *2010 IEEE International Conference on Systems, Man and Cybernetics*, pages 1703–1709. IEEE, 2010.

[38] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.

[39] Jules S Jaffe. Computer modeling and the design of optimal underwater imaging systems. *IEEE Journal of Oceanic Engineering*, 15(2):101–111, 1990.

[40] Qiuping Jiang, Yuese Gu, Chongyi Li, Runmin Cong, and Feng Shao. Underwater image enhancement quality evaluation: Benchmark dataset and objective metric. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):5959–5974, 2022.

[41] Ayoung Kim and Ryan M Eustice. Real-time visual slam for autonomous underwater hull inspection using visual saliency. *IEEE Transactions on Robotics*, 29(3):719–733, 2013.

[42] Deborah Levy, Amit Peleg, Naama Pearl, Dan Rosenbaum, Derya Akkaynak, Simon Korman, and Tali Treibitz. Seathru-nerf: Neural radiance fields in scattering media. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 56–65, 2023.

[43] Chongyi Li, Jichang Quo, Yanwei Pang, Shanji Chen, and Jian Wang. Single underwater image restoration by blue-green channels dehazing and red channel correction. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1731–1735. IEEE, 2016.

[44] Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao. An underwater image enhancement benchmark dataset and beyond. *IEEE Transactions on Image Processing*, 29:4376–4389, 2019.

[45] Chongyi Li, Saeed Anwar, and Fatih Porikli. Underwater scene prior inspired deep underwater image and video enhancement. *Pattern Recognition*, 98:107038, 2020.

[46] Zheng Liang, Xueyan Ding, Yafei Wang, Xiaohong Yan, and Xianping Fu. Gudcp: Generalization of underwater dark channel prior for underwater image restoration. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4879–4884, 2021.

[47] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[48] Guilhem Marre, Julie Deter, Florian Holon, Pierre Boissery, and Sandra Luque. Fine-scale automatic mapping of living posidonia oceanica seagrass beds with underwater photogrammetry. *Marine Ecology Progress Series*, 643:63–74, 2020.

[49] 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*, pages 4040–4048, 2016.

[50] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[51] Fabio Menna, Erica Nocerino, Mohammad Motasem Nawaf, Julien Seinturier, Alessandro Torresani, Pierre Drap, Fabio Remondino, and Bertrand Chemisky. Towards real-time underwater photogrammetry for subsea

metrology applications. In *OCEANS 2019-Marseille*, pages 1–10. IEEE, 2019.

[52] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[53] Curtis D Mobley. Light and water: radiative transfer in natural waters. *Academic Press*, 1994.

[54] David Nakath, Mengkun She, Yifan Song, and Kevin Köser. In-situ joint light and medium estimation for underwater color restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3731–3740, 2021.

[55] Srinivasa G Narasimhan and Shree K Nayar. Vision and the atmosphere. *International journal of computer vision*, 48:233–254, 2002.

[56] Soo-Chang Pei and Chia-Yi Chen. Underwater images enhancement by revised underwater images formation model. *IEEE Access*, 10:108817–108831, 2022.

[57] Lintao Peng, Chunli Zhu, and Liheng Bian. U-shape transformer for underwater image enhancement. *IEEE Transactions on Image Processing*, 2023.

[58] Yan-Tsung Peng and Pamela C Cosman. Underwater image restoration based on image blurriness and light absorption. *IEEE transactions on image processing*, 26(4):1579–1594, 2017.

[59] Paola Piazza, Vonda Cummings, Alice Guzzi, Ian Hawes, Andrew Lohrer, Simone Marini, Peter Marriott, Fabio Menna, Erica Nocerino, Andrea Peirano, et al. Underwater photogrammetry in antarctica: Long-term observations in benthic ecosystems and legacy data rescue. *Polar Biology*, 42: 1061–1079, 2019.

[60] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12240–12249, 2019.

[61] Ali M Reza. Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement. *Journal of VLSI signal processing systems for signal, image and video technology*, 38:35–44, 2004.

[62] J. Sauder, G. Banc-Praudi, A. Meibom, and **D. Tuia**. Scalable semantic 3d mapping of coral reefs with deep learning. *Methods Ecol. Evol.*, in press.

[63] Yoav Y Schechner and Nir Karpel. Clear underwater vision. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I. IEEE, 2004.

[64] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[65] Kurran Singh, Nick Rypkema, and John Leonard. Attention-based self-supervised hierarchical semantic segmentation for underwater imagery. In *OCEANS 2023-Limerick*, pages 1–6. IEEE, 2023.

[66] Wei Song, Yan Wang, Dongmei Huang, and Dian Tjondronegoro. A rapid scene depth estimation model based on underwater light attenuation prior for underwater image restoration. In *Advances in Multimedia Information Processing–PCM 2018: 19th Pacific-Rim Conference on Multimedia, Hefei, China, September 21-22, 2018, Proceedings, Part I 19*, pages 678–688. Springer, 2018.

[67] Yohay Swirski and Yoav Y Schechner. 3deflicker from motion. In *IEEE International Conference on Computational Photography (ICCP)*, pages 1–9. IEEE, 2013.

[68] Zachary Teed and Jia Deng. Deepv2d: Video to depth with differentiable structure from motion. *arXiv preprint arXiv:1812.04605*, 2018.

[69] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.

[70] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in Neural Information Processing Systems*, 34:16558–16569, 2021.

[71] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5038–5047, 2017.

[72] C Upstill. Light caustics from rippling water. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 365(1720):95–104, 1979.

[73] Nisha Varghese, Ashish Kumar, and AN Rajagopalan. Self-supervised monocular underwater depth recovery, image restoration, and a real-sea video dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12248–12258, 2023.

[74] Wenshan Wang, Yaoyu Hu, and Sebastian Scherer. Tartanvo: A generalizable learning-based vo. In *Conference on Robot Learning*, pages 1761–1772. PMLR, 2021.

[75] Yan Wang, Wei Song, Giancarlo Fortino, Li-Zhe Qi, Wenqiang Zhang, and Antonio Liotta. An experimental-based review of image enhancement and image restoration methods for underwater imaging. *IEEE access*, 7:140233–140251, 2019.

[76] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[77] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1164–1174, 2021.

[78] Haocheng Wen, Yonghong Tian, Tiejun Huang, and Wen Gao. Single underwater image enhancement with a new optical model. In *2013 IEEE In-*

ternational Symposium on Circuits and Systems (ISCAS), pages 753–756. IEEE, 2013.

[79] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[80] Hung-Yu Yang, Pei-Yin Chen, Chien-Chuan Huang, Ya-Zhu Zhuang, and Yeu-Horng Shiau. Low complexity underwater image enhancement based on dark channel prior. In *2011 Second International Conference on Innovations in Bio-inspired Computing and Applications*, pages 17–20. IEEE, 2011.

[81] Xuewen Yang, Xing Zhang, Nan Wang, Guoling Xin, and Wenjie Hu. Underwater self-supervised depth estimation. *Neurocomputing*, 514:362–373, 2022.

[82] 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*, pages 1983–1992, 2018.

[83] Matan Yuval, Iñigo Alonso, Gal Eyal, Dan Tchernov, Yossi Loya, Ana C Murillo, and Tali Treibitz. Repeatable semantic reef-mapping through photogrammetry and label-augmentation. *Remote Sensing*, 13(4):659, 2021.

[84] Mohua Zhang and Jianhua Peng. Underwater image restoration based on a new underwater image formation model. *IEEE Access*, 6:58634–58644, 2018.

[85] Huizhong Zhou, Benjamin Ummenhofer, and Thomas Brox. Deeptam: Deep tracking and mapping. In *Proceedings of the European conference on computer vision (ECCV)*, pages 822–838, 2018.

[86] 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*, pages 1851–1858, 2017.