CE-VAE: Capsule Enhanced Variational AutoEncoder for Underwater Image Enhancement

Rita Pucci*
Naturalis Biodiversity Center
Leiden, Netherlands (NL)
rita.pucci@naturalis.nl

Niki Martinel*
University of Udine
Udine, Italy (IT)
niki.martinel@uniud.it

Figure 1. Results of our approach for underwater compressed image reconstruction and enhancement. The first row shows the effects of water light refraction in underwater images introducing blurriness and cold greenish or bluish tone (among others) issues. The second row shows the enhanced results obtained by the proposed approach.

Abstract

Unmanned underwater image analysis for marine monitoring faces two key challenges: (i) degraded image quality due to light attenuation and (ii) hardware storage constraints limiting high-resolution image collection. Existing methods primarily address image enhancement with approaches that hinge on storing the full-size input. In contrast, we introduce the Capsule Enhanced Variational AutoEncoder (CE-VAE), a novel architecture designed to efficiently compress and enhance degraded underwater images. Our attention-aware image encoder can project the input image onto a latent space representation while being able to run online on a remote device. The only information that needs to be stored on the device or sent to a beacon is a compressed representation. There is a dual-decoder module that performs offline, full-size enhanced image generation. One branch reconstructs spatial details from the compressed latent space, while the second branch utilizes a capsule-clustering layer to capture entity-level structures and complex spatial relationships. This parallel decoding strategy enables the model to balance fine-detail preservation with context-aware enhancements. CE-VAE achieves state-of-the-art performance in underwater image enhancement on six benchmark datasets, providing up to 3× higher compression efficiency than existing approaches. Code available at https://github.com/iN1k1/ce-vae-underwater-image-enhancement.

1. Introduction

Marine exploration is crucial for monitoring and protecting underwater ecosystems, with image analysis providing essential data for scientific research and environmental conservation. Recent advancements in unmanned and au-

*Equal contribution

tonomous visual sensing systems have enhanced our ability to capture environmental data, enabling researchers to monitor [48], explore [50], and analyze [6] ocean depths while minimizing human risk.

However, underwater imagery poses significant challenges, such as severe color distortion and loss of detail due to light absorption, resulting in hazy images with greenish or bluish tinges (Fig. 1, first row). Addressing these degradation issues is critical for various marine applications, including ecological studies, underwater archaeology, and marine resource management.

In addition to image quality challenges, hardware storage limitations are a significant constraint for unmanned devices deployed in long-duration missions. These systems must operate autonomously for extended periods with limited capacity for storing high-resolution imagery [20, 21]. Efficient image compression becomes essential to allow longer data collection campaigns without sacrificing the ability to perform high-quality image reconstruction and analysis offline.

Addressing both image degradation and storage efficiency issues is of paramount importance for autonomous marine monitoring systems, yet our community has mostly focused on the former problem. Existing image enhancement methods can be categorized into: (i) traditional image processing techniques and (ii) machine learning-based methods. The former, including non-physics-based [11, 25] and physics-based [14, 15, 31] approaches, often lack generalization across diverse underwater environments. The latter [10, 12, 15, 17, 18, 34, 53, 57] offer superior generalization but are computationally intensive, limiting their integration into autonomous systems.

We introduce the Capsule Enhanced Variational AutoEncoder (`CE-VAE`), a novel architecture that synergizes the generative power of variational autoencoders with the strengths of capsule networks in capturing high-level image semantics [7, 38–41, 46]. `CE-VAE` comprises an encoder, a novel capsule layer, and a dual-decoder module, carefully designed to tackle the complex challenges of underwater image enhancement.

The novel attention-aware online encoder is designed to project the input image onto a highly compact low-dimensional latent representation. This allows us to (i) achieve efficient storage of underwater imagery while (ii) also forcing the model to learn a compact, informative representation of the image, ensuring that irrelevant information and noise are discarded.

For offline full-size image enhancement, our architecture introduces a dual-decoder module, each designed with a specific role to ensure high-quality image reconstruction. The first decoder reconstructs the enhanced image from the compressed latent space, ensuring the preservation of fine spatial details without requiring the full-size input. The second decoder leverages high-level features captured by a capsule layer, which provides a more abstract understanding of the scene. The use of capsules is motivated by their ability to capture spatial hierarchies and part-whole relationships, making them robust to distortions in underwater imagery. Capsules also dynamically cluster similar features, enhancing the model's generalization across diverse underwater environments. This dual-decoder design balances the retention of fine details with the ability to model complex structures, leading to superior image enhancement.

The key contributions of this work are:

- A novel attention-aware encoder that projects input images onto a highly compressed latent space, enabling efficient storage and real-time processing for underwater image enhancement.
- A dual-decoder architecture that reconstructs enhanced images by leveraging both compressed latent representations and high-level capsule features, balancing spatial detail preservation with context-aware reconstruction.
- State-of-the-art performance on six benchmark datasets, achieving superior image quality and generalization while offering $3\times$ improved storage efficiency by eliminating the need to store full-size input images.

2. Related Works

Traditional methods focus on the estimation of global background and water light transmission to perform image enhancement. In [2, 4], independent image processing steps have been proposed to correct non-uniform illumination, suppress noise, enhance contrast, and adjust colors. Other methods introduced edge detection operations to implement object-edge preservation during filtering operations for color enhancement [29]. In [24], it has been observed that the image channels are affected differently by the disruption of light: red colors are lost after a few meters from the surface while green and blue are more persistent. These differences introduced enhancement methods that act differently on each color channel and sacrifice generalization in favour of ad-hoc filters based on environmental parameters [33, 55]. Other approaches estimated the global background light parameters [33, 36] to apply specific color corrections (*i.e.*, to reduce the blueish and greenish effects). These models use the principles of light and color physics to account for various underwater conditions. Despite being more accurate, their application is limited due to the challenges of obtaining all the necessary variables that impact underwater footage. Efforts have been made to improve the estimation of the global background light [1] at the cost of increasing algorithm complexity and overfitting experimental data with poor generalization on new test data.

Machine learning-based methods for underwater im-

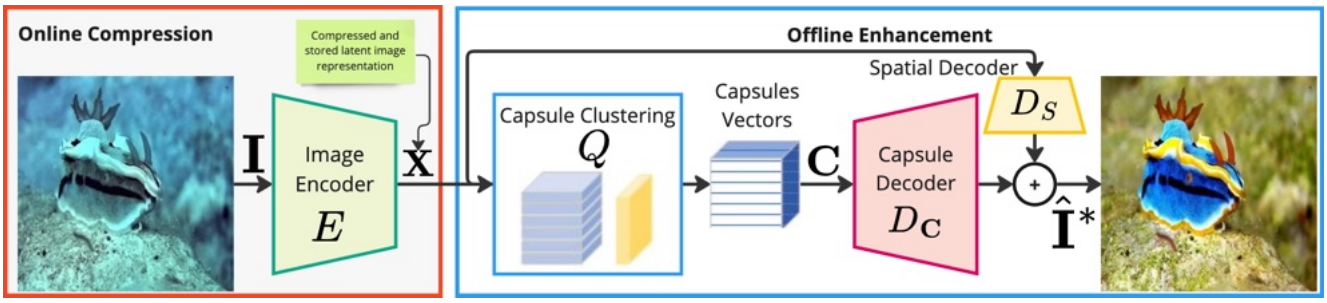

Figure 2. Proposed CE−VAE architecture with the new capsule vector latent space clusterization mechanism.

age enhancement made extensive use of a U-Net-like structure [44] to enhance the input image while preserving the spatial information and relationship between objects. Skip connections are often used to propagate the raw inputs to the final layers to preserve spatial relationships [23, 51] also with special attention and pooling layers [42]. Other methods explored the emerging application of Transformers via channel-wise and spatial-wise attention layers [35] or through customized transformer blocks leveraging both the frequency and the spatial-domains as self-attention inputs [22]. Generative Adversarial Networks (GANs) training schemes have also been explored for the task [12] along with approaches improving the information transfer between the encoder and decoder via multiscale dense blocks [18] or hierarchical attentions modules [13].

Our approach falls in the latter category. In contrast with such methods, we propose a novel architecture that removes the need for skip connections between the raw input and decoder layers. Our encoder projects the full-size input image into a highly compact, low-dimensional latent space that captures all relevant information for both enhancement and reconstruction. The dual-decoder module operates exclusively on this latent representation, fully independent of the full-size raw input. This design allows for real-time feature extraction during data collection, enabling efficient storage by retaining only the latent compressed representation, which is crucial for resource-constrained environments.

3. Proposed Method

Figure 2 illustrates our architecture, composed of two main phases. The online compression phase features an image encoder (E) that models the degraded input image $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$ and compresses it into a latent feature space carrying relevant information for enhancement. The compressed representation can then be stored and later used in the offline enhancement phase. This includes a capsule clustering module (Q), capturing entity-level features, followed by the capsule decoder ($D_{\mathbf{C}}$) and the spatial decoder (D_S) that jointly collaborate to generate the full-size enhanced image, *i.e.*, $\widehat{\mathbf{I}}^* \in \mathbb{R}^{3\times H\times W}$.

3.1. Image Encoder (E)

Our encoder architecture is designed to extract a compact yet informative latent representation while preserving crucial spatial information. The design follows a hierarchical structure that balances computational efficiency with feature richness.

We begin by computing $\mathbf{H}_0 = \texttt{Conv2D}_{3\times3}(\mathbf{I})$ to capture low-level features such as edges, textures, and color variations, which are often distorted in underwater environments. This is followed by N encoding blocks, each comprising a residual block and a self-attention mechanism with skip connections. The residual block computes $\mathbf{H}_l^{\text{res}}$ where $l \in [1, N]$

$$\mathbf{H}_l^{\text{res}} = \texttt{ResnetBlock}(\mathbf{H}_{l-1}) \in \mathbb{R}^{C_l\times H_l\times W_l} \quad (1)$$

ensuring effective information propagation through deeper layers for preserving and enhancing subtle underwater textures and colors, while mitigating vanishing gradients.

The subsequent self-attention mechanism further refines the extracted features to get

$$\mathbf{H}_l = \mathbf{H}_l^{\text{res}} + \texttt{SelfAttention}(\mathbf{H}_l^{\text{res}}) \quad (2)$$

The attention block allows the model to focus on salient regions in the feature map, which is particularly important for underwater images where certain areas may be more affected by scattering, absorption, or color distortion than others.

Between every two encoding blocks, we also add a Conv2D to half feature resolution spatial dimensions, optimizing computational efficiency while allowing the model to capture high-level abstract features. The balance between resolution and abstraction is essential for processing large underwater images efficiently while preserving critical information about the global color cast and lighting conditions.

At the output of the N encoder blocks, we add a convolutional block-computing

$$\mathbf{X} = \texttt{Conv2D}_{3\times3}(\texttt{Swish}(\texttt{BN}(\mathbf{H}_N))) \in \mathbb{R}^{C_X\times H_X\times W_X} \quad (3)$$

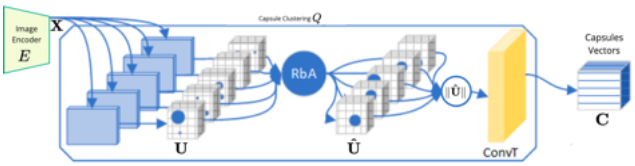

Figure 3. Proposed capsule vector clustering approach. It consists of a capsule layer and a convolutional transpose layer. The capsules extract $\mathbf{U}$ features which are clusterized by the RbA procedure, to obtain $\hat{\mathbf{U}}$. We aggregate the matrices and upsample them by a transposed convolution layer.

that refines the learned representation and produces the desired compact yet informative latent space.

3.2. Capsule Clustering (Q)

Following the encoder, the capsule layer processes the compressed latent representation to model entity-level relationships within the image (see Figure 3). Given L as the first capsule layer and $L+1$ as the consecutive one. First, β_L parallel convolutional layers (Conv2D) process the encoder output $\mathbf{X}$, generating a tensor $\mathbf{U} \in \mathbb{R}^{\beta_L \times C_\mathbf{U} \times H_\mathbf{U} \times W_\mathbf{U}}$, where $C_\mathbf{U}$ is the number of channels and $H_\mathbf{U}, W_\mathbf{U}$ are the spatial dimensions. For each spatial location, $\mathbf{u}_i \in \mathbb{R}^{C_\mathbf{U}}$ represents the output of capsule i at level L. Capsule $j \in [0, k]$ at the next level, $L+1$, receives information from all capsules at L, and computes the affine transformation of $\mathbf{u}_i$:

$$\hat{\mathbf{u}}_{j|i} = \mathbf{W}_{ij}\mathbf{u}_i \tag{4}$$

where $\mathbf{W}_{ij} \in \mathbb{R}^{C_\mathbf{U} \times C_{\hat{\mathbf{U}}}}$ is a weight matrix defining how capsule i contributes to capsule j. The vector $\hat{\mathbf{u}}_{j|i}$ estimates the relevance of capsule i for activating capsule j. Since not all parent capsules are equally important for higher-level entities, we apply a coupling coefficient c_{ij} to weigh their contributions. The softmax function is used to compute c_{ij}:

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})} \tag{5}$$

where b_{ij} is iteratively updated during the Routing-by-Agreement (RbA) process. The next step is to compute a weighted sum of the prediction vectors:

$$\mathbf{s}_j = \sum_i c_{ij}\hat{\mathbf{u}}_{j|i} \tag{6}$$

The squashing function is then applied to obtain the activity vector $\mathbf{v}_j$, representing the likelihood of entity presence:

$$\mathbf{v}_j = \frac{\|\mathbf{s}_j\|^2}{1+\|\mathbf{s}_j\|^2}\frac{\mathbf{s}_j}{\|\mathbf{s}_j\|} \tag{7}$$

Finally, b_{ij} is updated based on the agreement between $\mathbf{v}_j$ and $\hat{\mathbf{u}}_{j|i}$, where agreement strengthens the contribution of capsule i to capsule j.

The activity vector $\mathbf{v}_j$ carries information about the presence of an entity using a probability representation that removes information about the precise (pixel-level) spatial location of each relevant object. Such information, which is very relevant for precise enhanced image reconstruction, is however contained in the related prediction vectors. To bring the spatial information to the next layers, while preserving the information about entities, we first weight every prediction vector from level L (*i.e.*, $\hat{\mathbf{u}}_{j|i}$) by the corresponding coupling coefficient c_{ij} (estimated at the last iteration of the RbA procedure) to collectively obtain $\hat{\mathbf{U}} \in \mathbb{R}^{\beta_{L+1} \times C_{\hat{\mathbf{U}}} \times H_{\hat{\mathbf{U}}} \times W_{\hat{\mathbf{U}}}}$. Then, entity presence at a specific location is then captured through the ℓ_2-norm computation over β_{L+1}. Finally, a TransposedConv layer is exploited to obtain the capsule vectors $\mathbf{C} \in \mathbb{R}^{C_\mathbf{C} \times H_\mathbf{U} \times W_\mathbf{U}}$.

The motivation for using the capsule layer in this setting lies in its ability to capture spatial hierarchies and model part-to-whole relationships, which are key to representing entities within degraded underwater images. CNNs struggle to maintain precise entity-based information across layers, often losing finer details necessary for effective reconstruction. By employing the capsule networks mechanism while preserving the spatial pixel relationships via $\hat{\mathbf{U}}$, CE-VAE maintains both spatial and entity-level information. This is particularly important in underwater imagery, where preserving the structure of small objects and understanding spatial relationships is essential for accurate enhancement.

3.3. Decoding

To reconstruct the enhanced image, we introduce two parallel decoders: the capsule decoder ($D_\mathbf{C}$) and the spatial decoder (D_S).

$D_\mathbf{C}$ reconstructs the image leveraging the information about the presence of entities identified by the capsules vectors, *i.e.*, $\mathbf{C}$. It increases the input spatial feature map resolution to produce $\widehat{\mathbf{I}}^*_{D_\mathbf{C}} \in \mathbb{R}^{3 \times H \times W}$ by a sequence of 4 blocks, each consisting of a ResnetBlock and an UpSampleBlock [9]. This decoder works on an input that contains information about the presence of entities in the image but might not contain precise (*i.e.*, to pixel-level) information about their displacement. Since the enhancement must generate an output that preserves all the spatial details but removes the effects of underwater degradation, this information would be very relevant for reconstruction.

To mitigate such a limitation, we introduced the spatial decoder (D_S). This takes as input the low-resolution feature map, $\mathbf{X}$, that preserves all the spatial details about the input, and gradually increases its resolution to match the input image size, with a process that resembles image super-resolution works. Similarly to $D_\mathbf{C}$, such a module is com-

posed of 4 blocks, each consisting of a `TransposedConv` and a `ResnetBlock`, that emit $\widehat{\mathbf{I}}^*_{D_S} \in \mathbb{R}^{3\times H\times W}$.

The decoders produce the model output $\widehat{\mathbf{I}}^* = \widehat{\mathbf{I}}^*_{D_\mathbf{C}} + \widehat{\mathbf{I}}^*_{D_S}$.

3.4. Optimization Objective

The proposed architecture utilizes only the compressed latent representation extracted by E for both decoders D_C and D_S. This design enables edge computation in E of $\mathbf{X}$, and subsequent offline reconstruction via D_C and D_S. This approach effectively functions as a data compression method, facilitating extended underwater acquisition campaigns. However, maximizing compression efficiency necessitates learning a highly informative latent image representation. To achieve this, we introduce a composite loss function that encapsulates the essential aspects of underwater image enhancement: preservation of the spatial structure, improved color perception with artifact suppression, and overall image realism.

Reconstruction Loss. To ensure spatial coherence between the noise-free ground truth (*i.e.*, $\mathbf{I}^*$) and reconstructed image (*i.e.*, $\widehat{\mathbf{I}}^*$), we compute:

$$\mathcal{L}_{rec} = |\mathbf{I}^* - \widehat{\mathbf{I}}^*| \tag{8}$$

However, the model may generate blurry or overly smooth images, as the network tries to minimize the pixel-wise differences without necessarily capturing the high-level features and structures of the original image.

Perceptual Loss. We employ the Learned Perceptual Image Patch Similarity (LPIPS) metric, which has been shown to correlate well with human judgments of image quality [56]. This computes:

$$\mathcal{L}_{lpips} = ||\phi(\mathbf{I}^*) - \phi(\widehat{\mathbf{I}}^*)||_2 \tag{9}$$

where $\phi(\cdot)$ denotes a pre-trained model extracting features relevant to human perception.

Adversarial Loss. To further improve the realism of the generated images, we adopted an adversarial training procedure with a patch-based discriminator $\psi(\cdot)$ [19]. We follow the original formulation of [9] to define

$$\mathcal{L}_{GAN} = \lambda \left(log\psi(\mathbf{I}^*) + log(1 - \psi(\widehat{\mathbf{I}}^*)) \right) \tag{10}$$

where its contribution to the final objective is controlled by

$$\lambda = \frac{\nabla_{D_\mathbf{C}}(\mathcal{L}_{rec})}{\nabla_{D_\mathbf{C}}(\mathcal{L}_{GAN}) + \delta} \tag{11}$$

$\nabla_{D_\mathbf{C}}(\cdot)$ is the gradient of its input at the last layer of $D_\mathbf{C}$, and δ is used for numerical stability [9].

Structural Similarity Loss. To address the structural distortions common in underwater image degradation, we consider the Structural Similarity Index Measure (SSIM) [49] loss function:

$$\mathcal{L}_{SSIM} = \frac{1}{M}\sum_{i=1}^{M} \frac{2\mu_{\mathbf{I}^*_i}\mu_{\widehat{\mathbf{I}}^*_i} + \kappa_1}{\mu^2_{\mathbf{I}^*_i} + \mu^2_{\widehat{\mathbf{I}}^*_i} + \kappa_1} \frac{2\sigma_{\mathbf{I}^*_i\widehat{\mathbf{I}}^*_i} + \kappa_2}{\sigma^2_{\mathbf{I}^*_i} + \sigma^2_{\widehat{\mathbf{I}}^*_i} + \kappa_2} \tag{12}$$

where $\mathbf{I}^*_i$ and $\widehat{\mathbf{I}}^*_i$ are 11×11 non-overlapping image patches, μ and σ represent the mean and standard deviation operators. κ_1 and κ_2 are small constants added for stability.

Combined Loss. Our optimization objective is

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{lpips} + \mathcal{L}_{GAN} + \mathcal{L}_{SSIM} \tag{13}$$

4. Experimental Results

4.1. Datasets

Training. For a fair comparison with [10,17,18], we pretrained our model on 1.3M ImageNet training split [45], to reconstruct the input image (*i.e.*, we set $\mathbf{I} = \mathbf{I}^*$). The pretrained `CE-VAE` is then fine-tuned for underwater image enhancement on the LSUI Train-L split proposed in [35].

Validation. We validate our method on six benchmark datasets to assess its generalization across diverse underwater conditions. To perform a comparison between the enhanced image and the available ground truth, we considered the following full-reference datasets: (i) the LSUI-L400 dataset [35] comes with images featuring different water types, lighting conditions, and target categories[1]; (ii) the EUVP dataset [18] comprises 1970 validation image samples of varying quality; and (iii) the UFO-120 dataset [17] contains 120 full-reference images collected from oceanic explorations across multiple locations and water types.

To assess enhanced image quality in a broader context, we further analyzed our model performance on the following non-reference datasets: (i) the UCCS dataset [28] consists of 300 genuine underwater images captured across diverse marine organisms and environments, specifically designed to evaluate color cast correction in underwater image enhancement. (ii) the U45 [26] and (iii) SQUID [5] datasets contain 45 and 57 raw underwater images showing severe color casts, low contrast, and haze.

4.2. Metrics

We followed recent works [22, 26, 35, 43], and assessed our model performance considering the Peak Signal-to-Noise Ratio (PSNR), the Structural Similarity (SSIM) [49], and the Learned Perceptual Image Patch Similarity (LPIPS) [54] for full-reference datasets.

For non-reference datasets, we considered the Underwater Color Image Quality Evaluation Metric (UCIQE) [52], the Underwater Image Quality Measure (UIQM) [32], the Natural Image Quality Evaluator (NIQE) [30], and the Inception Score (IS) [47].

[1] The evaluation considers the Test-L 400 split proposed in [35].

	LSUI-L400			EUVP			UFO-120		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
RGHS [16]	18.44	0.80	0.31	18.05	0.78	0.31	17.48	0.71	0.37
UDCP [8]	13.24	0.56	0.39	14.52	0.59	0.35	14.50	0.55	0.42
UIBLA [37]	17.75	0.72	0.36	18.95	0.74	0.33	17.04	0.64	0.40
UGAN [10]	19.40	0.77	0.37	20.98	0.83	0.31	19.92	0.73	0.38
FUnIE-GAN [18]	-	-	-	23.53	0.84	0.26	23.09	0.76	0.32
Cluie-Net [27]	18.71	0.78	0.33	18.90	0.78	0.30	18.43	0.72	0.36
DeepSESR [17]	-	-	-	24.22	0.85	0.25	**23.38**	**0.78**	**0.29**
TWIN [43]	19.84	0.79	0.33	18.91	0.79	0.32	18.21	0.72	0.37
UShape-Transformer [35]	**23.02**	**0.82**	**0.29**	**27.59**	**0.88**	**0.23**	22.82	0.77	0.33
Spectroformer [22]	20.09	0.79	0.32	18.70	0.79	0.32	18.03	0.71	0.37
CE-VAE	**24.49**	**0.84**	**0.26**	**27.75**	**0.89**	**0.20**	**24.38**	**0.79**	**0.28**

Table 1. Quantitative comparison of CE-VAE and state-of-the-art methods across the three considered full-reference datasets: LSUI-L400, EUVP, and UFO-120. (↑ higher is better, ↓ lower is better). For each metric/dataset the best method is in red, second best in blue.

4.3. Implementation Details

We run the experimental evaluation with $\mathbf{I} \in \mathbb{R}^{3\times H=256\times W=256}$. E, having $N = 5$ encoding blocks, yields $\mathbf{X} \in \mathbb{R}^{256\times 16\times 16}$. Q starts with $\beta_L = 32$ to get $\mathbf{U} \in \mathbb{R}^{32\times 16\times 9\times 9}$. From this, we get $32\times 9\times 9$ tensors each (of dimensionality $C_{\mathbf{U}} = 16$) representing a capsule point of view. In RbA , we set $\alpha = 3$ to obtain $\hat{\mathbf{U}} \in \mathbb{R}^{64\times 32\times 9\times 9}$, where each vector obtained through the clusterization has $\beta_{L+1} = 64$ dimensions. The normalization and following transposed convolution layers output $\mathbf{C} \in \mathbb{R}^{256\times 16\times 16}$. We used the same settings as in [9] for the pre-trained CNN considered in (9). To optimize our loss function, we set $\delta = 10^{-6}$ following [9]. We ran ImageNet pretraining for 25 epochs, with a batch size of 6 using the Adam optimizer with a learning rate of $4.5e^{-6}$. Using the same optimization settings, we fine-tuned the resulting model on the LSUI Train-L dataset for 600 epochs using the adversarial strategy proposed in [9]. For both training processes, random cropping and horizontal flipping were applied.

4.4. State-of-the-art Comparison

We compare the performance of our CE-VAE model with existing traditional methods like RGHS [16], UDCP [8], and UIBLA [37] as well as state-of-the-art machine learning-based works including UGAN [10], FUnIE-GAN [18], DeepSESR [17], Cluie-Net [27], TWIN [43], UShape-Transformer [35], and Spectroformer [22]. We report on the results published in the corresponding papers or by running the publicly available codes.

Full-reference datasets. Table 1 shows that across diverse underwater datasets, our method consistently outperforms existing underwater image enhancement approaches. On the LSUI-L400 dataset, we achieve the best results considering all metrics, with a significant improvement in PSNR ($+1.47dB$) and SSIM ($+2\%$) compared to the best-performing method. The evaluation of the EUVP dataset further highlights the capabilities of our approach with substantial improvements in SSIM ($+1\%$) and LPIPS (-3%), showcasing its ability to restore image quality and perceptual fidelity. On the UFO-120 dataset, our method demonstrates notable improvements in PSNR ($+1dB$), SSIM ($+1\%$), and LPIPS (-1%). All such results substantiate the capabilities of our approach in leveraging a compressed latent space to precisely reconstruct the spatial relation between entities with great details under different water types, locations, lighting conditions, and multiple targets.

Non-reference datasets. Results presented in Table 2 show that our method demonstrates competitive performance across multiple non-reference datasets for underwater image enhancement. This is particularly evident when we consider the IS metric –computed to evaluate how realistic the enhanced images are using a model pre-trained on natural images (*i.e.*, ImageNet). In such a case our method obtains the best overall results on all the datasets (*e.g.*, $+0.04$ and $+0.02$ with respect to the best existing method on the UCCS and U45 datasets). More in detail, on the UCCS dataset, we rank in second place in almost all metrics, except IS, for which we are at the top of the leaderboard. A similar result is shown for the U45 dataset, where TWIN [43] has the highest scores, and we ranked either second or third place, yet reaching the 1st place on the IS metric. Similarly, on the SQUID dataset, our approach demonstrates competitive performance, achieving scores comparable to or higher than most existing methods across all metrics (*e.g.*, 0.01 difference between our method and the best performing one on UIQM and UCIQE metrics).

4.5. Ablation Study

Through the ablation study, we want to answer different questions that would help us understand the importance of each proposed component of our architecture.

How Efficient is the Encoder? Data storage and transmission are crucial for underwater data collection campaigns. We designed our online encoder to store only the

[2] https://planet-ocean.co.uk/surface-and-underwater-vehicles/

	UCCS				U45				SQUID			
	UIQM ↑	UCIQE ↑	NIQE ↓	IS ↑	UIQM ↑	UCIQE ↑	NIQE ↓	IS ↑	UIQM ↑	UCIQE ↑	NIQE ↓	IS ↑
RGHS [16]	2.97	0.55	5.14	2.23	2.57	**0.62**	4.34	2.21	1.46	**0.56**	9.25	2.02
UDCP [8]	2.06	0.55	5.72	2.51	2.09	0.59	4.83	2.37	0.97	**0.56**	8.69	2.01
UIBLA [37]	2.58	0.53	5.69	2.09	1.67	0.59	6.03	2.11	1.08	0.52	9.58	1.98
UGAN [10]	2.84	0.51	6.85	2.38	3.04	0.55	6.56	2.40	2.38	0.52	8.81	2.05
Cluie-Net [27]	3.02	0.55	5.19	2.28	3.19	0.59	4.41	2.30	2.12	0.51	7.13	2.18
TWIN [43]	**3.23**	**0.59**	**4.45**	2.22	**3.36**	**0.62**	**4.16**	2.30	2.31	**0.57**	**6.44**	**2.21**
UShape-Transformer [35]	3.16	**0.56**	4.69	**2.61**	3.11	0.59	4.91	2.31	2.21	0.54	8.33	2.09
Spectroformer [22]	**3.21**	0.55	4.80	2.36	3.21	**0.61**	**4.22**	**2.42**	**2.45**	**0.56**	**6.56**	2.11
CE−VAE	**3.21**	**0.56**	**4.65**	**2.65**	**3.23**	**0.61**	4.29	**2.44**	**2.44**	**0.57**	6.58	**2.19**

Table 2. Quantitative comparison of CE−VAE and state-of-the-art methods across three non-reference datasets: UCCS, U45, and SQUID (↑ higher is better, ↓ lower is better). For each metric/dataset the best method is in red, second best in blue.

Method	PSNR ↑	On-device Storage [MB] ↓	Maximum Recording Duration [h] ↑	Transmission time [s] ↓
Cluie-Net [27]	18.71	1.57	0.39	0.01256
TWIN [43]	19.84	1.57	0.39	0.01256
UShape-T. [35]	**23.02**	**1.57**	**0.39**	**0.01256**
Spectr. [22]	20.09	1.57	0.39	0.01256
CE−VAE	**24.49**	**0.52**	**1.17**	**0.00416**

Table 3. Enhancement performance and storage/transmission capabilities. Considering the real-world LSUI-L400 dataset, for a 256 × 256 input image, we computed the corresponding device storage space required for enhanced image generation and the related transmission time on a 1 GBbps bandwidth beacon. We also report on the maximum recording duration that would fit an off-the-shelf commercial device[2](↑ higher is better, ↓ lower is better).

Method	EUVP	LSUI-L400	UFO120
VQ−VAE [9]	20.58	19.36	19.62
CE−VAE w/o D_S	**22.20**	**20.27**	**20.71**
CE−VAE	**27.75**	**24.49**	**24.38**

Table 4. PSNR performance comparison between our three architecture variants and the VQ−VAE baseline.

resulting latent image representation on the autonomous device. This is the only information exploited by the offline dual-decoder module for full-size reconstruction and enhancement. To assess the benefits of such a solution, we computed the results in Table 3. Considering 30 samples to be enhanced, our method requires 15.6MB of storage and takes only 124.8ms to be transmitted through a 1Gbps bandwidth beacon. All other methods considered for comparison required 47.1MB and 376ms, respectively. These results demonstrate that our method, which takes approximatively 0.06 seconds to encode a single full-size input, offers a 3× more efficient solution while also delivering the highest PSNR compared to state-of-the-art methods.

Is Capsule Clustering Good at Modeling the Latent Space? To explore the effectiveness of different latent information modeling methods, we replaced our capsule clustering procedure with the codebook learning process from VQ-VAE [9]. Table 4 shows that our novel capsule-based approach (CE−VAE w/o D_S, excluding the spatial decoder

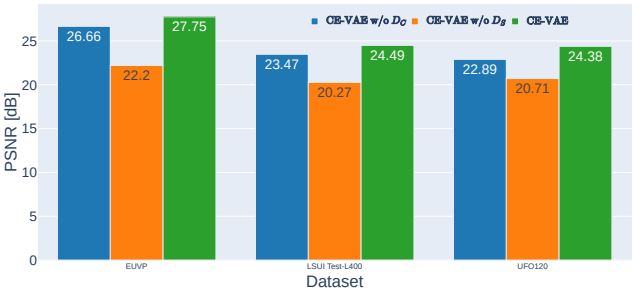

Figure 4. Analysis of the decoder components. Results are shown for our architecture (i) without the spatial decoder (CE−VAE w/o D_S), (ii) without the capsule decoder (CE−VAE w/o D_c), (iii) and for the complete CE−VAE.

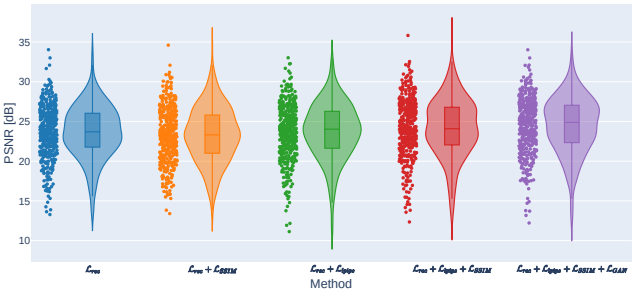

Figure 5. Evaluation of the different loss components and their PSRN impact on the LSUI TestL-400 dataset.

for equivalence) demonstrates a significant performance advantage over the VQ-VAE codebook variant, achieving an average improvement of more than1.5dB across the three full-reference datasets. These results highlight the effectiveness of capsules, offering improved performance over discrete quantization methods that also suffer from non-differentiability issues.

What is the Impact of Capsule and Spatial Decoders? Figure 4 illustrates the PSNR performance of our model with ablations of the capsule and spatial decoders. Our CE−VAE model achieves superior PSNR scores across all datasets: 27.75dB on EUVP, 24.49dB on LSUI Test-L400, and 24.38dB on UFO120. While the capsule decoder alone (CE−VAE w/o D_S) yields results on par with existing methods (*e.g.*, 22.2dB on EUVP), the spatial decoder (CE−VAE

w/o $D_{\mathbf{C}}$) demonstrates significantly higher performance (*e.g.*, 26.66dB on EUVP), demonstrating its importance for pixel-level reconstruction. These results substantiate our intuition on designing these two encoders that, leveraging their complementary strengths, effectively model entity-related and spatially precise details for image enhancement.

How Relevant Are the Loss Terms? Figure 5 shows the results obtained considering the LSUI Test-L 400 dataset by our approach when loss components are turned on and off (the analysis on the other validation datasets is in the supplementary). The performance shows a PSNR increase as more loss terms are included. Notably, including the adversarial loss (*i.e.*, $\mathcal{L}_{GAN}$) concentrates the distribution of PSNR values around the mean, highlighting the robustness of the method towards edge cases. This is likely to be due to the GAN discriminator role that might prevent the generation of unfeasible results containing hallucinated artifacts or values that are close to the real one in the RGB space, but result in different semantics.

4.6. Qualitative Analysis

To showcase the robustness and accuracy of our method for underwater color correction, we report on the qualitative performance comparison between our method and existing solutions on the Color-Checker7 dataset [3]. The dataset contains 7 underwater images captured with different cameras, each including a Macbeth Color Checker. This allows us to assess our method's performance across diverse imaging devices and its ability to accurately restore true colors. Figure 6 shows that our method provides a neat color restoration for all items in the Macbeth Color Checker while also yielding a realistic color for the skin.

5. Conclusions

We introduced a novel architecture for underwater image enhancement featuring a highly efficient encoder module that yields $3\times$ compression efficiency compared to existing approaches while running online. To decode the encoded, compressed, representation we proposed a dual-decoder module that leverages spatial and entity-level information (captured by a capsule layer) to precisely reconstruct a full-size enhanced version of the input. Our approach demonstrates superior image quality across six benchmark datasets while potentially facilitating longer missions and more comprehensive data collection.

Figure 6. Enhanced images comparison on the Color-Check7 dataset.

References

[1] Derya Akkaynak and Tali Treibitz. Sea-thru: A method for removing water from underwater images. In *International Conference on Computer Vision and Pattern Recognitino (CVPR)*, pages 1682–1691, 2019. 2

[2] Codruta O. Ancuti, Cosmin Ancuti, Christophe De Vleeschouwer, and Philippe Bekaert. Color balance and fusion for underwater image enhancement. *IEEE Transactions on Image Processing*, 27(1):379–393, 2018. 2

[3] Codruta O. Ancuti, Cosmin Ancuti, Christophe De Vleeschouwer, and Philippe Bekaert. Color balance and fusion for underwater image enhancement. *IEEE Transactions on Image Processing*, 27:379–393, 1 2018. 8

[4] Stephane Bazeille, Isabelle Quidu, Luc Jaulin, and Jean-Philippe Malkasse. Automatic underwater image preprocessing. In *CMM'06*, page xx, 2006. 2

[5] 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. 5

[6] Brian Bingham, Brendan Foley, Hanumant Singh, Richard Camilli, Katerina Delaporta, Ryan Eustice, Angelos Mallios, David Mindell, Christopher Roman, and Dimitris Sakellariou. Robotic tools for deep water archaeology: Surveying an ancient shipwreck with an autonomous underwater vehicle. *Journal of Field Robotics*, 27(6):702–717, 2010. 2

[7] Fei Deng, Shengliang Pu, Xuehong Chen, Yusheng Shi, Ting Yuan, and Shengyan Pu. Hyperspectral image classification with capsule network using limited training samples. *Sensors*, 18(9):3153, 2018. 2

[8] Paulo L.J. Drews, Erickson R. Nascimento, Silvia S.C. Botelho, and Mario Fernando Montenegro Campos. Underwater depth estimation and image restoration based on single images. *IEEE Computer Graphics and Applications*, 36:24–35, 3 2016. 6, 7

[9] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *International Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021. 4, 5, 6, 7

[10] Cameron Fabbri, Md Jahidul Islam, and Junaed Sattar. Enhancing underwater imagery using generative adversarial networks. In *ICRA*, pages 7159–7165, 2018. 2, 5, 6, 7

[11] Ahmad Shahrizan Abdul Ghani and Nor Ashidi Mat Isa. Underwater image quality enhancement through integrated color model with rayleigh distribution. *Applied soft computing*, 27:219–230, 2015. 2

[12] Yecai Guo, Hanyu Li, and Peixian Zhuang. Underwater image enhancement using a multiscale dense generative adversarial network. *IEEE Journal of Oceanic Engineering*, 45(3):862–870, 2019. 2, 3

[13] Jie Han, Jian Zhou, Lin Wang, Yu Wang, and Zhongjun Ding. Fe-gan: Fast and efficient underwater image enhancement model based on conditional gan. *Electronics*, 12(5):1227, 2023. 3

[14] Pingli Han, Fei Liu, Kui Yang, Jinyu Ma, Jianjun Li, and Xiaopeng Shao. Active underwater descattering and image recovery. *Applied Optics*, 56(23):6631–6638, 2017. 2

[15] Kai Hu, Yanwen Zhang, Chenghang Weng, Pengsheng Wang, Zhiliang Deng, and Yunping Liu. An underwater image enhancement algorithm based on generative adversarial network and natural image quality evaluation index. *Journal of Marine Science and Engineering*, 9(7):691, 2021. 2

[16] 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 *International Conference on Multimedia Modeling*, pages 453–465, 2018. 6, 7

[17] Md Jahidul Islam, Peigen Luo, and Junaed Sattar. Simultaneous enhancement and super-resolution of underwater imagery for improved visual perception. *arXiv:2002.01155*, 2020. 2, 5, 6

[18] Md Jahidul Islam, Youya Xia, and Junaed Sattar. Fast underwater image enhancement for improved visual perception. *IEEE Robotics and Automation Letters*, 5(2):3227–3234, 2020. 2, 3, 5, 6

[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *International Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017. 5

[20] Asif Hussain Khan, Christian Micheloni, and Niki Martinel. IDENet: Implicit Degradation Estimation Network for Efficient Blind Super Resolution . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 6065–6075, Los Alamitos, CA, USA, June 2024. IEEE Computer Society. 2

[21] Asif Hussain Khan, Christian Micheloni, and Niki Martinel. Lightweight prompt learning implicit degradation estimation network for blind super resolution. *IEEE Transactions on Image Processing*, 33:4556–4567, 2024. 2

[22] Raqib Khan, Priyanka Mishra, Nancy Mehta, Shruti S Phutke, Santosh Kumar Vipparthi, Sukumar Nandi, and Subrahmanyam Murala. Spectroformer: Multi-domain query cascaded transformer network for underwater image enhancement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1454–1463, 2024. 3, 5, 6, 7

[23] 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, 2020. 3

[24] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1731–1735. IEEE, 2016. 2

[25] Chong-Yi Li, Ji-Chang Guo, Run-Min Cong, Yan-Wei Pang, and Bo Wang. Underwater image enhancement by dehazing with minimum information loss and histogram distribution prior. *IEEE Transactions on Image Processing*, 25(12):5664–5677, 2016. 2

[26] Hanyu Li, Jingjing Li, and Wei Wang. A fusion adversarial underwater image enhancement network with a public test dataset. *arXiv preprint arXiv:1906.06819*, 2019. 5

[27] Kunqian Li, Li Wu, Qi Qi, Wenjie Liu, Xiang Gao, Liqin Zhou, and Dalei Song. Beyond single reference for training: Underwater image enhancement via comparative learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 33:2561–2576, 6 2023. 6, 7

[28] Risheng Liu, Xin Fan, Ming Zhu, Minjun Hou, and Zhongxuan Luo. Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(12):4861–4875, 2020. 5

[29] Huimin Lu, Yujie Li, and Seiichi Serikawa. Underwater image enhancement using guided trigonometric bilateral filter and fast automatic color correction. In *IEEE international conference on image processing*, pages 3412–3416. IEEE, 2013. 2

[30] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 5

[31] László Neumann, Rafael Garcia, József Jánosik, and Nuno Gracias. Fast underwater color correction using integral images. *Instrumentation Viewpoint*, (20):53–54, 2018. 2

[32] Karen Panetta, Chen Gao, and Sos Agaian. Human-visual-system-inspired underwater image quality measures. *IEEE Journal of Oceanic Engineering*, 41(3):541–551, 2016. 5

[33] Dubok Park, David K Han, and Hanseok Ko. Enhancing underwater color images via optical imaging model and non-local means denoising. *IEICE Transactions on Information and Systems*, 100(7):1475–1483, 2017. 2

[34] Jaihyun Park, David K Han, and Hanseok Ko. Adaptive weighted multi-discriminator cyclegan for underwater image enhancement. *Journal of Marine Science and Engineering*, 7(7):200, 2019. 2

[35] Lintao Peng, Chunli Zhu, and Liheng Bian. U-shape transformer for underwater image enhancement. *IEEE Transactions on Image Processing*, 32:3066–3079, 2023. 3, 5, 6, 7

[36] 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. 2

[37] Yan Tsung Peng and Pamela C. Cosman. Underwater image restoration based on image blurriness and light absorption. *IEEE Transactions on Image Processing*, 26:1579–1594, 4 2017. 6, 7

[38] Rita Pucci, Christian Micheloni, Gian Luca Foresti, and Niki Martinel. Deep interactive encoding with capsule networks for image classification. *Multimedia Tools and Applications*, 79(43):32243–32258, 2020. 2

[39] Rita Pucci, Christian Micheloni, Gian Luca Foresti, and Niki Martinel. Fixed simplex coordinates for angular margin loss in capsnet. In *ICPR*, pages 3042–3049, 2021. 2

[40] Rita Pucci, Christian Micheloni, Gian Luca Foresti, and Niki Martinel. Pro-ccaps: Progressively teaching colourisation to capsules. In *Winter Applications of Computer Vision (WACV)*, pages 2271–2279, 2022. 2

[41] Rita Pucci, Christian Micheloni, and Niki Martinel. Collaborative image and object level features for image colourisation. In *International Conference on Computer Vision and Pattern Recognition*, pages 2160–2169, 2021. 2

[42] Nianzu Qiao, Lu Dong, and Changyin Sun. Adaptive deep learning network with multi-scale and multi-dimensional features for underwater image enhancement. *IEEE Transactions on Broadcasting*, 2022. 3

[43] Huzhou Yang Risheng Liu, Zhiying Jiang and Xin Fan. Twin adversarial contrastive learning for underwater image enhancement and beyond. In *IEEE Transactions on Image Processing*. IEEE, 2022. 5, 6, 7

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 3

[45] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 5

[46] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in neural information processing systems*, pages 3856–3866, 2017. 2

[47] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 5

[48] Florian Shkurti, Anqi Xu, Malika Meghjani, Juan Camilo Gamboa Higuera, Yogesh Girdhar, Philippe Giguere, Bir Bikram Dey, Jimmy Li, Arnold Kalmbach, Chris Prahacs, et al. Multi-domain monitoring of marine environments using a heterogeneous robot team. In *IROS*, pages 1747–1753, 2012. 2

[49] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 5

[50] Louis Whitcomb, Dana R Yoerger, Hanumant Singh, and Jonathan Howland. Advances in underwater robot vehicles for deep ocean exploration: Navigation, control, and survey operations. In *Robotics Research*, pages 439–448. 2000. 2

[51] Zhengyu Xing, Meng Cai, and Jianxun Li. Improved shallow-uwnet for underwater image enhancement. In *2022 IEEE International Conference on Unmanned Systems (ICUS)*, 2022. 3

[52] Miao Yang and Arcot Sowmya. An underwater color image quality evaluation metric. *IEEE Transactions on Image Processing*, 24(12):6062–6071, 2015. 5

[53] Huiqing Zhang, Luyu Sun, Lifang Wu, and Ke Gu. Dugan: An effective framework for underwater image enhancement. *IET Image Processing*, 15(9):2010–2019, 2021. 2

[54] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *International Con-*

ference on Computer Vision and Pattern Recognition, pages 586–595, 2018. 5

[55] Weidong Zhang, Yudong Wang, and Chongyi Li. Underwater image enhancement by attenuated color channel correction and detail preserved contrast enhancement. *IEEE Journal of Oceanic Engineering*, 47(3):718–735, 2022. 2

[56] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57, 2017. 5

[57] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *International Conference on Computer Vision*, pages 2223–2232, 2017. 2