UnDIVE: Generalized Underwater Video Enhancement Using Generative Priors

Suhas Srinath[1] Aditya Chandrasekar[1,2] † Hemang Jamadagni[3] Rajiv Soundararajan[1]
Prathosh A P[1]
[1] Indian Institute of Science [2] Qualcomm [3] National Institute of Technology Karnataka

Abstract

With the rise of marine exploration, underwater imaging has gained significant attention as a research topic. Underwater video enhancement has become crucial for real-time computer vision tasks in marine exploration. However, most existing methods focus on enhancing individual frames and neglect video temporal dynamics, leading to visually poor enhancements. Furthermore, the lack of ground-truth references limits the use of abundant available underwater video data in many applications. To address these issues, we propose a two-stage framework for enhancing underwater videos. The first stage uses a denoising diffusion probabilistic model to learn a generative prior from unlabeled data, capturing robust and descriptive feature representations. In the second stage, this prior is incorporated into a physics-based image formulation for spatial enhancement, while also enforcing temporal consistency between video frames. Our method enables real-time and computationally-efficient processing of high-resolution underwater videos at lower resolutions, and offers efficient enhancement in the presence of diverse water-types. Extensive experiments on four datasets show that our approach generalizes well and outperforms existing enhancement methods. Our code is available at `github.com/suhas-srinath/undive`.

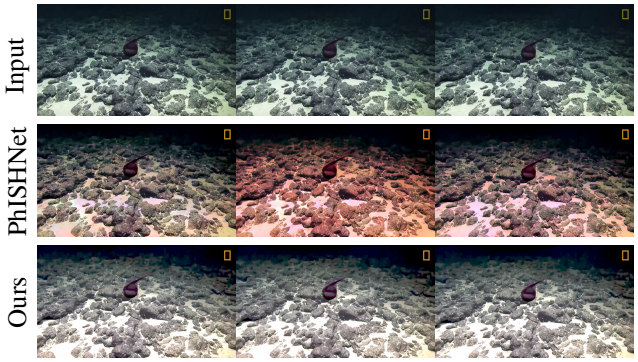
1. Introduction

The goal of underwater enhancement is to reduce artifacts and recover lost colors from the water scattering effect [30] in images and videos. Underwater enhancement finds applications in areas such as coral reef monitoring [25], archaeology [65] and underwater robotics [76]. Underwater video enhancement (UVE) is often challenging due to multiple reasons. Collecting high-quality videos amidst distortions like blur, reduced illumination, complex channel attenuation, and obtaining ground-truth video data to supervise learning-based algorithms are extremely cumbersome.

Recent works [27, 40, 52, 66, 80][†] try to solve UVE by training convolutional neural networks through the supervision of large-scale ground truth data, but often fail to generalize to diverse and unseen water types. While learning-based methods [40, 59, 86, 87] have been reasonably successful in achieving underwater image enhancement (UIE), they cannot be directly scaled to UVE since they do not account for object motion in videos. Moreover, variations in the illumination across enhanced frames from UIE methods cause flicker-like artifacts in the enhanced videos. Despite the greater requirement of videos than images in marine applications, very few UVE methods exist.

UIE methods have been very successful in restoring the color and contrast in underwater scenes. Earlier UIE methods attempted to solve enhancement through pixel adjustments and classical priors [50, 63]. With the development of deep learning, data-centric methods [16,40,44] leveraged paired image training to learn good enhancement. To generalize better, unsupervised methods [17] have also been developed for UIE. Despite numerous advancements in UIE, scaling these techniques to videos via frame aggregation remains challenging due to the lack of temporal alignment.

To address the aforementioned challenges, we propose to solve the UVE problem through the introduction of tem-



(a) Frame x_{t-1} (b) Frame x_t (c) Frame x_{t+1}

Figure 1. UnDIVE (bottom row) enhances contiguous video frames (top row) from the UOT32 [36] dataset (DeepSeaFish video), while maintaining consistent colors and illumination as opposed to image-based methods such as PhISH-Net (middle row).

[†]Work done while at Indian Institute of Science.

poral consistency into a physics-driven spatial enhancement network that mitigates artifacts that arise due to water scattering. Firstly, to learn efficient spatial enhancement, we propose to learn a generative prior through a self-supervised denoising diffusion probabilistic model (DDPM) [26] that learns robust representations of underwater images. To the best of our knowledge, this is the first work that learns a generative prior using diffusion for UVE. The encoder of the UNet learned by the DDPM is subsequently integrated into the UVE framework for efficient downstream enhancement.

In general, video enhancement [6, 51, 81] and restoration methods [43, 47] incorporate motion information during the learning process to make videos more temporally consistent through the alignment of objects or representations between consecutive frames. However, such transformer-based methods are prone to overfitting and do not generalize well (simultaneously perform well on diverse datasets and water-types). To tackle this issue, we propose to incorporate motion into UVE through an unsupervised optical flow loss so that objects in consecutive enhanced frames exhibit smoother motion and uniform illumination, colors, and contrast.

Leveraging the generalization capability of the generative prior and the unsupervised temporal consistency loss, we propose an **Un**derwater **D**omain **I**ndependent **V**ideo **E**nhancement (**UnDIVE**) framework that can efficiently process high-resolution videos with fairly low complexity and inference times. To summarize, our contributions are as follows:

- A two-stage training framework for UVE. The first stage learns a prior on underwater images, and the second stage utilizes this prior to learn spatial and temporal enhancement.
- A generative prior trained on carefully chosen underwater images learned through a DDPM to provide robust representations. The learned encoder is subsequently integrated into the UnDIVE network for downstream enhancement.
- An unsupervised temporal consistency loss to incorporate motion information into a physics-driven spatial enhancement network, enabling enhanced videos to maintain uniform illumination, accurate colors, and improved contrast.
- Through extensive experiments, we demonstrate the superiority and generalizability of UnDIVE over other enhancement methods on four diverse underwater video datasets on multiple no-reference (NR) visual quality metrics.

2. Related Work

2.1. Underwater Image/Video Enhancement

UIE methods can be broadly categorized into traditional (image-processing or model-based) and learning-based approaches. Traditional methods aim to restore degraded underwater images using prior visual characteristics or by treating UIE as an inversion problem, reversing the degradation caused by the imaging process. Popular methods employ various image processing techniques [4, 5, 24, 32, 53, 54, 84, 90], utilize priors [8, 12–14, 19, 60, 70, 75, 88], or rely on an underwater image formation model [1, 7, 29, 58, 79, 85, 89]. Learning-based methods, particularly deep learning methods, have become prominent in producing high-quality results.

Although some methods do not account for the physical process of underwater image formation, they learn to reverse the degradation using large-scale training. Notable CNN-based methods include WaterNet [40], UWCNN [39], SGUIE-Net [61], UICoE Net [62], LANet [49], PUIE-Net [18], Semi-UIE [31], PhISH-Net [9], UIE-Net [73], USUIR [17], and URanker [20]. While these methods are successful, they generally rely on large-scale training with ground-truth references, which is difficult and time-consuming to obtain. In the context of UVE, most of these works apply UIE on a frame-by-frame basis, often resulting in visual artifacts like inconsistent lighting across frames and flicker. We propose learning spatial enhancement along with temporal consistency to mitigate these issues.

2.2. Generative Models for Enhancement

With the success of generative adversarial networks (GANs), various GAN-based methods have been developed, enabling faster computation. GAN-based approaches include UGAN [16], UW-GAN [71], Spiral-GAN [23], FunIE-GAN [33], WaterGAN [44], Dense GAN [21], FE-GAN [22], TOPAL [35], CycleGAN [41] and CLUIE-Net [45]. PUIE-Net [18] uses a conditional variational autoencoder to generate an enhancement distribution.

Although generative models are shown to run with excellent computational speeds, they often trade-off performance due to undesired hallucinated colors. Such methods are unable to leverage the generalizability of generative models, and to this end, we propose to use a generative prior through a DDPM that learns good representations of underwater scenes. This prior enables efficient transfer learning for the downstream UVE task.

2.3. Temporal Consistency in Videos

Flickering tends to occur when single-frame-based methods are applied to video clips, leading to notable visual incoherence. Efforts to use image-based methods to enhance videos or enforce "temporal consistency" have been

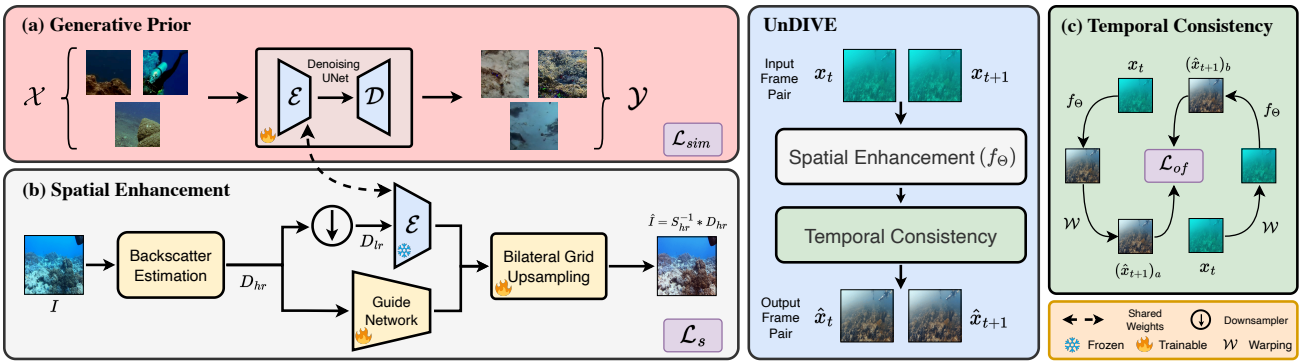


Figure 2. Overall framework of UnDIVE. (a) The first stage learns a generative prior on underwater images, where a denoising DDPM UNet is trained with the loss $\mathcal{L}_{sim}$. (b) The second stage utilizes the trained encoder, and learns the spatial enhancement (f_Θ) with loss $\mathcal{L}_s$. First, backscatter is removed, and the image D_{hr} is processed through a guide network to capture low-level local details, while the downsampled (by two) image is passed through $\mathcal{E}$ capturing global (high-level) information. Finally, both streams are fused and upsampled to match the input resolution. (c) A temporal consistency loss $\mathcal{L}_t$ enforces uniform illumination and colors in the enhanced frames.

explored in various contexts, such as video-to-video synthesis [72], video enhancement [83], style-transfer [64], depth estimation [46] and semantic segmentation [55], to mitigate flickering issues.

These methods typically employ self-consistency by enforcing the similarity of data pairs [10,15] or explicitly learn temporal consistency using additional modules [38] to improve the performance and stability of deep models. For instance, Zhang *et al.* [83] implicitly embed temporal consistency using optical flow [28] generated from single images. In this work, we leverage both forward and backward optical flow to learn visual consistency between a pair of consecutive frames of a video. This allows the model to learn consistent reconstructions of color, structure, and illumination in both directions to enable uniform enhancement.

3. UnDIVE

Our framework, UnDIVE, consists of learning a generative prior on underwater images, incorporating it into a spatial enhancement network and further introducing temporal consistency into the enhancements. An overview of the training framework is illustrated in Fig. 2.

3.1. Generative Prior for Underwater Images

DDPMs [26] are generative models that have shown tremendous success in various applications for their ability to learn generalizable feature encoders using unlabeled data. These encoders can then be downstreamed effectively for multiple transfer learning tasks. In our work, we leverage such an encoder to learn robust representations on underwater images. The training images are chosen such that the intensity histograms are closer to uniform to ensure the presence of significant objects. Since enhancement and generation are image-to-image tasks, we find it appropriate to train a UNet-based DDPM for generation and transfer it for the task of enhancement. The following paragraphs describe the training procedure of the DDPM.

Consider a UNet characterized by an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$ (refer Fig. 2) parameterized by θ. Let $\mathbf{x}_0 \in \mathcal{X}$ be an input image to $\mathcal{E}$, and $\mathbf{x}_t$ be the corresponding noisy image at time step $t = 1, 2, \ldots, T$. $\mathbf{x}_t$ is obtained from $\mathbf{x}_{t-1}$ according to the diffusion process:

$$\mathbf{x}_t = \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\epsilon_t, \tag{1}$$

where β_t is the noise schedule parameter at time t, and $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\forall t = 1, 2, \ldots, T$. The distributions of the forward diffusion process are denoted by $q(\mathbf{x}_t|\mathbf{x}_{t-1})$, which are assumed to follow a first-order Gaussian Markov process. The reverse (decoding process) is modeled using a parametric family of distributions denoted by $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$.

The formulation in (1) allows us to sample $\mathbf{x}_t$ directly from the original input image $\mathbf{x}_0$ as $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ and $\alpha_t = 1 - \beta_t$. If the noise schedule parameters $(\beta_t)_{t=1}^T$ are very small such that $\beta_T \to 0$, then the distribution of $\mathbf{x}_T$ can be well approximated by the standard Gaussian distribution i.e., $q(\mathbf{x}_T) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The goal of the DDPM is to estimate the parameters of p_θ by optimizing a variational lower bound on the log-likelihood of the data $\mathbf{x}_0$ under the model p_θ. A simplified loss function used to train the DDPM is as follows:

$$\mathcal{L}_{\text{sim}} = \sum_{t \geq 1} L_t \text{ , where } L_t = \mathop{\mathbb{E}}_{\mathbf{x}_0, \epsilon}\left[\|\epsilon - \epsilon_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\|^2\right], \tag{2}$$

where ϵ and $\epsilon_\theta(\mathbf{x}_t, t)$ correspond to the input (real) noise and the predicted noise at time step t, respectively. Once the DDPM is trained, the encoder $\mathcal{E}$ of the UNet at time step $t = 0$ (since there is no requirement of any noisy versions of $\mathbf{x}_0$) is used ($\mathcal{E}$ is kept frozen) in the subsequent spatial enhancement stage to provide robust representations.

3.2. Learning Spatial Enhancement

Since the colors and contrast in underwater scenes are lost due to the water scattering effect, it is necessary to mitigate this to recover objects present in the scene. The scattering effect towards the surface causes noisy artifacts, called backscatter, that proportionally increases with depth [2, 9].

Backscatter Estimation: A simple linear formulation [2,3] that models the backscatter was proposed as follows:

$$I_c = D_c + B_c, \tag{3}$$

where I_c is the c^{th} channel of an RGB image I, $c \in \{r, g, b\}$, D_c is the direct signal and B_c is the backscatter. B_c can be estimated as a function of wideband attenuation coefficients and the depth map. Since knowledge of depth is required for estimating the backscatter, we employ the SlowTV monocular depth estimator [67, 68] to generate robust depth maps in a computationally efficient manner. SlowTV, trained on a large dataset of images, including underwater scenes, is well-suited for depth estimation. Removing backscatter eliminates water from the input image, making the model independent of water-type and enhancing generalization..

Training Spatial Enhancement: Once the backscatter is removed from the input image I, we utilize the backbone of PhISH-Net [9] to obtain a high resolution illumination map S_{hr}. The key difference from PhISHNet is in the low-resolution feature encoder. We then use the encoder $\mathcal{E}$ to enhance the estimation of the illumination map by improving the structural representations of underwater scenes. The guide network captures low-level local features, which are fused with the high-level features from $\mathcal{E}$ resulting in a complete representation. The fused features are then bilinearly upsampled (with learnable weights) to match the input resolution. The final enhanced image $\hat{I}$ can be estimated from the high-resolution image I_{hr} as:

$$\hat{I} = \frac{I_{hr}}{S_{hr} + \epsilon}, \tag{4}$$

a pixel-wise division where $\epsilon > 0$ is a constant that ensures numerical stability. We utilize three spatial losses to train UnDIVE on a set of paired underwater images $\{(\hat{I}, I^{gt})\}$ (I^{gt} is the corresponding ground truth for I).

Reconstruction loss: To reconstruct the structure, color, illumination, and sharpness in the enhanced image, we employ a loss function consisting of a combination of structural similarity measure (SSIM) and the L_1 distance at a pixel level. The loss is expressed as

$$\mathcal{L}_r = 0.85(1 - \text{SSIM}(\hat{I}, I_{gt})) + 0.15 \sum_j |\hat{I}_j - I_j^{gt}|, \tag{5}$$

where j corresponds to the pixel locations in each image. This loss function has shown tremendous success in learning unsupervised optical flow [48] and in multiple view synthesis applications.

Smoothness Loss: To maintain uniform illumination in the enhancements, we impose a smoothness loss $\mathcal{L}_{sm}$ which is a sum of weighted L_2 norm of the gradients of the illumination map S_{hr} as follows:

$$\mathcal{L}_{sm} = \sum_j w_j \|\nabla_j S_{hr}\|_2, \tag{6}$$

where j denotes the pixel location and w_j is a weight corresponding to each spatial location.

Color Loss: Additionally, a color loss is introduced to keep the reconstructed intensities consistent with that of the ground truth image. The loss is expressed as the angle between the pixel intensities vectorized across the channels of the input and ground truth image as

$$\mathcal{L}_c = \sum_j \arccos \frac{\hat{I}_j \cdot I_j^{gt}}{\|\hat{I}_j\| \|I_j^{gt}\|}. \tag{7}$$

Overall Spatial Loss: The overall loss is a weighted combination of the aforementioned losses as follows:

$$\mathcal{L}_s = \lambda_1 \mathcal{L}_r + \lambda_2 \mathcal{L}_{sm} + \lambda_3 \mathcal{L}_c. \tag{8}$$

3.3. Learning Temporal Consistency

Image-based methods for underwater video enhancement (UVE) typically aggregate individually enhanced frames into a single video. However, this approach often results in temporal artifacts such as flickering, uneven illumination, and stabilization issues. We propose to learn temporally consistent enhanced frames through a loss based on optical flow [28]. Consider a pair of spatially enhanced frames $\hat{x}_t$ and $\hat{x}_{t+1}$. The optical flow $\mathbf{U}_{t,t+1}$ is a dense estimate of vectors describing the motion of pixels from frame $\hat{x}_t$ to the adjacent frame $\hat{x}_{t+1}$. This flow map can then be used to construct $\hat{x}_{t+1}$ from $\hat{x}_t$ via a warping operation $\mathcal{W}$ described as

$$\hat{x}_t(\mathbf{p}) = \hat{x}_{t+1}(\mathbf{p} + \mathbf{U}_{t,t+1}(\mathbf{p})) = \mathcal{W}(\hat{x}_{t+1}, \mathbf{U}_{t,t+1}), \tag{9}$$

where $\mathbf{p}$ denotes the pixel coordinates. Let the spatial enhancement model be denoted by f_Θ (parameterized by Θ). The optical flow-based loss is optimized in an unsupervised manner given by

$$\begin{aligned}\mathcal{L}_{of}(x_t, x_{t+1}) =& \|\mathcal{W}(f_\Theta(x_t), \mathbf{U}_{t+1,t}) \\ & - f_\Theta(\mathcal{W}(x_t, \mathbf{U}_{t+1,t}))\|_2^2.\end{aligned} \tag{10}$$

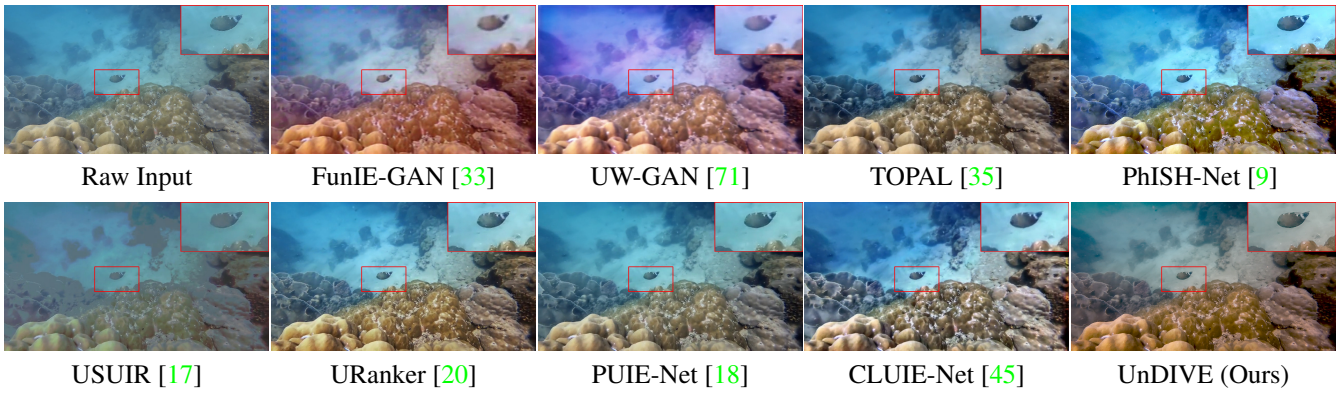


Figure 3. Results of different enhancement methods on frame 54 of the PhuQuoc1_Jun2022.mp4 video from the MVK [69] dataset. The blue hue in the scene is efficiently reduced by UnDIVE, while also improving the contrast in the enhanced image.

We consider the above loss in both forward and backward directions and utilize the total consistency loss $\mathcal{L}_t$ as

$$\mathcal{L}_t = 0.5\mathcal{L}_{of}(x_t, x_{t+1}) + 0.5\mathcal{L}_{of}(x_{t+1}, x_t). \quad (11)$$

This ensures that the optical flow warps the outputs from f_Θ in the same way that f_Θ enhances the warped frames. Moreover, it also introduces temporal stability in the reconstructions since the L_2 loss penalizes large motion outliers resulting in high errors in intensities between the pixels. We obtain the optical flow maps for video frame pairs using an off-the-shelf FastFlowNet [37], which provides real-time and efficient flow estimates.

Training UnDIVE: After the generative prior is learned, we train UnDIVE first using N_s epochs on underwater images using Eqn. (8) followed by N_t epochs of training with temporal consistency in Eqn. (11).

4. Experiments and Results

4.1. Training and Implementation Details

All experiments were carried out on three NVidia Tesla v100 GPUs with 32GB VRAM each. For training the DDPM, we utilize about 100,000 carefully chosen random crops of size 256 from the UIEB [40] dataset and train the DDPM for 100 epochs with a learning rate of 10^{-4} and a batch size of $N_D = 24$. To train the spatial enhancement, the network is first trained using 890 paired images from UIEB using the spatial loss for 100 epochs with a batch size of $N_s = 64$. This is followed by 100 epochs of training on $34,000$ frame pairs from the UVE-38k [62] dataset with both spatial and the unsupervised temporal consistency loss with a batch size of $N_t = 24$. The loss weights in Eqn. (8) are chosen as $(\lambda_1, \lambda_2, \lambda_3) = (1.0, 0.2, 0.1)$.

Datasets: For evaluating the enhancements, we consider four underwater video datasets: **VDD-C** [11] - a dataset of images of divers, drawn from videos taken in pool and field environments, and primarily meant for diver detection and tracking. **Brackish** [57] - a dataset consisting of 9 video sequences with multiple synthetic distortions. Brackish was created to study marine organisms in a brackish strait. **UOT32** [36] - 20 videos requiring enhancement from a dataset for benchmarking object tracking algorithms. **MVK** [69] - 10 videos from a large-scale high-resolution (UHD) dataset collected from seas around the world.

Metrics: We consider multiple image and video no-reference (NR) quality metrics. The image metrics consist of **UCIQE** (Underwater Colour Image Quality Evaluation) [82], **CCF** (Colorfulness Contrast Fog density index) [74], **UIQM** (Underwater Image Quality Measure), **UICM** (Underwater Image Colorfulness Measure), **UISM** (Underwater Image Sharpness Measure) and **UIConM** (Underwater Image Contrast Measure) [56].

The video metrics consist of **VSFA** [42], **FastVQA** (Fast Video Quality Assessment) [77] and **DOVER** (Disentangled Objective Video Quality Evaluator) (the technical quality score of DOVER) [78]. All metrics indicate higher quality when their corresponding values are higher.

4.2. Results and Comparisons

Comparing Methods: We compare UnDIVE with multiple state-of-the-art enhancement methods: (1) **FunIE-GAN** [33] - a fast enhancement method based on GANs, (2) **UW-GAN** [71] - an unsupervised GAN-based method that performs dehazing, (3) **TOPAL** [35] - a perceptual adversarial fusion network for UIE, (4) **PhISH-Net** [9] - a physics-inspired UIE network that operates on an image formation model, (5) **USUIR** [17] - an unsupervised underwater image restoration method using a homology constraint, (6) **URanker** [20] - a ranking-based image quality assessment method that utilizes a histogram prior

Dataset	Method	Image Quality Metrics						Video Quality Metrics		
		UCIQE (↑)	UIQM (↑)	UIConM (↑)	UISM (↑)	UICM (↑)	CCF (↑)	VSFA (↑)	FastVQA (↑)	DOVER (↑)
VDD-C [11]	FunIE-GAN [33]	0.5412	0.7228	0.5774	2.1575	4.1292	14.7077	0.6212	0.3252	2.2951
	UW-GAN [71]	0.5340	0.6758	0.5607	1.8062	3.4402	11.9872	0.6330	0.4720	4.3301
	TOPAL [35]	0.5282	0.7325	0.5997	2.1860	2.3578	15.7300	0.7027	0.6588	7.2169
	PhISH-Net [9]	0.6256	1.2666	0.8945	5.0350	8.9778	52.6171	0.7395	0.5520	5.2110
	USUIR [17]	0.4806	0.5095	0.4242	1.5376	0.5568	6.6373	0.6169	0.4029	4.0153
	URanker [20]	0.5639	0.9126	0.7507	2.5199	4.6189	14.0485	0.7323	0.6711	7.2502
	PUIE-Net [18]	0.5281	0.8216	0.6677	2.4717	3.0629	12.9824	0.7306	0.6914	7.7974
	CLUIE-Net [45]	0.5209	0.8431	0.6867	2.5228	3.0920	16.9909	0.6946	0.3508	6.9858
	UnDIVE	0.5768	1.2222	0.8659	5.2803	5.8755	47.1940	0.7490	0.6989	7.9712
Brackish [57]	FunIE-GAN [33]	0.5978	0.6084	0.5189	1.4411	3.2392	14.8027	0.5126	0.2920	3.1907
	UW-GAN [71]	0.5857	0.4960	0.4331	1.0128	3.0605	9.3128	0.5677	0.4397	3.6762
	TOPAL [35]	0.5584	0.6888	0.5919	1.7384	1.9933	28.5244	0.5850	0.3545	3.1957
	PhISH-Net [9]	0.5480	0.5056	0.4515	0.9529	2.6692	13.5727	0.6421	0.3932	3.0815
	USUIR [17]	0.5414	0.6724	0.5798	1.6155	2.5423	7.2011	0.6495	0.4235	4.8267
	URanker [20]	0.5489	0.5754	0.5369	0.8297	2.7944	8.7077	0.6366	0.2620	2.7925
	PUIE-Net [18]	0.5450	0.5039	0.4714	0.7404	2.1551	11.1319	0.6280	0.4754	3.5308
	CLUIE-Net [45]	0.5766	0.7480	0.6778	1.3819	3.0139	16.8279	0.6275	0.3508	3.2155
	UnDIVE	0.5734	0.5466	0.5086	0.9128	2.6213	14.4568	0.6516	0.5151	3.9462
UOT32 [36]	FunIE-GAN [33]	0.5584	1.0009	0.7291	3.8875	5.2395	24.5786	0.6145	0.1928	2.2915
	UW-GAN [71]	0.5476	0.7475	0.6028	2.1883	4.0130	12.6571	0.6285	0.3116	3.3599
	TOPAL [35]	0.5627	0.8673	0.7087	2.4709	4.1826	22.7635	0.6843	0.5345	6.0123
	PhISH-Net [9]	0.6068	1.0523	0.7400	4.2980	6.6543	40.1159	0.7051	0.4841	5.0983
	USUIR [17]	0.4950	0.8299	0.6512	2.8675	2.1537	9.1856	0.6742	0.5620	5.9473
	URanker [20]	0.5642	0.8622	0.7034	2.4719	4.6904	16.0681	0.6952	0.6711	5.7439
	PUIE-Net [18]	0.5540	0.7422	0.6039	2.1370	3.6716	13.9865	0.7074	0.5731	6.4301
	CLUIE-Net [45]	0.5609	0.9919	0.7627	3.4334	4.4822	22.9237	0.6946	0.6052	6.5477
	UnDIVE	0.5870	1.0454	0.7585	4.2328	5.4106	39.3131	0.7165	0.5782	6.3157
MVK [69]	FunIE-GAN [33]	0.5857	0.9507	0.8204	1.9898	6.5807	15.8532	0.6503	0.3320	4.2376
	UW-GAN [71]	0.5841	0.9470	0.8296	1.8691	6.1803	15.9746	0.6332	0.4973	5.2086
	TOPAL [35]	0.5919	1.1485	0.9792	2.7174	6.1778	18.5322	0.7503	0.7500	8.1786
	PhISH-Net [9]	0.9528	1.4684	1.0858	5.2185	10.7098	59.5795	0.7547	0.7486	8.2122
	USUIR [17]	0.5171	0.7412	0.6633	1.5070	2.6042	8.7099	0.7127	0.6671	6.1041
	URanker [20]	0.5973	1.2178	1.0197	3.0411	7.2402	18.2012	0.7871	0.7805	8.5637
	PUIE-Net [18]	0.5776	1.0748	0.9244	2.4490	5.7571	17.0048	0.7821	0.7655	8.3190
	CLUIE-Net [45]	0.5909	1.2080	1.0319	2.8679	6.1420	26.7405	0.7319	0.7542	7.1492
	UnDIVE	0.6226	1.4537	1.0864	5.2944	7.8019	56.1459	0.7689	0.7910	9.0325

Table 1. Performance comparison of UnDIVE with other enhancement methods on four datasets: VDD-C [11], Brackish [57], UOT32 [36] and MVK [69] using various image and video quality metrics. The best and second best are highlighted with purple and blue respectively.

to learn UIE, (7) **PUIE-Net** [18] - an uncertainty inspired UIE method that generates maximally probable enhancement outputs, and (8) **CLUIE-Net** [45] - a UIE model that learns from multiple available enhancement candidates.

Quantitative Results: We compare UnDIVE with the aforementioned methods and report all the quality metrics in Table 1. To ensure fairness in all evaluations, we compute all metrics on enhanced videos with a fixed frame resolution of 1024×576. We consistently see that UnDIVE outperforms other methods in terms of video metrics, but it is among the top two in image metrics. Since most of the UIE methods are trained on images, they naturally perform better on image metrics. We note that UnDIVE performs

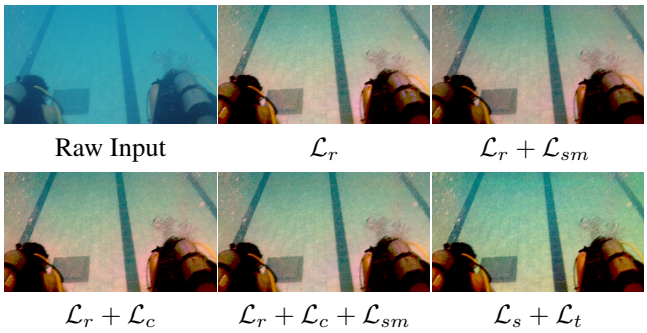


Figure 4. Results from the ablation study on the effect of different loss components. The model that uses $\mathcal{L}_s$ and $\mathcal{L}_t$ effectively removes the spurious reddish hue.

$\mathcal{L}_r$	$\mathcal{L}_{sm}$	$\mathcal{L}_c$	$\mathcal{L}_t$	FastVQA	DOVER
✓				0.6329	5.9192
✓	✓			0.6351	6.1339
✓		✓		0.6346	6.0749
✓	✓	✓		0.6464	6.3021
✓	✓	✓	✓	0.6989	7.9712

Table 2. Effect of different loss components of UnDIVE on UVE performance.

well specifically on UISM, VSFA, FastVQA, and DOVER. This could be attributed to better-restored sharpness in the enhanced frames along with reduced temporal artifacts like non-uniform illumination across frames. On VDD-C and MVK, UnDIVE consistently offers the best enhancement. However, for Brackish, image metrics show a drop and do not align with visual observations (detailed analysis in supplementary). For UOT32, while PhISH-Net performs best on image metrics, URanker and CLUIE-Net provide better temporal consistency. Overall, the metrics (video metrics in particular) confirm UnDIVE's strong generalization across diverse underwater scenes.

Qualitative Evaluation: Although the NR quality metrics provide a reasonable evaluation of different methods, they do not perfectly correlate with subjective human opinions. Moreover, the lack of ground-truth data makes the quantitative evaluation challenging, making it important to observe the results from a visual perspective. Methods apart from UnDIVE often generate color-based artifacts or reduced illumination in the enhanced images, as seen in Fig. 3. UnDIVE significantly reduces water scattering effects and enhances colors, unlike other methods that fail to reduce the blue hue in the input.

DDPM Prior	Image Pretraining	FastVQA	DOVER
		0.4830	4.7603
	✓	0.6876	7.7389
✓		0.6678	7.0849
✓	✓	0.6989	7.9712

Table 3. Effect of generative prior and pre-training on underwater image data on UVE performance.

4.3. Ablation Experiments

Effect of Loss Components: We analyze the effect of each component of the spatial loss in Eqn. (8) and the temporal loss in Eqn. (11). From Table 2, we observe that introducing each loss consistently improves the performance of the framework. From the bottom two rows in Table 2, we observe that the temporal consistency loss gives a significant improvement in both video metrics, validating its efficacy in UVE. We also provide some visual enhancement results from using the loss components in training UnDIVE in Fig. 4. We observe that using the reconstruction loss along with smoothness loss creates reddish coloration around the divers. With the introduction of all the spatial loss components (collectively denoted by $\mathcal{L}_s$) along with the temporal loss $\mathcal{L}_t$ we observe the resolution of the color issue as well as a reduction in the blue color in the input image.

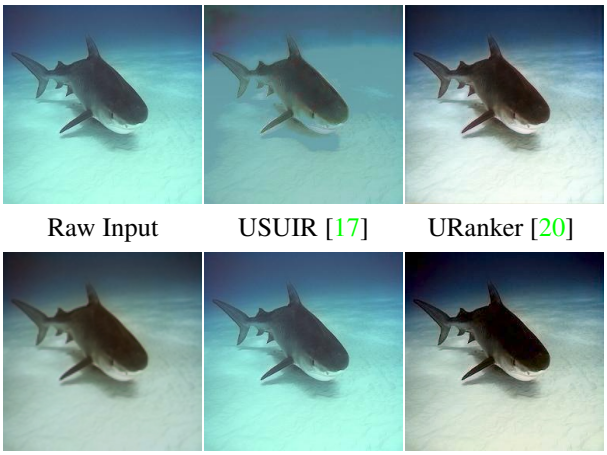
Effect of Generative Prior and Image Pre-training: We study the effectiveness of learning the generative prior and pre-training the spatial enhancement module on images from UIEB [40]. In cases where the prior is not used, the encoder $\mathcal{E}$ is trained from scratch along with the rest of the trainable parameters of f_Θ. Table 3 shows that using the prior significantly improves performance compared to an encoder trained from scratch, indicating that learning robust image representations enhances the results. Additionally, pre-training on UIEB before training on video frame-pairs provides a substantial performance boost by leveraging diverse underwater scene information. Visual results in Fig. 5 show that models trained without the prior or image pre-training reconstruct colors inaccurately. With only image pre-training, the enhanced image suffers from poor illumination, while with only the prior, it shows better contrast but a greenish hue. When both components are used, the colors are more vivid and accurate. More detailed analyses and experiments are provided in the supplementary.

4.4. UnDIVE for Underwater Image Enhancement

We enhance images from the EUVP [34] and UIEB [40] datasets in a cross-dataset setting and report the results in Table 4. We compare UnDIVE with four recent state-of-

Raw Input | No prior or pre-training | Only pre-training | Only prior | Prior and pre-training

Figure 5. Effect of the generative prior and the image pre-training on enhancement.



Raw Input | USUIR [17] | URanker [20]

PUIE-Net [18] | CLUIE-Net [45] | UnDIVE (Ours)

Figure 6. Cross-dataset results of various UIE methods on EUVP.

the-art UIE methods and notice that UnDIVE achieves best or second best performances consistently, validating its generalization capability beyond the task of UVE. From Fig. 6, we notice that UnDIVE is able to efficiently enhance the input image. URanker is able to provide better illumination while also reducing the blue hue. However, other methods like USUIR and CLUIE-Net are unable to mitigate backscatter from the water scattering.

4.5. Runtime Analysis

Table 5 presents a runtime analysis, including model and computational complexity for all enhancement methods. While GAN-based methods FunIE-GAN and UW-GAN offer the fastest computation, they yield lower-quality enhancements. Per-frame computation times are shown for resolutions of 256×256 for GAN-based methods and 512×512 for the others.

Limitations: The lack of clear ground-truth references makes training with annotated frame-pairs challenging. Most quality assessment methods correlate poorly with human perception, making quantitative results somewhat misleading, particularly for image metrics. Video-based quality metrics, however, are more reliable for performance comparison (see supplementary for details).

Method	EUVP [34]			UIEB [40]		
	PSNR (↑)	SSIM (↑)	UCIQE (↑)	PSNR (↑)	SSIM (↑)	UCIQE (↑)
USUIR	14.4970	0.7093	0.5111	12.3075	0.6073	0.5432
URanker	18.6961	0.9315	0.5781	20.4140	0.9074	0.6113
PUIE-Net	17.8718	0.7199	0.5720	19.4576	0.8560	0.5993
CLUIE-Net	17.8986	0.8886	0.5895	13.7158	0.6833	0.6031
UnDIVE	18.0893	0.9002	0.5884	20.4598	0.8717	0.6167

Table 4. UIE performances on the EUVP (cross-dataset) and UIEB (intra-domain) datasets.

Method	Runtime (s) (↓)	GFLOPs (↓)	Parameters (M) (↓)
FunIE-GAN	0.00122	10.24	7.019
UW-GAN	0.00037	0.004	1.925
TOPAL	1.98956	111.6	36.67
PhISH-Net	0.47244	0.090	0.556
USUIR	0.06322	10.33	0.225
URanker	0.03189	14.74	3.146
PUIE-Net	0.21560	33.64	1.401
CLUIE-Net	0.09292	24.68	13.39
UnDIVE	0.20908	7.153	6.723

Table 5. Runtime and Complexity Analysis of UnDIVE.

5. Conclusion

We propose a UVE method that effectively enhances underwater videos across various water types and degradations. This work is the first to use a generative prior from a self-supervised DDPM to guide the learning of a downstream enhancement task. By incorporating temporal consistency into the enhancement model, we demonstrate that UnDIVE can seamlessly generalize across multiple underwater video datasets (diverse water types) in real time. Furthermore, we show that UnDIVE's capabilities extend beyond UVE (for instance, UIE), potentially opening new avenues for research in other marine applications.

Acknowledgements

Suhas Srinath acknowledges the Ministry of Education, India. Prathosh A. P. acknowledges support from Infosys foundation and IISc startup grant for computational resources.

References

[1] 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. 2

[2] 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. 4

[3] Derya Akkaynak, Tali Treibitz, Tom Shlesinger, Yossi Loya, Raz Tamir, and David Iluz. What is the space of attenuation coefficients in underwater computer vision? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4931–4940, 2017. 4

[4] Cosmin Ancuti, Codruta Orniana Ancuti, Tom Haber, and Philippe Bekaert. Enhancing underwater images and videos by fusion. In *2012 IEEE conference on computer vision and pattern recognition*, pages 81–88. IEEE, 2012. 2

[5] 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, 2017. 2

[6] Eric P Bennett and Leonard McMillan. Video enhancement using per-pixel virtual exposures. In *ACM SIGGRAPH 2005 Papers*, pages 845–852. Association for Computing Machinery, 2005. 2

[7] 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. 2

[8] 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. 2

[9] Aditya Chandrasekar, Manogna Sreenivas, and Soma Biswas. Phish-net: Physics inspired system for high resolution underwater image enhancement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1506–1516, 2024. 2, 4, 5, 6

[10] Chen Chen, Qifeng Chen, Minh N Do, and Vladlen Koltun. Seeing motion in the dark. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 3185–3194, 2019. 3

[11] Karin de Langis, Michael Fulton, and Junaed Sattar. An analysis of deep object detectors for diver detection. *arXiv preprint arXiv:2012.05701*, 2020. 5, 6

[12] 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. 2

[13] Paulo LJ Drews, Erickson R Nascimento, Silvia SC Botelho, and Mario Fernando Montenegro Campos. Underwater depth estimation and image restoration based on single images. *IEEE computer graphics and applications*, 36(2):24–35, 2016. 2

[14] Paulo LJ Drews, Erickson R Nascimento, Silvia SC Botelho, and Mario Fernando Montenegro Campos. Underwater depth estimation and image restoration based on single images. *IEEE computer graphics and applications*, 36(2):24–35, 2016. 2

[15] Gabriel Eilertsen, Rafal K Mantiuk, and Jonas Unger. Single-frame regularization for temporally stable cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11176–11185, 2019. 3

[16] Cameron Fabbri, Md Jahidul Islam, and Junaed Sattar. Enhancing underwater imagery using generative adversarial networks. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 7159–7165. IEEE, 2018. 1, 2

[17] Zhenqi Fu, Huangxing Lin, Yan Yang, Shu Chai, Liyan Sun, Yue Huang, and Xinghao Ding. Unsupervised underwater image restoration: From a homology perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 643–651, 2022. 1, 2, 5, 6, 8

[18] Zhenqi Fu, Wu Wang, Yue Huang, Xinghao Ding, and Kai-Kuang Ma. Uncertainty inspired underwater image enhancement. In *European Conference on Computer Vision (ECCV)*, pages 465–482, 2022. 2, 5, 6, 8

[19] Adrian Galdran, David Pardo, Artzai Picón, and Aitor Alvarez-Gila. Automatic red-channel underwater image restoration. *Journal of Visual Communication and Image Representation*, 26:132–145, 2015. 2

[20] Chunle Guo, Ruiqi Wu, Xin Jin, Linghao Han, Zhi Chai, Weidong Zhang, and Chongyi Li. Underwater ranker: Learn which is better and how to be better. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. 2, 5, 6, 8

[21] 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

[22] 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. 2

[23] Ruyue Han, Yang Guan, Zhibin Yu, Peng Liu, and Haiyong Zheng. Underwater image enhancement based on a spiral generative adversarial framework. *IEEE Access*, 8:218838–218852, 2020. 2

[24] Najmul Hassan, Sami Ullah, Naeem Bhatti, Hasan Mahmood, and Muhammad Zia. The retinex based improved underwater image enhancement. *Multimedia Tools and Applications*, 80:1839–1857, 2021. 2

[25] John D Hedley, Chris M Roelfsema, Iliana Chollett, Alastair R Harborne, et al. Remote sensing of coral reefs for monitoring and management: a review. *Remote Sensing*, 8(2):118, 2016. 1

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3

[27] Pooja Honnutagi, YS Laitha, and VD Mytri. Underwater video enhancement using manta ray foraging lion optimization-based fusion convolutional neural network. *International Journal of Image and Graphics*, 23(04):2350031, 2023. 1

[28] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981. 3, 4

[29] Guojia Hou, Nan Li, Peixian Zhuang, Kunqian Li, Haihan Sun, and Chongyi Li. Non-uniform illumination underwater image restoration via illumination channel sparsity prior. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 2

[30] Shirui Huang, Keyan Wang, Huan Liu, Jun Chen, and Yunsong Li. Contrastive semi-supervised learning for underwater image restoration via reliable bank. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18145–18155, 2023. 1

[31] Shirui Huang, Keyan Wang, Huan Liu, Jun Chen, and Yunsong Li. Contrastive semi-supervised learning for underwater image restoration via reliable bank. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18145–18155, 2023. 2

[32] 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. 2

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

[34] 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. 7, 8

[35] Zhiying Jiang, Zhuoxiao Li, Shuzhou Yang, Xin Fan, and Risheng Liu. Target oriented perceptual adversarial fusion network for underwater image enhancement. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2022. 2, 5, 6

[36] Landry Kezebou, Victor Oludare, Karen Panetta, and Sos S Agaian. Underwater object tracking benchmark and dataset. In *2019 IEEE International Symposium on Technologies for Homeland Security (HST)*, pages 1–6. IEEE, 2019. 1, 5, 6

[37] Lingtong Kong, Chunhua Shen, and Jie Yang. Fastflownet: A lightweight network for fast optical flow estimation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021. 5

[38] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 170–185, 2018. 3

[39] Chongyi Li, Saeed Anwar, and Fatih Porikli. Underwater scene prior inspired deep underwater image and video enhancement. *Pattern Recognition*, 98:107038, 2020. 2

[40] 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. 1, 2, 5, 7, 8

[41] Chongyi Li, Jichang Guo, and Chunle Guo. Emerging from water: Underwater image color correction based on weakly supervised color transfer. *IEEE Signal processing letters*, 25(3):323–327, 2018. 2

[42] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *Proceedings of the 27th ACM international conference on multimedia*, pages 2351–2359, 2019. 5

[43] Dasong Li, Xiaoyu Shi, Yi Zhang, Ka Chun Cheung, Simon See, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. A simple baseline for video restoration with grouped spatial-temporal shift. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9822–9832, 2023. 2

[44] Jie Li, Katherine A Skinner, Ryan M Eustice, and Matthew Johnson-Roberson. Watergan: Unsupervised generative network to enable real-time color correction of monocular underwater images. *IEEE Robotics and Automation letters*, 3(1):387–394, 2017. 1, 2

[45] 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(6):2561–2576, 2023. 2, 5, 6, 8

[46] Siyuan Li, Yue Luo, Ye Zhu, Xun Zhao, Yu Li, and Ying Shan. Enforcing temporal consistency in video depth estimation. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1145–1154, 2021. 3

[47] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *IEEE Transactions on Image Processing*, 2024. 2

[48] Liang Liu, Jiangning Zhang, Ruifei He, Yong Liu, Yabiao Wang, Ying Tai, Donghao Luo, Chengjie Wang, Jilin Li, and Feiyue Huang. Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation. In *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2020. 4

[49] Shiben Liu, Huijie Fan, Sen Lin, Qiang Wang, Naida Ding, and Yandong Tang. Adaptive learning attention network for underwater image enhancement. *IEEE Robotics and Automation Letters*, 7(2):5326–5333, 2022. 2

[50] Huimin Lu, Yujie Li, Lifeng Zhang, and Seiichi Serikawa. Contrast enhancement for images in turbid water. *JOSA A*, 32(5):886–893, 2015. 1

[51] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. Mbllen: Low-light image/video enhancement using cnns. In *BMVC*, volume 220, page 4, 2018. 2

[52] Janarthanan Mathiazhagan, Sabitha Gauni, and Rajesvari Mohan. Underwater video transmission with video enhancement using reduce hazing algorithm. *Journal of Optical Communications*, 45(2):379–388, 2024. 1

[53] Monika Mathur and Nidhi Goel. Enhancement of underwater images using white balancing and rayleigh-stretching. In *2018 5th International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 924–929. IEEE, 2018. 2

[54] Sangeetha Mohan and Philomina Simon. Underwater image enhancement based on histogram manipulation and multiscale fusion. *Procedia Computer Science*, 171:941–950, 2020. 2

[55] David Nilsson and Cristian Sminchisescu. Semantic video segmentation by gated recurrent flow propagation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6819–6828, 2018. 3

[56] Karen Panetta, Chen Gao, and Sos Agaian. Human-visual-system-inspired underwater image quality measures. *IEEE Journal of Oceanic Engineering*, 41(3):541–551, 2015. 5

[57] Malte Pedersen, Joakim Bruslund Haurum, Rikke Gade, Thomas B. Moeslund, and Niels Madsen. Detection of marine animals in a new underwater dataset with varying visibility. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 5, 6

[58] Soo-Chang Pei and Chia-Yi Chen. Underwater images enhancement by revised underwater images formation model. *IEEE Access*, 10:108817–108831, 2022. 2

[59] Lintao Peng, Chunli Zhu, and Liheng Bian. U-shape transformer for underwater image enhancement. *IEEE Transactions on Image Processing*, 32:3066–3079, 2023. 1

[60] Yan-Tsung Peng, Keming Cao, and Pamela C Cosman. Generalization of the dark channel prior for single image restoration. *IEEE Transactions on Image Processing*, 27(6):2856–2868, 2018. 2

[61] Qi Qi, Kunqian Li, Haiyong Zheng, Xiang Gao, Guojia Hou, and Kun Sun. Sguie-net: Semantic attention guided underwater image enhancement with multi-scale perception. *IEEE Transactions on Image Processing*, 31:6816–6830, 2022. 2

[62] Qi Qi, Yongchang Zhang, Fei Tian, QM Jonathan Wu, Kunqian Li, Xin Luan, and Dalei Song. Underwater image co-enhancement with correlation feature matching and joint learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. 2, 5

[63] 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. 1

[64] Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. Artistic style transfer for videos. In *Pattern Recognition: 38th German Conference, GCPR 2016, Hannover, Germany, September 12-15, 2016, Proceedings 38*, pages 26–36. Springer, 2016. 3

[65] Hanumant Singh, Jonathan Adams, David Mindell, and Brendan Foley. Imaging underwater for archaeology. *Journal of Field Archaeology*, 27(3):319–328, 2000. 1

[66] Jitendra P Sonawane, Mukesh D Patil, and Gajanan K Birajdar. Adaptive rule-based colour component weight assignment strategy for underwater video enhancement. *The Imaging Science Journal*, pages 1–22, 2023. 1

[67] Jaime Spencer, Chris Russell, Simon Hadfield, and Richard Bowden. Deconstructing self-supervised monocular reconstruction: The design decisions that matter. *Transactions on Machine Learning Research*, 2022. Reproducibility Certification. 4

[68] Jaime Spencer, Chris Russell, Simon Hadfield, and Richard Bowden. Kick back & relax: Learning to reconstruct the world by watching slowtv. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 4

[69] Quang-Trung Truong, Tuan-Anh Vu, Tan-Sang Ha, Jakub Lokoč, Yue-Him Wong, Ajay Joneja, and Sai-Kit Yeung. Marine video kit: a new marine video dataset for content-based analysis and retrieval. In *International Conference on Multimedia Modeling*, pages 539–550. Springer, 2023. 5, 6

[70] Yosuke Ueki and Masaaki Ikehara. Weighted generalization of dark channel prior with adaptive color correction for defogging. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 685–689. IEEE, 2021. 2

[71] Nan Wang, Yabin Zhou, Fenglei Han, Haitao Zhu, and Yaojing Zheng. Uwgan: Underwater gan for real-world underwater color restoration and dehazing, 2019. 2, 5, 6

[72] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *Advances in Neural Information Processing Systems*, 31, 2018. 3

[73] Yang Wang, Yang Cao, Jing Zhang, Feng Wu, and Zheng-Jun Zha. Leveraging deep statistics for underwater image enhancement. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(3s):1–20, 2021. 2

[74] Yan Wang, Na Li, Zongying Li, Zhaorui Gu, Haiyong Zheng, Bing Zheng, and Mengnan Sun. An imaging-inspired no-reference underwater color image quality assessment metric. *Computers & Electrical Engineering*, 70:904–913, 2018. 5

[75] Yi Wang, Hui Liu, and Lap-Pui Chau. Single underwater image restoration using adaptive attenuation-curve prior. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 65(3):992–1002, 2017. 2

[76] Louis L Whitcomb. Underwater robotics: Out of the research laboratory and into the field. In *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065)*, volume 1, pages 709–716. IEEE, 2000. 1

[77] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. In *European conference on computer vision*, pages 538–554. Springer, 2022. 5

[78] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20144–20154, 2023. 5

[79] Jun Xie, Guojia Hou, Guodong Wang, and Zhenkuan Pan. A variational framework for underwater image dehazing and deblurring. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(6):3514–3526, 2021. 2

[80] Yaofeng Xie, Lingwei Kong, Kai Chen, Ziqiang Zheng, Xiao Yu, Zhibin Yu, and Bing Zheng. Uveb: A large-scale benchmark and baseline towards real-world underwater video enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22358–22367, 2024. 1

[81] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127:1106–1125, 2019. 2

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

[83] Fan Zhang, Yu Li, Shaodi You, and Ying Fu. Learning temporal consistency for low light video enhancement from single images. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4965–4974, 2021. 3

[84] Weidong Zhang, Peixian Zhuang, Hai-Han Sun, Guohou Li, Sam Kwong, and Chongyi Li. Underwater image enhancement via minimal color loss and locally adaptive contrast enhancement. *IEEE Transactions on Image Processing*, 31:3997–4010, 2022. 2

[85] Weidong Zhang, Peixian Zhuang, Hai-Han Sun, Guohou Li, Sam Kwong, and Chongyi Li. Underwater image enhancement via minimal color loss and locally adaptive contrast enhancement. *IEEE Transactions on Image Processing*, 31:3997–4010, 2022. 2

[86] Jingchun Zhou, Lei Pang, Dehuan Zhang, and Weishi Zhang. Underwater image enhancement method via multi-interval subhistogram perspective equalization. *IEEE Journal of Oceanic Engineering*, 48(2):474–488, 2023. 1

[87] Jingchun Zhou, Jiaming Sun, Weishi Zhang, and Zifan Lin. Multi-view underwater image enhancement method via embedded fusion mechanism. *Engineering applications of artificial intelligence*, 121:105946, 2023. 1

[88] Jingchun Zhou, Tongyu Yang, Weishen Chu, and Weishi Zhang. Underwater image restoration via backscatter pixel prior and color compensation. *Engineering Applications of Artificial Intelligence*, 111:104785, 2022. 2

[89] Peixian Zhuang, Jiamin Wu, Fatih Porikli, and Chongyi Li. Underwater image enhancement with hyper-laplacian reflectance priors. *IEEE Transactions on Image Processing*, 31:5442–5455, 2022. 2

[90] Karel Zuiderveld. Contrast limited adaptive histogram equalization. In *Graphics gems IV*, pages 474–485. 1994. 2