U2NeRF: Unifying Unsupervised Underwater Image Restoration and Neural Radiance Fields

Mukund Varma T[1†*], Vinayak Gupta[1†*], Manoj S[1†*], [*], Kaushik Mitra[1†]
[1]Indian Institute of Technology Madras
{mukundvarmat, vinayakguptapokal, manoj.s.2908}@gmail.com,
kmitra@ee.iitm.ac.in

Abstract

*Underwater images suffer from colour shifts, low contrast, and haziness due to light absorption, refraction, scattering and restoring these images has warranted much attention. In this work, we present Unsupervised Underwater Neural Radiance Field (**U2NeRF**), a transformer-based architecture that learns to render and restore novel views conditioned on multi-view geometry simultaneously. Due to the absence of supervision, we attempt to implicitly bake restoring capabilities onto the NeRF pipeline and disentangle the predicted color into several components - scene radiance, direct transmission map, backscatter transmission map, and global background light, and when combined reconstruct the underwater image in a self-supervised manner. In addition, we release an Underwater View Synthesis (**UVS**) dataset consisting of 12 underwater scenes, containing both synthetically-generated and real-world data. Our experiments demonstrate that when optimized on a single scene, U2NeRF outperforms several baselines by as much LPIPS ↓11%, UIQM ↑5%, UCIQE ↑4% (on average) and showcases improved rendering and restoration capabilities. The multi-view inductive prior enables U2NeRF to be trained on multiple scenes, and successfully transfer to unseen scenes with or without per-scene fine-tuning. Code will be made available upon acceptance.*

1. Introduction

Underwater images suffer from degradation due to poor, complex lighting conditions in water - more specifically due to light scattering, absorption and refraction [11]. Therefore, it is important to develop methods that can enhance underwater images, so that they are more suitable for visualization and downstream tasks like detection, tracking, etc. Recent advancements in deep learning has enabled great performance in several computer vision tasks [9, 13], including underwater image enhancement [6, 12]. Most of these methods [8, 45] rely on synthetic training data due to the absence of large-scale real-world underwater image datasets with corresponding ground truth restored images. However, the synthesized data may not capture complex real-world degradation and thus suffer from domain shifts [6]. More recently, "zero-shot" methods [3, 16] train a small image-specific network during test time and do not use any supervision other than the input image itself. However, they are not suitable for real-world applications due to a large number of optimization iterations at test time.

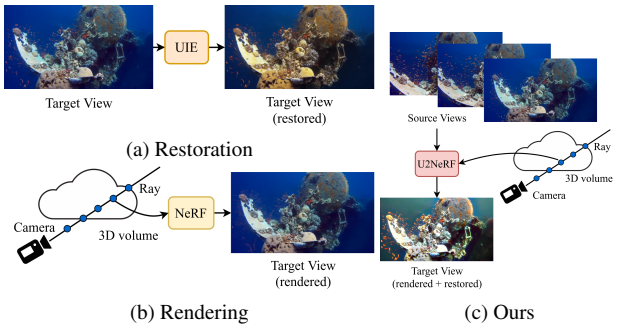


Figure 1. Unlike standalone methods like Radiance Fields (for rendering) and Image Enhancement (for restoration), our method U2NeRF simultaneously restores and renders an underwater scene.

Neural Radiance Fields [23] and its follow-up works [1, 5, 24] have achieved remarkable success on novel view synthesis, generating photo-realistic, high-resolution, and view-consistent scenes. However, all these methods are trained on scenes containing clean, high-resolution images. Since NeRFs integrates information from multiple views, we hypothesize that these methods have a strong potential to be leveraged in multi-frame image restoration tasks - and in the context of this paper, underwater image enhancement.

In this paper, we attempt to use NeRFs for simultaneous

[*]Equal contribution.
[†]Correspondence to: Mukund Varma T, and Kaushik Mitra.

novel view rendering and restoration. However, most methods operate at a pixel level, limiting its capacity to automatically restore the predicted color. We demonstrate that by predicting image patches (rather than pixels), we provide sufficient spatial context for restoration. Motivated by [3], our method disentangles the predicted image patch into 4 components, namely scene radiance, direct transmission map, back scatter transmission map, and global background light. These components are later combined to reconstruct the original image, and along with suitable regularization enables our network to be trained in a self-supervised manner in the absence of clean ground truth image. Towards this end, we adapt the recently proposed Generalizable NeRF Transformer (GNT) [36], which is composed of a *view transformer* to aggregate multi-view information and render a novel view by composing colors along a ray using *ray transformer*. Our method dubbed **U2NeRF** (Unsupervised Underwater Neural Radiance Field), trained in a fully unsupervised setting, learns to simultaneously renders and restore a novel view. Our primary contributions can be summarized as follows:

1. We extend the idea of radiance fields for the novel task of simultaneously rendering and restoring novel views, more specifically for underwater scenes.

2. Our proposed method U2NeRF, augments existing radiance fields with spatial awareness, and when combined with a physics-informed image formation model can successfully restore underwater images.

3. We contribute novel UVS Dataset consisting of 12 underwater scenes, containing both synthetically-generated and real-world data for novel view synthesis. Our proposed approach achieves best performance across perceptual (LPIPS $\downarrow$11%) and color restoration metrics (UIQM $\uparrow$5%, UCIQE $\uparrow$4%) in single scene rendering. We additionally showcase generalization capabilities of the network.

4. Our results indicate that U2NeRF implicitly learns to generate physically meaningful image components, bringing us one step closer to using transformers as a universal modeling tool for graphics.

2. Related Work

Neural Radiance Fields. NeRF introduced by [23] synthesizes consistent and photorealistic novel views by fitting each scene as a continuous 5D radiance field parameterized by an MLP. Since then, several works have improved NeRFs further. For example, Mip-NeRF [1, 2] efficiently addresses scale of objects in unbounded scenes, Nex [41] models large view dependent effects, others [25, 38, 44] improve the surface representation, extend to dynamic scenes [27, 28, 32] , introduce lighting and reflection modeling [4, 37], or leverage depth to regress from few views [7, 42]. A recent work [29] demonstrates the ability of NeRFs for burst denoising. Unlike other methods, our work aims to simultaneously render and restore a novel view, more specifically in the context of underwater scenes.

Underwater Image Restoration. The existing UIE approaches can roughly be divided in three groups, which are non-physical model based, physical model based and learning based methods. The first set of methods involve enhancing pixel intensity directly to achieve a better image quality without using any physics-based constraints. These set of methods include White balance (WB) [19], Retinex [33], and conventional contrast limited adaptive histogram equalisation [31]. The second set of methods are prior-based approaches, that treats underwater image enhancement as an inverse problem and restores the images by utilising physical imaging models. The Jaffe–McGlamery [14, 21] underwater image model is one of the well-known physics-based model that models the observed image as a function of back-scattered light, distance between the camera and the scene and light attenuation coefficient. Many images priors were also explored later, these include Generalized Dark Channel Prior (GDCP) [30] and Underwater Dark Channel Prior (UDCP) [10] which makes use of the red channel information. The third set of methods consist of deep learning-based techniques that develop an enhanced mapping based on a large amount of paired and unpaired training data and automatically extract representations. These methods needed a large set of high-quality aerial-underwater image pairs for supervised learning. In an attempt to generate the counterparts, WaterGAN [17] was introduced, where the generator synthesizes realistic underwater image by modelling light attenuation, light back scattering and camera model. For our work on U2NeRF, We draw inspiration from [3], which proposes an unsupervised CNN model that estimates the scene radiance and transmission maps, and utilizes the physics-based equation to compute the actual underwater image.

3. Unsupervised Underwater Neural Radiance Fields

Our method U2NeRF extends GNT for the task for rendering and restoring novel views in underwater scenes. In this section, we first describe the preliminary of radiance fields, GNT, followed by a detailed description of our proposed method.

3.1. Preliminary

Neural Radiance Fields. NeRFs [23] converts multi-view images into a radiance field and interpolates novel views by re-rendering the radiance field from a new an-

gle. Technically, NeRF models the underlying 3D scene as a continuous radiance field $\mathcal{F} : (\boldsymbol{x}, \boldsymbol{\theta}) \mapsto (\boldsymbol{c}, \sigma)$ parameterized by a Multi-Layer Perceptron (MLP) $\boldsymbol{\Theta}$, which maps a spatial coordinate $\boldsymbol{x} \in \mathbb{R}^3$ together with the viewing direction $\boldsymbol{\theta} \in [-\pi, \pi]^2$ to a color $\boldsymbol{c} \in \mathbb{R}^3$ plus density $\sigma \in \mathbb{R}_+$ tuple. To form an image, NeRF performs the ray-based rendering, where it casts a ray $\boldsymbol{r} = (\boldsymbol{o}, \boldsymbol{d})$ from the optical center $\boldsymbol{o} \in \mathbb{R}^3$ through each pixel (towards direction $\boldsymbol{d} \in \mathbb{R}^3$), and then leverages volume rendering [15] to compose the color and density along the ray between the near-far planes:

$$\boldsymbol{C}(\boldsymbol{r}|\boldsymbol{\Theta}) = \int_{t_n}^{t_f} T(t)\sigma(\boldsymbol{r}(t))\boldsymbol{c}(\boldsymbol{r}(t), \boldsymbol{d})dt, \quad (1)$$
$$\text{where } T(t) = \exp\left(-\int_{t_n}^{t} \sigma(s)ds\right)$$

where $\boldsymbol{r}(t) = \boldsymbol{o} + t\boldsymbol{d}$, t_n and t_f are the near and far planes respectively. In practice, the Eqn. 1 is numerically estimated using quadrature rules [20]. Given images captured from surrounding views with known camera parameters, NeRF fits the radiance field by maximizing the likelihood of simulated results. Suppose we collect all pairs of rays and pixel colors as the training set $\mathcal{D} = \{(\boldsymbol{r}_i, \widehat{\boldsymbol{C}}_i)\}_{i=1}^N$, where N is the total number of rays sampled, and $\widehat{\boldsymbol{C}}_i$ denotes the ground-truth color of the i-th ray, then we train the implicit representation $\boldsymbol{\Theta}$ via the following loss function:

$$\mathcal{L}(\boldsymbol{\Theta}|\mathcal{R}) = \mathbb{E}_{(\boldsymbol{r}, \widehat{\boldsymbol{C}}) \in \mathcal{D}} \|\boldsymbol{C}(\boldsymbol{r}|\boldsymbol{\Theta}) - \widehat{\boldsymbol{C}}(\boldsymbol{r})\|_2^2, \quad (2)$$

Generalizable NeRF Transformer. GNT [36] considers the problem of novel view synthesis as a two stage information process: the multi-view image feature fusion, followed by the sampling-based rendering integration. It is composed of (i) *view transformer* to aggregate pixel-aligned image features from corresponding epipolar lines to predict coordinate-wise features, (ii) *ray transformer* to compose coordinate-wise point features along a traced ray via attention mechanism. More formally, the entire operation can be summarized as follows:

$$\mathcal{F}(\boldsymbol{x}, \boldsymbol{\theta}) = \text{View-Transformer}(\boldsymbol{F}_1(\Pi_1(\boldsymbol{x}), \boldsymbol{\theta}), \cdots, \boldsymbol{F}_N(\Pi_N(\boldsymbol{x}), \boldsymbol{\theta})), \quad (3)$$

where $\text{View-Transformer}(\cdot)$ is a transformer encoder, $\Pi_i(\boldsymbol{x})$ projects position $\boldsymbol{x} \in \mathbb{R}^3$ onto the i-th image plane by applying extrinsic matrix, and $\boldsymbol{F}_i(\boldsymbol{z}, \boldsymbol{\theta}) \in \mathbb{R}^d$ computes the feature vector at position $\boldsymbol{z} \in \mathbb{R}^2$ via bilinear interpolation on the feature grids. The multi-view aggregated point features are fed into the ray transformer, then pooled to extract a single ray feature to predict the target pixel color.

$$\boldsymbol{C}(\boldsymbol{r}) = \text{MLP} \circ \text{Mean} \circ \text{Ray-Transformer}(\mathcal{F}(\boldsymbol{o} + t_1\boldsymbol{d}, \boldsymbol{\theta}), \cdots, \mathcal{F}(\boldsymbol{o} + t_M\boldsymbol{d}, \boldsymbol{\theta})), \quad (4)$$

where $t_1, \cdots, t_M$ are uniformly sampled between near and far plane, Ray-Transformer is a standard transformer encoder.

3.2. Baking Restoration Capabilities onto U2NeRF

NeRF represents 3D scene as a radiance field $\mathcal{F} : (\boldsymbol{x}, \boldsymbol{\theta}) \mapsto (\boldsymbol{c}, \sigma)$, where each spatial coordinate $\boldsymbol{x} \in \mathbb{R}^3$ together with the viewing direction $\boldsymbol{\theta} \in [-\pi, \pi]^2$ is mapped to a color $\boldsymbol{c} \in \mathbb{R}^3$ plus density $\sigma \in \mathbb{R}_+$ tuple. However, a single pixel does not provide sufficient context for automatic restoration. In our work, we first adapt GNT to render an image patch of size p. The final ray feature obtained from the ray transformer block is passed on to a sequence of convolution and upsampling layers. Motivated by [3], we disentangle the underwater image into several components - scene radiance (J), global background light (A) and degradation components - direct and back scatter transmission maps (T_D, T_B) that account for attenuation and light reflection respectively. These individual components can be combined to reconstruct the original image I at pixel i as:

$$I(i) = J(i)T_D(i) + (1 - T_B(i))A \quad (5)$$

This enables our network to be trained in a fully self-supervised manner in the absence of ground truth image. To predict J, T_D, and T_B, we initialize separate output heads to project the final ray feature to the desired patch size. Since A is independent of the input image content, we pass the nearest source image from the target view direction onto a Variational AutoEncoder (VAE) to estimate global background light. In addition to the photometric loss given in Eqn. 2 ($\mathcal{L}_{\text{rec}}$), we (1) minimize the difference between encoded feature z and latent code sampled from Gaussian $\hat{z}$ in the vae ($\mathcal{L}_{\text{kl}}$), (2) minimize the difference between the saturation and brightness of a predicted scene radiance to reduce haze ($\mathcal{L}_{\text{con}}$), (3) minimize the potential color deviations ($\mathcal{L}_{\text{col}}$), (4) ensure constant back-scatter coefficients ($\mathcal{L}_{\text{trans}}$) across channels, (5) enforce constant global background light by minimizing variance within each local neighbourhood ($\mathcal{L}_{\text{glob}}$) as proposed in the original paper [3]. Together, the network is trained to optimize:

$$\mathcal{L} = \lambda_1\mathcal{L}_{rec}, +\lambda_2\mathcal{L}_{con} + \lambda_3\mathcal{L}_{col} + \lambda_4\mathcal{L}_{kl} + \lambda_5\mathcal{L}_{trans} + \lambda_6\mathcal{L}_{glob} \quad (6)$$

where λ indicates the weight for each loss term.

4. Experiments

We conduct extensive experiments to compare U2NeRF with several baseline methods for novel view synthesis and restoration in the context of underwater scenes. We first provide qualitative and quantitative results in the single scene training setting, followed by extending our approach for generalization to unseen scenes.

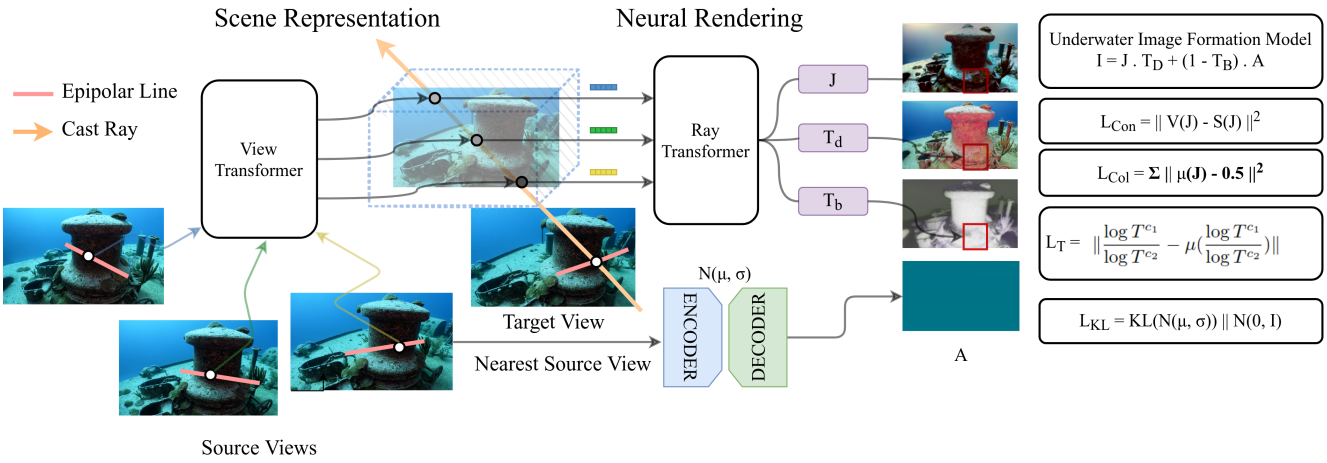


Figure 2. Overview of U2NeRF: 1) Identify source views for a given target view, 2) Extract features for epipolar points using a trainable U-Net like model, 3) For each ray in the target view, sample points and directly predict a target patch disentangled into scene radiance, direct and backscatter transmission maps, and global background light. 4) The individual components are combined based on the image formation model to reconstruct the underwater image which is used as a self-supervision loss.

4.1. Implementation Details

Source and Target view sampling. As described in [39], we construct a training pair of source and target views by first selecting a target view, then identifying a pool of k × N nearby views, from which N views are randomly sampled as source views. This sampling strategy simulates various view densities during training and therefore helps the network generalize better. During training, the values for k and N are uniformly sampled at random from [1-3] and [8-12], respectively.

Network Architecture. To extract features from the source views, we use a U-Net like architecture with a ResNet34 encoder, followed by two up-sampling layers as decoder [39]. Each view transformer block contains a single-headed cross-attention layer while the ray transformer block contains a multi-headed self-attention layer with four heads. The outputs from these attention layers are passed onto corresponding feedforward blocks with a Rectified Linear Unit (RELU) activation and a hidden dimension of 256. A residual connection is applied between the pre-normalized inputs (LayerNorm) and outputs at each layer. For all our single scene experiments, we alternatively stack 4 view and ray transformer blocks while our larger generalization experiments use 8 blocks each. All transformer blocks (view and ray) are of dimension 64. We set the patch size p as 4 for all our experiments to arrive at a balance between performance and network complexity. The VAE network contains 4 convolution layers with dimensions [16, 32, 64, 128] in the encoder, each followed by relu activation. The encoded input is projected to 100 dimension feature vector before Gaussian re-sampling. The sampled Gaussian latent is then passed onto a 3-layer decoder network with dimensions [128, 64, 32] to predict global background light A.

Training / Inference Details. We train both the feature extraction network and U2NeRF end-to-end on datasets of multi-view posed images using the Adam optimizer to minimize the loss given in Eqn. 6. We empirically set the values of λ to $\lambda_1 = 1$, $\lambda_2 = 0.1$, $\lambda_3 = 1$, $\lambda_4 = 1$, $\lambda_5 = 0.1$, $\lambda_6 = 0.1$. The base learning rates for the feature extraction network and U2NeRF are 10^{-3} and 5×10^{-4} respectively, which decay exponentially over training steps. During finetuning, we optimize the feature extraction network and U2NeRF using a smaller learning rate of 5×10^{-4} and 2×10^{-4}. For the single scene and cross scene generalization experiments, we train U2NeRF for 250,000 steps with 512 rays sampled in each iteration while during fine-tuning on each scene, the pre-trained network is only fine-tuned for 50,000 steps with 256 rays sampled in each iteration. Unlike most NeRF methods, we do not use separate coarse, fine networks and therefore to bring GNT to a comparable experimental setup, we sample 192 coarse points per ray across all experiments (unless otherwise specified).

Metrics. To evaluate our method's rendering and restoration quality, we use widely adopted metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [40], Learned Perceptual Image Patch Similarity (LPIPS) [46], Underwater Image Quality Measurement (UIQM) [26], and the Underwater Color Image Quality Evaluation Metric (UCIQE) [43]. We report the averages of each metric over different views in each scene and

Models	PSNR↑	SSIM↑	LPIPS↓
UPIFM	12.883	0.329	0.399
UIESS	18.818	0.790	0.174
NeRF	12.283	0.558	0.360
NeRF + Clean	17.948	0.741	0.297
U2NeRF	13.978	0.594	0.230

(a) Easy Split

Models	UIQM↑	UCIQE↑	LPIPS (gray)↓
UPIFM	1.424	32.940	-
UIESS	1.136	30.534	-
NeRF	0.501	31.622	0.208
NeRF + Clean	0.865	31.054	0.223
Clean + NeRF	0.858	30.336	0.198
U2NeRF	1.570	32.556	0.174

(b) Medium Split

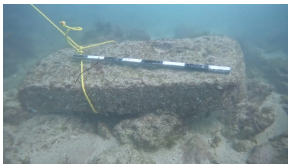
Models	UIQM↑	UCIQE↑	LPIPS (gray)↓
UPIFM	1.182	28.537	-
UIESS	0.649	27.161	-
NeRF	0.463	18.370	0.334
NeRF + Clean	0.486	27.453	0.328
Clean + NeRF	0.456	26.530	0.292
U2NeRF	1.100	26.788	0.260

(c) Hard Split

Table 1. Comparison of U2NeRF against baseline methods for single-scene rendering on the UVS dataset

across multiple scenes in each dataset. Following previous works [23, 36], we compute the PSNR, SSIM, LPIPS scores between the rendered and ground truth restored views in the case of synthetic scenes, and the UIQM, UCIQE scores to evaluate real underwater scenes, with no reference restored image. In the case of real world data, we additionally report LPIPS scores between the gray scale image of rendered and restored target views using [6] (to remove color differences) and quantitatively measure the rendering capabilities of different methods.

4.2. Underwater View Synthesis Dataset



(a) Scene1 from hard split

(b) Coral from medium split

(c) Fern from easy split

Figure 3. Illustrative examples of scenes from the UVS Dataset, one scene from each split is shown here.

Due to the absence of multi-view underwater scene datasets suitable to evaluate novel-view rendering, we establish a new benchmark Underwater View Synthesis (UVS) Dataset containing 12 scenes, equally split into *easy* (synthetic underwater scenes), *medium* (real-world high quality), *hard* (real-world low quality). The easy split contains 4 scenes from the LLFF dataset [22], namely "fern", "fortress", "flower", "trex" that were synthetically corrupted to simulate underwater scenes [8]. For real-world data, we hand-pick 4 scenes from high quality youtube videos to form the medium split, while the hard split is composed of low-quality noisy real-world captures obtained during a diving expedition. For each scene from the medium and hard splits, we select roughly 100-150 images and calibrate them using Structure-from-Motion (SfM) algorithm in an open-source software package COLMAP [34, 35]. For COLMAP, we use a "simple radial" camera model with a single radial distortion coefficient and a shared intrinsic for all images. We use a "sift feature guided matching" option in the exhaustive matcher step of SfM and also refine principle points of the intrinsic during the bundle adjustment. Fig. 3 provides an illustration of the scenes present in the easy, medium, hard splits.

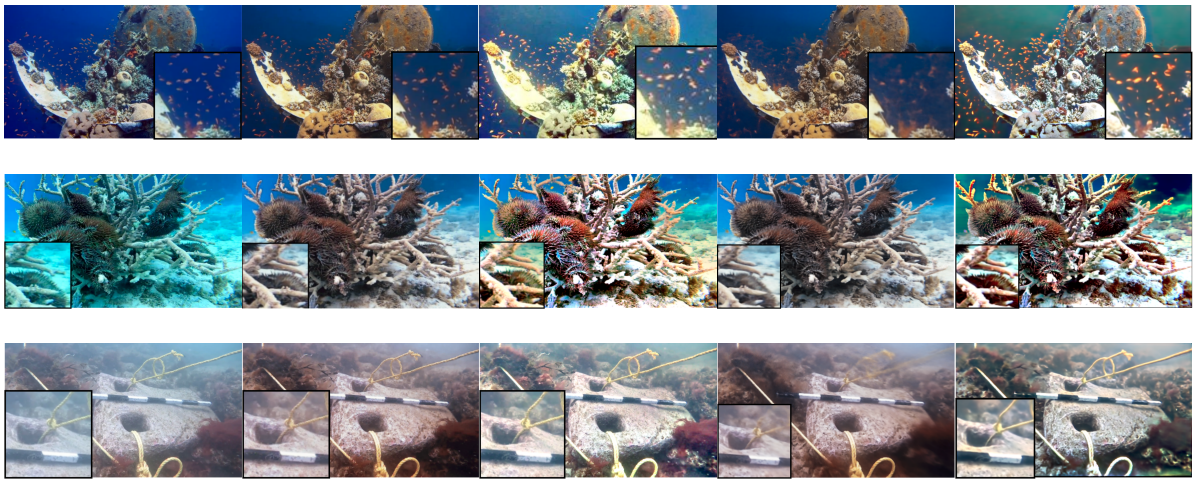
4.3. Baselines

The task of simultaneous restoration and rendering is novel and hence we establish several baselines to compare U2NeRF against in the UVS benchmark. We select NeRF as our neural renderer and identify different strategies to automatically "restore" the rendered views. As an initial baseline (labelled *NeRF*), we train a vanilla NeRF model on the original underwater scenes, append state-of-the-art (SOTA) restoration methods as a post processing strategy (labelled *NeRF + Clean*), and even train a NeRF model on restored images (labelled *Clean + NeRF*). Additionally, we also consider non-rendering baselines where we assume access to the target view and attempt to restore it. We leverage SOTA underwater image restoration pipelines - UIESS [6], Underwater Physics-informed Image Formation Model (which we label *UPIFM*) [3].

4.4. Single Scene Results

Datasets. To evaluate the single scene view generation capacity of U2NeRF, we perform experiments on the easy, medium and hard splits from the UVS dataset. We report average scores across all scenes within each split - easy: [Fern, Fortress, Flower, Trex], medium: [Starfish, Coral, Debris, Shipwreck], hard: [scene1, scene2, scene3, scene4] in Tables. 1a, 1b, 1c respectively.

Discussion. We compare U2NeRF against the baselines discussed in Sec. 4.3. On the easy split, our proposed method achieves moderate PSNR scores but best LPIPS scores when compared to other rendering baselines by 20%. This could be because PSNR fails to measure



(a) Underwater Image (b) UIESS (c) UPIFM (d) NeRF + Clean (e) U2NeRF

Figure 4. Qualitative results for single-scene rendering. In the Debris scene (row-1), U2NeRF is able to successfully recover and restore fishes, and enhance its visibility. In the Starfish scene (row-2), U2NeRF reconstructs edges with greater detail and even comparable to the non-rendering baselines. In scene2 from 'hard' split (row-3), U2NeRF renders complex, moving structures (rope) with higher visual quality.

Models	UIQM↑	UCIQE↑	LPIPS (gray)↓
U2NeRF Single Scene	1.570	32.556	0.174
U2NeRF Generalized	1.426	32.293	0.279
U2NeRF Generalized + Finetuned	1.856	34.113	0.222

(a) Medium Split

Models	UIQM↑	UCIQE↑	LPIPS (gray)↓
U2NeRF Single Scene	1.100	26.788	0.260
U2NeRF Generalized	1.000	23.548	0.290
U2NeRF (Generalized + Finetuned)	1.093	23.530	0.265

(b) Hard Split

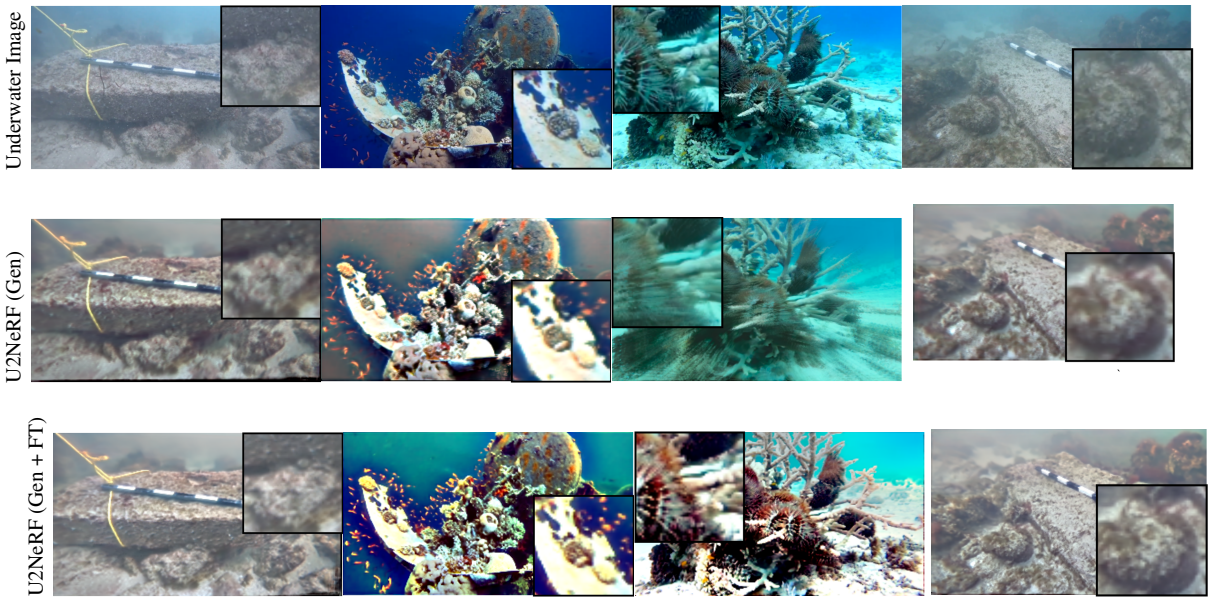
Table 2. Comparison of U2NeRF (with and without fine-tuning) against baseline methods for cross-scene rendering on the UVS dataset

structural distortions, blurring, has high sensitivity towards brightness, and hence does not effectively measure visual quality. Similar inferences are discussed in [18] regarding discrepancies in PSNR scores and their correlation to rendered image quality. In the case of more complex scenes present in the medium, hard data splits, we can clearly see the superiority of U2NeRF both in terms of rendering (LPIPS $\downarrow 11\%$), and color restoration quality (UIQM $\uparrow 5\%$, UCIQE $\uparrow 4\%$). More interestingly, we find that U2NeRF even outperforms no rendering baselines, that is, those algorithms that assume direct access to the target view and perform only restoration. Although our method extends upon UPIFM, we still manage to outperform the 'only restoration' baseline with sufficient margin. This signifies the relevance of multi-view geometry to automatically restore a target view. We show qualitative results in Fig. 4, and can clearly see that U2NeRF renders and restores images with greater visual quality when compared to other methods. In the case of debris, U2NeRF successfully recovers the fishes and enhances its visibility to improve restoration quality, while in the case of scene2, U2NeRF is able to render complex, moving structures like ropes while still maintaining higher detail along the surface of the rock.

4.5. Cross Scene Results

Datasets. U2NeRF leverages multi-view features complying with epipolar geometry, enabling generalization to unseen scenes. We randomly select scenes from video data captured during our diving expedition, and we use a total of 45 scenes for training. Table. 2 discusses results of the trained network on all 8 scenes from the medium and hard splits in the UVS dataset. Please note that the scenes from the UVS dataset are held out during training to gauge the model's generalization performance.

Discussion. We compare U2NeRF's generalization performance with a corresponding network trained only on a single scene. Although the network is only trained on data captured during our diving expedition, it still manages to generalize to unseen objects present in the medium split. Once finetuned (with as little as 50k training steps), U2NeRF manages to perform as well or even outperform a vanilla NeRF trained on each scene. Fig. 5 visualizes qualitative results on the UVS dataset. We can clearly see that the



(a) Scene 1(Hard) (b) Debris(Medium) (c) Starfish(Medium) (d) Scene 4(Hard)

Figure 5. Qualitative results for cross-scene rendering. We visualize the underwater scene (row-1), novel views rendered using the pre-trained network (row-2), novel views rendered using the finetuned network across different scenes (from left to right). U2NeRF successfully generalizes across scenes and when finetuned captures more intricate details.

pre-trained model can successfully generalize across several scenes and when finetuned, further improves performance.

4.6. Ablation Studies

Models	UIQM↑	UCIQE↑	LPIPS (gray)↓
U2NeRF ($p = 2$)	1.964	34.234	0.249
U2NeRF ($p = 4$)	2.222	35.120	0.187
U2NeRF ($p = 8$)	2.096	34.011	0.202

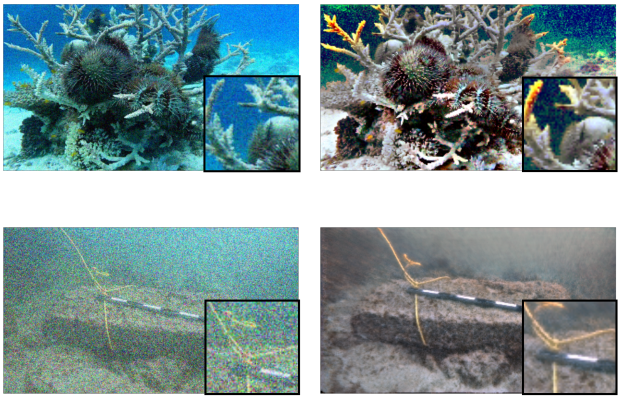
Table 3. Effect of Patch Size

Effect of patch size. To verify the effect of patch size on the rendered image quality, we train U2NeRF on the starfish scene with varying patch sizes (2, 4, and 8). From Table. 3, we can clearly see that $p = 4$ yields the best results across all three metrics. A larger patch size requires more information (beyond just epipolar points), for accurate reconstruction, while a smaller patch size does not act as a useful prior for restoration. Therefore, patch size = 4 strikes the ideal balance between performance and network complexity.

Models	UIQM↑	UCIQE↑	LPIPS (gray)↓
U2NeRF ($N = 3$)	2.208	35.084	0.191
U2NeRF ($N = 5$)	2.214	35.106	0.189
U2NeRF ($N = 8$)	2.220	35.116	0.188
U2NeRF ($N = 10$)	2.222	35.120	0.187

Table 4. Effect of number of source views

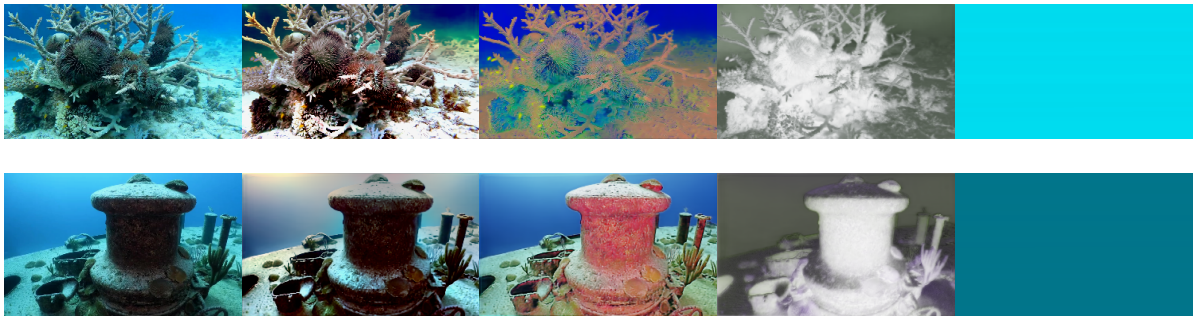
Effect of sparse source views. To test the performance of U2NeRF in the presence of sparse views, we evaluate a trained model on the starfish scene but with fewer source views. From Table. 4, we can see that as the number of source views increase, the model performs better. However, there is almost no significant drop in performance even when only 3 source views are given as input. This verifies the suitability of U2NeRF even in the presence of sparse input views.



(a) Underwater Image (b) U2NeRF

Figure 6. Denoising results of U2NeRF

Effect of Gaussian noise. Unlike standard NeRF methods, U2NeRF predicts an image patch rather than a single



(a) Degraded Image (b) Restored Image (c) Direct Transmission (d) Backscatter Trans. (e) Global Light

Figure 7. Visualisations of the predicted image components (scene radiance, transmission maps, global light).

pixel color. Therefore, we hypothesize that without explicit training, U2NeRF can denoise small perturbations present in the scene. To verify this claim, we evaluate a trained U2NeRF model on scenes corrupted using a Gaussian noise (unseen) with mean 0, and standard deviation 0.05. We show qualitative results in Fig 6.

4.7. Physical Interpretation of U2NeRF

U2NeRF attempts to implicitly predict individual image components J, T_D, T_B and A which when combined restores the underwater image. Fig 7 visualizes examples of underwater scenes with their corresponding predicted image components. We can clearly see that visualized T_D and T_B maps closely simulate depth which is consistent with the physics-informed image formation model [3]. A indicates the global background light which corresponds to the brightest pixel in the scene, and as seen from Fig. 7 corresponds to the color removed to enhance it. Therefore, with no explicit supervision, U2NeRF learns to physically ground its learnable operations.

5. Conclusion

We present Unsupervised Underwater NeRF (U2NeRF), that extends radiance fields to simultaneously render and restore novel views, more specifically in the context of underwater images. We demonstrate that by augmenting existing radiance fields with spatial awareness, and when combined with a physics-informed underwater image formation model can successfully restore underwater images. Additionally, we contribute a novel Underwater View Synthesis Dataset (UVSDataset) consisting of 12 underwater scenes, containing both synthetically generated, and real-world data. Extensive experiments reveal that U2NeRF outperforms existing baselines and achieves best perceptual metric scores (LPIPS $\downarrow$ 11%, UIQM $\uparrow$ 5%, UCIQE $\uparrow$ 4%. These results demonstrate that transformers can be successfully used to model the underlying physics in 3D vision.

References

[1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 1, 2

[2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022. 2

[3] Shu Chai, Zhenqi Fu, Yue Huang, Xiaotong Tu, and Xinghao Ding. Unsupervised and untrained underwater image restoration based on physical image formation model. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2774–2778. IEEE, 2022. 1, 2, 3, 5, 8, 11

[4] Hao Chen, Bo He, Hanyu Wang, Yixuan Ren, Ser Nam Lim, and Abhinav Shrivastava. Nerv: Neural representations for videos. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[5] Tianlong Chen, Peihao Wang, Zhiwen Fan, and Zhangyang Wang. Aug-nerf: Training stronger neural radiance fields with triple-level physically-grounded augmentations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15191–15202, 2022. 1

[6] Yu-Wei Chen and Soo-Chang Pei. Domain adaptation for underwater image enhancement via content and style separation. *IEEE Access*, 10:90523–90534, 2022. 1, 5, 11

[7] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022. 2

[8] Chaitra Desai, Badduri Sai Sudheer Reddy, Ramesh Ashok Tabib, Ujwala Patil, and Uma Mudenagudi. Aquagan: Restoration of underwater images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 296–304, 2022. 1, 5

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner,

Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 1

[10] 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

[11] Zhenqi Fu, Wu Wang, Yue Huang, Xinghao Ding, and Kai-Kuang Ma. Uncertainty inspired underwater image enhancement. In *European Conference on Computer Vision*, pages 465–482. Springer, 2022. 1

[12] 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. 1

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

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

[15] J KAJITA. Ray tracing volume densities. In *SIGGRAPH'84*, volume 18, pages 165–174, 1984. 3

[16] Aupendu Kar, Sobhan Kanti Dhara, Debashis Sen, and Prabir Kumar Biswas. Zero-shot single image restoration through controlled perturbation of koschmieder's model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16205–16215, 2021. 1

[17] 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. 2

[18] Kai-En Lin, Lin Yen-Chen, Wei-Sheng Lai, Tsung-Yi Lin, Yi-Chang Shih, and Ravi Ramamoorthi. Vision transformer for nerf-based view synthesis from a single input image. In *WACV*, 2023. 6

[19] Yung-Cheng Liu, Wen-Hsin Chan, and Ye-Quang Chen. Automatic white balance for digital still camera. *IEEE Transactions on Consumer Electronics*, 41(3):460–466, 1995. 2

[20] Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108, 1995. 3

[21] BL McGlamery. A computer model for underwater camera systems. In *Ocean Optics VI*, volume 208, pages 221–231. SPIE, 1980. 2

[22] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 5

[23] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, pages 405–421, 2020. 1, 2, 5

[24] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022. 1

[25] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5589–5599, 2021. 2

[26] 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. 4

[27] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *ICCV*, 2021. 2

[28] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. In *ACM Conference and Exhibition on Computer Graphics and Interactive Techniques in Asia (SIGGRAPH Asia)*, 2021. 2

[29] Naama Pearl, Tali Treibitz, and Simon Korman. Nan: Noise-aware nerfs for burst-denoising. In *CVPR*, 2022. 2

[30] 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

[31] Stephen M Pizer. Contrast-limited adaptive histogram equalization: Speed and effectiveness stephen m. pizer, r. eugene johnston, james p. ericksen, bonnie c. yankaskas, keith e. muller medical image display research group. In *Proceedings of the first conference on visualization in biomedical computing, Atlanta, Georgia*, volume 337, page 1, 1990. 2

[32] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 2

[33] Zia-ur Rahman, Daniel J Jobson, and Glenn A Woodell. Multi-scale retinex for color image enhancement. In *Proceedings of 3rd IEEE international conference on image processing*, volume 3, pages 1003–1006. IEEE, 1996. 2

[34] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[35] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 5

[36] Mukund Varma T, Peihao Wang, Xuxi Chen, Tianlong Chen, Subhashini Venugopalan, and Zhangyang Wang. Is attention

all nerf needs? *arXiv preprint arXiv:2207.13298*, 2022. 2, 3, 5

[37] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. *arXiv preprint arXiv:2112.03907*, 2021. 2

[38] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 2

[39] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 4

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

[41] Suttisak Wizadwongsa, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. Nex: Real-time view synthesis with neural basis expansion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

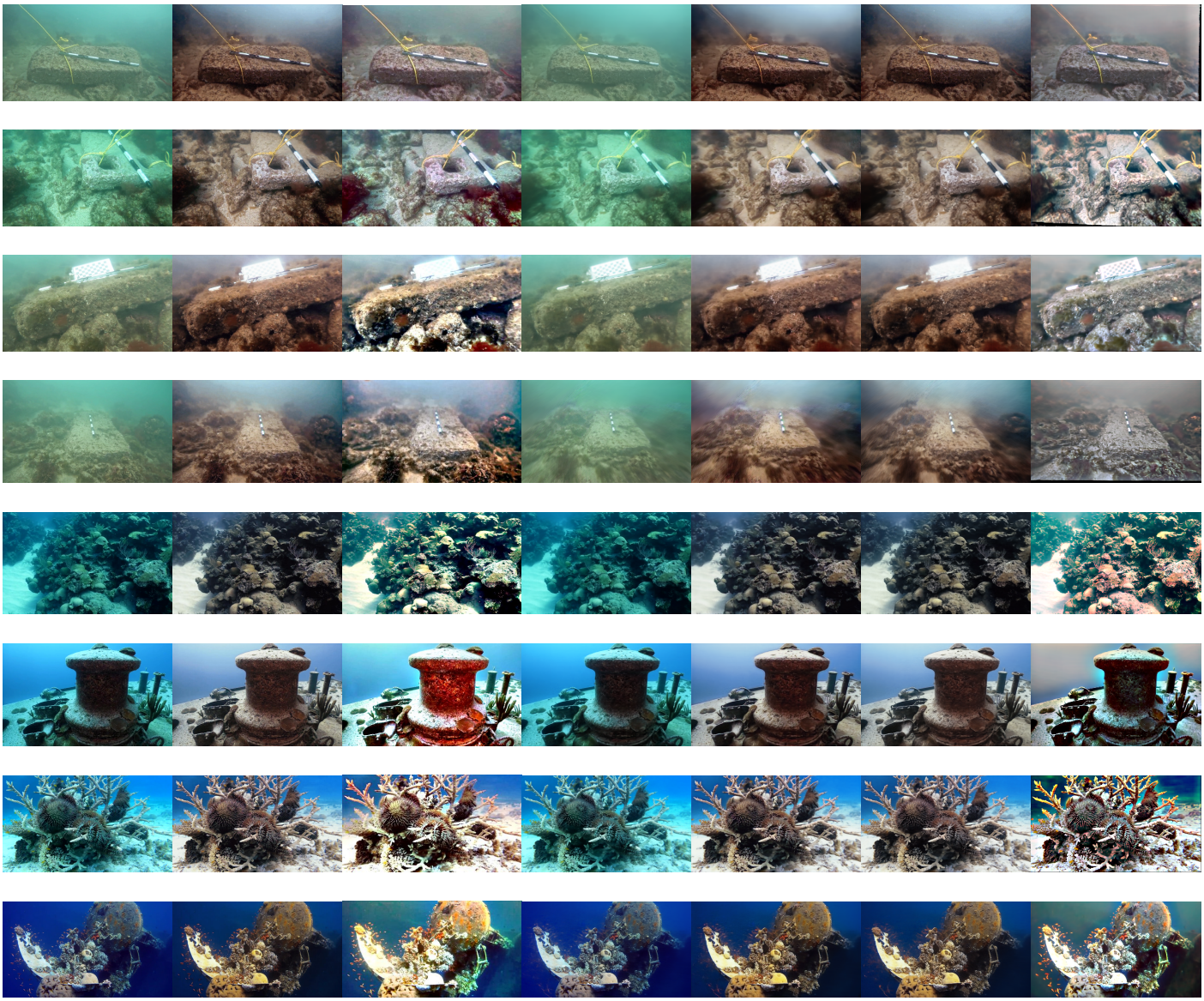
[42] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. *arXiv preprint arXiv:2204.00928*, 2022. 2

[43] Miao Yang and Arcot Sowmya. An underwater color image quality evaluation metric. *IEEE Transactions on Image Processing*, 24(12):6062–6071, 2015. 4

[44] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34, 2021. 2

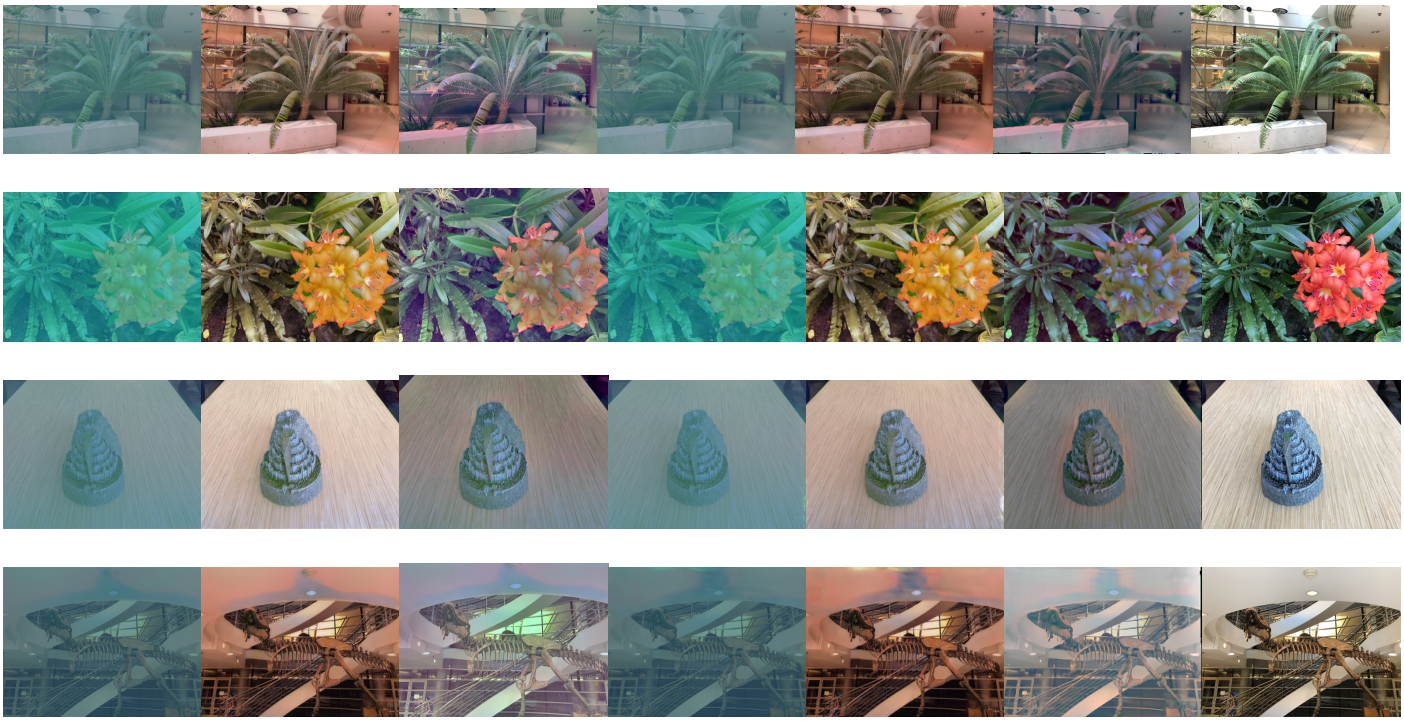
[45] Xiangyu Yin, Xiaohong Liu, and Huan Liu. Fmsnet: Underwater image restoration by learning from a synthesized dataset. In *International Conference on Artificial Neural Networks*, pages 421–432. Springer, 2021. 1

[46] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 4



(a) Underwater Image (b) UIESS (c) UPIFM (d) NeRF (e) NeRF + Clean (f) Clean + NeRF (g) U2NeRF

Figure 8. Qualitative results on single-scene rendering for Medium and Hard Scenes. The top 4 rows depict the scenes from the hard split(Scene 1, Scene 2, Scene 3 and Scene 4 respectively) and the last 4 rows depict the scene from the medium split(coral, shipwreck, starfish and debris respectively). (a) represents the actual underwater image from the scene, (b) & (c) represents the no rendering baseline methods ([6] [3]), (d), (e) & (f) refers to the renderings from NeRF on raw underwater image, restored view after NeRF rendering and NeRF rendering on restored input underwater images respectively, and (g) refers to results from our method: U2NeRF. U2NeRF is able to render better high-quality images when compared to other rendering+restoring methods.



(a) Degraded Image (b) UIESS (c) UPIFM (d) NeRF (e) NeRF + Clean (f) U2NeRF (g) Actual Image

Figure 9. Qualitative results for single-scene rendering for Easy Category. Similar to medium and hard splits, we show qualitative results on rendering+restoring baselines as well as no rendering methods. Our method U2NeRF is able to restore on-par with the rendering+restoring baselines.

Setting	Models	Scene 1			Scene 2			Scene 3			Scene 4		
		UIQM↑	UCIQE↑	LPIPS↓	UIQM↑	UCIQE↑	LPIPS↓	UIQM↑	UCIQE↑	LPIPS↓	UIQM↑	UCIQE↑	LPIPS↓
No Rendering	UIESS	0.63	25.19	-	0.68	28.06	-	0.57	29.85	-	0.70	25.53	-
	UPIFM	0.89	23.03	-	1.19	29.80	-	1.23	30.75	-	1.41	30.56	-
Rendering	NeRF	0.32	14.01	0.26	0.68	22.91	0.32	0.32	20.48	0.31	0.51	16.06	0.43
	NeRF+Clean	0.50	**25.82**	0.25	0.53	27.84	0.32	0.34	**30.06**	0.30	0.57	**26.08**	0.42
	Clean+NeRF	0.44	24.37	**0.21**	0.49	27.62	0.30	0.35	29.42	0.30	0.52	24.69	**0.34**
	U2NeRF	**0.84**	23.33	0.22	**1.32**	**30.42**	**0.21**	**1.04**	29.60	**0.22**	**1.18**	23.80	0.37

Table 5. Comparison of U2NeRF against SOTA methods for single scene rendering on the UVS Dataset, Hard Split (scene-wise).

Setting	Models	Coral			Debris			Starfish			Shipwreck		
		UIQM↑	UCIQE↑	LPIPS↓	UIQM↑	UCIQE↑	LPIPS↓	UIQM↑	UCIQE↑	LPIPS↓	UIQM↑	UCIQE↑	LPIPS↓
No Rendering	UIESS	1.11	28.11	-	0.84	31.83	-	1.61	32.55	-	0.98	29.63	-
	UPIFM	1.30	32.05	-	1.11	32.09	-	1.91	33.85	-	1.35	33.75	-
Rendering	NeRF	0.19	28.19	0.22	0.50	**34.45**	0.175	0.94	31.60	0.21	0.35	**32.23**	0.21
	NeRF+Clean	0.71	28.01	0.23	0.70	32.57	0.20	1.34	32.96	0.23	0.70	30.65	0.22
	Clean+NeRF	0.69	27.93	0.21	0.68	31.51	0.175	1.33	32.50	0.20	0.71	29.39	0.20
	U2NeRF	**1.34**	**31.17**	**0.16**	**1.17**	32.08	**0.17**	**2.22**	**35.12**	**0.18**	**1.54**	31.83	**0.17**

Table 6. Comparison of U2NeRF against SOTA methods for single scene rendering on the UVS Dataset, Medium Split (scene-wise).