

Hybrid Physics-Guided CNN-Based Underwater Image Enhancement Using Multi-Scale Pyramid Fusion and Comprehensive Quality Assessment

Chukka. Ramakrishna ¹  K. Mamatha ² 

¹. PG Scholar, Department of ECE, JNTUA College of Engineering(Autonomous), India.

². Assistant Professor Ad-hoc, Department of ECE, JNTUA College of Engineering(Autonomous), I

¹ <https://orcid.org/0009-0001-6339-0312> ² <https://orcid.org/0000-0001-9911-4021>

Email: ramakrishnachukka2001@gmail.com , k.mamatha53@gmail.com.

Abstract—Underwater photographs usually look green or blue, flat and hazy. Water absorbs and scatters light differently at different wavelengths, and this limits how useful the images are in marine biology, underwater archaeology, pipeline inspection and vision for autonomous vehicles. In this paper we present a hybrid enhancement framework that is guided by physics and assisted by a CNN, organised as a 14-stage pipeline. Classical restoration methods estimate scene transmission with the dark-channel prior (DCP). We instead replace the transmission/depth ramp with a dense monocular depth map from a MiDaS/DPT network, which gives a physically consistent depth cue that is largely unaffected by the strong colour cast of underwater scenes. The depth map drives a physics-based decomposition and colour-restoration stage, and a lightweight three-layer convolutional residual-refinement network then corrects the artefacts this stage leaves behind. A second, independent branch handles illumination estimation, depth-based colour compensation, scattering compensation and contextual sharpening. We merge the two branches using multi-cue weight maps (Laplacian contrast, saliency and saturation) and a three-level Laplacian pyramid fusion, which gives an output with high contrast and natural colour. The pipeline is evaluated with a five-metric module made up of two no-reference measures (UIQM, UCIQE) and three full-reference measures (PSNR, SSIM, LPIPS). On the representative test image, UIQM rose from -1.3716 to 2.5098 and UCIQE from 0.4226 to 0.6138 . Relative to the original, the enhanced output reached PSNR = 12.8511 dB, SSIM = 0.8289 and LPIPS = 0.0007 . Colourfulness, sharpness and contrast improve markedly, while the structure and perceptual content of the source image are largely retained.

Keywords—Underwater image enhancement, monocular depth estimation, MiDaS, DPT, physics-based image restoration, convolutional neural network, residual refinement, Laplacian pyramid fusion, CLAHE, guided filter, image quality assessment, UIQM, UCIQE, LPIPS.

I. INTRODUCTION

Underwater imaging is central to many kinds of work: marine ecology surveys, coral-reef monitoring, underwater archaeology, inspection of subsea pipelines and cables, and vision for autonomous underwater vehicles (AUVs). Light behaves very differently in water than in air. It is absorbed selectively by wavelength and scattered many times on its way to the camera [1], [2]. Red is attenuated fastest, then green and then blue, so raw underwater images tend to carry a green or blue cast, with low contrast and reduced sharpness. In turbid water they also look hazy and hard to see through. Artificial light from dive lamps or strobes is rarely even, which makes things worse, and the degradation grows non-linearly with scene depth and with the distance between camera and object. That is why a single correction formula cannot fix an entire scene.

Traditional restoration methods borrow the atmospheric dark-channel prior (DCP) [3] and adapt it to water [4]. They estimate a transmission map and invert the image-formation model to recover scene radiance [5], [6]. This works reasonably well in mild haze, but DCP-derived priors were designed for fog on land and are often violated underwater. The problem is worst in the blue and green channels, where the “dark channel” assumption breaks down, leading to colour artefacts and regions

that are over- or under-corrected. Fusion-based methods take a different route. They pre-process the same image in several independent ways (for example white balancing and contrast enhancement), weight the results with perceptual maps such as Laplacian contrast, local contrast, saliency and exposedness, and rebuild the image by multi-scale pyramid blending [7] [8] [9]. These methods are robust and need no reference image, but they do not reason explicitly about scene depth. Learning-based methods, including convolutional and generative-adversarial architectures trained on paired or synthetic underwater datasets [10] [11] [12], can learn complex colour and contrast corrections directly from data. Their generalisation, however, depends heavily on how diverse the training data are, and they can struggle with water types, depths or lighting conditions that the training set never covered.

Recent monocular depth-estimation networks, notably MiDaS [13] and the Dense Prediction Transformer (DPT) [14], are trained on large and varied in-air datasets and generalise well to many kinds of scenes in a zero-shot setting. They estimate relative depth from image structure rather than colour statistics, so they are largely unaffected by the colour cast that undermines DCP-style transmission estimation in water. This observation is the central idea of our work. Inside a physics-based underwater

restoration pipeline, we replace the depth/transmission ramp with a dense monocular depth estimate. Colour restoration, backscatter removal and scattering compensation are then driven by a depth cue that is physically meaningful and robust to colour, rather than by a proxy derived from the image itself.

Building on this idea, we propose a 14-stage hybrid pipeline with four parts: (i) physics-based restoration guided by monocular depth, (ii) a lightweight CNN residual-refinement stage that corrects the colour and artefact residuals the physical model leaves behind, (iii) an independent illumination and scattering-compensation branch, and (iv) multi-cue, multi-scale Laplacian pyramid fusion of the two branches. A five-metric quality-evaluation module then scores the result. Our contributions are as follows:

1) **Depth-aware restoration:** a MiDaS/DPT-estimated dense depth map [13], [14] takes the place of the linear or dark-channel-derived depth ramp used in earlier underwater restoration pipelines. This improves physical consistency in scenes where foreground and background are clearly separated in depth.

2) **CNN residual refinement:** a lightweight three-layer convolutional network learns a residual correction that suppresses the over-amplified, yellow-shifted artefacts that purely physics-based colour restoration typically produces.

3) **Dual-branch multi-cue pyramid fusion:** Laplacian-contrast, saliency-contrast and saturation weight maps are computed for two independently enhanced branches, which are then fused through a three-level Laplacian pyramid [9]. This avoids the halos and edge artefacts common in single-scale blending.

4) **Comprehensive evaluation:** we score the enhancement with five metrics: two no-reference perceptual measures (UIQM [15], UCIQE [16]) and three full-reference measures (PSNR, SSIM [17], LPIPS [18]). Together they show how much colourfulness and contrast improve and how well structural and perceptual fidelity is kept.

The rest of the paper is organised as follows. Section II reviews related work on physical-model, fusion-based and learning-based underwater enhancement, monocular depth estimation and image quality assessment. Section III describes the 14-stage methodology in detail with reference to the system block diagram. Section IV presents qualitative and quantitative results. Section V discusses strengths, limitations and future work, and Section VI concludes.

II. RELATED WORK

A. Physical Model-Based Restoration

He et al. introduced the dark-channel prior (DCP) for single-image haze removal [3]. Drews et al. later adapted it to water (UDCP) [4], limiting the dark-channel computation to the green and blue channels so that it better matches underwater

attenuation statistics. Berman et al. [5] proposed a restoration method based on haze-lines, which clusters pixels of similar reflectance to estimate transmission and veiling light together. Akkaynak and Treibitz [2] revisited the classical underwater image-formation model. They showed that the widely used single attenuation coefficient is physically inaccurate and proposed a revised formulation that depends on depth and reflectance. Two points follow from these works. First, transmission/depth estimation is the main bottleneck in physics-based restoration. Second, image-derived depth proxies such as the dark channel are unreliable under strong colour casts. This is why our pipeline uses an independent, learning-based depth estimator.

B. Fusion-Based Enhancement

Ancuti et al. [7] proposed a fusion strategy that blends a white-balanced version and a contrast-enhanced version of the input. The blend uses Laplacian-contrast, local-contrast, saliency and exposedness weight maps combined through multi-scale (Laplacian) pyramid blending [9]. They later added an explicit colour-balance pre-processing stage [8]. Fusion is attractive because it needs no training data and is robust across many water types. However, the weight maps come purely from image statistics and ignore scene depth, which can bias the blend when object-to-camera distance varies strongly across the scene. Our pipeline also draws on established enhancement tools: contrast-limited adaptive histogram equalisation (CLAHE) [19] for local contrast, the guided filter [20] for edge-preserving smoothing, and ideas related to Retinex theory [21] for the reflectance/illumination decomposition and local illumination estimation. We keep the multi-scale pyramid-fusion principle, but the two input branches come from depth-aware physical restoration and depth-based scattering compensation respectively. In this way we combine the strengths of fusion-based blending with explicit reasoning about depth.

C. Learning-Based Enhancement

Li et al. [10] built the UIEB benchmark, a dataset of paired real-world underwater images, and proposed WaterNet, a CNN that predicts fusion weights for gamma-corrected, white-balanced and histogram-equalised inputs. Islam et al. [11] proposed FUnIE-GAN, a conditional generative-adversarial network trained for fast, perceptually driven enhancement that suits real-time robotic vision. Li et al. [12] also proposed UWCNN, which uses underwater imaging models to synthesise training pairs for ten water types and trains a lightweight CNN end to end. These learning-based methods can give visually convincing results, but their performance depends on how representative the training data are, and end-to-end “black-box” correction offers little physical interpretability. In our method the CNN is only a residual refinement step on top of an

interpretable physics-based backbone, so that data-driven correction and physical consistency work together.

D. Monocular Depth Estimation

Ranftl et al. proposed MiDaS [13], a robust monocular depth-estimation network trained on a mixture of datasets with a scale- and shift-invariant loss, which gives strong zero-shot generalisation across datasets. They later proposed DPT [14], which swaps the convolutional encoder for a vision-transformer backbone to improve dense-prediction accuracy. Neither network is trained on underwater imagery. Both estimate depth from monocular structural and textural cues such as relative size, occlusion and texture gradient, not from colour statistics. This makes them much less sensitive to the green/blue cast that corrupts DCP-based depth proxies. We exploit exactly this property in Step 2.0 of our pipeline, where the MiDaS/DPT depth estimate replaces the conventional depth ramp.

E. Underwater Image Quality Assessment

Panetta et al. [15] proposed UIQM, a no-reference measure inspired by the human visual system that combines colourfulness (UICM), sharpness (UISM) and contrast (UIConM) sub-measures. Yang and Sowmya [16] proposed UCIQE, another no-reference metric, computed in CIE Lab colour space as a weighted combination of chroma standard deviation, luminance contrast and mean saturation. Wang et al. [17] introduced the structural similarity index (SSIM) as a full-reference alternative to PSNR that correlates better with

perceived structural fidelity. Zhang et al. [18] proposed the Learned Perceptual Image Patch Similarity (LPIPS) metric, which compares deep convolutional-network feature activations instead of raw pixel differences. We use all five metrics together (Section III-C). UIQM and UCIQE capture the gain in perceptual colour and contrast, and PSNR, SSIM and LPIPS capture structural and perceptual fidelity to the source image.

III. PROPOSED METHODOLOGY

A. System Overview

Each underwater image goes through a 14-stage pipeline, shown in the system block diagram of Fig. 1. The pipeline is best seen as two coupled halves that meet in a final fusion stage. The first half (Steps 1–5) covers pre-processing, monocular-depth-guided physical decomposition, physics-based colour restoration, contrast enhancement and CNN residual refinement. It produces a colour-corrected branch with suppressed artefacts. The second half (Steps 6–9) works independently. It estimates local illumination from the pre-processed input, then applies depth-based colour compensation, scattering compensation and contextual sharpening, giving a complementary branch that preserves detail. Steps 10–12 compute multi-cue weight maps for the two branches, normalise them and fuse the branches through a three-level Laplacian pyramid, which yields the final 600×800 enhanced output. Step 13 evaluates that output with five complementary quality metrics.

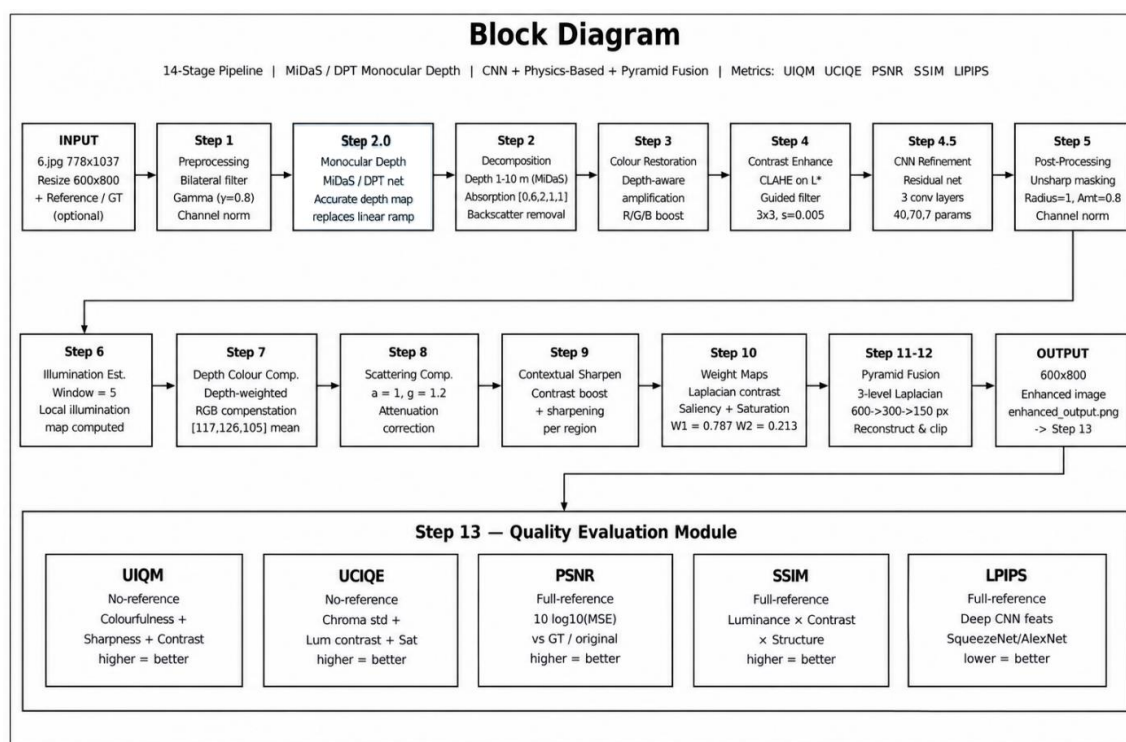


Fig. 1. System block diagram of the proposed 14-stage hybrid physics-guided, CNN-assisted underwater image enhancement pipeline, showing the monocular-depth-guided restoration branch (Steps 1–5), the illumination/scattering-compensation branch (Steps 6–9), multi-cue weighted pyramid fusion (Steps 10–12) and the five-metric quality-evaluation module (Step 13).

B. Stage-Wise Description

1) Preprocessing (Step 1)

The raw underwater image is first resized to 600×800 pixels and, where one is available, paired with a reference/ground-truth image for later benchmarking. Bilateral filtering then removes sensor and backscatter noise while keeping edges intact (Fig. 2(a)–(b)). Gamma correction with $\gamma = 0.8$ lifts the mid-tones without saturating highlights (Fig. 2(c)). Finally, per-channel normalisation equalises the dynamic range of the R, G and B channels before depth estimation (Fig. 2(d)).

2) Monocular Depth Estimation (Step 2.0)

A pretrained MiDaS/DPT dense-depth network estimates a relative depth map directly from the pre-processed RGB image. We normalise the predicted depth to the 1–10 m range, which is typical of imaging distances in recreational and survey diving. This depth map replaces the dark-channel-derived transmission map and drives every later depth-dependent operation. Fig. 3 shows the result. The diver in the foreground (green/yellow) is correctly separated from the background scene (blue), so the network recovers a physically plausible depth ordering even though the input has a strong green colour cast.

3) Decomposition (Step 2)

Using the estimated depth, we split the observed image into a reflectance (scene-radiance) component and an illumination (backscatter/veiling-light) component, using depth-scaled, wavelength-dependent absorption coefficients. Backscatter is estimated and subtracted from the observed radiance before colour restoration.

4) Colour Restoration (Step 3)

The decomposed reflectance is amplified by inverting the depth-aware image-formation model, with per-channel R/G/B gain boosting to compensate for the much stronger attenuation of red light over distance. This stage recovers the global colour balance. Because it is a purely physical inversion, though, it can over-amplify noise and leave a yellow-tinted intermediate result, as visible before CNN refinement.

5) Contrast Enhancement (Step 4)

Contrast-limited adaptive histogram equalisation (CLAHE) is applied to the L^* channel of the CIELab-converted image, using 3×3 tiles and a clip limit of $s = 0.005$. This boosts local contrast while limiting noise amplification. An edge-preserving guided filter then smooths flat regions and keeps sharp structural boundaries, which reduces the halo artefacts typical of unconstrained histogram equalisation.

6) CNN Residual Refinement (Step 4.5)

A lightweight three-layer convolutional network (40–70–7 channel configuration) takes the physically restored, contrast-enhanced image and predicts a residual correction, which is added back to the input. This data-driven correction targets the residual colour casts, banding and local artefacts that the

physical model cannot resolve on its own. In effect it acts as a learned post-corrector for the interpretable physics-based backbone.

7) Post-Processing (Step 5)

Unsharp masking (radius = 1, amount = 0.8) restores fine detail that the earlier smoothing operations attenuated, and a final channel normalisation prepares the output of the first branch for fusion.

8) Illumination Estimation (Step 6)

In parallel, a local illumination map is computed from the pre-processed input with a 5×5 sliding window. It captures the spatially varying brightness caused by non-uniform artificial lighting (for example, dive-light hotspots).

9) Depth-Based Colour Compensation (Step 7)

Each colour channel is compensated according to the estimated scene depth, pushing the image statistics toward a grey-world reference mean of [117, 126, 105]. This counteracts the systematic green/blue bias, and because it follows the true local depth it is more precise than a single global correction factor.

10) Scattering Compensation (Step 8)

Forward-scattering and veiling-light effects are attenuated further with an exponential correction (parameters $a = 1$ and $g = 1.2$). This sharpens object boundaries that multiple-scattering blur had softened.

11) Contextual Sharpening (Step 9)

Local, region-adaptive contrast boosting and sharpening are applied region by region. Fine texture (for example skin and equipment detail) is enhanced, while low-texture background areas are not over-sharpened globally.

12) Weight Map Computation and Normalisation (Step 10)

For each of the two enhancement branches we compute three complementary weight maps: a Laplacian-contrast weight (LCW) that favours high-frequency detail, a saliency-contrast weight (SCW) that favours perceptually salient regions, and a saturation weight (SatW) that favours well-saturated, natural colour. Fig. 6 shows LCW1/2, SCW1/2 and SatW1/2 for the two branches. The six maps are then normalised so that, at every pixel, the weights of the two branches sum to one. This gives the normalised fusion weights NW1 and NW2 shown in Fig. 7.

13) Multi-Scale Pyramid Fusion (Steps 11–12)

Each branch image and its normalised weight map are decomposed into a three-level Laplacian pyramid (600→300→150 px). The two branches are blended level by level with their normalised weights, and the fused pyramid is collapsed to reconstruct the final enhanced image. This image is clipped to the valid intensity range and saved as the 600×800 output (enhanced_output.png).

C. Quality Evaluation Module (Step 13)

The final output is scored with five complementary metrics. $UIQM = c_1 \cdot UICM + c_2 \cdot UISM + c_3 \cdot UIConM$ combines colourfulness, sharpness and contrast sub-measures inspired by the human visual system, and higher values mean better perceptual quality. $UCIQE = c_1 \cdot \sigma_C + c_2 \cdot con_1 + c_3 \cdot \mu_s$ combines chroma standard deviation, luminance contrast and mean saturation in CIELab space, and higher values mean a better-balanced colour distribution. $PSNR = 10 \cdot \log_{10}(MSE^{-1})$ (scaled by the peak signal level) measures pixel-level fidelity to a reference image. SSIM evaluates the product of luminance, contrast and structural similarity between two images. LPIPS computes a weighted distance between deep convolutional (SqueezeNet/AlexNet) feature activations of the two images, and it tracks human perceptual judgement more closely than pixel-wise metrics. The two branches converge into one enhanced output, and we had no independent ground-truth photograph to compare against. In this study PSNR, SSIM and LPIPS are therefore computed between the enhanced output and the original input. They tell us how large and what kind of transformation the pipeline applies, not how closely the result agrees with an external reference, and they complement the no-reference UIQM/UCIQE scores.

IV. EXPERIMENTAL RESULTS

A. Implementation Details

We implemented and evaluated the pipeline on one representative real-world underwater photograph (778×1037 px, resized to 600×800 px) showing a corroded bronze statue and a scuba diver. We chose it because it combines near-field foreground detail, a mid-field subject and a hazy far-field background, conditions that test depth-dependent restoration particularly hard. All classical image-processing stages (bilateral filtering, CLAHE, guided filtering, weight-map computation and Laplacian pyramid fusion) use standard numerical image-processing routines. The monocular depth map comes from a pretrained MiDaS/DPT network, and the residual-refinement stage uses a compact three-layer convolutional network. Quality metrics follow the formulations summarised in Section III-C.

B. Qualitative Results

Fig. 2 shows what the preprocessing stage does step by step. Bilateral denoising removes fine sensor noise, gamma correction brightens the mid-tones, and channel normalisation slightly rebalances a palette that is still strongly green. Low-level preprocessing alone clearly cannot correct the underlying colour cast. Fig. 3 shows the depth-guided physical decomposition and restoration. The MiDaS/DPT depth map (a) clearly separates the near-field diver and statue from the hazy background, and the estimated reflectance (b) and illumination (c) components separate scene content from veiling light.

Physics-based colour restoration (d) recovers warm skin and metal tones, but at the cost of a strong yellow global cast and reduced dynamic range, which is why the later refinement stages are needed. Fig. 4 covers contrast enhancement and CNN refinement. CLAHE (a) and guided filtering (b) progressively restore local contrast while suppressing halos, and CNN residual refinement (c) with post-processing (d) tightens the colour balance and sharpens fine detail. A residual warm cast remains at this point, and the fusion stage corrects it later. Fig. 5 shows the second branch, based on illumination and scattering compensation. Local illumination estimation (a) supports depth-based colour compensation (b), which greatly reduces the green cast and reveals natural flesh and coral-like statue tones. Scattering compensation (c) sharpens edges further, and contextual contrast and sharpening (d) gives a well-balanced, naturally coloured intermediate image that is much closer to the final result than the first branch alone. This suggests the two branches are complementary rather than redundant.

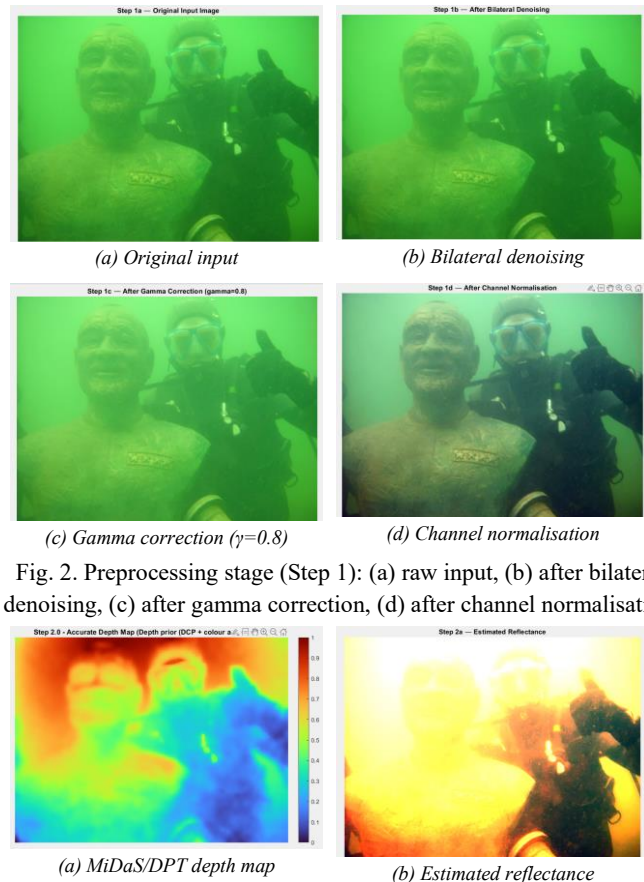


Fig. 2. Preprocessing stage (Step 1): (a) raw input, (b) after bilateral denoising, (c) after gamma correction, (d) after channel normalisation.

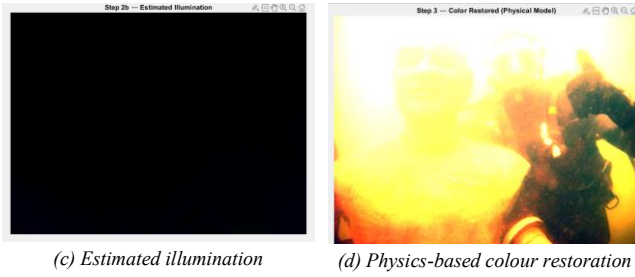


Fig. 3. Monocular-depth-guided decomposition and physical colour restoration (Steps 2.0, 2 and 3).

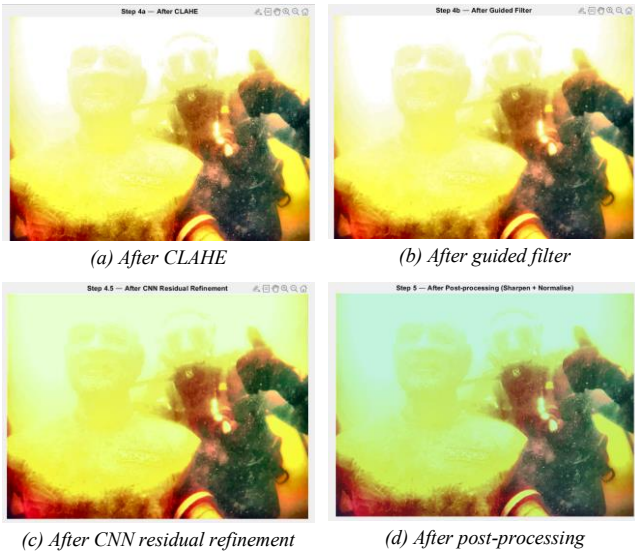


Fig. 4. Contrast enhancement, CNN residual refinement and post-processing (Steps 4, 4.5 and 5).

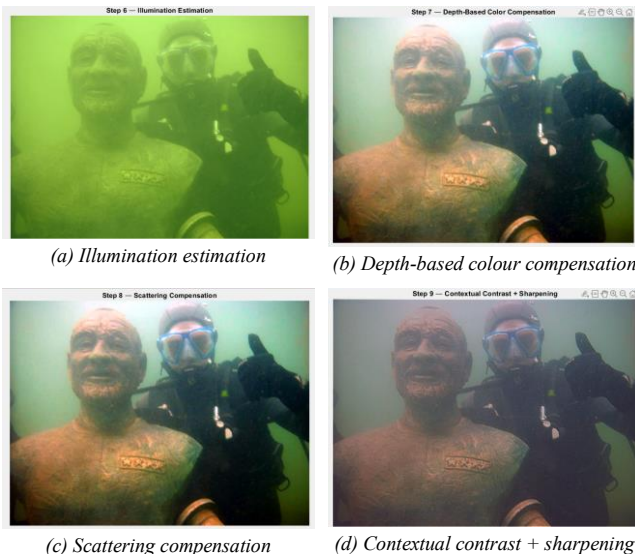


Fig. 5. Independent illumination/scattering-compensation branch (Steps 6–9).

Fig. 6 shows the multi-cue weight maps for the two branches at Step 10. The Laplacian-contrast (LCW), saliency-contrast (SCW) and saturation (SatW) maps mark the diver and statue as high-weight regions, while the low-texture haze in the background gets lower weight. This matches the aim of the

fusion, which is to preserve foreground detail. Fig. 7 shows the corresponding normalised fusion weights (NW1, NW2). NW1 is close to one over most of the frame, so Branch 1 (depth-guided physical restoration plus CNN refinement) dominates the fused output in high-contrast foreground regions. NW2 contributes more strongly in localised regions, which agrees with the complementary role of the illumination/scattering-compensation branch noted above.

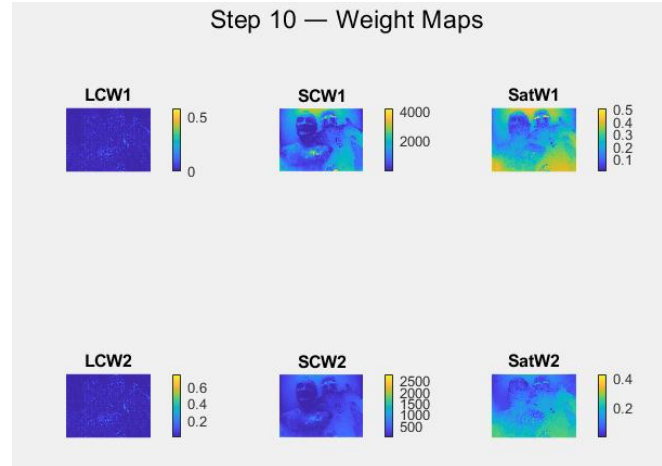


Fig. 6. Multi-cue weight maps (Step 10): Laplacian-contrast (LCW), saliency-contrast (SCW) and saturation (SatW) weights for Branch 1 (top row) and Branch 2 (bottom row).

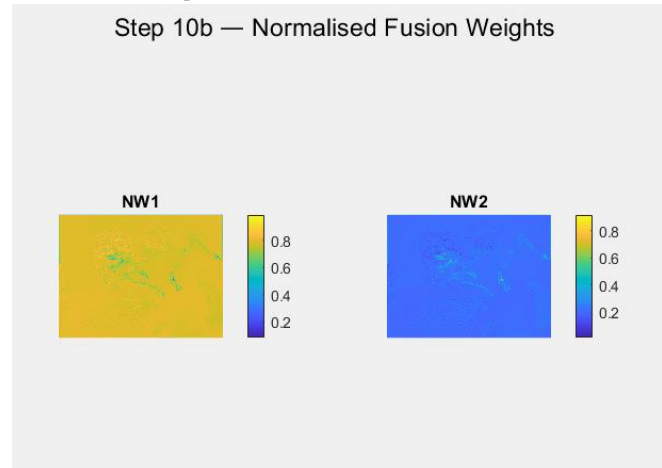


Fig. 7. Normalised fusion weights (Step 10b): NW1 (Branch 1) and NW2 (Branch 2) after per-pixel normalisation.

C. Quantitative Results

Table II reports the five quality metrics of Section III-C for the original input and the final enhanced output. UIQM rose from -1.3716 to 2.5098 , an absolute gain of 3.88 , which reflects a large increase in the combined colourfulness, sharpness and contrast score. UCIQE rose from 0.4226 to 0.6138 , a relative increase of 45.2% , meaning a better-balanced distribution of chroma, luminance contrast and saturation that fits the visible removal of the green cast. Both no-reference measures therefore support the colour-correction and contrast-enhancement effects seen in Figs. 2–5. The full-reference measures are computed

between the enhanced output and the original input. PSNR = 12.8511 dB and SSIM = 0.8289 indicate a large but controlled transformation. An SSIM above 0.8 means that roughly four-fifths of the luminance, contrast and structure information of the original scene is preserved even though colour, contrast and sharpness have changed a lot. The enhancement is therefore structure-preserving and does not rebuild the image content wholesale. LPIPS = 0.0007 is very small on the metric's [0, 1] scale, so the enhanced output stays close to the original in deep perceptual feature space. This is expected, since LPIPS is more sensitive to structural and textural changes than to the removal of a global colour cast. It also supports the view that the pipeline

applies a targeted colour and contrast correction and does not disrupt the structure of the image.

TABLE II

Metric	Original (Input)	Enhanced (Output)
UIQM $\uparrow$	-1.3716	2.5098
UCIQE $\uparrow$	0.4226	0.6138
PSNR (dB) $\uparrow$	—	12.8511
SSIM $\uparrow$	—	0.8289
LPIPS $\downarrow$	—	0.0007

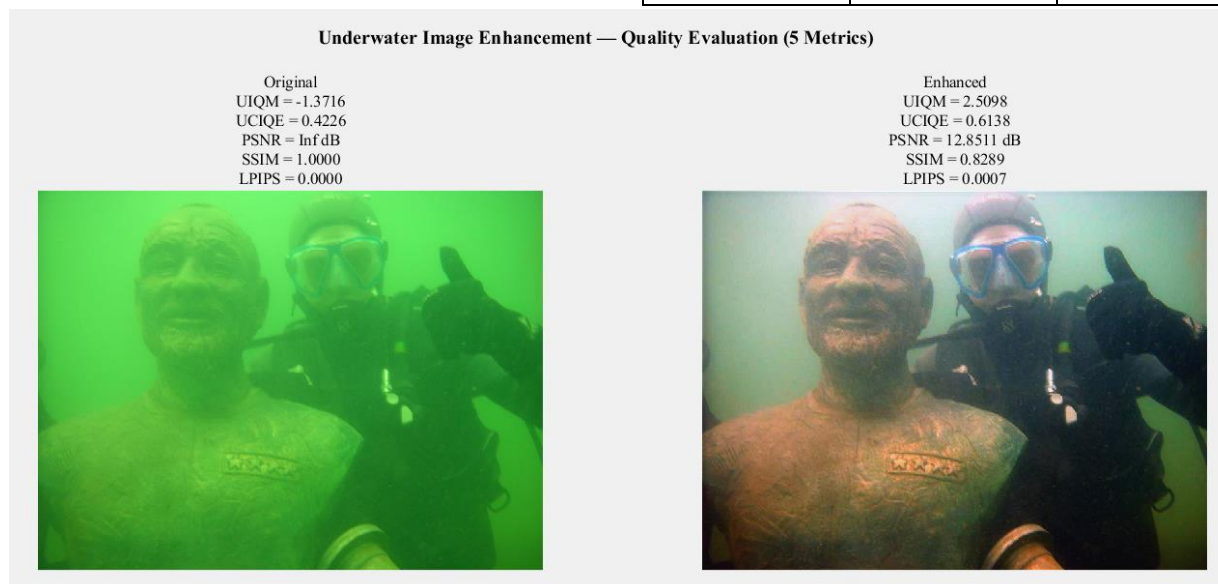


Fig. 8. Side-by-side qualitative and quantitative comparison of the original (left) and final enhanced (right) images, annotated with all five quality metrics.

Fig. 8 puts the original and enhanced images side by side with all five metrics annotated, and Fig. 9 shows the same metrics as paired bar charts. The visual comparison agrees with the numbers. The enhanced image has natural skin and metal tones, a clearer green-blue water background instead of an opaque green haze, and much better local contrast and edge definition around the diver's mask, regulator and the statue's facial features. The overall composition, pose and scene layout of the original photograph are kept. Fig. 10 shows the final enhanced image alone at full resolution for closer inspection.

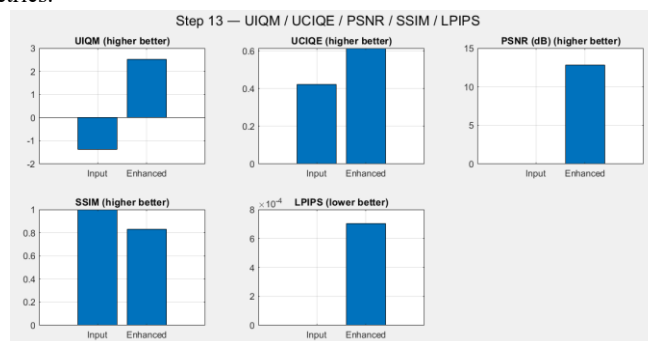


Fig. 9. UIQM, UCIQE, PSNR, SSIM and LPIPS scores for the original input and enhanced output (Step 13).



Fig. 10. Final enhanced underwater image produced by the proposed pipeline.

V. DISCUSSION

The results show that replacing the conventional dark-channel/linear depth ramp with the output of a monocular depth network is a practical and effective choice for underwater colour restoration. The MiDaS/DPT depth map (Fig. 3(a)) correctly separates the near-field diver and statue from the hazy background, even with the strong colour cast that usually defeats colour-derived depth proxies. The depth-scaled absorption, backscatter removal and colour restoration stages can therefore work on a physically meaningful signal. We added a lightweight CNN residual-refinement stage on top of the physics-based backbone instead of replacing the backbone with an end-to-end network. This keeps the method interpretable and still corrects the over-saturated, yellow-tinted artefacts typical of purely physical inversion (compare Fig. 3(d) with Fig. 4(c)–(d)). The dual-branch design fuses an independent illumination/scattering-compensation branch (Fig. 5) with the depth-guided restoration branch through multi-cue, multi-scale pyramid fusion. The normalised fusion weights in Fig. 7 support this design: both branches contribute noticeably across the frame and neither dominates everywhere. This suggests that the two branches capture complementary information, namely global depth-aware colour balance on one side and local illumination and scattering detail on the other.

Several limitations should be acknowledged. First, we evaluated on a single representative test image. That is useful for validating the pipeline and inspecting every intermediate stage visually, but statistically robust conclusions need benchmarking on standard underwater datasets such as UIEB, EUVP or RUIE, and comparison against established baselines: UDCP, DCP-based restoration, fusion-based enhancement, and learning-based methods such as WaterNet and FUNIE-GAN. Second, no independent ground-truth photograph was available for this scene, so PSNR, SSIM and LPIPS were computed

against the original degraded input. This is a common and informative way to quantify the size and perceptual character of the enhancement, but it is not the same as full-reference fidelity against true radiometric ground truth. Future work should use paired or synthetically generated ground-truth data where possible. Third, the fourteen sequential stages include several iterative filtering steps (bilateral filtering, CLAHE, guided filtering) and a separate depth-network forward pass. The computational cost is therefore higher than that of a single end-to-end CNN, which may matter for real-time AUV deployment. Fourth, the numerical parameters used here (for example $\gamma = 0.8$, CLAHE clip $s = 0.005$, the absorption coefficients, and scattering parameters $a = 1$ and $g = 1.2$) were tuned for the test scene and its lighting and turbidity. Generalising to very different water types, turbidities and lighting setups would benefit from adaptive or learned parameterisation of these stages.

Future work includes: (i) end-to-end joint training of the CNN residual-refinement network and the pyramid-fusion weight predictors on large-scale underwater datasets; (ii) systematic benchmarking against DCP/UDCP, fusion-based and GAN-based baselines on UIEB/EUVP-style paired data with true reference images; (iii) domain adaptation or fine-tuning of the monocular depth network on underwater imagery, to improve depth accuracy in highly turbid or artificially lit scenes; and (iv) computational optimisation of the pipeline (for example stage fusion and network quantisation) for real-time operation on embedded AUV hardware.

VI. CONCLUSION

This paper presented a hybrid underwater image enhancement pipeline that is guided by physics and assisted by a CNN. It brings monocular depth estimation (MiDaS/DPT) into a physically motivated restoration model, refines the result with a lightweight residual convolutional network, and fuses it with an independent illumination/scattering-compensation branch through multi-cue, multi-scale Laplacian pyramid fusion. Under a five-metric quality assessment, the pipeline raised UIQM from -1.3716 to 2.5098 and UCIQE from 0.4226 to 0.6138 , and reached PSNR = 12.8511 dB, SSIM = 0.8289 and LPIPS = 0.0007 relative to the original image. Colourfulness, contrast and sharpness improve clearly, and the structure of the scene is preserved. These results support our main hypothesis: a colour-robust monocular depth estimate improves the physical consistency of underwater restoration compared with colour-derived transmission proxies. They also suggest that interpretable physics-based processing, targeted CNN refinement and depth-aware multi-branch pyramid fusion together form an effective and extensible approach to practical underwater image enhancement. Future work will extend the

evaluation to large-scale benchmark datasets and comparative baselines, to test how well the approach generalises.

REFERENCES

- [1] J. Y. Chiang and Y.-C. Chen, "Underwater image enhancement by wavelength compensation and dehazing," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1756–1769, 2012.
- [2] D. Akkaynak and T. Treibitz, "A revised underwater image formation model," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6723–6732.
- [3] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, 2011.
- [4] P. Drews Jr., E. do Nascimento, F. Moraes, S. Botelho, and M. Campos, "Transmission estimation in underwater single images," in *Proc. IEEE Int. Conf. Computer Vision Workshops (ICCVW)*, 2013, pp. 825–830.
- [5] D. Berman, D. Levy, S. Avidan, and T. Treibitz, "Underwater single image color restoration using haze-lines and a new quantitative dataset," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2822–2837, 2021.
- [6] C. Dai and M. Lin, "Single underwater image restoration using variational framework guided by imaging model with noise," *IEEE Access*, vol. 12, pp. 82427–82442, 2024.
- [7] C. Ancuti, C. O. Ancuti, T. Haber, and P. Bekaert, "Enhancing underwater images and videos by fusion," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 81–88.
- [8] C. O. Ancuti, C. Ancuti, C. De Vleeschouwer, and P. Bekaert, "Color balance and fusion for underwater image enhancement," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 379–393, 2018.
- [9] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. 31, no. 4, pp. 532–540, 1983.
- [10] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao, "An underwater image enhancement benchmark dataset and beyond," *IEEE Trans. Image Process.*, vol. 29, pp. 4376–4389, 2020.
- [11] M. J. Islam, Y. Xia, and J. Sattar, "Fast underwater image enhancement for improved visual perception," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 3227–3234, 2020.
- [12] C. Li, S. Anwar, and F. Porikli, "Underwater scene prior inspired deep underwater image and video enhancement," *Pattern Recognit.*, vol. 98, p. 107038, 2020.
- [13] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1623–1637, 2022.
- [14] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2021, pp. 12179–12188.
- [15] K. Panetta, C. Gao, and S. Agaian, "Human-visual-system-inspired underwater image quality measures," *IEEE J. Ocean. Eng.*, vol. 41, no. 3, pp. 541–551, 2016.
- [16] M. Yang and A. Sowmya, "An underwater color image quality evaluation metric," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 6062–6071, 2015.
- [17] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.
- [18] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 586–595.
- [19] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*, San Diego, CA, USA: Academic Press, 1994, pp. 474–485.
- [20] K. He, J. Sun, and X. Tang, "Guided image filtering," in *Proc. European Conf. Computer Vision (ECCV)*, 2010, pp. 1–14.
- [21] E. H. Land, "The Retinex theory of color vision," *Scientific American*, vol. 237, no. 6, pp. 108–128, 1977.