Just Scratching the Surface: Exploring Underwater Image Enhancement with CLIP

Jingwen Wang
1004817362
zoejingwen.wang@mail.utoronto.ca

Zixuan Hu
1005142579
zixuanh.hu@mail.utoronto.ca

Abstract – This project explores the integration of advanced neural networks to enhance underwater imagery, addressing challenges like light distortion and quality degradation. Our approach involves assessing images using CLIP to provide context-driven insights, employing contrastive loss to steer the enhancement process. While WaterNet serves as a baseline model, our CLIP-enhanced model focuses on achieving richer contextual outcomes. This approach has led to improvements in colorfulness, white balance, and exposure, as demonstrated through both quantitative assessments and qualitative evaluations. Despite some trade-offs in traditional image quality metrics, our model excels in enhancing visually perceptible attributes.

Index Terms—**Computational Photography, Denoising, Contrastive Language-Image Pretraining (CLIP), Underwater Image Enhancement, Guided Diffusion Model, WaterNet Model**

I. Introduction And Motivation

UNDERWATER photography provides a wide range of benefits to life, in areas as diverse as science, agriculture, industry, and entertainment. It also facilitates the exploration and protection of the marine ecosystem, the monitoring of the impact of climate change, and the investigation of archaeological sites. It is also critical for the maintenance of underwater facilities and the conducting of search-and-rescue operations.

Additionally, it also supports the sustainable management of fisheries, encourages the marine tourism industry, and contributes to the educational media, as well as assisting scientific research by recording the changes in the marine environment.

Unfortunately, enhancing underwater images is challenging due to light absorption and scattering in water, leading to degraded image quality with issues like low contrast, blurring, and color distortion. These problems not only affect visual quality but also hinder the performance of underwater information processing systems [1] [2]. Therefore, it is important to investigate the effective underwater optical photography reconstruction approaches to mitigate seawater's impact on photography and improve the performance of intelligent processing systems.

Neural network applications, especially in image enhancement, have become increasingly popular for their ability to learn from large datasets without requiring handcrafted features [3]. This trend is especially relevant in complex areas like underwater imaging. The advent of the Underwater Image Enhancement Benchmark (UIEB), with its extensive collection of 950 real underwater images, has greatly propelled this field forward. The UIEB dataset aids in evaluating and improving underwater image enhancement algorithms, moving beyond the limitations of specific scenes. The development and training of the CNN model, Water-Net, on this benchmark, underscores the effectiveness of deep learning approaches in enhancing underwater image quality [4]. Additionally, Contrastive Language–Image Pretraining (CLIP) which has shown remarkable capabilities in understanding and processing images in conjunction with textual descriptions, offers innovative perspectives in image enhancement; Its ability to interpret and apply semantic concepts to visual data makes it an ideal candidate for enhancing the accuracy and relevance of data-driven image enhancement methods [5].

In light of these advancements, this project aims to leverage the strengths of data-driven enhancement methods, combining the cutting-edge capabilities of the CLIP Language Model with proven deep learning networks like WaterNet. This approach seeks not only to enhance underwater image quality but also to harness the power of semantic inputs, thereby revolutionizing the field of underwater photography.

II. Related Work

WaterNet [4], introduced with the UIEB dataset, is adept at closely aligning underwater images with reference images. This model utilizes a fusion-based approach, applying White Balance, Histogram Equalization, and Gamma Correction to generate three inputs tailored to underwater image characteristics. These inputs are then combined using learned confidence maps in a gated fusion network architecture, effectively addressing color casts, contrast issues, and dark regions. However, these approaches are primarily recovery-focused, lacking the flexibility to adapt enhancements based on specific user needs or contextual variations.

The CLIP model signifies a major advancement in the contextual comprehension of images. trained on a vast collection of image-text pairs, it possesses the unique capability to associate visual content with corresponding textual descriptions. As shown in Figure 1, this model leverages the image-text relationship for image evaluation.

Recent research has emerged showcasing the use of CLIP for innovative image-text pair applications. This progression highlights how CLIP is being utilized to bridge the gap between visual content and contextual understanding, opening up new frontiers in image enhancement. **VQGAN-CLIP [6]:**

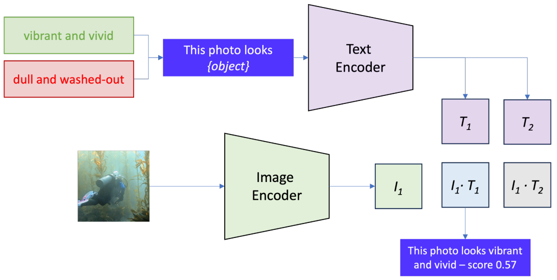


Fig. 1: CLIP model: Assessment of image quality by contrasting pairs.

This novel methodology allows for the generation and editing of high-quality images from text prompts, showcasing CLIP's flexibility and superiority over previous models in handling complex semantic tasks. **CLIP-LIT [7]:** it underscores CLIP's potential in unsupervised backlit image enhancement. It optimizes enhancement networks by learning initial prompt pairs, fine-tuning the process through iterative prompt learning, and significantly outperforming current state-of-the-art methods. **BioViL-T [8]:** In biomedical vision-language processing, BioViL-T employs CLIP in a CNN-Transformer hybrid setup, adeptly leveraging temporal content for advanced clinical predictions and analysis.

Drawing inspiration from these applications, our project endeavors to utilize CLIP's contextual understanding for enhancing underwater images. We explore the potential of CLIP as a guiding tool to augment image-enhancement neural networks. This involves employing contrastive evaluations by CLIP to steer the enhancement process. Our goal is to achieve a more controllable and context-sensitive enhancement method, aiming to improve the usability and quality of underwater imagery.

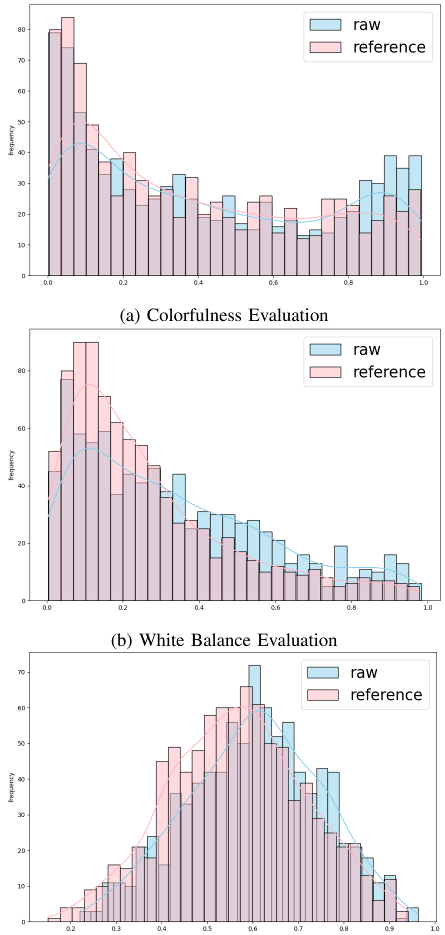
III. Method

A. Contrastive Pair Selection

In our method, we designed prompts in alignment with WaterNet's fusion-based capabilities to address key aspects of underwater image enhancement. For colorfulness, we used the prompt ”Vibrant and Vivid” to restore colors often subdued by water. ”Accurate Color Representation” was chosen to correct the blue/green color cast which is a common issue WaterNet's White Balance algorithm targets. For exposure, ”Well-lit and clear” was selected to ensure clarity and detail and it complements WaterNet's ability to enhance contrast and illuminate dark regions through Histogram Equalization and Gamma Correction. These prompts guide the enhancement process, aiming for a contextually aware improvement in line with WaterNet's confidence map-driven fusion strategy.

B. Score distributions of UIEB dataset

In our initial analysis, we examined the distribution of positive scores for contrastive pairs on colorfulness, white balance, and exposure in the UIEB dataset, as depicted in Figure 2. The comparison between raw and reference images revealed a notable similarity in their score distributions, with overlapping curves suggesting minimal variation in the range and central tendencies. Quantitative assessments further substantiated this, showing marginal mean differences of 0.043, 0.073, and 0.037 for colorfulness, white balance, and exposure, respectively, and standard deviations consistently below 0.2, as detailed in Table I. These tight clusters of differences around the means indicate that the reference images in the UIEB dataset may not exhibit substantial improvements from the view of CLIP model in the aspects of colorfulness, white balance, and exposure, signaling an opportunity for further enhancement.



(a) Colorfulness Evaluation

(b) White Balance Evaluation

(c) Exposure Quality Assessment

Fig. 2: The distribution of positive scores for contrastive pairs, comparing raw and reference images from the UIEB dataset.

Attribute Type	Mean	Std
Colorfulness	0.043	0.163
White balance	0.073	0.182
Exposure	0.037	0.150

TABLE I: Compare References to Raws: Differential Mean and Standard Deviation of Positive Scores for Visual Attributes in UIEB Dataset

C. Model Architecture

Our enhancement pipeline builds on a pre-trained WaterNet model, which was initially trained on the UIEB dataset for 400 epochs with an input resolution of 112x112. In our architecture, as illustrated in Figure 3, we incorporate two CLIP

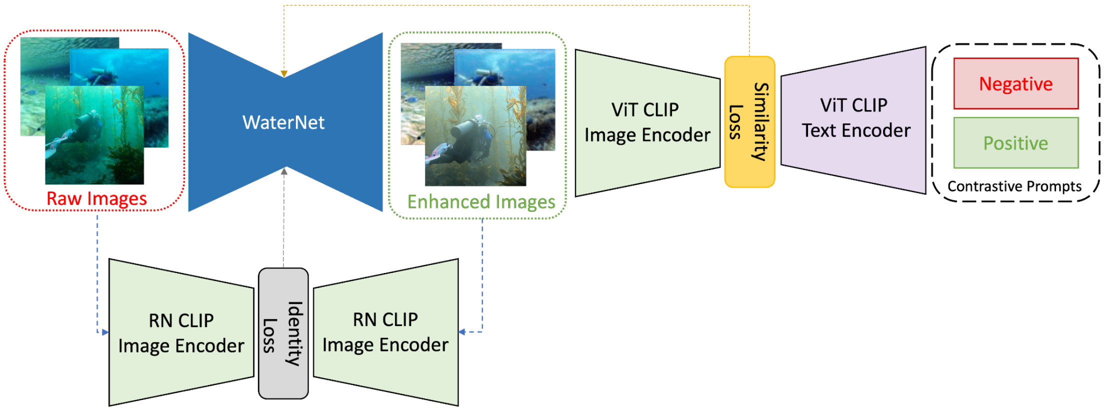


Fig. 3: Enhanced WaterNet: Integration of CLIP contrastive loss and CLIP perceptual loss

models — ViT CLIP and RN101 CLIP. The ViT CLIP model is employed to compute the contrastive loss, assessing how well the enhanced images align with desired visual outcomes based on contrastive pairs. Simultaneously, the RN101 CLIP model is tasked with evaluating identity loss, ensuring the enhanced images retain their original identity. This dual-CLIP integration allows us to harness the models' robust image-text correlation capabilities to quantitatively measure and steer image enhancements, particularly focusing on improving colorfulness, white balance, and exposure in underwater imagery.

D. *Training and Evaluations*

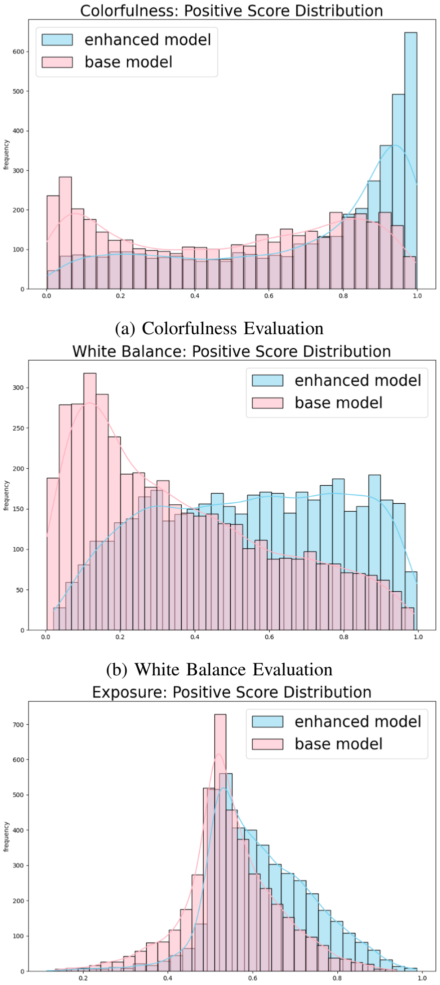
In the training phase, we employed distinct sets of prompts for colorfulness, white balance, and exposure, to fine-tune three separate WaterNet enhancements—each tailored to one of these visual attributes. Each specialized network underwent training for 20 epochs, resulting in models dedicated to color enhancement, white balance correction, and exposure adjustment. Subsequently, we merged these prompt sets to train a comprehensive model, referred to as the all-enhanced WaterNet. This training exclusively utilized the UIEB dataset, maintaining consistency with the pretraining conditions of the base network.

For evaluation, we turned to the LSUI dataset [9], which boasts 4,279 pairs of raw and reference images and it offers a comprehensive range of scenes, lighting conditions, water types, and target categories. Validation involved computing Peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) metrics against reference images for outputs from each model variant. Additionally, we calculated colorfulness, white balance, and exposure scores by averaging the positive scores from the relevant contrastive pairs, thereby quantitatively assessing the performance improvements in these key areas.

IV. Experiments Results

In our results, we present a detailed analysis of the enhancements achieved by our models. We conducted a quantitative assessment using the LSUI dataset, focusing on key metrics such as PSNR and SSIM for image quality, along with colorfulness, white balance, and exposure scores. Additionally, we performed a qualitative review of the enhanced images to visually confirm the improvements.

A. *Performances analysis of Baseline Model and Enhanced Model*



(a) Colorfulness Evaluation

(b) White Balance Evaluation

(c) Exposure Quality Assessment

Fig. 4: The distribution of positive scores for contrastive pairs, comparing raw and reference images from the UIEB dataset.

In evaluating our enhanced models against the baseline on the LSUI dataset, we observed a distinct shift in the positive score distributions for the enhanced images (Figure 4). This divergence is particularly evident when comparing the base model's output to that of the WaterNet models enhanced for colorfulness, white balance, and exposure. The table II illustrates these differences. The mean values indicate a significant improvement in colorfulness and white balance attributes, with both showing an approximate increase of 0.15 over the baseline. While the increase in exposure is less pronounced, it still demonstrates a positive trend.

Attribute type	Mean	Std
Colorfulness	0.193	0.166
White balance	0.194	0.188
Exposure	0.071	0.128

TABLE II: Compare Enhanced Model to Baseline: Differential Mean and Standard Deviation of Positive Scores for Visual Attributes in LSUI Dataset

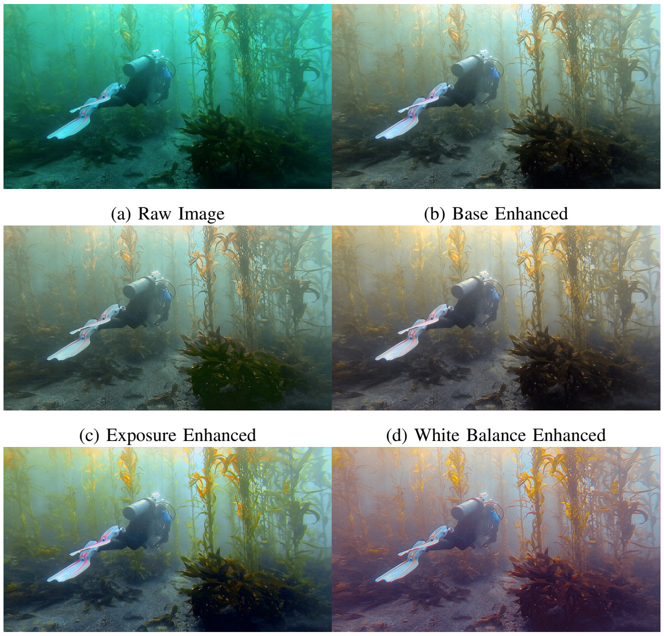
B. Overview of Perfomrances

Table III provides a comprehensive overview of the performance metrics for various image enhancement models on the LSUI dataset. The baseline model achieves the highest PSNR and SSIM values, indicating strong overall image quality preservation. However, when considering the specific attributes of colorfulness, white balance, and exposure, the specialized enhancement models demonstrate notable improvements.

The Color Enhanced model significantly outperforms the baseline in colorfulness, reflecting its effectiveness in enriching the vibrancy of underwater images. The White Balance Enhanced model scores highest in white balance, suggesting it successfully corrects color casts more effectively than the baseline or other models. Similarly, the Exposure Enhanced model leads in exposure scores, indicating its strength in rendering well-lit underwater scenes.

C. Visualized Enhanced Images

Visually, these models improved color, white balance, and exposure compared to the base WaterNet model. An example of Figure 5 showcases a series of underwater images, illustrating the effectiveness of our enhancement models. Image 5a displays the original raw capture, with noticeable color distortion and limited visibility. Image 5b reveals the improvements made by the base WaterNet enhancement. Image 5c demonstrates the exposure-enhanced model's effectiveness. It moderates the image's lighting extremes, drawing out hidden details in darker regions while tempering the brighter areas. The White Balance Enhanced model 5d neutralizes the bluish tint to deliver a natural color temperature further, and the Colorfulness Enhanced model 5e augments the scene's hues and makes the underwater landscape more vivid. Finally, image 5f shows the All Enhanced model's result, combining all enhancement aspects to deliver a comprehensive improvement in vibrancy, color accuracy, and lighting, creating a more beautiful view. More images are available in the Appendix.



(a) Raw Image

(b) Base Enhanced

(c) Exposure Enhanced

(d) White Balance Enhanced

(e) Colorfulness Enhanced

(f) All Enhanced

Fig. 5: Underwater Clarity: From Raw Capture to Enhanced Reality

V. Discussion

This project demonstrates the CLIP model's capacity to serve as a bridge between textual descriptions and visual content, effectively 'understanding' and evaluating image quality. By integrating the CLIP model into our existing framework, we have successfully developed an enhancement model that not only produces beautifully enhanced images but also steers enhancements in the desired directions. The CLIP model's power can be extended to benchmarking which can offer a means to score and assess image visual quality quantitatively. The CLIP model thereby can be particularly advantageous when working with limited datasets, as the CLIP model aids in refining and augmenting models to achieve better results with fewer data requirements. This capability underscores the potential of using sophisticated image-text correlation models to enhance and evaluate visual content in a targeted and controlled manner.

Our examination of the enhanced images (Figure 5) and quantitative data (Table III) elucidates the effectiveness of various WaterNet enhancements on underwater imagery. Side-by-side visual comparisons offer insights into the impact of each model variant—base, color, white balance, exposure, and the integrative all-enhanced—on the images' aesthetic attributes. These comparative analyses highlight the trade-offs between maintaining image fidelity and achieving perceptual enhancement in color and exposure, informing future refinements in model design and application. Notably, the models enhance images in a manner that aligns with our perceptual preferences.

The all-enhanced model, despite registering lower PSNR and SSIM values, surpasses the base model in colorfulness, white balance, and exposure, suggesting a more pronounced enhancement of these attributes. This apparent discrepancy

model	psnr	ssim	colorfulness	white balance	exposure
baseline	**22.154**	**0.850**	0.489	0.354	0.548
Color Enhanced	20.421	0.822	**0.682**	0.310	0.558
White Balance Enhanced	20.490	0.839	0.332	**0.549**	0.559
Exposure Enhanced	21.202	0.830	0.407	0.230	**0.618**
All Enhanced	18.985	0.771	0.622	0.704	0.573

TABLE III: Performances of the image enhancement models in LSUI dataset

underscores that lower fidelity metrics are not necessarily indicative of inferior quality. Rather, they may signify a deliberate emphasis on enhancing certain visual properties that are more aligned with human perceptual biases than with objective similarity to the original image.

Our project advances underwater image enhancement but encounters limitations. Over-enhancement remains a concern, potentially introducing unrealistic elements while adjusting color, white balance, and exposure. Additionally, the model's performance varies under diverse underwater lighting conditions. This highlights the need for more adaptive models to handle complex lighting without over-processing. Furthermore, our use of CLIP models introduces variability, as similar contrastive pairs can yield different scores that affect enhancement consistency. The influence of preprocessing, especially image resolution, on CLIP's scoring also suggests the importance of fine-tuning these aspects for more reliable results.

VI. Conclusion

This project has described the creation and testing of an improved underwater image processing model. The combination of the WaterNet and CLIP models, together with the application of contrastive learning, has resulted in considerable improvements in the color, white balance, and exposure of underwater photos. Despite significant picture quality tradeoffs, quantitative tests confirmed the model's superior performance in improving visual qualities. Qualitative evaluations confirmed these findings, emphasizing the model's ability to generate perceptually attractive visuals. To create even more robust and flexible underwater picture improvement, future studies will try to optimize these techniques further, addressing issues such as over-enhancing, fluctuating illumination circumstances, and improving the CLIP model's adaptability.

References

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

[2] Martin Ludvigsen, Bjørn Sortland, Geir Johnsen, and Hanumant Singh. Applications of geo-referenced underwater photo mosaics in marine biology and archaeology. *Oceanography*, 20(4):140–149, 2007.

[3] Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8:1–74, 2021.

[4] 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.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[6] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision*, pages 88–105. Springer, 2022.

[7] Zhexin Liang, Chongyi Li, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Iterative prompt learning for unsupervised backlit image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8094–8103, 2023.

[8] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15016–15027, 2023.

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

Appendix

More enhanced Images are available on the following pages. A video demo is available at HERE

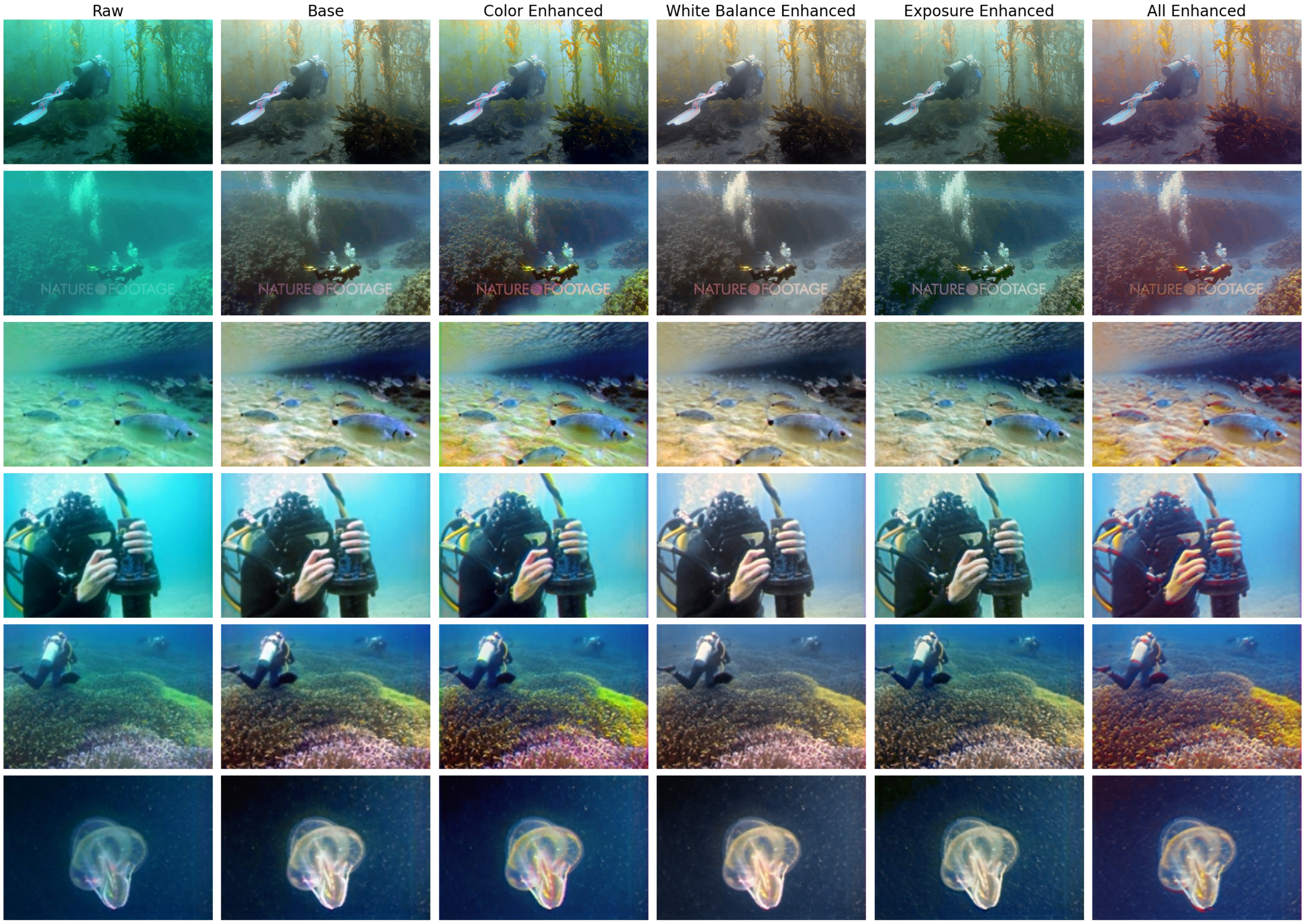


Fig. 6: Comparison image for all models

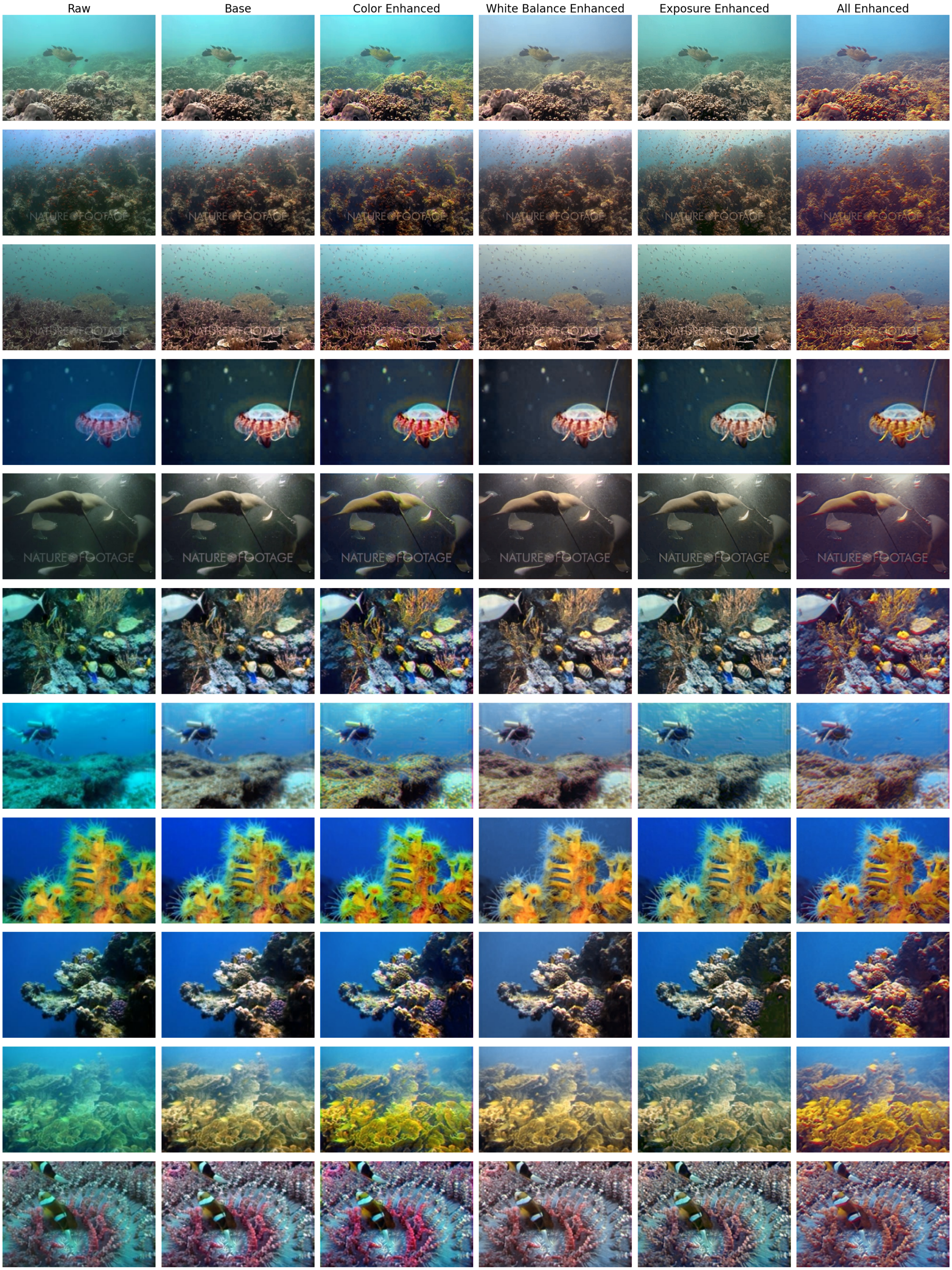


Fig. 7: More Comparison image for all models