Using Vision Transformers (ViT) For Low-Level Vision Enhancement

By

Muneeba Daud

Supervisor

Dr. Hammad Afzal

A thesis submitted to the Department of Computer Software Engineering, Military College of Signals, National University of Sciences and Technology, Islamabad, Pakistan, in partial fulfilment of the requirement for the degree of MS in Software Engineering.

January 2023

Using Vision Transformers (ViT) For Low-Level Vision Enhancement

By

Muneeba Daud

00000330087

Supervisor

Dr. Hammad Afzal

A thesis submitted in partial fulfilment of the requirement for the degree of Mater of Science in Software Engineering MSSE

In

Department of Computer Software Engineering, Military College of Signals (MCS), National University of Sciences and Technology, Islamabad, Pakistan

(January 2023)

Thesis Acceptance Certificate

Certified that final copy of MS/MPhil thesis entielted **"Using Vision Transformers (ViT) For Low-Level Vision Enhancement"** written by **Muneeba Daud**, (Registration No**. 00000330087**), of Department Computer Softwre Engineering Military College of Signals (MCS), has been vetted by undersigned, found complete in all respect as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial, fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the student have been also incorporated in the said thesis.

Signature: _____

Name of Supervisor: **Dr. Hammad Afzal**

Date: _____

Signature (HoD): _____

Date: _____

Signature (Dean): _____

Date: _____

Declaration

I, *Muneeba Daud* declare that this thesis "Using Vision Transformers (ViT) For Low-Level Vision Enhancement" and the work presented in it are my own and have been generated by me as a result of my original research.

I confirm that:

1) This work was done wholly or mainly while in candidature for a Master of Science degree at NUST.
2) Where any part of this thesis has previously been submitted for a degree or any other qualification at NUST or any other institution, this has been clearly stated.
3) Where I have consulted the published work of others, this is always clearly attributed.
4) Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my work.
5) I have acknowledged all main sources of help.
6) Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

_____

Muneeba Daud,
NUST00000330087 MSSE27

Copyright Notice

Dedication

“In the name of Allah, the most Beneficent, the most Merciful"

I dedicate this thesis to my parents.

Abstract

Restoring and enhancing underwater images is a significant issue in image processing and computer vision. Poor underwater imaging quality is caused by the scattering and absorption of light by underwater contaminants. Images taken underwater frequently suffer from quality issues, such as low contrast, poor sight (due to the absorption of natural light), blurred details, changing colors, additive noise, blurred effects, and uneven illumination, etc.

The study of underwater image analysis has gained a lot of attention and achieved substantial advancements during the past few decades. The current techniques can broaden the application of underwater photography while improving image contrast and resolution. Traditional image enhancement techniques have some drawbacks when applied directly to underwater optical environments; hence, some specific algorithms, such as histogram-based, retinex-based, and picture fusion-based algorithms, are proposed. Deep learning has recently shown a strong potential for creating results that are satisfying and have the right colors and details, but these methods significantly increase the size of the image processing inference models and therefore cannot be applied or deployed directly to the edge devices.

Recently, Vision Transformers (ViT)-based architectures are producing incredible results. In recent years, there has been more interest in transformers. Their interactions between image content and attention weights can be thought of as a convolution that changes in space, and their self-attention mechanism is good at simulating long-distance dependencies and global features.

The suggested approach is a pipeline based on a context-aware lightweight vision transformer with the goal of improving image quality without sacrificing the naturalness of the image, as well as reducing the inference time and size of the model. In this study, we trained a deep network-based transformer model on two standard datasets, i.e., Large-Scale Underwater Image (LSUI) and Underwater Image Enhancement Benchmark Dataset (UIEB), so that the network becomes more generalized, which subsequently improved the performance. Our real-time underwater image enhancement system shows superior results on edge devices. Also, we provide a comparison with other transformer-based methods. Overall findings indicate that the suggested method has produced underwater images of higher quality than the original input underwater images, which had a high noise ratio and more color disruption.

Keywords: Tokenization, Feature Extraction, Image Enhancement, Underwater Image Restoration, ViTs, Computer Vision.

Acknowledgments

All praises to Allah for the strengths and His blessing in completing this thesis. I would like to convey my gratitude to my supervisor, Dr. Hammad Afzal, PhD, for his supervision and constant support. His priceless help of constructive comments and suggestions throughout the experimental and thesis works are major contributions to the success of this research. Also, I would thank my teacher Lt. Col. Khawir Mahmood for his support and guidance throughout and committee members; Assoc. Prof. Dr. Naima Iltaf, and Assoc. Prof Dr. Ihtesham ul Islam for their support and knowledge regarding this topic. Lastly, I am highly thankful to my parents for their constant support. I would like to thank them for their patience, cooperation and motivation in times of stress and hard work.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

CNN	***Convolutional Neural Network.***
LSTM	***Long Short Term Memory.***
GAN	***Generative adversarial network.***
LSUI	***Large-scale underwater image.***
NLP	***Natural language processing.***
SQUID	***Stereo quantitative underwater image dataset.***
UIE	***Underwater image enhancement.***
UIEB	***Underwater image enhancement benchmark.***
ViT	***Vision Transformer.***
DCP	***Dark Channel Prior***
PSNR	***Peak Signal to Noise Ratio***
DuGAN	***Dual Generative Adversarial Network***
IPT	***Image Processing Transformer***
CMSFFT	***Channel-wise Multi-Scale Feature Fusion Transformer***
SGFMT	***Spatial-wise Global Feature Modelling Transformer***
AGA	***Adaptive Group Attention***
WB	***White Balance***
GC	***Gamma Correction***
SSIM	***Structural Similarity Index Measure***
MSE	***Mean Square Error***
HVS	***Human Visual System***

Chapter 1

1. Introduction

1.1 Overview

Underwater light propagation is hampered by scattering and absorption, much like light passing through the air. However, there is a great deal of absorption in water as compared to air. While in the air, the light reduction coefficients are estimated in inverse kilometers, they are measured in inverse meters in an underwater environment. When light is severely degraded, it is extremely difficult for optical sensors to collect data from a target underwater location. Water, in contrast to air, is opaque to all other wavelengths and only permeable to the visible portion of the electromagnetic spectrum. The visible spectrum's component wavelengths are also absorbed at various rates, with longer wavelengths being absorbed more quickly. It is actually amazing how quickly light energy degrades in water. By 150 meters in depth, less than 1% of incident light is still present in the middle ocean's extremely clear waters. As a result, the object is harder to see beyond a 20-meter distance, and in muddy coastal waters, the visibility drops below the 5-meter threshold, as shown below. [27]

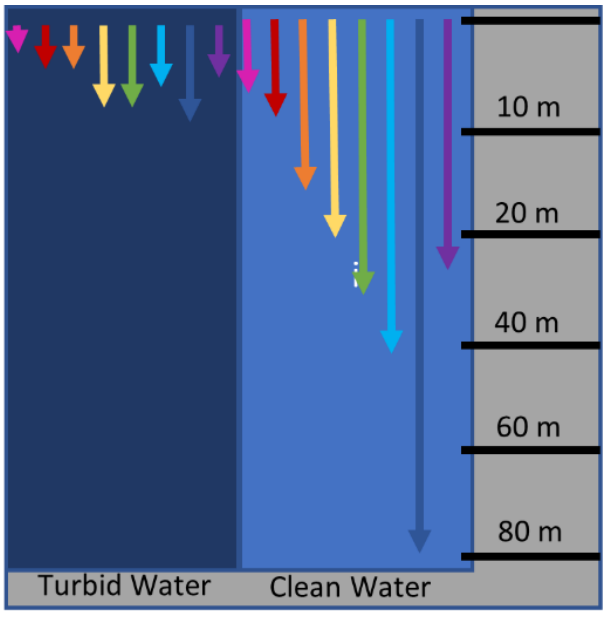

Figure 1: Attenuation of Light in Clean and Turbid Water

Underwater lighting is typically scarce due to two inevitable realities. One is that light under water loses some of its true luminosity, and two is that there is a good probability that light will disperse within water media that are full of suspended particles. The portion of light energy that reaches the water is quickly absorbed and transformed into heat, which in turn energizes the water molecules, causes them to warm up, and causes them to have a tendency to evaporate. [30] Additionally, some of the sunlight energy is consumed by the microscopic organisms found in plants that use it for photosynthesis. As already said, the part of the light that doesn't get taken up by water molecules may not move in a straight line but instead moves randomly in a Brownian motion. As shown below, this happens because there are particles in the water that move around.

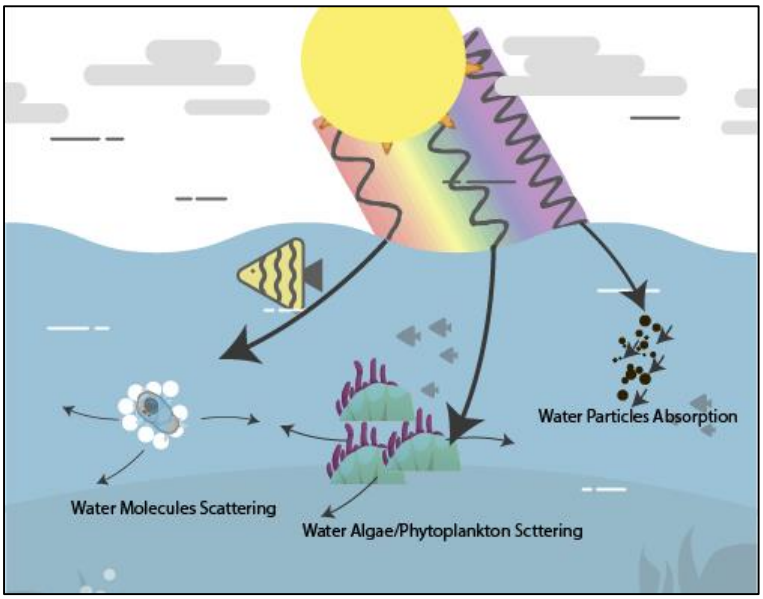

Figure 2: Underwater Light Scattering and Absorption

The light beam is reflected and deflected by the dissolved salts, organic and inorganic substances in water, especially in sea water. The light beam can also leave the ocean surface and disperse back into the atmosphere. There are three basic parts to the light that the camera detects. The direct portion of light that is directly reflected from an object when it is not scattered by water. The light beam that has a flaw after striking the target item and before it reaches the image sensor is said to have undergone forward scattering, i.e., the second component. Typically, this kind of scattering causes an image to appear blurry. [9] The third component is backscattering, which occurs when a light beam strikes an imaging system without first returning from the object. It only serves to further reduce an image's contrast.

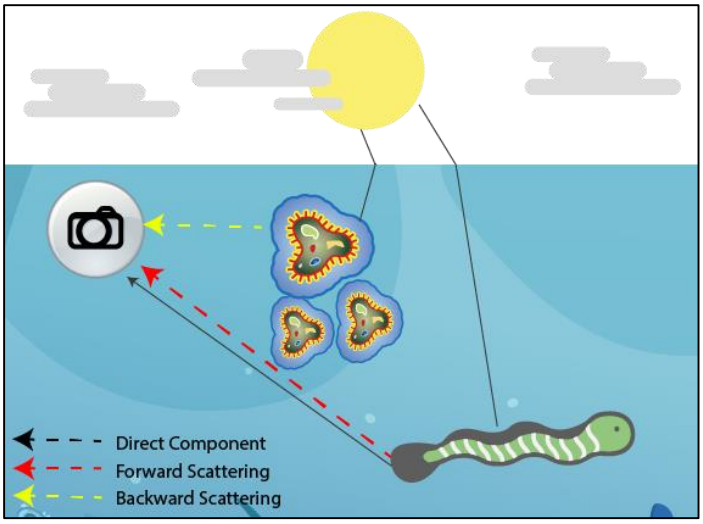

Figure 3: Components of Camera Light

In order to evaluate underwater visibility, optical and acoustic imaging methods have traditionally been used. Although acoustic sensors have a lesser spatial resolution than optical systems, they have the important advantage of being able to penetrate water even more quickly. However, when looking for high resolution outputs, acoustic sensors are very big. On the contrary, optical systems have lately been used by evaluating the physical impacts of visibility loss, despite some drawbacks such low underwater visibility. [23] Adding an artificial light source is one technique to increase visibility but this solution has its own drawbacks.

Aside from the issues with light dispersion and attenuation noted above, artificial light has a tendency to illuminate the subject of interest unevenly and typically results in a bright point in the middle of the image with darker tones surrounding it. The lighting apparatus is also expensive and bulky. Additionally, they needed a steady supply of electricity, either connected to surface ship or in the form of batteries. As a result, the traditional image processing methods that work well for enhancing terrestrial imaging must be altered or abandoned entirely, and new solutions must be developed.

1.2 Motivation and Problem Statement

It's a general statement that an image captured in water will always have worse quality. The contrast and genuine tone quality required for identifying the subject of interest in the image are lost. The efficacy of the algorithms being used to obtain information from the photos is negatively impacted by this condition, making it extremely difficult to retrieve near features from the data. Despite the

existence of various image improvement algorithms, when used on underwater image regions with poor lighting, existing techniques typically provide inaccurate results and unavoidably deteriorate some visual artefacts of the image. Researchers have always been impressed by the process of constructing and training neural networks, yet there is still considerable opportunity for advancement. For instance, the generalizability of existing approaches is likely to be constrained by their tendency to bias towards a narrow range of brightness values and scenarios. These networks' growing popularity lately sparked the development of novel deep-learning-based models for improving underwater images. Therefore, the suggested research intends to increase the quality of underwater images using deep learning models.

1.3 Aims and Objectives

The following objectives are the focus of the research:

- Create a cutting-edge algorithm that uses lightweight Vision Transformers to lessen the inconsistent attenuation problem that affects underwater photos in many color channels and spatial regions.
- Offer a range of supplementary resources and comparative experimental investigations on common datasets using state-of-the-art methodologies to illustrate the capability of our proposed paradigm.
- Carry out a study to assess the effectiveness and efficiency of our methodology by assessing the outcomes of improved photographs using objective evaluation.

1.4 Research Contribution

To the finest of our knowledge, the methodology and framework presented in this thesis have not previously been used in the process of improving the quality of underwater images. The primary points of this thesis are:

- Implementing a novel real-time Underwater Image Enchantment Benchmark (UIEB) model based on deep learning methods. The base pipeline of the model is contingent on Vision Transformers (ViT) that are structure-aware and are able to capture long range dependencies between image patches.

- Authenticating the viability of the proposed model by conducting experimental analysis of common datasets on the proposed method as so that the network is accustomed to better generalization for performance improvement.
- In addition, the suggested model is trained on numerous versions to explore the versatility. By altering the transformer module's base parameter and preloading the images into the model with various analysis strategies, an ablation study is carried out.

1.5 Thesis Organization

The thesis has been organized as follows:

- Chapter 2 gives an overview regarding the related work in the domain of Underwater Image Enchantment (UIE). The section 2.1 thoroughly discusses the Physical Model-Based techniques used in UIE. Section 2.2 gives briefing about Non-Physical Model Enhancement. Section 2.3 gives overview of Underwater Image Enhancement Analysis Using Data-Driven Methods and Section 2.4 provides summary of Underwater Image Enhancement Analysis Using Deep Learning. The chapter also gives systematic review of the models used in the current study.
- Chapter 3 discusses materials and methodology used for conducting the analysis. It gives an overview regarding dataset collection, preprocessing, feature extraction techniques, baseline models and discusses the proposed framework.
- Chapter 4 is results and discussion which presents results of the best baseline models applied and their limitations. It also provides an insight of the results improved by applying proposed framework for the particular dataset. The comparative analysis of previous studies is also presented which shows how classification has improved with application of deep learning models.
- Chapter 5 is conclusion and future work which summarizes the research work, presents the limitations of the study and the proposed framework with respect to the UIE, it also suggests future direction in the corresponding domain.

Chapter 2

2. Related Work

This chapter provides a thorough analysis of the literature with an emphasis on several methods for improving underwater images and restoring their clarity and quality. The existing UIE techniques can generally be divided into three groups: physical model-based i.e., the image restoration techniques, non-physical based method, and data-driven or deep learning-based techniques. [29] Nonphysical model enhancement methods and physical model-based enhancement algorithms are examples of traditional underwater image enhancement techniques. The following discussion classifies and groups the various technologies and techniques utilized for underwater image enhancement according to their distinctive characteristics.

2.1 Underwater Image Enhancement Analysis Using Physical Model-Based or Image Restoration Techniques

In order to enhance the quality of the image from the imaging principle, an algorithm using the physical model examination uses the reverse process of the imaging paradigm to obtain a clear image. It is also known as the "image restoration method." These approaches incorporate integral imaging, polarization, and dark channel priors.

While using physical model-based improvement techniques, underwater imaging models are extremely important. A particularly popular recovery model is the Jaffe-McGlamery underwater imaging model. Using the Jaffe-McGlamery underwater imaging model, which is depicted in Figure 3, the light obtained by the camera E_t was divided into three parts: the light reflected straight from an object, E_d; the forward scattered portion, which is small-angle light reflected from a target, E_f; and the backscattered light, which is non-target reflected light E_b.

$$E_t = E_d + E_f + E_b \qquad (1)$$

A self-calibrating filter based on a condensed Jaffe-McGlamery model was proposed by Trucco and Olmos. Trucco et al. [25] suggested a more straightforward version of the filter to automatically modify the image restoration. The ideal filter parameter value is automatically determined by gauging global contrast requirements for picture quality evaluation. Based on less backscatter, a simplified model could produce results that are more ideal. The polarization was estimated by

Ferreira et al. [20] by using the unreferenced mass measure and optimizing the settings using the particle swarm algorithm and subsequently improved visual quality and greater adaptability as the cost function for restoration. However, the parameter optimization procedure makes the operation more time-consuming.

Meng et al. [16] exploited the color balance and volume approaches for color correction and image sharpening. In undersea pictures the color balance changes when the red channel value is close to the blue channel. Otherwise, the Dark Channel Prior (DCP) based recovery depending on the sharpening method's maximum a posteriori probability (MAP) improves visibility, lessens fuzziness, and enhances foreground retention textures but too many new settings added.

In spite of the significant scattering effect of murky water, integral imaging technology has a tremendous impact since it can combine signals from several images. Single-photon imaging with a threshold was suggested by Li et al. [14]. A detecting method to separate photon signals from the chaotic undersea environment. By using this technique, photos captured in a high-loss underwater environment are reconstructed. The Peak Signal to Noise Ratio (PSNR) is theoretically improved by applying photon-limited computational techniques and in the high-noise environment. However, this technology is expensive to adopt and is dependent on the development of an imaging system.

Physical model-based UIE can only work well in complex and varied underwater scenes found in real life. It is difficult to evaluate numerous parameters at once, and model hypotheses are not always reasonable in the complex and dynamic undersea environment.

2.2 Underwater Image Enhancement Analysis Using Non-Physical Model Enhancement

There are a number of targeted techniques proposed, including histogram-based, retinex-based, and visual fusion-based algorithms, due to the limitations of applying standard picture enhancing techniques to the special underwater environment.

It is clear that retinex has a limited number of direct applications for improving underwater image quality. The issue with the enhanced image is either too little contrast or too much exposure. To change the color and illumination, it is customary to blend RGB with HSV or other color schemes. Additionally, it can be used with other preprocessing or post-processing techniques like filtering, contrast stretching, and color correction. This may result in plainly improved visuals. The fact that

the best models of this kind of approach frequently have too many parameters is an inherent drawback. Various undersea conditions require different parameter settings.

Underwater image quality can be effectively increased using the fusion technique. These techniques, however, call for the acquisition of several fusion weights and images. The solution to the issue is in the adoption of effective tactics to get the best fusion weight.

A Bayesian combined with retinex framework was created by Zhuang et al. [35] for multi-order gradient prior's enhancement of a single underwater picture of illumination and reflection. An underwater image formulation with a maximum posteriori color-corrected image is enhanced by applying a multi-order gradient, prioritizing lighting and reflectance. This algorithm performs well as technique for color correction, maintains naturalness, promotes structures and details. But the breakdown and alternative optimization of sub problems take too long to solve.

Song et al. [22] suggested a technique based on the global stretching and multiscale fusing of dual models. White-balancing was used to get rid of the unwanted color variation and show an updated image. The model used contrast and spatial signals in combination with the saliency weight coefficient method. The red, green, and blue channels are simultaneously stretched globally. There are still issues with this method regarding the depth of color model.

Li et al. [13] proposed a paradigm for underwater hybrid systems. Stretching the histogram while using an improved underwater white balance method. Contrast and saturation are increased, scattering-related blur is eliminated, color adjustment, haze reduction, and clarity of details are all enhanced by creating a variable brightness and saturation enhancement model.

2.3 Underwater Image Enhancement Analysis Using Data-Driven Methods

Images can be degraded because of a variety of factors and correcting just one of these factors might affect the other. For example, a technique for improving an image's brightness or contrast may not be appropriate for images with high saturated areas. As a result, when using image enhancement algorithms, many important factors must be considered like sharpness, dynamic range, and distortion etc. Recently, data-driven methods that may be considered as machine learning technologies in the UIE domain have demonstrated excellent performance on UIE tasks. Several models for improving underwater images have already been put forth. Here, we give a brief overview recent efforts based on Data-Driven approaches that scientists have looked into thus far.

WATER-NET, a gated fusion network, was suggested by Li et al. [12]. The underwater image is enhanced using white balance, histogram equalization, and gamma correction algorithms, and the final image is produced by integrating the confidence graphs of various enhancement techniques. Despite the data' poor quantitative analysis, the reference model performs well in terms of generalization and has opportunity for improvement.

In order to process underwater images, Wang et al. [26] presented a parallel convolutional neural network with two parallel branches, a transmission estimation network, and a global ambient light estimate network. To avoid the halo effect and maintain edge properties, the network uses multiscale estimation and cross-layer connectivity. The contrast improvement, however, is not strong enough.

The adaptive color correction algorithm was implemented by Ding et al. [4] to correct color distortion. The color-corrected image was instantly turned into transmission image for repair after CNN network was utilized to calculate the depth map. It is necessary to enhance the algorithm's real-time performance and robust adaptation.

Lack of a large dataset with a variety of underwater settings and high-fidelity reference photos is a problem for the current data-driven based underwater image enhancement (UIE) algorithms. Additionally, the uneven attenuation in various color channels and space regions is not fully taken into account for enhanced images.

2.4 Underwater Image Enhancement Analysis Using Deep Learning Approaches

Deep learning approaches immediately learn the translation relationship between the source input images and the clean underwater image without being constrained by model assumptions or previous conditions.

For the purpose of enhancing underwater images, Guo et al. [7] presented the UWGAN, a new multiscale dense generated adversarial network that incorporates residual multiscale dense blocks into the generator. Multiscale manipulation, dense cascading, and residual learning, respectively were applied to enhance performance, render more detail, and fully exploit features. The discriminant uses the spectral normalization calculation method to stabilize discriminant training. The algorithm currently lacks real-time and adaptive capabilities.

An end-to-end dual generative adversarial network (DuGAN) for improving underwater images is proposed by Zhang et al. [32]. In which two discriminators are utilized to complete adversarial training toward various portions of images using various training procedures after segmenting the images into clear and unclear parts. However, this solution relied on a user-guided way to gather reference photos, making it challenging to train with fresh images.

By learning many samples, the deep learning-based approach can lessen the effect of the challenging undersea environment on the outcomes. However, the dataset is crucial, as the existing dataset's coverage is still constrained. However, the majority of deep learning-based techniques just concentrate on improvement rather than fully integrating the underwater image model.

Although compared to conventional UIE approaches, deep learning-based UIE methods significantly improved. Its further advancement is still constrained by two factors: (1) the uniform convolution kernel cannot adequately describe the uneven attenuation of underwater images in various color channels and spatial regions; and (2) the CNN-GAN architecture is more focused on local features than long-term dependence and global feature modelling.

Computer vision is undergoing a revolution because of introduction of a new architecture called Vision Transformers (ViT). It is based on the self-attention process, which projects vectors created by splitting the input image into patches into linear space. One of the most promising methods in computer vision today, the Vision Transformers-based architectures are producing incredible results. Vision Transformers have attracted increasing amounts of interest recently; their content-based interactions between image content and attention weights can be understood as spatially varying convolution; and their self-attention mechanism is effective at simulating long-distance dependencies and global features. Transformers provide a number of advantages over recurrent networks, including the ability to simulate lengthy dependencies between input sequence parts and support for simultaneous processing of sequence. Transformers, as opposed to Long Short-Term memory (LSTM) and convolutional networks, are perfectly suited as set-functions and only need minor inductive biases for their design.

Image Processing Transformer (IPT) is the name of a pre-trained model that Chen et al. [8] proposed based on the Transformer architecture. It is capable of restoring images in a variety of ways, including super-resolution, denoising, and deraining. IPT has a shared encoder-decoder Transformer body as well as many heads and tails that can each do a distinct task independently.

As demonstrated in a recent study, image transformer networks have emerged as a formidable rival to conventional CNNs in many applications. Denoising, dehazing, and two degrees of up-scaling are the four key tasks that the pre-trained image transformer network has been trained to complete. The Image Processing Transformer (IPT) network's results demonstrated that the trained IPT consistently beat the other, more focused CNNs in the aforementioned tasks.

A large-scale underwater image (LSUI) dataset was initially created by Peng et al. [18], which covers more underwater scenes and better visual quality reference photos than current underwater datasets. The collection includes 5004 actual underwater photographs, and as comparison references, the matching clear photographs are created. Additionally, based on the attention mechanism, the researchers created a channel-wise multi-scale feature fusion transformer (CMSFFT) and a spatial-wise global feature modelling transformer (SGFMT), which they then integrated into the U-shape Transformer. They also created a multi-color space loss function that includes RGB, LAB, and LCH in accordance with the color space selection experiment.

An innovative underwater picture improvement technique called UDAformer, developed by Shen et al. [21], is based on the Dual Attention Transformer Block (DATB), which also includes the Channels Self-Attention Transformer (CSAT) as well as Shifted Window Pixel Self-Attention Transformer (SW-PSAT). In particular, the extreme and uneven loss of underwater images makes conventional underwater image enhancement depending solely on channel self-attention insufficient. The effective storing and decoding of underwater picture information is therefore proposed using a unique fusion method that combines channel and pixel self-attention. Then, in order to increase computational efficiency, the shifted window approach for pixel self-attention is suggested.

Huang et al. [10] unique Adaptive Group Attention (AGA) proposal allows for the dynamic selection of visually comparable channels based on dependencies, hence requiring fewer additional attention parameters. End-to-end underwater image improvement network is designed using the AGA, which is utilized inside the Swin Transformer (ST) module. The multiresolution cascade component as well as the channel attention technique are also introduced by the network. However, their approach has the following drawbacks. The complexity of the calculation brought on by stepwise reinforcement learning results in the average speed performance, and it also produces less-than-ideal details in areas where black pixels are concentrated.

A fusion neural network was presented by Sun et al. [24], which is based on the merging of two images that are created from the white-balanced and color-corrected version of the original underwater image. In this study, they created two inputs from an underwater image by using the white balance (WB) & gamma correction (GC) methods, respectively. The SwinMT module, which has two components: a low-frequency feature extraction unit and a high-frequency feature to recover high-quality which extracts features in turn.

Learning-based techniques have advanced significantly in the domain of photo enhancement in recent years. The improvement techniques, however, rely on intricate network structures and use an excessive amount of processing power.

Chapter 3

3. Methodology and Framework

Many software-based solutions have been proposed and improved over the years to address two of the main problems with degraded underwater photos. But even systems already in place are still susceptible to the underwater picture attenuation problem, as well as lack the adaptability to deal with the diversity of real-world settings. [19]

This chapter introduces an enhanced underwater image enhancement method that outperforms the existing UIE method. To begin, patches of the source images are created, and these patches are tokenized to create token embedding, which are similar to word embedding. The model actively learns token-wise dependencies for picture patches rather than directly computing pixel-wise connections. This architecture enhances an image with exceptional efficiency. It can also give answers that are more semantically relevant than CNNs due to its great efficiency and ability to implicitly understand the semantic structure. We carry out a number of ablation investigations in order to determine the most effective Context-Aware Vision Transformer (CAViT) settings.

3.1 Overview of Vision Transformers (ViT) Vision Transformers

Transformer is a Seq2Seq framework that replaces conventional recurrent neural network used in Natural Language Processing (NLP) nearly entirely by introducing a self-attention strategy and using position embedding to account for the position information. When Transformers were discovered to be so successful in the domain of natural language interpretation in 2020, Google experts questioned: "How would they perform with images?". Knowing that the transformer accepts word vectors as input, how do we convert an image to a word vector? Using all of the image's pixels and placing them "inline" to create a vector is the initial option. Attention has intrinsic complexity of $O(N^2)$, which means that to evaluate the complexity of every pixel in relation to every other pixel in any low-resolution images like 256x256 pixels, the quantity of calculations would be enormous and completely beyond the capabilities of today's technology. [34] Therefore, to make this strategy effective it is suggested in the paper "A picture is worth 16x16 words" [6] to split the image into patches, then transform each piece into a vector that used a linear projecting that would map the patches in a dimensional space.

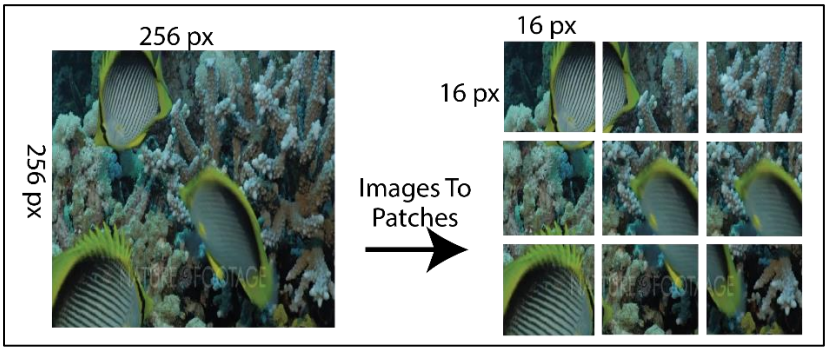

Figure 4: Underwater Image Divided Into Patches

The patches are projected linearly to produce vectors, which are then combined with knowledge about the patch's location within the picture and fed into a traditional Transformer Encoder. The insertion of information regarding the patch's original position inside the image is essential since, despite being crucial to completely comprehending the image's content, this knowledge would be lost during the linear projection. The result relating to this patch being the one that is taken into account and fed into a Multi-Layer Perceptron (MLP). An additional vector is inserted that is unrelated of the picture being analyzed and is utilized to get global data on the entire image.

Learning-based techniques have advanced significantly with in field of photo improvement in recent years. However, the implementation of the enhancement approaches on light-weight devices becomes significantly more challenging because they depend on complicated network architectures and use a lot of Computational resources. The approaches also perform poorly in real-time when processing photos with very high resolution. In contrast to earlier research on creating structurally varied CNN networks, photo augmentation can be accomplished using a simple self-attentive approach for global-local tuning.

Numerous studies, demonstrate that using global functions alone is unable to offer sufficient and versatile enhancing capabilities. A semantic-aware prediction approach is another type of image enhancement techniques that teaches the CNN model to calculate translation or transformation functions using semantic masks. Because the prediction results in these methods are conditional to semantically calculated components like other pixel-wise techniques, models in such methods are typically adaptable.

3.2 Proposed Underwater Image Enhancement using Context-Aware Lightweight Vision Transformers

We create the CAViT model, which is focused on image improvement tasks and can operate without stacking convolution to extract structural data more effectively. Similar to word embedding in NLP, patches of the photograph are tokenized and turned into token embedding in our model. CAViT actively understands token-wise dependencies for input images rather than directly computing pixel-wise connections. CAViT enhances the image with excellent efficiency. Along with being highly effective, CAViT can intuitively learn the semantic information and hence produce results that are more semantically meaningful than CNNs. Nevertheless, obtaining comparable performance with CNN often requires a large amount of training data or extra supervision else cannot perform as expected due to the lack of inductive biases.

Suggested model's overall layout is shown in Figure 7. To start, we flatten an image $I \epsilon R^{H\ X\ W\ X\ C_I}$ into a series of tokens $I_T \epsilon R^{L\ X\ C_T}$, wherein C_I & C_T are the channel counts, respectively. The Context-Aware Vision Transformer module will receive the created tokens as inputs and produce structural map $S_I \epsilon R^{L\ X\ L\ C_S}$. The predicted structural map will be utilized to estimate additional transformations for underwater image improvement.

3.2.1 Feature Maps Extraction (2D-Flattened Patches)

Given an unprocessed underwater image that was shrunk into the form $I \epsilon R^{H\ X\ W\ X\ C_I}$, first the image was divided into patches of the form $X_{patch} \epsilon R^{(H\ /P)(W/P)3}$, where P is the patch's size. The shallow embedding feature F_s is then obtained using a linear projection, and C is the dimensionality of shallow features. The Context-Aware Transformer module will receive the shallow embedded feature to produce deep features $F_d\ \varepsilon\ R^{(H\ /2P)\ (W/2P)\ 2C}$ with down sampling. The down sample doubles the features while reducing distortion and preserving the structural integrity of the image.

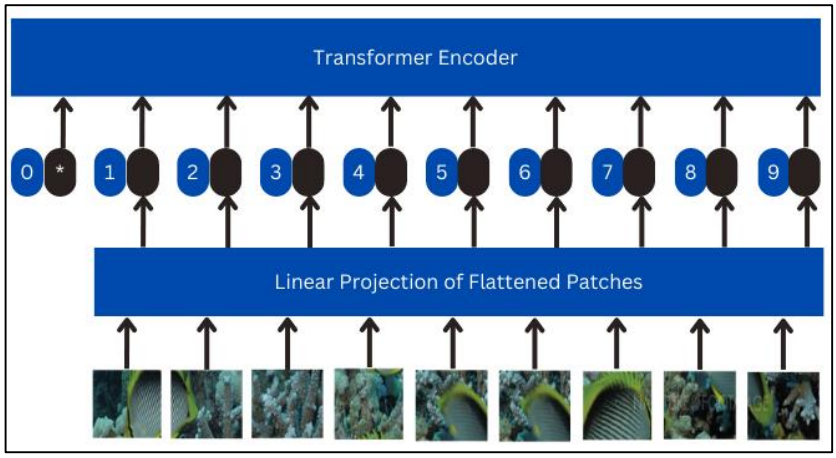

Figure 5: Underwater Image Patches fed into the Vision Transformer Encoder

A primitive method of turning an image into tokens is to flatten it into raw patches, as described earlier. The features $I \epsilon R^{H\ X\ W\ X\ C_I}$, in this case are reshaped into a series of patches and treated as tokens. This approach, meanwhile, will use up a lot of memory. In particular, the input token vectors are designed to have a big dimension $t_i \epsilon R^{P^2\ X\ C_I}$, $i = 1,2,3, \dots, N$, necessitating high training parameters (e.g., 33 M parameters in [8]). An alternate approach is to derive input for a sequence using a CNN's feature maps. After spatially down sampling, the patches size in this instance can be considered to be 1x1, and tokens are extracted using a stacked convolution technique.

3.2.2 Tokenization Strategy

With the subsequent process, we put the tokenization pipeline into action to deliver efficient real-time image enhancement. As shown in Figure 5, we first flatten all of the resolution features into a collection of patches. Behind that, each Patch goes a cascading dimension reduction procedure. Then, we extract tokens from each patch in more detail using linear embedding learning. The Mean Head method, where Adaptive Average Pooling immediately reduces the spatial size followed by the Linear Head Embedding, is used to lower memory use with dimension reduction. This was inspired by the squeeze-and-excitement block [11].

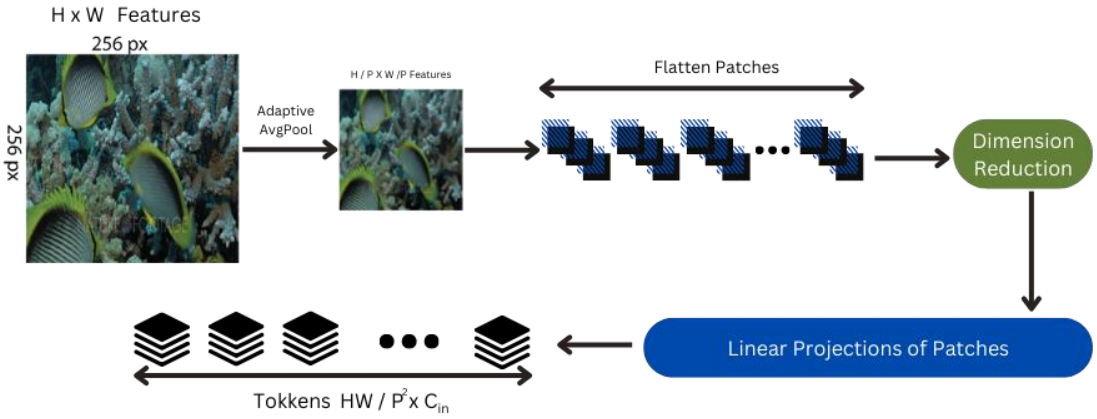

Figure 6: Mean Head / Squeeze-and-Excitation Tokenization Strategy

Input spatial resolution is shown by the H x W features, while Transformer dimension is indicated by the C_{in} features. In Mean heads strategy adaptive average pooling is applied to reduce the spatial size before implementing the linear head technique. In Linear Head Feature maps are split into patches directly using the linear head technique, which is followed by projection and embedding. Similar to [31], we apply a 7 x 7convolution with stride 4 and output channels 16 to the picture and then feed output features to the aforementioned tokenization modules for more informative representations. By using Mean Head, we can decrease the tokenization complexity as much as possible.

3.2.3 Attention mechanism

We employed a local global spatial module of the Transformer used in the area of computer vision; instead of a two-branch Transformer design, we used a single branch with spatial local-global attention to process token sequences. We employ a standard transformer module, which includes an MLP with a skip connection and a multi-head self-attention module. We choose GELU as the non-linearity function and LayerNorm as the normalization. In order to keep position information intact we also add a 1D learning position embedding $p \in R^{C_T/2}$ to Transformer inputs.

$$Attention(Q,K,V) = Softmax\left(\frac{QK'}{C} + p\right)V \qquad (2)$$

$$MLP\big(LN(T_n)\big) + T_n, \quad n = 1 \ldots N \qquad (3)$$

Specifically, MSA, MLP, and LN stand for multi-head self-attention, multilayer perceptron, and layer normalization where N represents the Transformer's depth (number of basic transformer blocks).

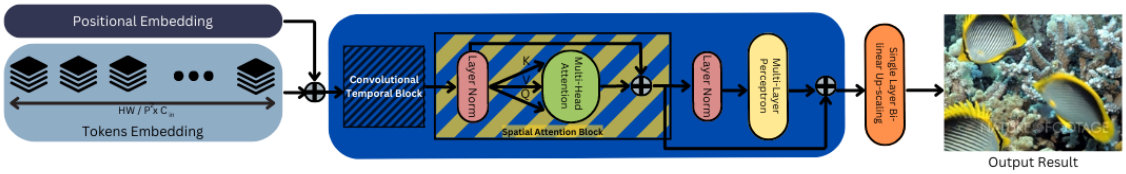

Figure 7: A Long-Short Range Transformer module is shown in an illustration. We use a single branch design that explicitly divides global and local context extraction modules to lessen model complexity. We use position embedding and the conventional Transformer design as recommended by [24].

In a typical Transform block, a linear layer creates the projections from the input features, Query (Q), Key (K), & Value (V), but only accomplishes global spatial interactions. It seems sense to substitute a convolution with a kernel size of 3 x 3 for the linear layer in order to employ more local information, as this simultaneously reinforces the channel and spatial augmentation. In order to cover neighboring tokens for the convolutions 2-D Block convolutions are utilized to analyze the rearranged picture tokens as opposed to 1-D convolutions, which are used for processing sentences in natural language processing (NLP). We embedded the convolution branch using the entire Transformer module rather than embedding the convolution inside the inner Transformer module.

3.3 Underwater Image Datasets

There are two datasets utilized in this study:

3.3.1 The UIEB dataset.

The underwater Image Enhancement Benchmark (UIEB) consists of two subsets of 950 real-world underwater photos. 890 pairs of raw underwater photographs and the associated high-quality reference images, and 60 difficult images without reference. The dataset was annotated from the different Internet sites, relevant papers, and video footage and, contains a variety of underwater scenes and aquatic animals. To create the high-quality reference images, 12 image enhancement techniques were applied to the training dataset. Volunteers choose the final, high-quality reference photos.

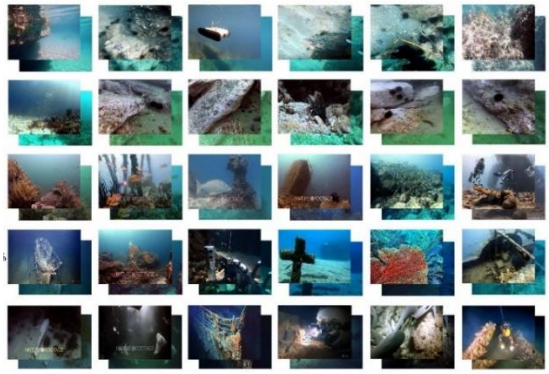

Figure 8: The LSUI Dataset.

3.3.2 The LSUI Dataset.

5004 pairs of natural underwater photographs are included in the Large-Scale Underwater Image (LSUI) dataset. Compared to the current underwater datasets, it has better reference photographs with more varied underwater habitats. Compared to the current ones, the images in large-scale underwater image (LSUI) dataset pairings feature richer underwater settings (lighting conditions, water kinds, and target categories).

3.3.3 Preprocessing for CAViT

According to [3], the training and assessment images are both downsized to 1200 x 900 pixels depending on their longest side. In order to achieve comparability, the particular dataset determines the proportion of test data. To identify the best trained model, the data set was split into 80% training examples and 20% validation data. Additionally, random cropping, resizing, flipping, and rotating are used to enhance training data.

3.4 Experimental Analysis of Proposed Framework

3.4.1 Implement Details

Pytorch is used to implement the suggested CAViT-UIE along with an NVIDIA NVIDIA-SMI 460.32.03 GPU and CUDA Version 11.2 without pre-trained networks. The Adam [3] optimizer is used, processing 30 epochs with a batch size of 8, with a preset learning rate of $1e^{-4}$. The default setting was set to use Transformer depth 1 as concluded with multiple experiments that incurring the depth of the transformers did not increase any artifacts of the suggested pipeline.

3.4.2 Hyper parameters Details

The path embedding size for the network is 32, and the skip path dropping ratio is set to 0.1. The scale factor in Q and K is 8, the ratio multiplied in MLP is 4, the head number of Transformer is also 8. Each image will be separated into 32 by 32 tokens. The network has a CNN with 7 layers and a maximum of 24 output channels (21.87 k parameters). The Transformer dimension (C_{in}) was initially set to be the same 32. Additionally, the internal MLP dimensions remain unchanged. And models both predict 24 curves (totaling 8 iterations for 3 channels).

3.4.3 Loss function

To objectively evaluate the model performance of the model, we use gradient loss that uses the straightforward L1-norm as the loss function. The L1-norm between generated and ground truth patches is minimized by employing MAE (Mean Absolute Error) losses during network training. It is advantageous to sharpen the edge of the improved image because the gradient loss not only collects low-frequency information, such as the L1 loss, but also by adding a second-order constraint it acquires the high-frequency information. L1 gradient loss is stated as follows: Let $\hat{G}$and G denote the gradient map of $\hat{X}$ and X, restored images $\hat{X}$ and the real images X, respectively.

$$L_{gd} = E_{\hat{G}\sim Q(r),G\sim Q}\left\| \hat{G} - G \right\|_1 \qquad (4)$$

Where Q(r) and Q(g) are the distribution of $\hat{G}$and G, respectively. We recommend minimizing overall patch-wise absolute measure of cosine similarity across various patch representations as a simple method. We modify the training objective to include the patch-wise cosine loss provided the final-layer patch approximation h[L]of an input x:

$$L_{cos} = \rho(h^{[l]}) \qquad (5)$$

Specifically, this regularization loss reduces overall pairwise cosine similarity among various patches. The linear summation of the loss functions for each task results in the total loss function, which would be written as:

$$L_{sum} = L_{coss} + L_{gd} \qquad (6)$$

3.5 Ablation Study – Experimental Analysis

The White Balancing and Gamma Correction branch tries to improve the appearance of underwater images by eliminating undesired color cast brought on by various illuminants. It is employed because the underwater images suffer noticeably when water depths are greater than 30 feet. The goal of the White Balancing and Gamma Correction branch is to improve the overall contrast and brighten up dark areas of the underwater images. The evaluation comparison is explained in the next chapter.

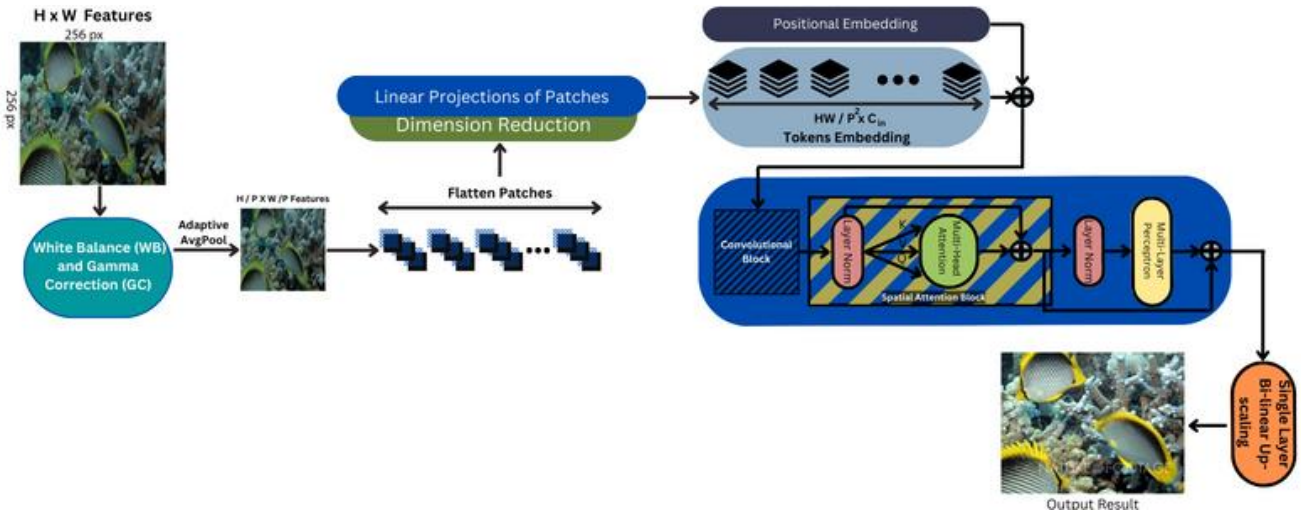

Figure 9: An overall architecture of the proposed Context-Aware Light weight Vision Transformer with White Balancing and Gamma Correction

Chapter 4

4. Results and Discussion

To ensure that our model may concentrate on obtaining global contexts and reduce computations, we use a specific two-module architecture named long-short Range Transformer rather than using single section for general information. According to experimental findings, CAViT can frequently improve these tasks' performance while using models that are much less sophisticated, which has major benefits for real-time computation on edge devices.

4.1 Objective Evaluation Metrics

Non-reference evaluation and fully-reference evaluation are the two basic categories of objective assessment methods. A complete reference assessment involves a group of images in "true color" and "ideal contrast." The non-reference methods are more appealing because it can be difficult to find ground-truth underwater photography photos. Due to the necessity to understand overall Human Visual System (HVS) as well as how it perceives the quality as a whole, this task becomes more challenging. In this work, we will evaluate our model on both full-reference and no-reference assessment techniques.

4.1.1 Full-Reference Evaluation

We carry out the assessment using Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) measures, that reflect the similarity to the reference, in order to objectively evaluate the recovered images with paired reference images provided on the dataset. The more similar the structure between images is, the higher the PSNR and SSIM values.

For comparison purposes, both mean square error & peak signal to noise ratio are typical measures. The mean square error measures the total square difference between the original input image and the final output image. The equation below displays the mean square.

$$MSE = \frac{1}{M\ X\ N}\sum_{M,N}[l_{1(m,n)} - l_{2(m,n)}]^2 \qquad (7)$$

$l_{1(m,n)} - l_{2(m,n)}$ stands for the original and improved images, respectively. The image's sides are shown in the format MxN, where m and n stand for pixel's x & y values in the image's dimensions.

The greatest practicable signal-to-noise ratio is known as the PSNR. A peak signal-to-noise ratio is given by the following equation.

$$PSNR = 2log_{10}\left(\frac{L-1}{RMSE}\right) \qquad (8)$$

Root mean squared error is referred to as RMSE. Lower mean square error as well as a high peak signal to noise ratio denote the best final image, respectively.

Three crucial elements are extracted from an image via the Structural Similarity Index (SSIM) metric; Structure, Contrast and Luminance. These three elements serve as the foundation for the comparison of the two photos. And finally, the SSIM score is given by,

$$SSIM(x,,y) = [l(x,y)]^{\alpha}.[c(x,y)]^{\beta}.[s(x,y)]^{\gamma} \qquad (9)$$

where $\alpha > 0$, $\beta > 0$, $\gamma > 0$ denote the relative importance of each of the three components.

4.1.2 No-Reference Evaluation

It is a challenging undertaking to fairly and thoroughly evaluate UIEs for non-reference testing data. UIQM focuses on underwater image color measure (UICM), underwater image sharpness measure (UISM), and underwater image contrast measure (UIConM). Better visual perception is indicated by a higher UIQM score.

The equation for the underwater image quality measure (UIQM), which combines the measures of color, sharpness, and contrast, is provided by:

$$\text{UIQM} = \alpha \cdot \text{U ICM} + \beta \cdot \text{U ISM} + \gamma \cdot \text{U IConM} \qquad (10)$$

The weight coefficients used to balance the values of the three measures are α, β and γ. These settings are often configured to be 0.0282, 0.2953, and 3.5753.

- Underwater Image Colorfulness Measure (UICM)

The majority of underwater photographs suffer from color casting, which worsens as depth increases and exhibits variable attenuating ratios depending on the color. Due to the fact that red has the shortest wavelength and generally disappears first, while blue and green wavelengths decay more slowly, underwater scenes frequently exhibit a green or blue tinge. In addition, as was already noted, the light attenuation greatly diminishes an image's hues. Both Red-Green (RG) and Yellow-Blue (YB) color components are thus evaluated by the UICM in order to gauge the effectiveness of color correcting algorithms.

- Underwater Image Sharpness Measure (UISM)
  The corners of a picture are reflected in sharpness, and finely caught photographs are likely to exhibit superior sharpness. However, because of backscatter and absorption, photographs taken underwater suffer from extreme blurring and distortion. The Operator is first used on each color component to create edge maps, which are then used to measure the sharpness. Then, to determine the grayscale edge maps, the resulting edge maps are multiplied by the original color component.
- Underwater Image Contrast Measure (UiconM)
  Contrast is an underwater visual performance factor. Backscattering is typically to blame for the contrast reduction in underwater photos.

4.2 Experimental Analysis of Proposed Transformer Models

On both UIEB dataset and LUSI dataset, we carried out extensive trials. The ability of Vision Transformers to learn important information even at the lowest layers, as opposed to CNN, allows them to stand out from the competition. CNN can only extract high-level information from the last layers. As a result, the datasets are shrunk additionally by eliminating redundant scenes. For visual comparison, we selected a sampling of each type's most representative photographs.

For the UIEB 300 corresponding original and reference images denoted as Train-U are prepared as the training dataset in the training process of the model. Additionally, remaining 90 photos from the UIEB dataset, designated as Test-U90, are the testing data. These testing dataset Test-U90 are used as the Full-Reference testing dataset as the images pair are included in the dataset originally.

For the Non-Reference testing dataset, the Challenging Set of UIEB dataset is being utilized. The Challenging Dataset contain 60 images for which the reference images are not included due to the complexity of the Underwater scenes. The Dataset is denoted as the Test-U60 testing dataset. There are five different underwater environments in Test-U60 scenes that exhibit high backscattering and color variations include those that are reddish, yellowish, greenish, bluish, and hazy.

The LSUI collection includes 5004 actual underwater pictures with more numerous undersea sceneries (water kinds, lighting conditions and target groups) as well as comparison references, are presented. Moreover, it offers the intermediate transmission map and semantic segmentation for each unprocessed underwater photo. The LUSI dataset containing 1500 images denoted as Train- L is being utilized as the training images for the proposed transformer models. And the rest 70 images demoted as Test-L70 are utilized as the Full-Reference testing dataset for the proposed dataset. The all obtained statistical results are discussed in the following sections.

4.2.1 Evaluation on Multiple Dataset

Training Parameters for LUSI training dataset Train-L and UIEB training dataset Train-U					
	Tokenization	**Branches**	**Epochs**	**Batch Size**	**No. of Features**
CAViT	Adaptive Average Pooling (Mean Head Strategy)	Single Branch Model	30	8	24
CAViT$_G$	Adaptive Average Pooling (Mean Head Strategy)	Single Branch Model	30	8	24

Table 1: Training Parameters

Comparisons of CAViT and CAViT$_G$ denote the model with/without gamma correction. We report the model size and corresponding average PSNR and SSIM on LUSI and UIEB

evaluation sets. The best model must give its performance high in PSNR value, low MSE value, high SSIM value and high UIQM values.

LUSI Full-Reference Test on Test Set Test-L70				
	Training Time (s) ↓	**Parameters (K) ↓**	**PSNR (dB) ↑**	**SSIM ↑**
CAViT	2273.31 s	21.87 K	24.80	0.93
CAViT$_G$	10253.25 s	21.87 K	25.76	0.95

Table 2: Full-Reference Test

Figure 10: Enhancement results of CAViT and CAViT$_G$ trained on LUSI underwater datasets. (a): Input images. (b): Enhanced results using the model trained on the Train-L dataset. (c): Enhanced results using the model trained with Gamma Correction component on the Train-L dataset. (d): Reference images (recognized as ground truth (GT).

LUSI Non-Reference Test on Test Set Test-L70				
	UICM ↑	**UISM ↑**	**UIConM ↑**	**UIQM ↑**
CAViT	5.188	5.59284	0.1937	2.4904
CAViT$_G$	5.588	5.79	0.1913	2.69

Table 3: Non-Reference Test

LUSI Inference Results					
	Runtime Latency (s) ↓	**Parameters (K) ↓**	**PSNR (dB) ↑**	**SSIM ↑**	**MSE ↓**
CAViT	0.566 s	21.87 K	23.81	0.967	270.689
CAViT$_G$	0.659 s	21.87 K	23.89	0.969	265.188

Table 4: Inference Results

(a) (b) (c) (d)

Figure 11: Inference results. (a): Input images. (b): Enhanced results using the model trained on the Train-L dataset. (c): Enhanced results using the model trained with Gamma Correction component on the Train-L dataset. (d): Reference images (recognized as ground truth (GT).

UIEB Full-Reference Test on Test Set Test-U90				
	Time (s) ↓	**Parameters (K) ↓**	**PSNR (dB) ↑**	**SSIM ↑**
CAViT	1057.23 s	21.87 K	21.37	0.89

CAViT$_G$	4174.406 s	21.87 K	23.54	0.96

Table 5: Full-Reference Test

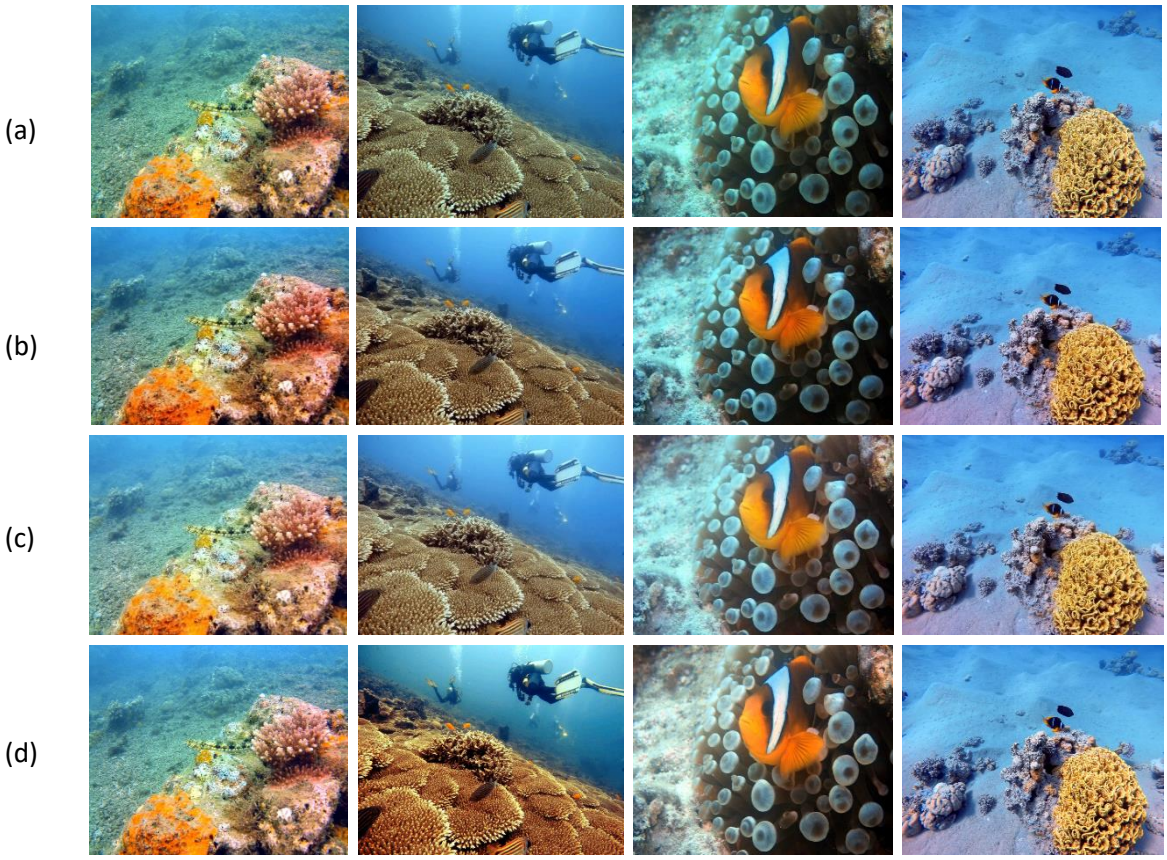

Figure 12: Enhancement results of CAViT and CAViT$_G$ trained on UIEB underwater datasets. (a): Input images. (b): Enhanced results using the model trained on the Train-U dataset. (c): Enhanced results using the model trained with Gamma Correction component on the Train-U dataset. (d): Reference images (recognized as ground truth (GT).

UIEB Non-Reference Test on Test Set Test-U90				
	UICM ↑	**UISM ↑**	**UIConM ↑**	**UIQM ↑**
CAViT	8.25	7.162	0.248	3.233
CAViT$_G$	7.89	7.893	0.281	3.290

Table 6: Non-Reference Test

UIEB Inference Results					
	Runtime Latency (s) ↓	**Parameters (K) ↓**	**PSNR (dB) ↑**	**SSIM ↑**	**MSE ↓**
CAViT	0.44 s	21.87 K	25.49	0.981	183.83
CAViT$_G$	0.77 s	21.87 K	26.36	0.981	150.178

Table 7: Inference Results

(a) (b) (c) (d)

Figure 13: Inference results. (a): Input images. (b): Enhanced results using the model trained on the Train-U dataset. (c): Enhanced results using the model trained with Gamma Correction component on the Train-U dataset. (d): Reference images (recognized as ground truth (GT).

4.3 Comparative Analysis of Various UIE Methods

To demonstrate our performance superiority, we compare the CAViT Model with 6 UIE approaches. It contains comparison of different data-driven techniques: WaterNet [12], U-Trans [18], Ucolor [2], FUnIE [15], UIE-DAL [17] and UGAN [1]. The top results are bold. This analysis includes both the Non-Reference and Full Reference evaluation techniques. As well as the visual results are presented at the end to fully understand the metrics and demerits of each method.

Methods	**Test U-90**		**No. of Parameters ↓**	**Time ↓**
	PSNR ↑	**SSIM ↑**		
WaterNet [12]	19.81	0.86	24.81M	0.61s
U-Trans [18]	22.91	0.91	65.6M	0.07s

Ucolor [2]	20.78	0.87	157.4M	2.75s
FUnIE [15]	19.45	0.85	7.019M	0.09s
UIE-DAL [17]	16.37	0.78	18.82M	0.07s
UGAN [1]	20.68	0.84	57.17M	**0.05s**
Ours	**23.54**	**0.96**	**21.87K**	0.44s

Table 8: Quantitative comparison of different UIE methods on the full-reference testing dataset.

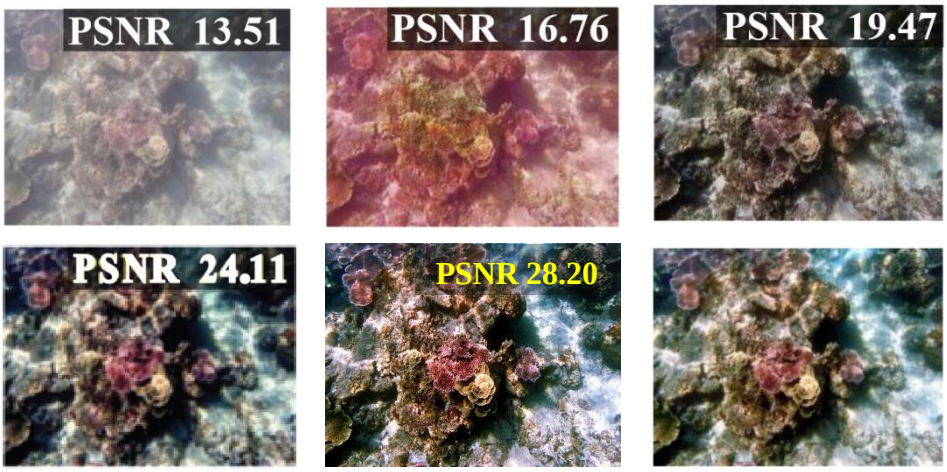

Figure 14: Visual comparison of enhancement results sampled from the Test-U90 (UIEB) dataset. From left to right are raw underwater images, FUnIE[15], UGAN[1], Ucolor[2], U-Trans [18] and our CAViT. the reference image recognized as ground truth (GT). The highest PSNR value is marked in yellow.

Methods	**Test U-60**
	UIQM ↑
WaterNet [13]	0.92
U-Trans [18]	0.85
Ucolor [2]	0.84
FUnIE [2]	1.03
UIE-DAL [17]	0.72

UGAN [1]	0.86
Ours	**2.37**

Table 9: Quantitative Comparison among different UIE methods on the non-reference testing dataset.

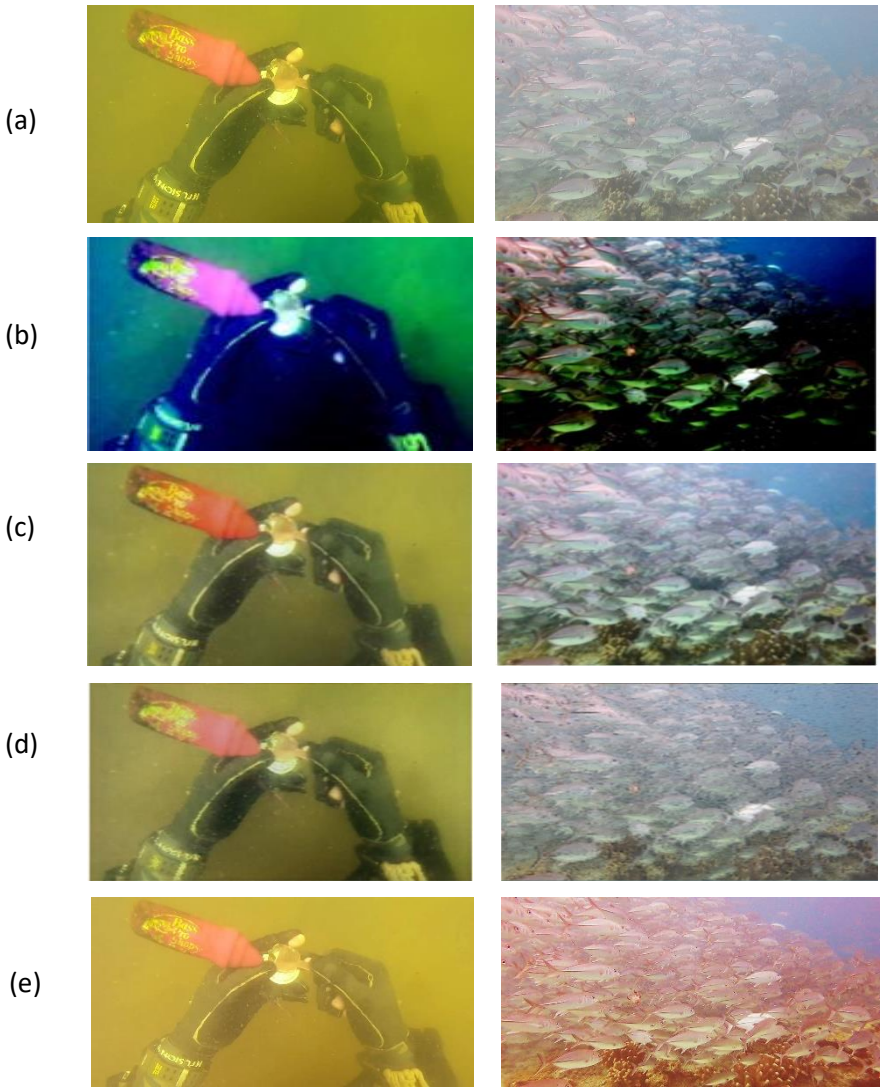

Figure 15: Enhancement results of different methods for Test-C60. The images represent underwater scenes of the yellowish, greenish- bluish colors. (a)Raw images. (b) WaterNet [12]. (c) Ucolor [2]. (d) U-Trans [18]. (e) Proposed Model.

Chapter 5

5. Conclusion and Future Work

5.1 Conclusion

This chapter, briefly summarizes the efforts, limitations, and recommendations for future studies. After explaining the conceptual approach, conducting experiments, and reviewing the findings the final observations and expositions are discussed. The goal of the project is to create a model that performs better than cutting-edge approaches. Underwater image improvement and repair are essential for many practical uses such as underwater tasks including exploration, monitoring, and recovery carried out by autonomous or semiautonomous robots, and this presents a significant challenge for computer vision and image processing. In this paper, we introduce Context-Aware Vision Transformer (CAViT), a unique and portable deep learning model for improving underwater image quality. The suggested techniques provide quick inference with less memory consumption. To assess Context-Aware Vision Transformer (CAViT) settings, we undertook numerous tests. The proposed Context-Aware Vision Transformer (CAViT) is shown to be efficient and effective when compared to previous state-of-the-art work through quantitative and qualitative results.

5.2 Limitations

By learning many samples, the methodology based on deep neural algorithms can lessen the effect of the complicated undersea environment. However, the dataset is crucial, as the existing dataset's coverage is still constrained. Deepest learning-based techniques, put more emphasis on improving full integration of the underwater imaging model. Therefore, maximizing the spatial features can bear good generalization performance.

5.3 Future Work

The network can be trained using perception-related loss function and introduce factors that are consistent with human interpretation, which will make the network more effective across wider range of scenarios.

Additionally, in order for the network to handle the loss of detailed data while taking into consideration speed, researchers can use the network's multi scale context features more frequently and specifying the step-wise reinforcement learning techniques while improving real-time performance and strengthen research on underwater video enhancement technology.

6. References

1. C. Fabbri, M. J. (2018). Enhancing underwater imagery using generative adversarial networks. *ICRA*, (pp. 7159–7165).

2. C. Li, S. A. (2021). Underwater image enhancement via medium transmission-guided multi-color space embedding. *IEEE T. Image Process*, 4985–5000.

3. Chunle Guo, C. L. (2020). Zero-reference deep curve estimation for low-light image enhancement. . *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 1780-1789).

4. Ding, X., Wang, Y., Zhang, J., & Fu, X. (2017). Underwater image dehaze using scene depth estimation with adaptive color correction. *In Proceedings of the OCEANS*.

5. Dong, L. Z. (2022). Underwater image enhancement via integrated RGB and LAB color models. *Signal Processing: Image Communication*, 104.

6. Dosovitskiy, A. B. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. . *preprint arXiv*.

7. Guo, Y. H. (2019). Underwater image enhancement using a multiscale dense generative adversarial network. *IEEE Journal of Oceanic Engineering* , 862-870.

8. Hanting Chen, Y. W. (2020). Pre-Trained Image Processing Transformer. *preprint arXiv*.

9. Hu K, W. C. (2022). An Overview of Underwater Vision Enhancement: From Traditional Methods to Recent Deep Learning. *Journal of Marine Science and Engineering.*, 241.

10. Huang, Z. L. (2022). Underwater Image Enhancement via Adaptive Group Attention-Based Multiscale Cascade Transformer. *IEEE Transactions on Instrumentation and Measurement*, 1-18.

11. Jie Hu, L. S. (2018). Squeeze-and-excitation networks. . *In Proceedings of the IEEE conference on computer vision and pattern recognition,*, (pp. 7132–7141).

12. Li, C. (2019). An underwater image enhancement benchmark dataset and beyond . *IEEE Transactions on Image Processing* , 4376-4389.

13. Li, X., Hou, G., Tan, L., & Liu, W. (2020). A hybrid framework for underwater image enhancement. *IEEE Access* , 197448–197462.

14. Li, Z., Zhou, H., Li, Z., Yan, Z., Hu, C., Gao, J., & Jin, X. (2021). Thresholded single-photon underwater imaging and detection. *Opt. Express* , 28124–28133.

15. M. J. Islam, Y. X. (2020). Fast underwater image enhancement for improved visual perception. *IEEE Robot. Autom*, 3227–3234.

16. Meng, H., Yan, Y., Cai, C., Qiao, R., & Wang, F. (2020). A hybrid algorithm for underwater image restoration based on color correction and image sharpening. . *Multimed. Syst.*, 1–11.

17. P. M. Uplavikar, Z. W. (2019). All-in-one underwater image enhancement using domain-adversarial learning. *CVPR Workshops*, (pp. 1–8).

18. Peng, L. C. (2021). "U-shape Transformer for Underwater Image Enhancement." . *preprint arXiv:*.

19. Salman Khan, M. N. (2021). Transformers in Vision: A Survey. *ACM Comput. Surv.*

20. Sánchez-Ferreira, C., Coelho, L., Ayala, H., Farias, M., & Llanos, C. (2019). Bio-inspired optimization algorithms for real underwater image restoration. . *Signal Process. Image Commun.* , 49–65.

21. Shen, Z. X. (2022). UDAformer: Underwater image enhancement based on dual attention transformer. *SSRN* , p. 4162641.

22. Song, H., & Wang, R. (2021). Underwater image enhancement based on multi-scale fusion and global stretching of dual-model. *Mathematics* , 595.

23. Soni, O. K. (2020). A survey on underwater images enhancement techniques. *In 2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT)* (pp. 333-338). IEEE.

24. Sun, J. D. (2022). Swin transformer and fusion for underwater image enhancement. *In International Workshop on Advanced Imaging Technology (IWAIT) 2022 (Vol. 12177), SPIE.*, (pp. 627-631).

25. Trucco E, O.-A. A. (2006). Self-tuning underwater image restoration. *Oceanic Engineering, IEEE Journal* , 511-519.

26. Wang, K. (2019). Underwater image restoration based on a parallel convolutional neural network. *Remote sensing*, 1591.

27. Wang, R. W. (2015). Review on underwater image restoration and enhancement algorithms. *In Proceedings of the 7th International Conference on Internet Multimedia Computing and Service*, (pp. 1-6).

28. Wang, Z. (2021). Uformer: A general u-shaped transformer for image restoration. *preprint arXiv*.

29. Yang, M. &. (2019). An In-Depth Survey of Underwater Image Enhancement and Restoration. *IEEE Access*, 1.

30. Yuan X, G. L. (2022). A Survey of Target Detection and Recognition Methods in Underwater Turbid Areas. *Applied Sciences*, 4898.

31. Yuan, L. C. (2021). Tokens-to-token vit: Training vision transformers from scratch on imagenet. *In Proceedings of the IEEE/CVF International Conference on Computer Vision* , (pp. 558-567).

32. Zhang, H., Sun, L., Wu, L., & Gu, K. D. (2021). An effective framework for underwater image enhancement. . *IET Image Process.*

33. Zhang, Z. J. (2021). Star: A structure-aware lightweight transformer for real-time image enhancement. . *In Proceedings of the IEEE/CVF International Conference on Computer Vision* , (pp. 4106-4115).

34. Zhou, J. W. (2022). Underwater image enhancement method with light scattering characteristics. *Computers and Electrical Engineering*.

35. Zhuang, P., Li, C., & Wu, J. (2021). Bayesian retinex underwater image enhancement. . *Eng. Appl. Artif. Intell.*, 104171.