Master's Thesis

Underwater Image Enhancement Based on Attention Mechanism and Multi-scale Generative Adversarial Network

Jinhua LIU

Supervisor: Kazunori Kotani

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

March 2022

Abstract

With the development of ocean exploration, in recent years, people have relied on underwater robots for resource exploration, environmental surveying and other activities. However, due to the underwater medium and a large number of suspended particles, the light is absorbed and scattered in the water. Therefore, the original underwater image has suffered serious degradation, such as color distortion, low contrast, and blurred images. In order to obtain a clear underwater scene, it is of great significance and value to adopt underwater image enhancement technology.

Underwater image enhancement technology mainly includes traditional methods and deep learning methods. Traditional methods perform overall modeling of underwater scenes and invert the degradation process based on physical models; deep learning methods are based on each pixel value of the whole image, learning the mapping relationship between underwater images and original images. But for most backgrounds, such as water, its color is often not our focus. Compared with the whole part, we focus on the more important parts.

In this paper, we propose a novel improvement on Generative Adversarial Networks that can be simply embedded in existing neural networks and can distinguish the foreground and background parts of an image without additional processing.

Through this optimization, which pay more attention to foreground and enhance the details. we make our generative adversarial network achieve better visual performance when the SSIM, PSNR and other indicators are the close or even inferior to other networks and achieves better performance in color correction and detail preservation. We also demonstrate the effectiveness of our proposed attention module and multi-scale module through ablation experiments.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Research Background

In recent years, with the gradual decrease of land resources, the ocean has become the target of human exploration. As an important information carrier, underwater images intuitively reflect underwater environmental information, and play an important role in the research of marine exploration, marine environmental monitoring, biological rescue, underwater robots, marine military applications, etc.

However, due to the influence of wavelength-dependent light absorption and scattering, underwater images are often accompanied by serious color shifts, blurring and loss of details. When light propagates underwater, due to the different attenuation degrees of light of different colors in the water body. In general, red light absorbs the most quickly in water, whereas blue light absorbs the least, resulting in underwater images captured mainly appearing blue green. In addition, light will have a scattering effect in water. Forward scattering and reverse scattering are two types of scattering. The deviation of light reflected by objects in the water is referred to as forward scattering due to the influence of scattering particles, resulting in low image clarity; back scattering refers when the light directly into the water, it is affected by the scattered particles, and some of the light will be absorbed by the camera, resulting in low image contrast [1]. Low-quality underwater images obtained in these situations are not conducive to target recognition, detection, and tracking are examples of image processing vision applications. Consequently, underwater image enhancement is extremely essential for underwater research.

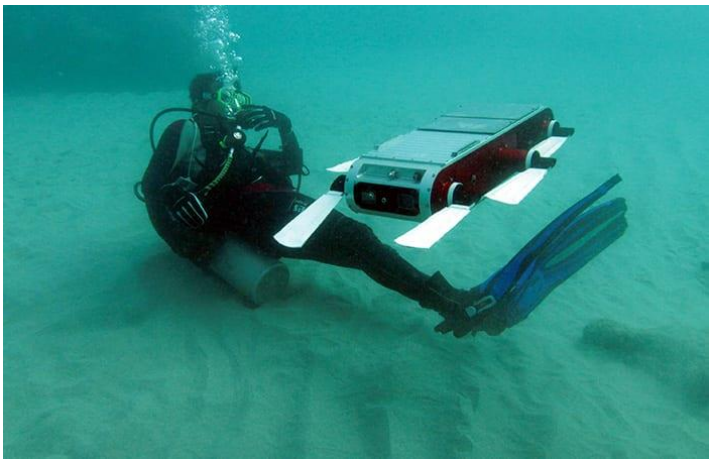


Figure 1.1: The underwater robot[67]

Figure 1.2: Comparison before and after underwater image processing [68]

1.2 Previous Work Background

At present, techniques based on non-physical models, methods based on physical models, and approaches based on deep learning are the three categories of underwater image improvement methods.

1) The method based on the non-physical model realizes the enhancement of the image by adjusting the pixel value of the image. For example, to increase the saturation and contrast of underwater photographs, Iqbal et al. [2] expanded the pixel range in RGB and HSV color spaces.

Problem: The calculation cost of this method is high, and it will cause the problem of over-enhancement or under-enhancement in the local area of the image.

2) The physical model-based method regards underwater image enhancement as the inverse problem of the underwater degradation process, and constructs a degradation model of the underwater degradation process, and restores high-quality underwater images by solving the degradation model.

Problems: However, the large number of physical parameters that this type of method needs to calculate is very complicated, and it also heavily relies on camera parameters and prior conditions, resulting in a particularly limited number and types of underwater images that can be enhanced.

3) Deep learning methods are based on understanding the mapping relationship between underwater images and original images by analyzing each pixel value of the entire image. In particular, the generation adversarial network (GAN)[3], in terms of unsupervised deep learning, can be used to capture the high-level correlation of data without target class label information. Underwater Generative Adversarial Networks were suggested by Fabbri et al. [4]. (UGAN), which used a coding and decoding framework similar to U-Net [5] to better improve the quality of underwater images. Multi-Scale Cycle Generative Adversarial Networks were suggested by Lu et al. [6]. (MCycleGAN), which introduced multi-scale to restore the color of the image. Gao et al. [7] proposed Self-Attention Underwater Image Enhancement by Data Augmentation (SAUIE), which introduced an attention mechanism to better improve the contrast.

Problems: All the features learned in UGAN and MCycleGAN are equally important and do not suppress useless information, such as water; while SAUIE focuses on important feature information, but due to the lack of water quality diversity in the data set, the generalization ability performance needs to be improved.

1.3 Purpose of Study

In order to solve the weak points of above methods, we propose a network based on U-net, which introduces two modules: attention mechanism and multi-scale convolution. Committed to learning not only to restore the overall color, but also to focus on the important part and detailed information in the image. Among them, the attention mechanism module enhances the contrast between the foreground and the background, and can focus on foreground information, such as underwater creatures; The multi-scale convolution module restores the detailed information on the basis of

paying attention to the key target, and makes the target visually clearer. In addition, in order to increase the diversity of underwater images, an underwater image data set generated based on the attenuation coefficients of 10 water qualities proposed by Jerlov [8] is used as the experimental data for this time.

1.4 Chapter Organization

- Chapter 1 introduces the significance of this research and the research background, then discussed previous research and their shortcomings, the research purpose of this thesis, and the organizational structure of the paper.

- Chapter 2 introduces the theory of underwater imaging, including the environment of underwater imaging and the factors that affect image quality, shows the degeneration model formula established for underwater scenes, and then mainly introduces the processing of underwater image enhancement from three kinds of methods, includes methods based on non-physical models, based on physical models and based on deep learning.

- Chapter 3 introduces some kinds of data sets to generate the paired data by using deep learning methods and proposed RGBD data set with rich water quality in this study for the lack of diversity of data sets in deep learning methods. Then explain the structure of the proposed network, it includes attention mechanism, multi-scale convolution and the skip connection, and the used loss functions are also discussed.

- Chapter 4 introduces the experience about this thesis, includes the settings of the network structure, the training loss and other indexes and the results of experience.

- Chapter 5 introduces the evaluation methods for estimating experimental results are introduced, including subjective evaluation methods and objective evaluation methods. Then, compare the results of the method in this paper with sea-thru, which as the representation of the degradation model method, and other deep learning methods. Finally, the results of the comparison are discussed.

- Chapter 6 introduces the results of this experiment and summarizes the problems existing in underwater images and the effectiveness of our method for their problems

Chapter 2

Related Works

2.1 Introduction

This part first discusses the principle of underwater imaging and the effects of light absorption and scattering in water. Then the influence of the attenuation coefficient of different water quality on the wavelength is introduced.

2.2 Principles of Underwater Imaging

Because of the unique water medium and the enormous amount of suspended particles, the light is absorbed and scattered in the water. Therefore, color distortion, poor contrast, and fuzzy images are all present in the underwater image as shown in Figure 2.1.

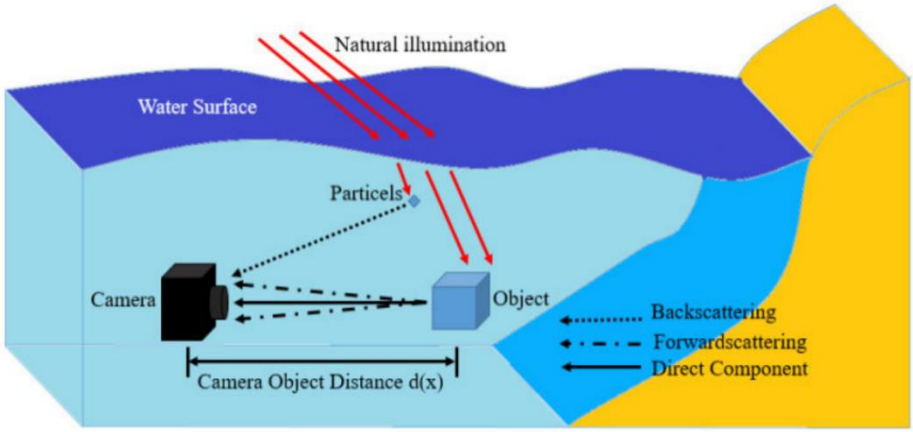


Figure 2.1: General underwater imaging model [9]

The scattering effect of light consists of three main components: forward scattering from the object, direct component from the object and backward scattering from the particle.

2.2.1 Underwater Imaging Environment

In underwater scenes, the water medium has a great influence on the transmission of light. The optical components in water can be mainly divided into water body dissolved matter, gravel, phytoplankton and other particulate matter. The above-mentioned various optical components make the underwater images taken seriously damaged. It includes about three aspects: one is color distortion, mainly because the light of different wavelengths is attenuated to different degrees in the water body. As shown in Figure 2.2, the red light disappears about 5 meters underwater, and the green light and blue light disappear after 30 meters and 60 meters underwater respectively. Therefore, underwater images tend to be blue green. The second is the decrease in contrast, mainly because the back scattered light hits suspended particles in the water and is reflected back to the camera lens. The third is that the image is blurred, producing a "fog"-like effect similar to outdoor weather, mainly because the presence of suspended particles in the underwater scene will cause the deviation of forward scattered light. Figure 2.3 lists common water degradation images, with varying degrees of color cast, low contrast, and blurred.

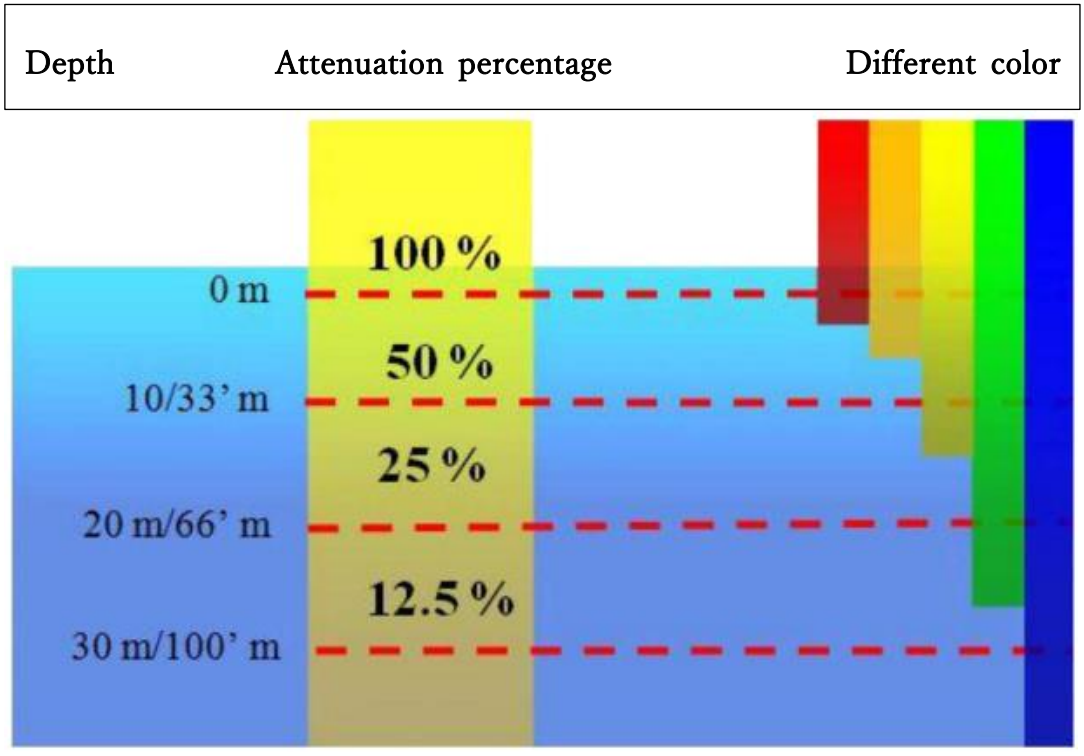


Figure 2.2: The absorption of light of different wavelengths [10]

Figure 2.3: The common raw underwater images picked from UIEB [11]

2.2.2 Attenuation of different wavelengths of light in water

The ocean is easy to pass blue and green light with a wavelength of 0.4～0.5μm, but has a strong absorption effect on purple light with a wavelength of 0.2～0.4μm and red light with a wavelength of 0.5～0.8μm, that is, red and violet light attenuate the most. Therefore, the ocean is often blue-green, and the attenuation coefficients with different wavelengths of light shown as below:

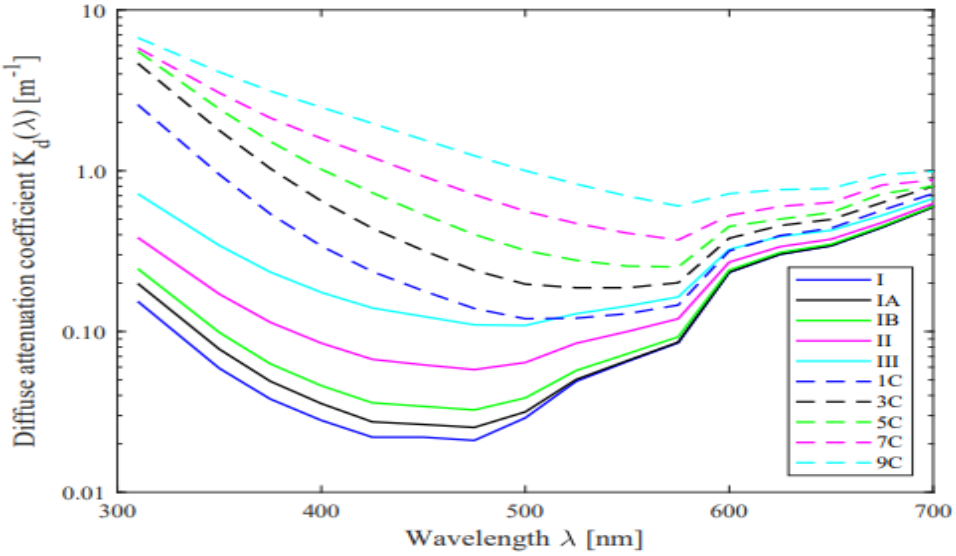


Figure 2.4: Different attenuation coefficients of different wavelengths in underwater [12]

The figure below shows the different wavelengths of visible light in Figure 2.5:

Color	Wavelength (nm)	Frequency (THz)	Photon energy (eV)
violet	380–450	670–790	2.75–3.26
blue	450–485	620–670	2.56–2.75
cyan	485–500	600–620	2.48–2.56
green	500–565	530–600	2.19–2.48
yellow	565–590	510–530	2.10–2.19
orange	590–625	480–510	1.98–2.10
red	625–750	400–480	1.65–1.98

Figure 2.5 Different wavelengths of Visible light [13]

2.3 Previous Work

In this part, Firstly, the traditional degradation model formula and its optimization formula proposed by previous studies based on the above environment are discussed. Secondly, introduce researches related with hardware methods and software methods of underwater image enhancement. Then, hardware methods to upgrade hardware platforms are discussed. Finally, software methods to improve software algorithms with three parts are reported later.

Table 2.1: Variables used in the part

Variable	Description
λ	wavelength
$E(z,\lambda)$	irradiance
$a(\lambda)$	coefficient of absorption of a beam
$b(\lambda)$	coefficient of scattering of a beam
$\beta(\lambda)$	beam attenuation coefficient:$a(\lambda)+b(\lambda)$
$S_c(\lambda)$	sensor spectral response
$\rho(\lambda)$	reflectance spectrum of the object
c	color channels R, G, B
β_c	attenuation coefficient
$B^{\infty}(\lambda)$	veiling light
B_c^{∞}	veiling light with a wide band
$\mathbf{I}_c$	Signal attenuation in RGB image
$\mathbf{J}_c$	Signal no-attenuation in RGB image
d	depth
z	geometric distance along line of sight
ξ	direction among the scene to camera
B	backscattered light
D	direct transmitted light

2.3.1 Underwater Imaging Model

At present, In the subject of underwater image enhancement, the most extensively used model is the Akkaynak model [14], the underwater image formation usually is controlled by:

$$L = D + B \qquad (2.1)$$

Where L is the total quantity of radiation that reaches the camera from the scene, B is the backscattered light, and D is the direct transmitted light, $\mathbf{I}_c$ is the RGB image with attenuated signal in the L situation, specifically expressed as below:

$$\mathbf{I}_c = \mathbf{J}_c t_c + B_c^{\infty}(1 - t_c) \qquad (2.2)$$

where $\mathbf{I}_c$ refers to the underwater image taken by the camera lens, $\mathbf{J}_c$ represents the desired enhanced high-quality image, t_c is the transmittance, which means the residual energy ratio of the scene radiation after the image is degraded, and B_c^{∞} is the estimation of the background light. c is the RGB color space channels. $\mathbf{J}_c t_c$ is the light from the surface of the observed object entering the camera lens, $B_c^{\infty}(1 - t_c)$ is the underwater ambient light reaching the camera lens. t_c can be further expressed as the following exponential decay term:

$$t_c = e^{-\beta_c d} \qquad (2.3)$$

In contrast to the image degradation model, d is the depth, β_c is wavelength dependent and is influenced by seasonal, geographic, and climatic fluctuations, causing scenes to seem blue, green, or yellow.

The Akkaynak model [14] is similar to the image defogging model, so the underwater image can be enhanced by image defogging. However, compared with the image defogging assuming that the RGB channels have the same degree of attenuation and uniform atmospheric illumination, the underwater RGB channels have different degrees of attenuation, and the underwater environment is mostly non-uniform illumination. Therefore, directly applying the image defogging algorithm to underwater does not produce satisfactory enhancement results.

We can see that B_c^{∞} and t_c are important prior information for underwater image restoration. A large number of researchers have improved and designed models to better estimate the above two model parameters to obtain enhanced results with better subjective

and objective effects.

2.3.2 Improved Underwater Imaging Model

Recent research has discovered that the widely used Akkaynak underwater imaging model misses certain essential aspects of the real-world underwater imaging process. The backscatter attenuation coefficient is influenced by the veiling light. Furthermore, unlike absorption in the outdoor defogging model, water absorption should not be overlooked. The most essential distinction is that the direct and dispersed signals have different attenuation coefficients. As a result, Akkaynak et al. [14] employed oceanographic measuring methods to determine the physical effective space of backscattered signals, as well as proving that backscattering coefficients differ from those of direct transmission. Finally, in version 2.3, an improved underwater imaging model was suggested.

$$\mathbf{I}_c = \mathbf{J}_c e^{-\beta_c^D(v_D)\cdot z} + B_c^{\infty}(1 - e^{-\beta_c^B(v_B)\cdot z}) \qquad (2.4)$$

Where the camera and the objects are separated by z range (distance) B_c^{∞} is the wideband veiling light, when obtaining the value of backscatter at infinity, also known as veiling light, is possible if the value of z is chosen to be sufficiently great. Thus as $z \to \infty$. The wideband attenuation coefficient β_c, the direct transmitted light D, and the backscattered light B are all used in this equation. The $\mathbf{v}_D$ and $\mathbf{v}_B$ vectors reflect coefficient dependency are defined as:

$$\mathbf{v}_D = \{z, \rho, E, S_c, \beta\} \qquad (2.5)$$

$$\mathbf{v}_B = \{E, S_c, b, \beta\} \qquad (2.6)$$

Where ρ represents the reflectance, E represents the illuminance, S_c represents the spectral response of the sensor, b represents the beam scattering coefficient. $\mathbf{J}_c$ is the clear restored image, $\mathbf{I}_c$ is the observed degraded image [14]. Furthermore, backscattering characteristics change depending on the kind of sensor, ambient light, and water quality. In general, the backscatter coefficient is not the same as the direct signal coefficient.

$$\mathbf{I}_c = \frac{1}{k}\int_{\lambda_1}^{\lambda_2} S_c(\lambda)\,\rho(\lambda)E(d,\lambda)e^{-\beta(\lambda)z}d\lambda + \frac{1}{k}\int_{\lambda_1}^{\lambda_2} S_c(\lambda)\,B^{\infty}(\lambda)(1 - e^{-\beta(\lambda)z})d\lambda \quad (2.7)$$

Where $\rho(\lambda)$ represents the of the object's reflectance spectrum, k is a scalar that

determines image exposure and camera pixel geometry [15], and λ_1 and λ_2 are the electromagnetic spectrum's integration bounds.
At depth d, the unattenuated image $\mathbf{J}_c$ is:

$$\mathbf{J}_c = \frac{1}{k}\int_{\lambda_1}^{\lambda_2} S_c(\lambda)\rho(\lambda)E(d,\lambda)d\lambda \qquad (2.8)$$

The veiling light B_c^{∞} , as acquired by the same sensor, is as follows:

$$B_c^{\infty} = \frac{1}{k}\int_{\lambda_1}^{\lambda_2} S_c(\lambda)\frac{b_c E_c}{\beta_c}d\lambda \qquad (2.9)$$

And β_c^D has been derived from the direct transmission (D) term, β_c^B has been derived from the direct transmission (B) term, The following are the equations connecting RGB coefficients β_c^D and β_c^B to wavelength-dependent physical quantities [14]:

$$\beta_c^D = ln\left[\frac{\int_{\lambda_1}^{\lambda_2} S_c(\lambda)\rho(\lambda)E(d,\lambda)e^{-\beta(\lambda)z}d\lambda}{\int_{\lambda_1}^{\lambda_2} S_c(\lambda)\rho(\lambda)E(d,\lambda)e^{-\beta(\lambda)(z)}d\lambda}\right]/z \qquad (2.10)$$

$$\beta_c^B = -ln\left[1\text{-}\frac{\int_{\lambda_1}^{\lambda_2} S_c(\lambda)B^{\infty}(\lambda)(1\text{-}e^{-\beta(\lambda)z})d\lambda}{\int_{\lambda_1}^{\lambda_2} B^{\infty}(\lambda)S_c(\lambda)d\lambda}\right]/z \qquad (2.11)$$

Where λ_1 and λ_2 are the visible light range's limits (400 and 700nm), respectively, while the spectrum of ambient light at depth d is denoted by E.
If $E(0,\lambda)$ is light at the sea surface, then $E(d,\lambda)$ at depth d is [16]:

$$E(d,\lambda) = E(0,\lambda)e^{-K_d(\lambda)d} \qquad (2.12)$$

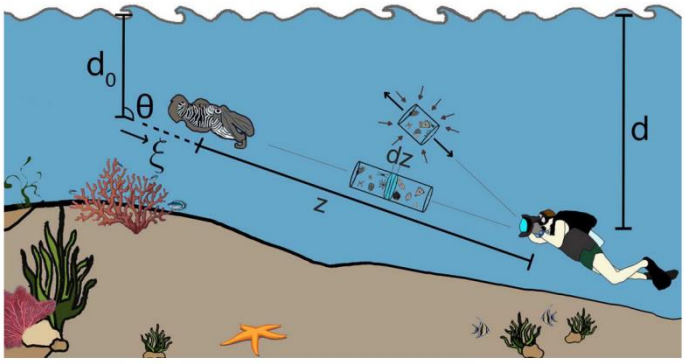


Figure 2.6: Ambient light attenuation under different distance parameters [14]

Where during the time it takes for ambient light to reach an object, its intensity diminishes exponentially with depth d, and its intensity diminishes exponentially between the object and the sensor in the direction ξ, over a distance of z.

However, the Akkaynak model cannot be applied to certain underwater situations, such as shallow waters with low backscatter. Therefore, Akkaynak et al. [14] proposed a physically accurate degradation model to further improve the underwater image imaging model. However, due to the complexity of its models and parameters, this improved model has hardly attracted much attention.

In order to solve above weak points of [14], sea-thru [17] is proposed to estimate the complex parameters and dependencies by some measures:

1) Previously, it was considered that the coefficients $\beta_c^D = \beta_c^B$ were equal and that they had a single value for a particular scene [18] but has been showed in [14] by sea-thru that they are different, and that they rely on diverse parameters.

2) To retrieve $\mathbf{J}_c$, the optical water type indicated by b and c; light E; the distance between the camera and the scene z; the shooting depth d; the reflectivity of each item in the scene; and the camera's spectral response S_c must all be known or estimated. However, these factors are seldom understood before to shooting. It is well established that z in [14] has the greatest effect on β_c^D, whereas optical water type and illumination E have the most effect on β_c^B. Therefore, the range map of the scene is generated for the purpose of calculating β_c^D utilizing the structure self-motion SFM. Where z requires an absolute value, but SFM gives merely a range of scaling, so objects of known size are put in the image.

3) It is considered that the coefficients cannot generally be communicated between images [14], and only the relevant parameters of a specific image are calculated from this image.

4) Divide the range map into ten evenly spaced clusters encompassing the lowest and largest values of z, before attempting to estimate backscatter. In $\mathbf{I}_c$, we look for the RGB triplet with the lowest one-hundredth percentile value, which we indicate with a. Therefore,$\widehat{\boldsymbol{B}}_c(\Omega) \approx \mathbf{I}_c(\Omega)$ is an estimation of backscatter, which the new model as:

$$\widehat{\boldsymbol{B}}_c = B_c^{\infty}(1 - e^{-\beta_c^B z}) + \mathbf{J}_c^{'} e^{-\beta_c^{D^{'}} z} \qquad (2.13)$$

where $\mathbf{J}_c^{'} e^{-\beta_c^{D^{'}} z}$ represents a residual term like the direct signal. And estimated the ranges of parameters B_c^{∞}, β_c^B, $\mathbf{J}_c^{'}$ and $\beta_c^{D^{'}}$, but the boundaries for β_c^D and β_c^B won't

adjusted further using the loci mentioned in [14] if information about the camera sensor, water type, and so on is not available. Besides, if some complex physical parameters are expected to be estimated, certain prior knowledge is required, and under certain conditions, assumptions need to be added. Therefore, the requirements for achieving image enhancement tasks through such complex degradation formulas will be very demanding.

2.3.3 Hardware Methods and Software Methods

Current underwater image enhancement methods include hardware methods to upgrade hardware platforms and software methods to improve software algorithms.The hardware method is mainly to improve the visibility of underwater images by upgrading optical imaging equipment and designing specific hardware platforms and cameras. Under dynamic, natural lighting, and murky circumstances, Roser et al. [19] suggested a platform for simultaneous underwater image quality evaluation, visibility augmentation, and parallax computation to improve stereo range resolution. Schechner et al. [20] proposed an adaptive filtering method based on polarization imaging is used to filter through the polarization filter in front of the camera, which not only significantly improves the visibility of the original image, but also filters noise. However, the cost of building a hardware platform is high, and different deployment environments have a great impact on the effect, so the application scenarios are limited. Therefore, people often improve software algorithms. Physical methods, degradation methods, and deep learning methods are the most common types of software methods.

2.3.4 Physical Methods

The underwater image enhancement approach based on the non-physical model eliminates complicated physical factors and directly modifies the image pixel value to correct the color shift of the image to increase the image contrast.

Researchers attempted to directly apply typical image enhancing algorithms to underwater photos in the early stages of underwater image research. Such methods include histogram equalization method [21-23] and its derived white balance method [22]. The histogram equalization approach distributes image pixel intensity uniformly, which can increase image quality to a degree, but it ignores the image's global structural

information, resulting in artifacts in the upgraded image. Equalization of histograms with adaptive histograms divides the image into several local blocks, calculates the histogram of each local block, and then redistributes the brightness of each area. This local processing method removes artifacts. However, the algorithm also has drawbacks, it makes the local noise of the image enhanced. In response to this problem, a contrast-limited adaptive histogram equalization [23] technique that limits the contrast of each local block while also speeding up the computation through interpolation might successfully restrict this type of adverse enhancement. Liu [22] et al. proposed an automatic white balance algorithm that uses fuzzy logic rules to determine color parameters, thereby minimizing the color temperature difference of various light sources.

With the development of underwater image enhancement technology, researchers also improved image enhancement methods based on the characteristics of underwater images. Ancuti et al. [24] used image fusion to enhance underwater image quality. They defined two inputs for image fusion, one is a color-corrected image, and the other is a contrast-enhanced image. In addition, four weight maps are defined for Laplacian fusion to determine which pixel is more suitable to appear in the restored output. In the research of Zhuang et al. [25], an algorithm combining Retinex and edge preservation filtering was developed. First, they still use the Retinex method to generate the reflected light image and the incident light image, then they use guided filtering to refine the edge features of the two images so as to obtain a higher-quality incident light image. These methods based on non-physical models have improved the contrast and clarity of underwater scenes to a certain extent, but the output images in some scenes may be over-enhanced or under-enhanced.

2.3.5 Degradation Methods

The physical model's underwater image improvement technique treats underwater image enhancement as an inverse issue, in which the image creation model's potential parameters are inferred from a damaged image. The procedure for most of these approaches is the same: To recover a clean underwater scene, first create a deteriorated physical model, then estimate the physical model parameters, and then solve the inverse issue.

The Akkaynak et al. [14] underwater imaging model describes the process of underwater image distortion. The image is restored by calculating model parameters, including underwater light, transmittance and other parameters, then inverting the degradation process. Complex underwater images are similar to foggy images (such as

backscattering) to a certain extent, so some researchers apply image defogging methods to underwater image enhancement. The dark channel priori method in image defogging assumes that in most partially fog-free outdoor images, at least one-color channel has some pixels with very low brightness. By using this a priori assumption to estimate the transmittance map and restore the foggy image [26]. Underwater Dark Channel Prior (UDCP) [27] based on the observation of the absorption rate of the red channel in a large number of underwater images and proposed a priori knowledge suitable for underwater to correct the color shift. However, the underwater dark channel priori algorithm is very sensitive to changes in the underwater scene, so its application is limited. Similarly, the red channel method [28] is another variant of the dark channel underwater, which corrects for short-wavelength-related colors. Due to the lack of abundant model training data, these methods based on dark channel priors perform poorly in ocean scenes.

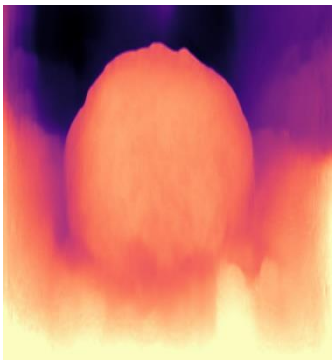
In 2017, Wang et al. [29] used a method of maximum attenuation recognition to defog and correct the color of underwater images. It is assumed that the attenuation effect composed of absorption and scattering is closely related to the depth of the image, so the depth map is first obtained according to the attenuation, and then the attenuation model is expressed as a simplified form of the underwater light transmission model to restore the underwater image. Then the transmittance will be estimated: the initial relative transmittance of the red channel is estimated and optimized, and then the attenuation factor is calculated, and the transmittance of the three-color channels is adjusted through saturation constraints. Finally, the image can be restored by underwater light and transmittance.

Most of the above prior-based methods can only be used in specific scenes. Once the prior knowledge is not satisfied, the restoration effect of underwater images will be unsatisfactory.

In 2019, Derya, Tali[17] proposed a method base on degradation model which can move the water from underwater images. Through certain prior knowledge and assumptions, it is proved that β_c^D and β_c^B are different, and the z dependence of β_c^D is crucial, and the visual effect of "remove water" is achieved. However, the result is affected by the distance between the scene and the camera. In addition to the prior knowledge and the complex parameters of the degradation formula that must be known, the two kinds of camera parameters used are also as important factors as raw image information, and p-dependent has not been resolved yet. Therefore, it is difficult to obtain original information in such a complex environment.

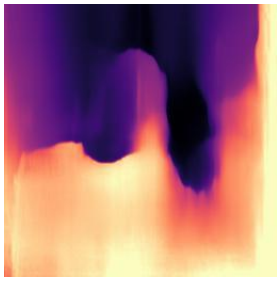
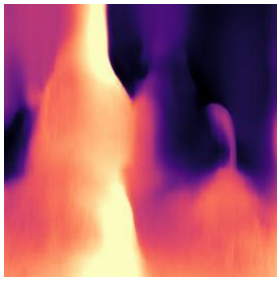
We experimented with Sea-thru method and obtained good recovery results from their original data as follows:

Table 2.2: Experiment with Sea-thru

Input image	Original Depth map	The result
		
Input image	Monodepth map	The result
		

But when without original depth map or using other datasets to verify the Sea-thru method, the effect is not very satisfactory, and the correct depth map information is lacking, so the result can only be obtained by generating the depth map through the mono-depth method provided by the Sea-thru paper. It shows that the original depth map is necessary and important for degradation methods.

Table 2.3: Experiment with Sea-thru on SUID data set

Input image	Depth map	The result	Original image (Ground truth)
			
			

In summary, due to the degradation model method requires a lot of original physical parameters and camera parameters and needs to ensure that an original depth map is obtained. So, the performance is not very good for other datasets without depth maps. The deep learning method aims to restore the image by learning the mapping relationship between the degraded image and the original image, and has a certain generalization ability, which can adapt to more data sets, which is more ideal than the traditional method.

2.3.6 Deep Learning Methods

Researchers have begun to use deep learning to underwater image restoration problems in recent years, after its success in advanced computer vision tasks, natural language processing, and other domains [30]. Ding [31] et al. enhanced the quality of the original image using an improved white balance technique, then utilized a convolutional neural network to estimate the BL and transmission map, and lastly completed image restoration using the underwater image's optical imaging model. The method lowers the impact of the complicated underwater environment on image quality and enhances it through certain preprocessing techniques, although it is prone to over saturation. Water-Net [32]

corrects the original image with white balance, histogram equalization, and gamma correction, which improves the image's contrast while also addressing color shift and uneven illumination issues. After that, a deep learning network is built to predict the preprocessed image and the confidence map of the original image, and the enhanced image is obtained. This approach produced a image that satisfied human visual perception. However, the procedure is time-consuming, and various datasets have a stronger influence on the model's training.

Deep learning's tremendous data processing skills have also been used to various image processing jobs in sophisticated computer vision applications. The popularity of GAN [33] has resulted in widespread support for image processing. Goodfellow et al. [3] established a unique approach for estimating the generative model using the adversarial process (Generative Adversarial Network, GAN). GAN has two pieces to its network structure: a generator network G that learns the data distribution and a probability discriminator network D that estimates the sample from the training data. G and D are, in general, playing a mutual game. G tries to deceive D by mixing the spurious with the genuine data created by the generator, while D tries to tell the difference between the phony data made by the generator and the actual data. To achieve the aim of being able to detect data that has been tampered with. GAN's architecture is as follows:

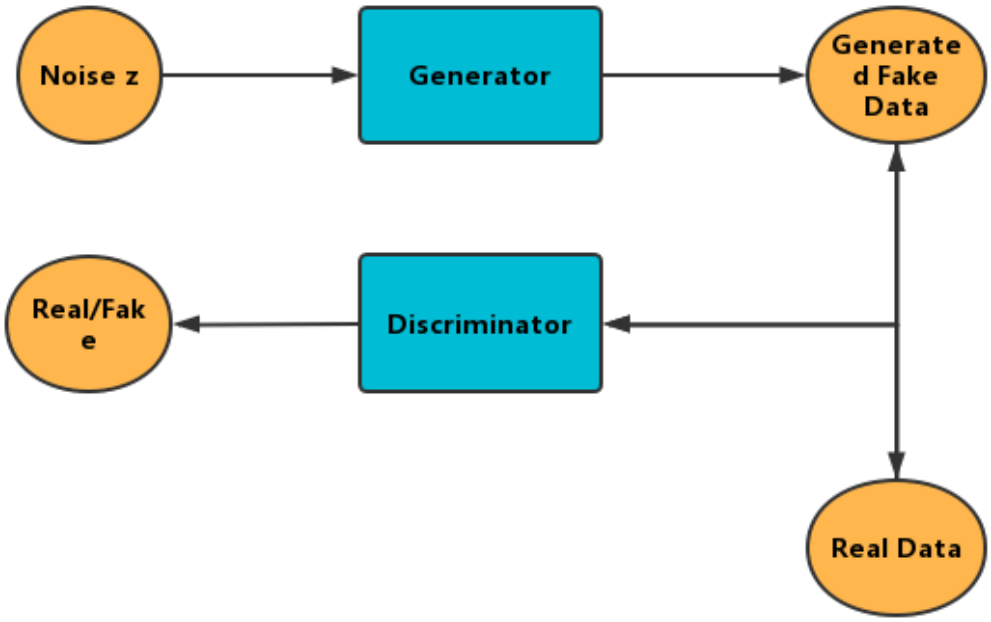


Figure 2.7: Architecture of GANs

The final result obtained by the discriminator D through a 2-classification function is used as the judgment basis. The generator G expects that the result will be close to 1, and the discriminator D hopes that the result will be close to 0. Through continuous training, the Nash balance is finally reached, that is, the generation ability of G is comparable to the discrimination ability of D, making the result probability approximately 0.5. The following is the loss function of the GAN network:

$$\min_{G}\max_{D} V(D,G) = E_{x\sim p_{(x)}}[log\, D\,(x)] + E_{z\sim p_{(z)}}[log(\,1 - D(G(z)))] \qquad (2.14)$$

Where $p(z)$ represents the noise variable of the input G, $p_{(x)}$ is the original data that needs to be fitted, $G(z)$ is the output of the generator, $D(x)$ is the probability of discriminating that x is the original data, $D(G(z))$ is the probability of D to discriminate $G(z)$ as original data. The original GAN has two loss functions: Minimax and Non-saturating. Goodfellow et al. [3] proved through theoretical analysis and experiments that the minimum-maximum form has better performance than the unsaturated form.

Arjovsky et al. [34] proposed in 2017 to utilize the Wasserstein distance to assess the difference between the produced distribution and the true distribution in order to tackle the collapse mode problem of the original GAN, such as the lack of variety in the generated samples (Wasserstein GAN, WGAN) WGAN It potentially overcomes the flaws in the original GAN while also introducing a new difficulty, the Lipschitz restriction. Weight clipping is used directly by WGAN to solve the Lipschitz constraint, however it can create low-quality samples and cause the loss function to fail to converge. Gulrajani et al. [35] presented a different weight clipping method (WGAN with Gradient Penalty, WGAN-GP). WGAN-GP outperforms WGAN and can virtually completely tune the training of a variety of GAN architectures [36][37]. Miyato et al. [38] presented a lightweight weight normalization approach called Spectral Normalization to stabilize the discriminator's training, as opposed to the gradient penalty, which requires more computer resources and time. There are many additional GAN variations that are not Lipschitz restrictions. For example, instead of judging "is an image more actual than another," the relative discriminator (The Relativistic Discriminator) [39] learns to assess "is an image more real than another."

Due to the complexity of the actual underwater situation, the traditional enhancement and restoration methods cannot show good generalization. Deep learning [40] relies on a large amount of data to learn the relationship between samples and has achieved good generalization and robustness.

Deep learning has shown promise in low-level tasks like image super-resolution

[41][42], image rain removal [43][44], and high-level computer vision tasks like target identification [39] and target segmentation [45] during the last 10 years. Therefore, more researchers apply CNN and GAN to underwater image enhancement.

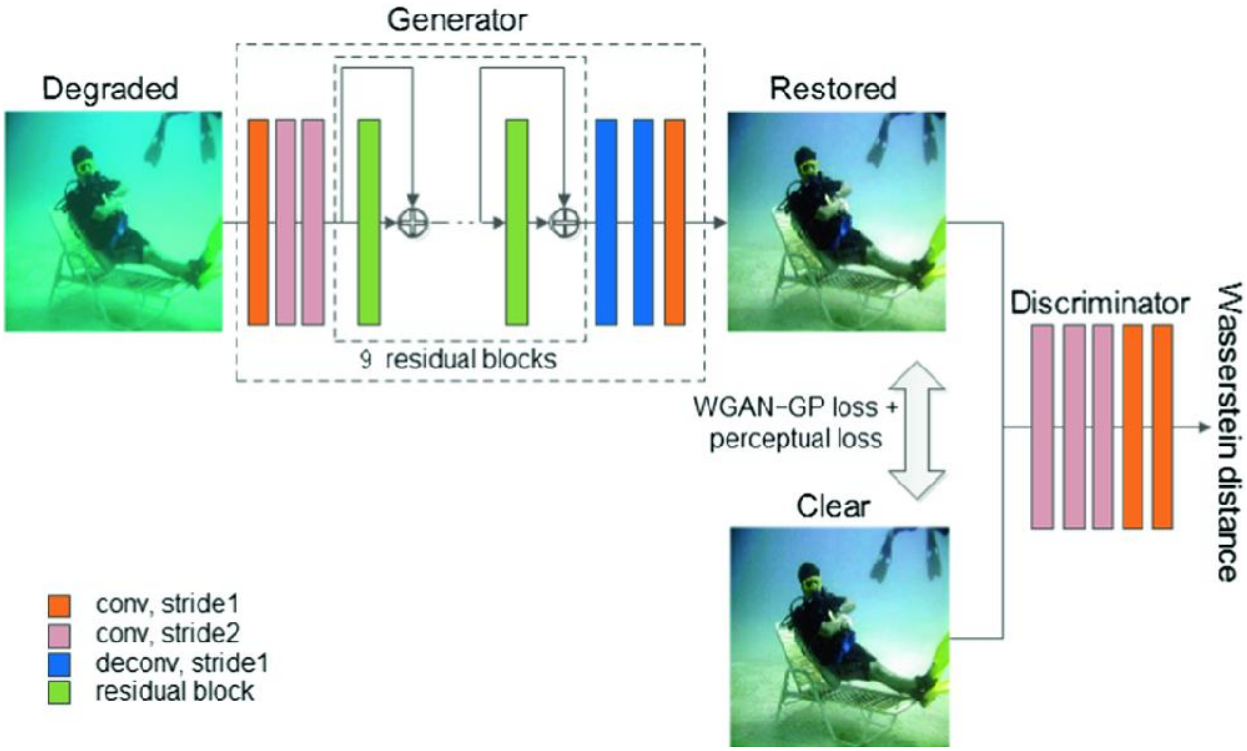


Figure 2.8: Underwater GAN underwater image enhancement model [69]

In WaterGAN [46], different underwater image datasets are synthesized using the attenuation model of atmospheric image and underwater image. Build a network model based on the synthetic datasets to perform color correction on the distorted image. Although this approach can produce relative results, it requires the training of many models for various undersea kinds and is unable to cope with the variable aquatic environment. Moreover, the data set obtained by the synthetic method cannot replace the original underwater image and only images with similar features have a better recovery effect. A poorly supervised learning model was suggested by Zhu et al. [47]. (CycleGAN), which eliminates the limitation of paired data sets during network model training. The network uses two image domains of clear images and low-quality images for style conversions. The Underwater Generative Adversarial Network was suggested by Fabbri et al. [4]. (UGAN), which uses an encoding and decoding framework similar to U-Net [5] to better improve the quality of underwater images. However, this type of method pays attention to the overall image characteristics, and does not suppress useless information,

for example, image information occupies a lot of water. The Multi-scale Cyclic Generative Adversarial Network was proposed by Lu et al. [6]. (MCycleGAN), which introduced multi-scale to restore the color of the image and better highlight the image details. However, this method does not pay more attention to the important part and allocates computing resources with more useless information. Gao et al. [7] proposed Self-Attention Underwater Image Enhancement by Data Augmentation (SAUIE), which introduced an attention mechanism to better improve contrast. But the generated image is not clear enough, and the detailed information is fuzzy. Moreover, the data set lacks water quality diversity, and the performance of non-generalization needs to be further improved.

In summary,

For traditional methods, the established degradation model can achieve the purpose of underwater image enhancement, but it usually requires calculation of a lot of complex parameters and requires the original camera parameters and certain prior knowledge, can only perform well in a part of the data.

For deep learning methods, the goal of underwater image improvement is to discover the mapping relationship between degraded and clean images. but it often lacks original images corresponding to underwater scenes, and the network structure also requires careful design.

This paper proposes a U-net-based network that introduces two modules: attention mechanism and multi-scale convolution. Committed to not only restoring the overall color, but also paying attention to the important parts and details of the image. Among them, the attention mechanism module enhances the contrast between the foreground and the background, and can focus on foreground information, such as underwater creatures; the multi-scale convolution module restores detailed information on the basis of focusing on key targets, making the target visually clearer. In addition, in order to ensure that the ground truth is original image, the indoor open-source data set is used as the ground truth, and the corresponding underwater image is generated by a degradation model. At the same time, in order to increase the diversity of underwater images, this experimental data uses the underwater image data set generated by 10 water quality attenuation coefficients proposed by Jerlov [12].

Chapter 3

Data set and Model

3.1 Introduction

This part first introduces the common methods of obtaining paired data sets in deep learning methods and analyzes why experiment data set of this thesis is used. Then introduces the proposed attention mechanism module and multi scale module. Finally, the proposed network structure based on above two modules is introduced.

3.2 Acquisition of paired datasets

First, the deep learning methods require paired datasets to learn the mapping relationship to better restore underwater images. And the paired data set includes two parts: one part is a degraded image taken in an underwater environment, and the other part is a clear and no-degraded image on the ground corresponding to the underwater scene. However, in reality, it is impossible to find a corresponding ground data set with underwater scenes.

As a result, various approaches must be used to construct paired data sets. There are two kinds of approaches used to produce underwater image data sets or data sets on the land that match to underwater scene.

3.2.1 Method of generating underwater image data

This method first ensures that the data set on the ground, that means ground truth is the original data. The data set on the ground uses an open-source outdoor data set or an indoor data set, and the corresponding underwater image is generated by image processing of the ground data or through a degradation model formula to restore underwater degradation factors, including blue and green light attenuation, low contrast and blurring and etc. The generated underwater image data and open-source ground data are used as the paired data set for the training set of the next image enhancement task. Paired open-source data sets generated based on this method such as SUID [48], etc.

The SUID data set adopts ground truth is a original image on the land, and the original image is simulated by an algorithm. Shown as follows:

(a)Building

(b)Mountain

(c)Soccer

(d)Statue

Figure 3.1: Some original images (ground truth) of SUID [48]

(a)Building-Bluish

(b)Mountain-Bluish

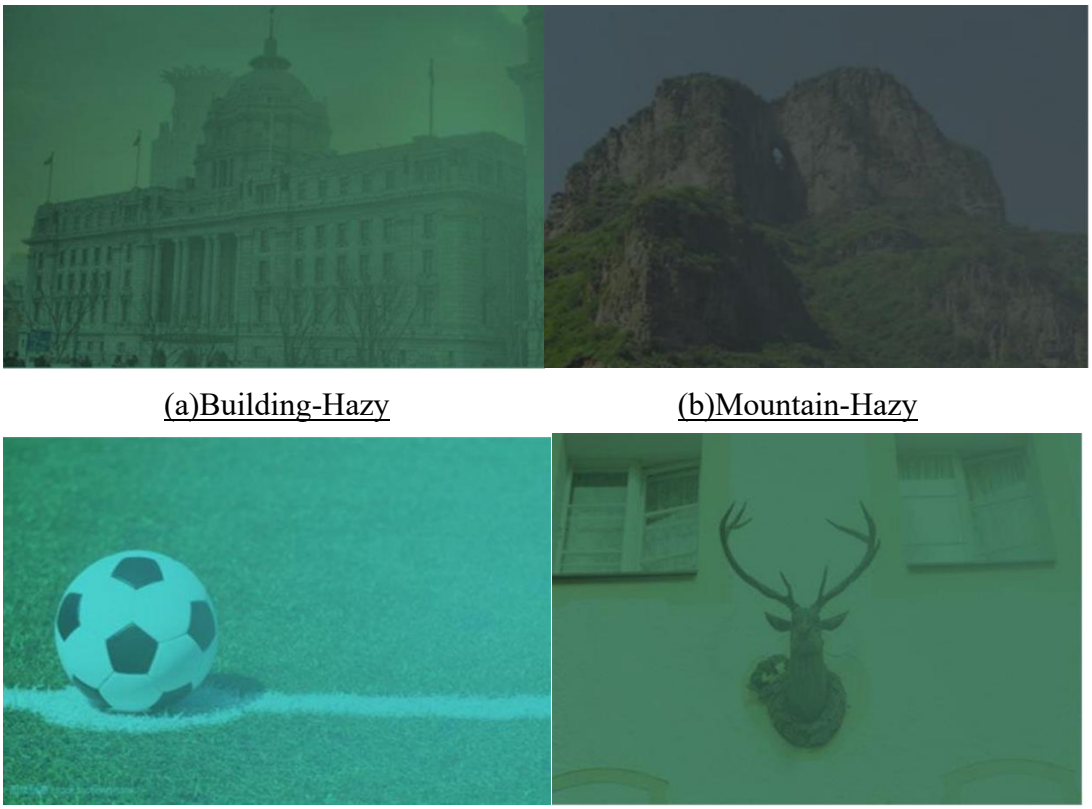
(c)Soccer-Bluish (d)Statue-Bluish

Figure 3.2: Synthesis underwater images with bluish of SUID [48]

(a)Building-Greenish (b)Mountain-Greenish

(c)Soccer-Greenish (d)Statue-Greenish

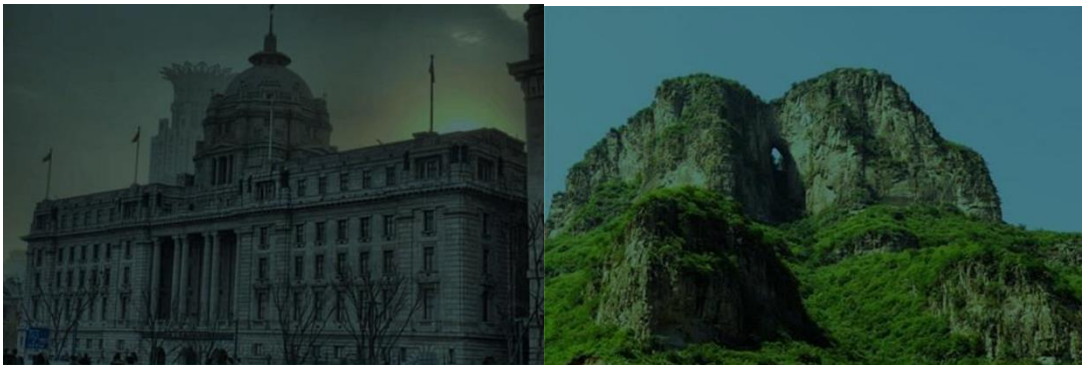
Figure 3.3: Synthesis underwater images with greenish of SUID [48]



(a)Building-Hazy (b)Mountain-Hazy

(c)Soccer-Hazy (d)Statue-Hazy

Figure 3.4: Synthesis underwater images with hazy of SUID [48]



(a)Building-Low light (b)Mountain-Low light

(c)Soccer-Low light (d)Statue-Low light

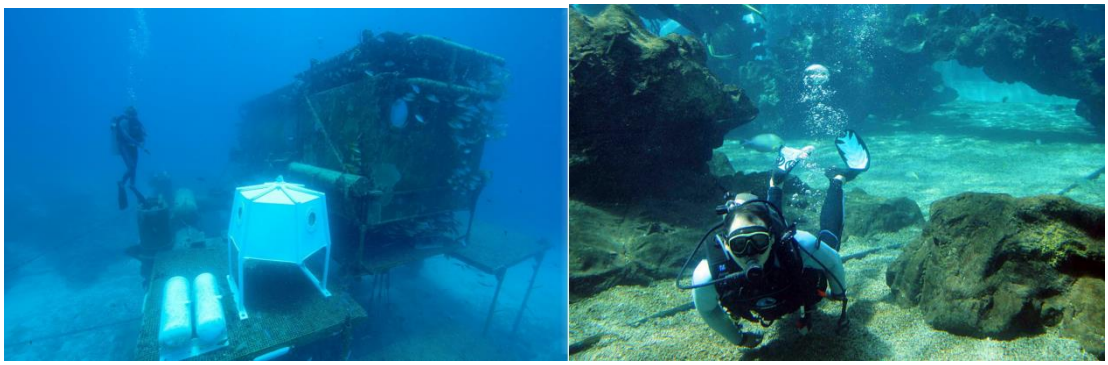
Figure 3.5: Synthesis underwater images with low light of SUID [48]

3.2.2 Method of generating ground data

This method firstly guarantees that the underwater image is actually captured. So far, some papers have provided original data taken in the water by professional high-definition cameras and other advanced equipment, which truly restores the underwater image. The corresponding ground scene data is the restored image obtained by some better performance degradation model methods as ground truth, or the fake image generated by the deep learning method as ground truth. That is, the results obtained by other methods are used as the corresponding underwater ground data, and then combined with the acquired original underwater images to form a paired data set, based on this method to generate a paired open-source data set such as: UIEB, etc.

The UIEB [49] data set adopts: the raw image is the original underwater image, and the ground truth is selected by a variety of traditional methods. Shown as follows:

(a)Turtle (b) Fish & Rock



(c)Container (d)Person

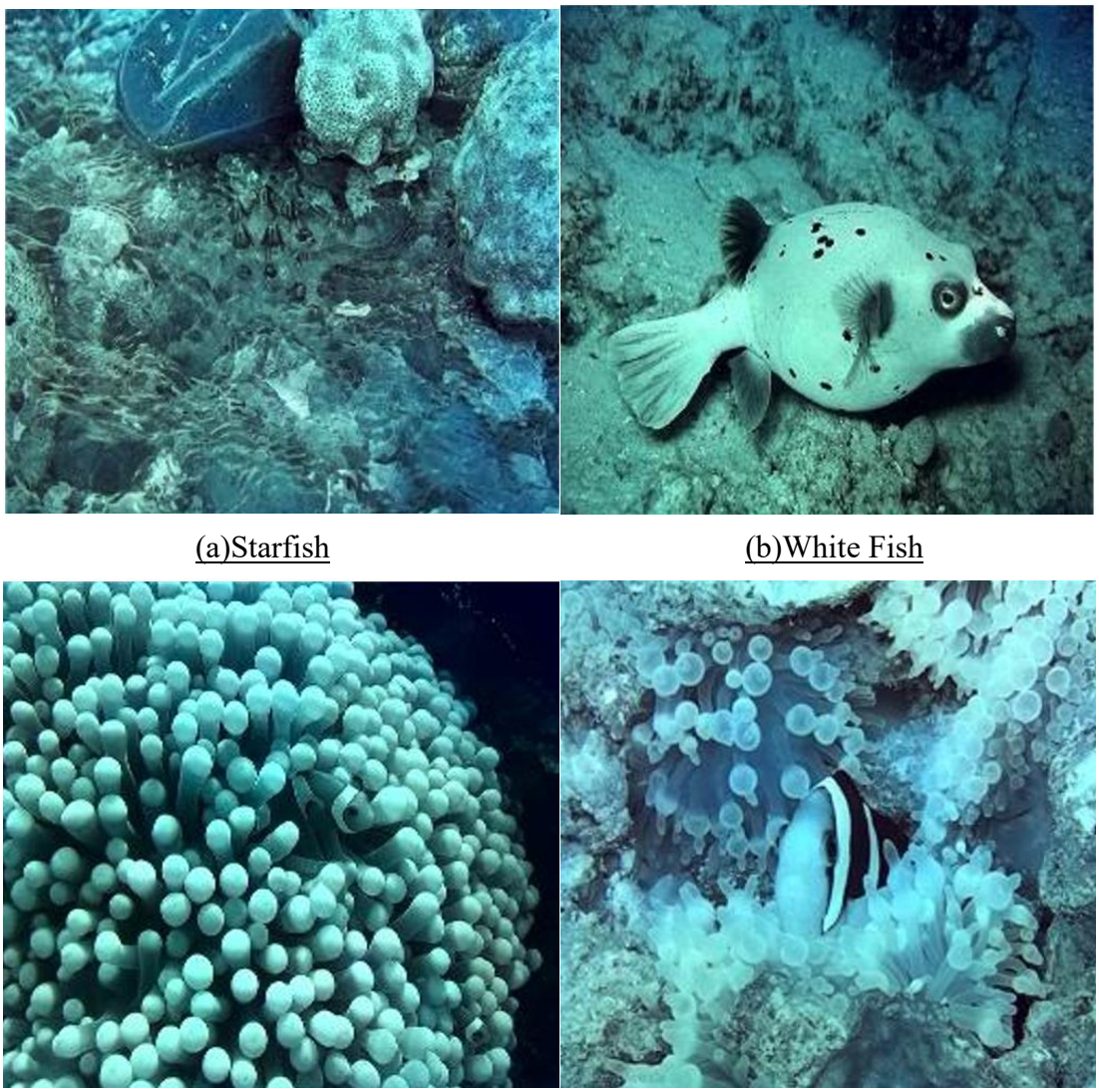
Figure 3.6: Captured original underwater images of UIEB [49]

(a)Turtle-Ground truth (b) Fish & Rock-Ground truth

(c)Container-Ground truth (d)Person-Ground truth

Figure 3.7: Corresponding ground truth by tradition methods of UIEB [49]

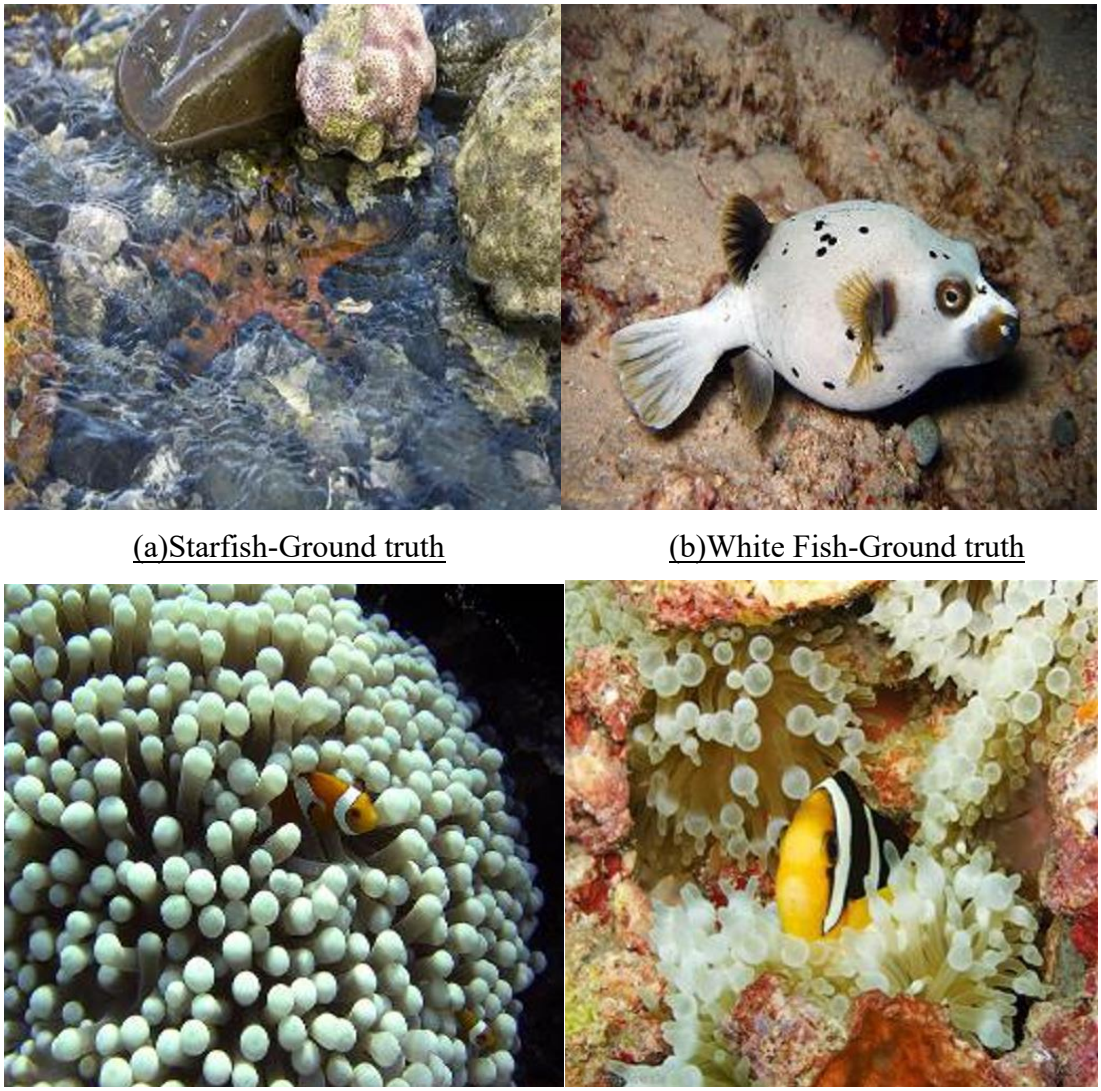
The EUVP [50] data set adopts: Underwater raw image uses seven different camera equipment to capture underwater images, ground truth is generated by the trained Cycle-GAN. Shown as follows:



(a)Starfish (b)White Fish

(c) Coral (d)Orange-Black Fish

Figure 3.8: Captured original underwater images of EUVP [50]



(a)Starfish-Ground truth (b)White Fish-Ground truth

(c) Coral-Ground truth (d)Orange-Black Fish-Ground truth

Figure 3.9: Corresponding ground truth by Cycle-GAN of EUVP [50]

3.2.3 Selection of two kinds of methods

By enumerating the operation process of the two methods, the first method is more suitable for the deep learning method. Since deep learning methods are used, one very important factor is ground truth. It will be very convincing to ensure that ground truth is original data, and the effect will be better for training. On the contrary, if ground truth uses the results of other people's methods as reference data to train the network, it is very dangerous and unconvincing. And to a certain extent, the generated results often cannot

exceed ground truth, that is, if the second such method is used, the generated results cannot be better than other methods, and the research task loses its meaning.

3.2.4 RGB-D data set

Depth image = ordinary RGB three-channel color image + Depth Map

Depth Map is an image or image channel in 3D computer graphics that carries information about the distance to the surface of the scene object of the perspective. Depth Map is one of them, and it looks like a grayscale image with the exception that each pixel value represents the actual distance between the sensor and the object. Typically, the RGB image and the Depth image are registered, resulting in one-to-one pixel correlation.

(a)Indoor original image (b)Corresponding depth map

Figure 3.10: Original image and corresponding depth map [51]

This Figure3.10 intuitively shows that the depth map reflects the distance between the scene and the object in the original image from the camera. The depth map shows brightness proportional to the distance from the camera. The closer the color is darker, the farther the color is lighter.

Therefore, RGBD is equipped with depth information, which refers to the distance between the target scene and the camera. Not only can more detailed information about the original scene be obtained through the use of this important parameter, but the low contrast of the underwater image can also be improved by knowing the depth information, which means that the contrast between the object in the target scene and the background can be improved. Additionally, we will get more detailed edge information.

3.3 Paired data with depth map data set

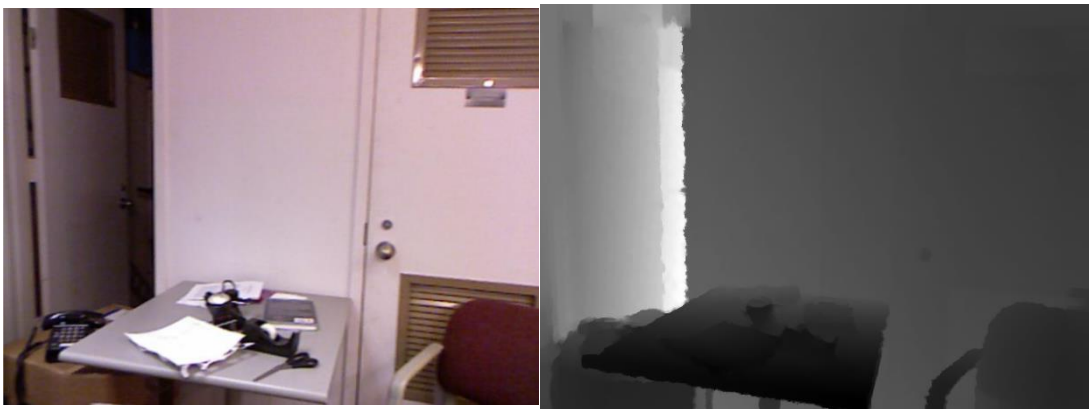
The RGBD data set based on the first method of generating the paired data set mentioned above is the most ideal data set for this research.
So, the acquisition process of this data set is as follows:

3.3.1 Select a suitable RGBD ground data set

This experiment selected the open source indoor RGBD data set NYU Depth V2[51], The NYU-Depth V2 data collection contains video sequences from a range of interior situations captured by the Microsoft Kinect's RGB and Depth cameras. It contains 1449 tightly labeled pairs of matched RGB and depth images, as well as 464 additional scenes from three cities and 407,024 new unidentified frames. A class and an instance number are assigned to each object (cup1, cup2, cup3, etc.)

There are multiple parts to the dataset: 1) Labeled: A portion of the video data with dense multi-class labels is labeled. This data has also been preprocessed to include depth labels where they are absent. 2) Raw: The RGB, depth, and accelerometer data supplied by the Kinect in its raw form. 3) Toolbox: A collection of useful utilities for working with data and labels. Pick some examples of NYU Depth v2 shown as follows:

(a)Indoor original image-1 (b)Corresponding depth map-1



(c)Indoor original image-2 (d)Corresponding depth map-2

(e)Indoor original image-3 (f)Corresponding depth map-3

Figure 3.11: Left are original images and right are corresponding depth map by NYU Depth v2[51]

3.3.2 Increase the diversity of water types

In order to obtain data sets of different water quality, we adopted 5 different ocean coefficients and 5 different coastal coefficients proposed by Jerlov and Colleagues [8], which represent the characteristics of different sea areas in subtropical, tropical and temperate regions. The degree of pollution in the ocean coefficients steadily rises from Type I to Type III, with Type I being the clearest and Type III being the most contaminated. The pollution degree of Type-1 to Type-9 steadily rises in the coastal coefficient, with Type 1 being the cleanest and Type 9 being the most turbid.

As discussed before, the absorption of violet (400~435nm) and red (605~700nm) wavelengths in water is strong, and the absorption of blue (450~480nm) and

green(500~560nm) are weak, so it tends to appear blue green in the ocean. However, in coastal, the light absorption degree of each wavelength is similar, so there will be more situations, such as yellow. The following shows the absorption coefficients for different wavelengths based on the five ocean water qualities and five coastal discussed above.

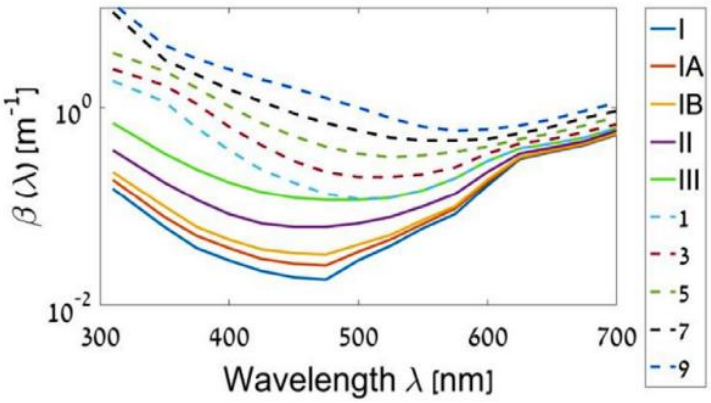


Figure 3.12: Wavelength-dependent light attenuation coefficients by 10 water types [4]

3.3.3 Generate the synthesis underwater images

Combining the open source indoor RGBD data set NYU Depth v2 and the attenuation coefficients of 10 different water types, through the proposed degradation model, 10 underwater images with different attenuation conditions will be synthesized. The flowchart is as follows:

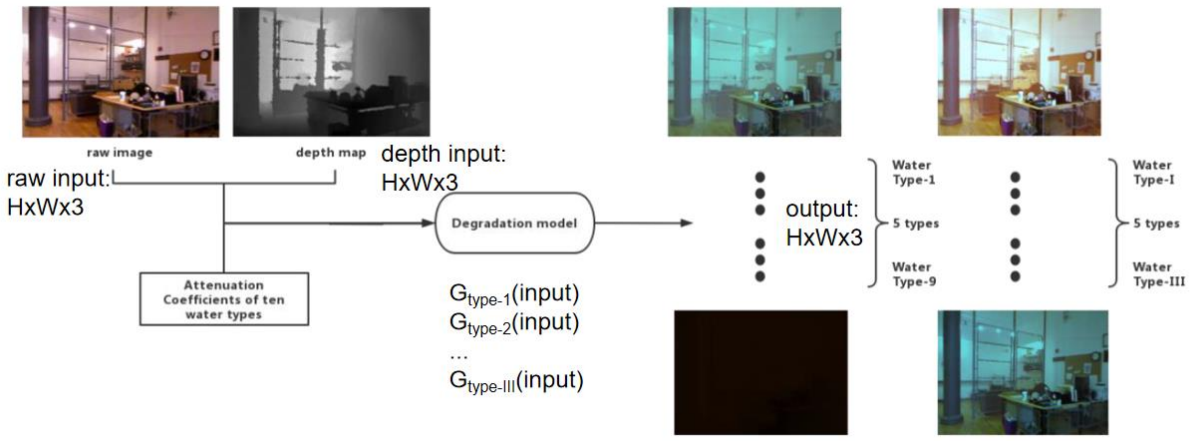


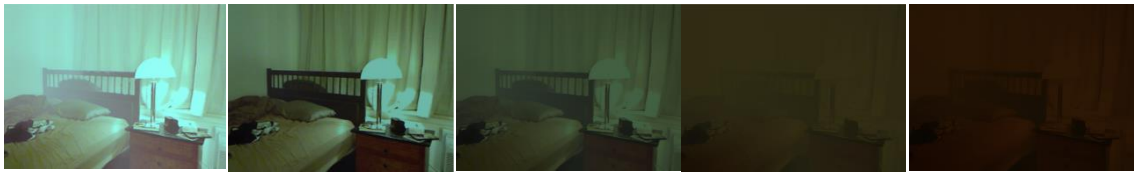
Figure 3.13: Flow chart of generating the synthesis underwater images

We select 1449 images in NYU Depth v2 and combine the degradation model of Formula 1 and Formula 2 with the attenuation coefficients of the 10 water types mentioned above to synthesize a rich variety of underwater images, including 10 categories, and there are 1449 images in each category, totally are 14490 underwater images.

Because coastal waters suffer from significant attenuation in deep water, such as type 3 above 10m and type 9 above 5m, things become nearly invisible in coastal seas. But the attenuation of certain water in shallow water is quite tiny. For instance, I, IA, and IB type water has an attenuation of around 1m to 5m and has essentially little influence on objects. In order to differentiate between water types, we defined distinct depth ranges: Water 5, 7, and 9 had their depth ranges adjusted to [0.5, 4.5], water 1 and 3 had their depth ranges set to [0.5, 14.5], and water I, IIA, and III had their depth ranges set to [5, 20]. Meanwhile, we choose a global backdrop light [0, 1] at random from a pool of options.

We cropped the original size 480x640 of NYU-v2 to 460x620 to enhance the quality of the datasets. The following is the final generation effect:

Raw image depth map



Type-1 Type-3 Type-5 Type-7 Type-9

Type-I Type-IA Type-IB Type-II Type-III

Figure 3.14: Ten types of underwater images

3.4 Unsupervised Implementation Model

3.4.1 Multiple Scale Convolution

Multi-scale operation can be simply understood as, in order to obtain more expressive feature information, some methods are implemented between the layers of the neural network and the interior of a single layer to achieve better detailed expression effects. It includes two aspects: the serial multi-branch structure between layers and the parallel multi-branch structure inside a single layer.

3.4.1.(1) Serial Multi-Branch Structure: (U-net)

Theoretically, the performance of a network increases as the number of network layers increases, but it will face the problem of gradient dispersion. The more layers pass, the previous information will gradually weaken and dissipate. Skip connection solves the problem of gradient disappearance, improves the utilization efficiency of features, and helps restore the loss of features caused by image down-sampling.

Taking U-net as an example. Its network structure looks like a U-shape as a whole. If each up-sampling and down-sampling is regarded as a layer operation, a total of 4 layers of down-sampling and 4 layers of up-sampling constitute the U-net network structure. . The serial multi-branch structure is embodied in the skip connection. From the Figure 3.15, it can be seen that the down-sampling of each layer and the up-sampling of each layer are connected through the skip connection operation. Note that each pair of connection operations here are performed on two feature maps of the same size, that is to say, the connection operation is the addition of the corresponding points.

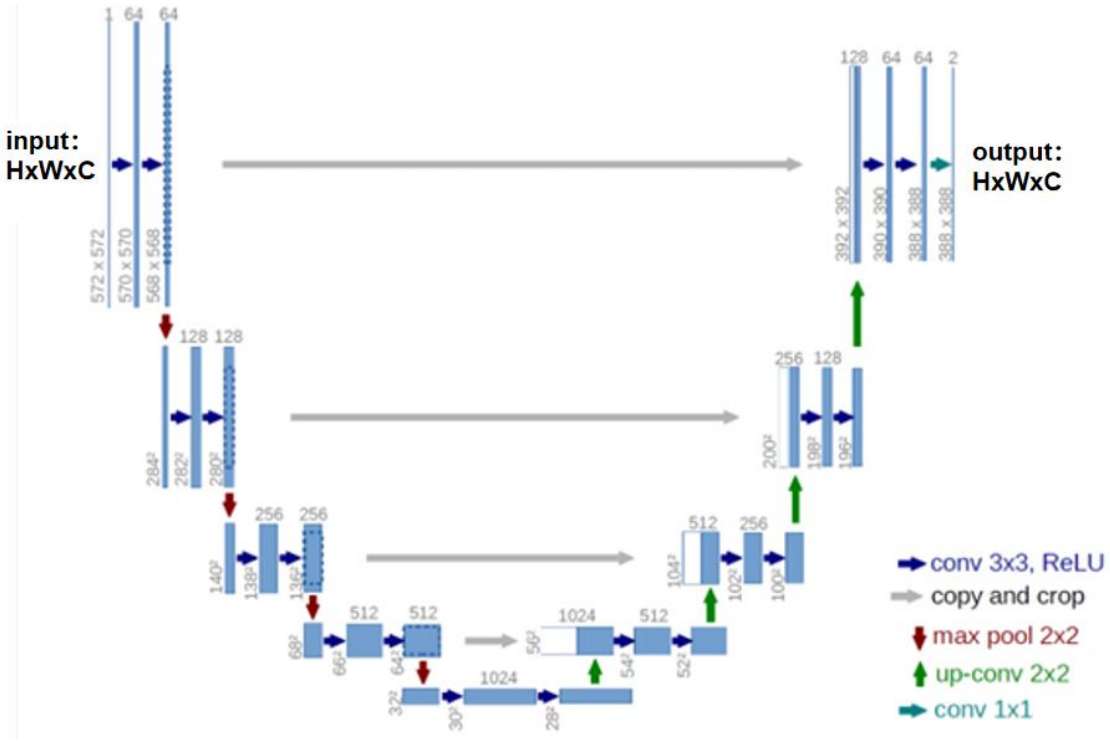


Figure 3.15: The network structure of U-net [5]

Each blue box corresponds to a multi-channel feature map. The white box represents the copied feature map. The gray arrow represents that the feature map of the down-sampling layer has undergone a skip connection operation with the corresponding up-sampling after being cropped.

Through such a serial multi-branch structure, such as skip connection, the shallow feature information is combined with the deep feature information, which solves the problem that the shallow feature information decreases or even disappears as the number of network layers is superimposed. U-Net performed a total of 4 times of up-sampling, and used skip connection in the same stage, instead of directly supervising and loss back propagation on high-level semantic features, so as to ensure that the finally restored feature map integrates more low-level features. Then, it makes information such as the edge recovery of the segmentation map in the image segmentation task more refined.

3.4.1.(2) Parallel Multi-Branch Structure: (Inception)

Taking the Inception network structure as an example, as a manifestation of a parallel multi-branch structure, it refers to the use of convolution kernels of different scales. For one layer of the network, for the incoming feature map from the upper layer, four different scale convolution kernels are used to convolute the same feature map, and four types of feature information with different degrees are obtained.

They represent different semantic interpretations of the same image. Compared with traditional convolution kernels of the same size, they get richer feature information. Then, the four types of feature maps carrying different feature information are fused, so that the final feature map obtained has more diversified semantic information, and then it is passed to the next layer.

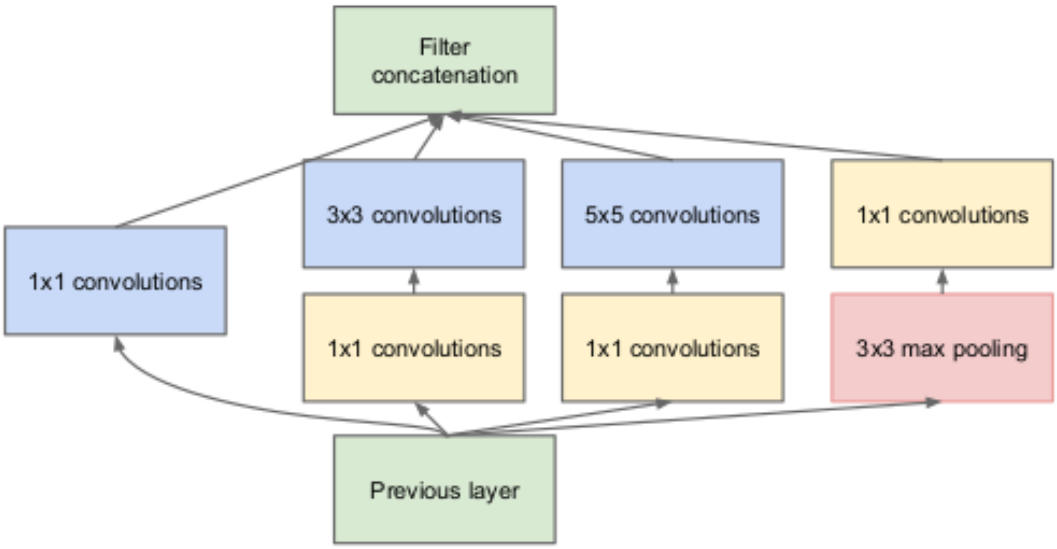


Figure 3.16: The network structure of Inception [52]

The function of the 1*1 convolution kernel is usually to control the number of channels, because the use of different scale convolution kernels (such as 5*5, or even 7*7) will increase a large number of parameter calculations. When the feature map is fused, a 1*1 convolution kernel is usually needed to control the number of parameters so that it can be calculated easily in the subsequent network. In view of the huge amount of parameter calculations, in the subsequent InceptionV2 structure, 5*5, 7*7 convolution kernels are also replaced with n*1 and 1*n, which not only increases the number of layers of the network, but also reduces the parameters calculation.

3.4.2 Attention Mechanism

The attention module can capture long-distance contextual information to obtain better feature representation. Through the self-attention module, the response of a location is calculated as the weighted sum of all the features from different spatial locations. Therefore, it connects the long-term dependence of any two locations in the feature map.

The attention mechanism in deep learning draws on human attention thinking. Therefore, we first briefly introduce the selective attention of human vision.

3.4.2.(1) Attention Mechanism of Human Vision

The visual attention mechanism is a type of brain signal processing that is only seen in humans. By swiftly scanning the global image, or the so-called focus of attention, human eyesight receives the target region that has to be focused on. Then devotes additional attention resources to this region in order to gain more comprehensive information about the target on which attention should be focused while suppressing irrelevant data. This is a method for people to swiftly screen out high-value information from a huge volume of data using their limited attention resources. It is a long-term survival technique developed by humans. The efficiency and precision of visual information processing are considerably improved by the human visual attention system.

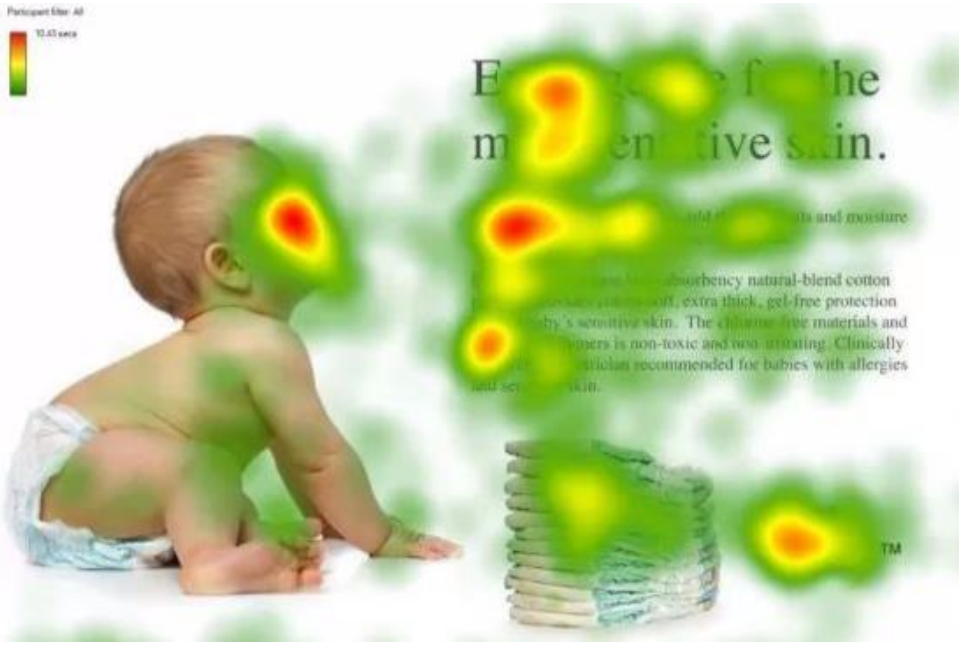


Figure 3.17: Different levels of attention when people receive information [53]

When people examine a image, Figure 3.18 graphically depicts how they efficiently deploy limited attention resources. The red region denotes the focus of the visual system, whilst the green area denotes the non-focused portion. People will naturally pay greater attention to the human face, the text title, and the first phrase of the article in the situation depicted in Figure 3.18.

Deep learning's attention method is quite similar to humans' selective visual attention mechanism. The main purpose is to choose from a big volume of data the information that is most important to the present job goal.

3.4.2.(2) Attention Mechanism of Deep Learning

The author of this research [54] investigated the role of attention in network architecture. Attention not only tells us where to focus, but it also helps us enhance our attention expression. The idea is to use attention processes to boost expressiveness by focusing on key elements and suppressing those that aren't. The author uses channel and spatial attention modules to learn what to pay attention to and where to pay attention in the channel and space dimensions, respectively, in order to stress the relevant aspects in the two dimensions of space and channel.

Channel attention is designed to teach the network “what to look”, which shows the correlation between different channels, and automatically obtain the importance of each feature channel through network learning, and finally assign different weight coefficients

to each channel to strengthen important features suppress non-important features.

Spatial attention aims to teach the network "where to look", which improves the feature expression of key areas. In essence, the spatial information in the original image is transformed into another space through the spatial conversion module and the key information is retained, and a weight mask is generated for each position. The output is weighted to enhance the specific target area of interest while weakening the irrelevant background area.

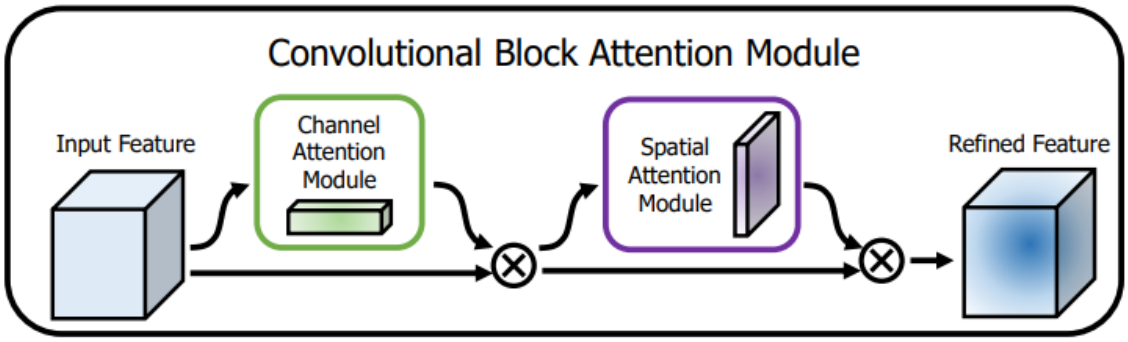


Figure 3.18: The network structure of CBAM [54]

The main network architecture is composed of a channel attention module and a spatial attention module in series, and more refined features are obtained through this integration method.

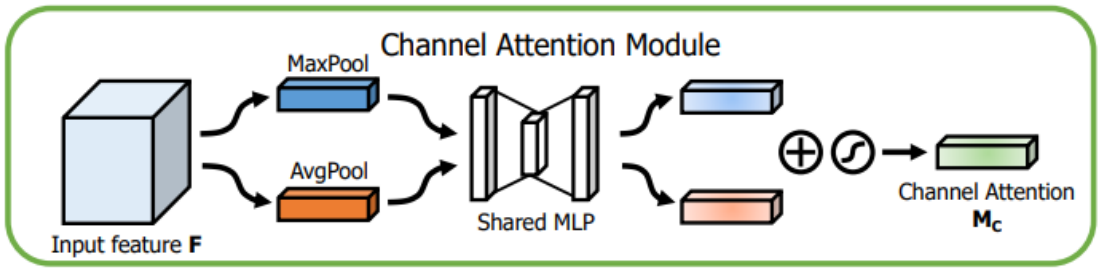


Figure 3.19: The network structure of channel attention module [54]

To construct our channel attention map $\boldsymbol{M_c}$, we first conduct average pooling and maximum pooling operations on the input feature $\boldsymbol{F}$, and then feed the two outcomes to the shared network. A multi-layer perceptron (MLP) with a hidden layer makes up the shared network.

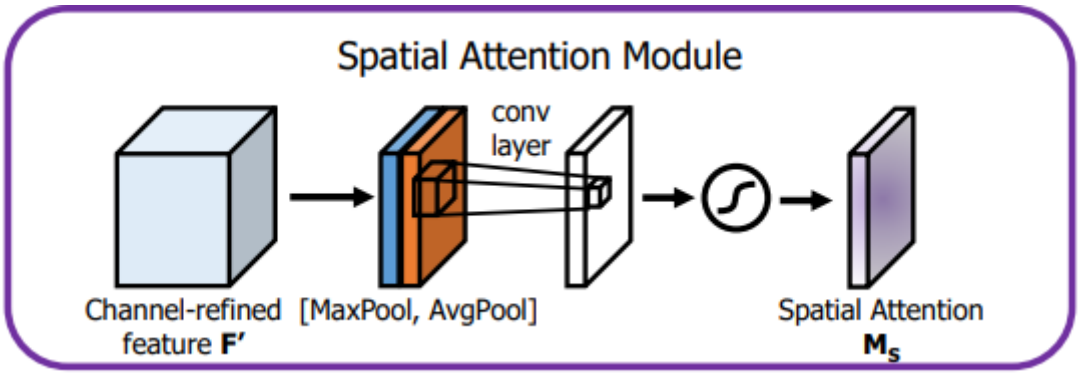


Figure 3.20: The network structure of spatial attention module [54]

To create spatial attention maps, we leverage the spatial connections of features. The "where" space attention is focused on is a supplement to channel attention, and it differs from channel attention in that it is an information portion. To acquire the spatial attention map $\boldsymbol{M_s}$, first execute maximum pooling and average pooling operations on the input $\boldsymbol{F'}$ created by the channel attention module, stitch the results together for convolution, and then use the sigmoid function to obtain the spatial attention map $\boldsymbol{M_s}$.

3.5 Proposed Modules and Overall GAN model

This part will introduce the network model UMA-GAN proposed in this study. Based on the GAN network discussed above, UMA stands for U-net network, multi-scale convolution and attention mechanism.

3.5.1 Proposed Multiple Scale Module

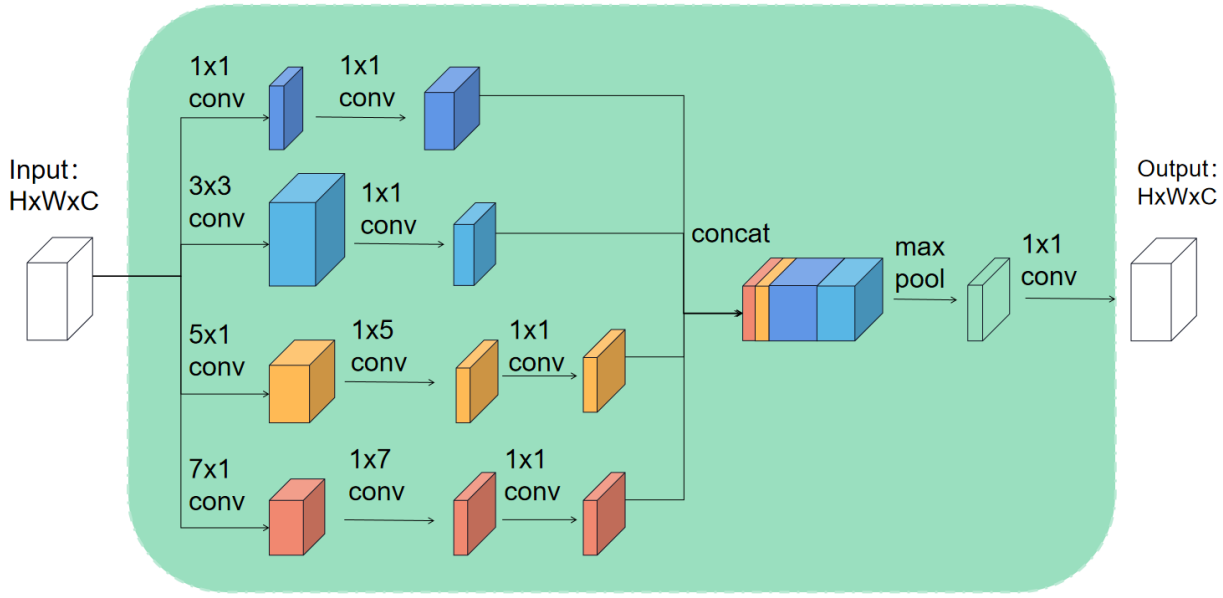


Figure 3.21: The Multiple Scale Module of our network structure

Which based on the network structure of InceptionV2, a 1*1 convolution is attached to the convolution kernel of each size, which reduces the number of parameters before concatenation. After concatenation, it first passes through a max pooling network layer. The linear relationship between different parameters is increased, and more parameter dependence relationships are obtained, and then after a 1*1 convolution, the number of channels is controlled again as the fusion result and passed to the next layer. And each convolutional layer is accompanied by an activation function layer Relu.

Finally, in our research, we use the combination of serial multi-branch structure and parallel multi-branch structure as the generator network structure. Not only can we obtain rich feature information in a single network layer, but also can integrate the shallow layer information and deep layer information between network layers. And then the experiment

proves that our model can get the ability to have more detailed feature information with good visible effect and high evaluation index.

3.5.2 Proposed Attention Mechanism Module

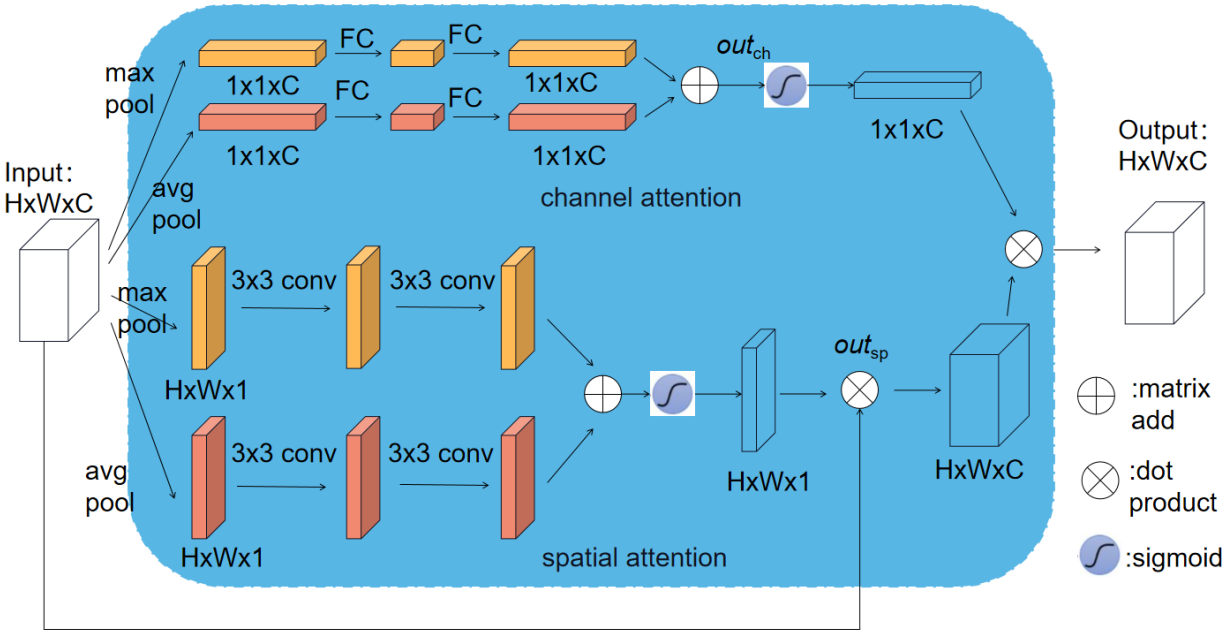


Figure 3.22: The attention mechanism module of this thesis

$$out_{ch} = \sigma((mp(in_{1\times1\times C}) + ap(in_{1\times1\times C})) \quad [3.1]$$

$$out_{sp} = [\sigma((mp(in_{H\times W\times1}) + ap(in_{H\times W\times1}))] \cdot in_{H\times W\times C} \quad [3.2]$$

$$out = out_{sp} \cdot out_{ch} \quad [3.3]$$

Where mp is the max pooling, ap is the average pooling, $in_{H\times W\times C}$ is the input feature, $in_{1\times1\times C}$ is the input with size $1\times1\times C$, $in_{H\times W\times1}$ is the input with size $H\times W\times1$, out_{ch} is the result of channel attention part, out_{sp} is the result of spatial attention part, and out is the result of the dot product operation between out_{ch} and out_{sp}.

Two-part processing is performed on the feature map of size $H\times W\times C$ passed in from the previous layer: the channel attention part and the spatial attention part. Each part is divided into two types of processing methods by the average pooling operation and the maximum pooling operation. In the channel attention part, the results of the two types of pooling are subjected to matrix addition and sigmoid function to obtain the result of

channel attention; In the spatial attention part, the results of the two types of pooling are subjected to sigmoid function to obtain the result of spatial attention. Finally, through the dot product operation of results of channel attention, spatial attention, then dot operation with input $in_{H\times W\times C}$, the output result out of the attention mechanism module is obtained.

3.5.3 Proposed new U-net model as Generator

Based on the attention mechanism module and multi-scale convolution module, we proposed an optimized U-net network model as the generator of this thesis, as follows:

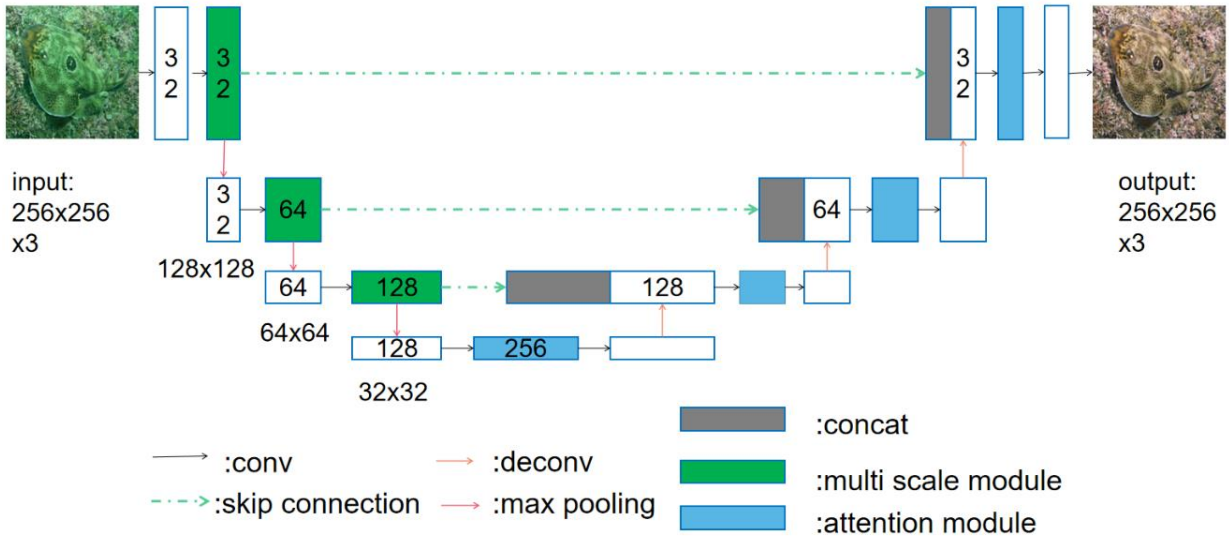


Figure 3.23: The network structure of proposed generator

The 256*256*3 input image is integrated into the multi-scale convolution module in the down-sampling feature extraction process. The output result is divided into two parts: one part is passed to the next layer after the maximum pooling operation, and the other part is passed to the corresponding up-sampling layers. After convolution and other operations of down-sampling to obtain a wealth of feature information, the attention mechanism module is added to make the network fully "pay attention" to the key information, and then the attention mechanism modules are added to the up-sampling network layers to attention the restored process of the image. And the input image and output image are the experimental results of this thesis.

3.5.4 Proposed new Convolution network as Discriminator

The discriminator uses PatchGAN [55], which is a special discriminator. PatchGAN is different from ordinary GAN discriminator. The ordinary GAN discriminator converts the input to a real number, indicating the probability that the input sample is actual data. PatchGAN converts the input into a N*N patch matrix. The probability that each patch is actual data is represented by the value of X. The discriminator's final output is the average value of X. The feature map from the convolution layers' output is referred to as X. We can examine the effect of this point on the final output result by tracing this feature map back to a specific position in the original image.

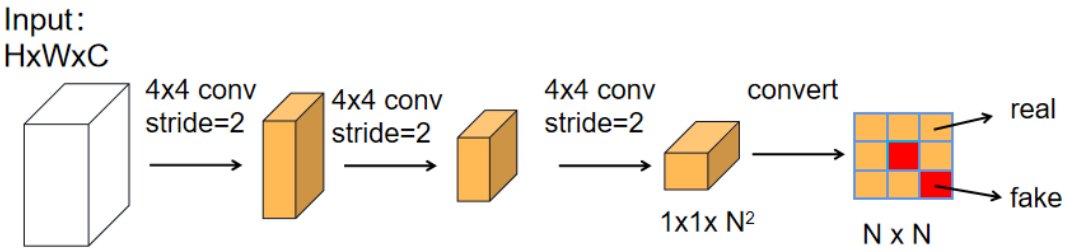


Figure 3.24: The network structure of PatchGAN

The advantages of using PatchGAN: ordinary GAN only outputs a probability value to indicate the similarity between the generated image and the original image, while PatchGAN outputs a matrix, and finally the results of each probability value of the matrix are summed and averaged. In other words, considering the influence of different parts of the image is like considering the suggestions of multiple people and then making a decision.

In fact, some studies have shown that ordinary GAN discriminators are not suitable for image fields that require high-resolution and high-definition details. As a result, PatchGAN is shown. Its receptive field corresponds to a tiny region in the input image, and its X relates to the discriminative output of a small section of the input image by the discriminator, allowing the model to pay greater attention to the image's features as a result of the training.

3.5.5 Overall GAN model Structure

Then, combine the multi-scale module and attention module of generator and the PathchGAN of discriminator, the totally network structure of our GAN as below:

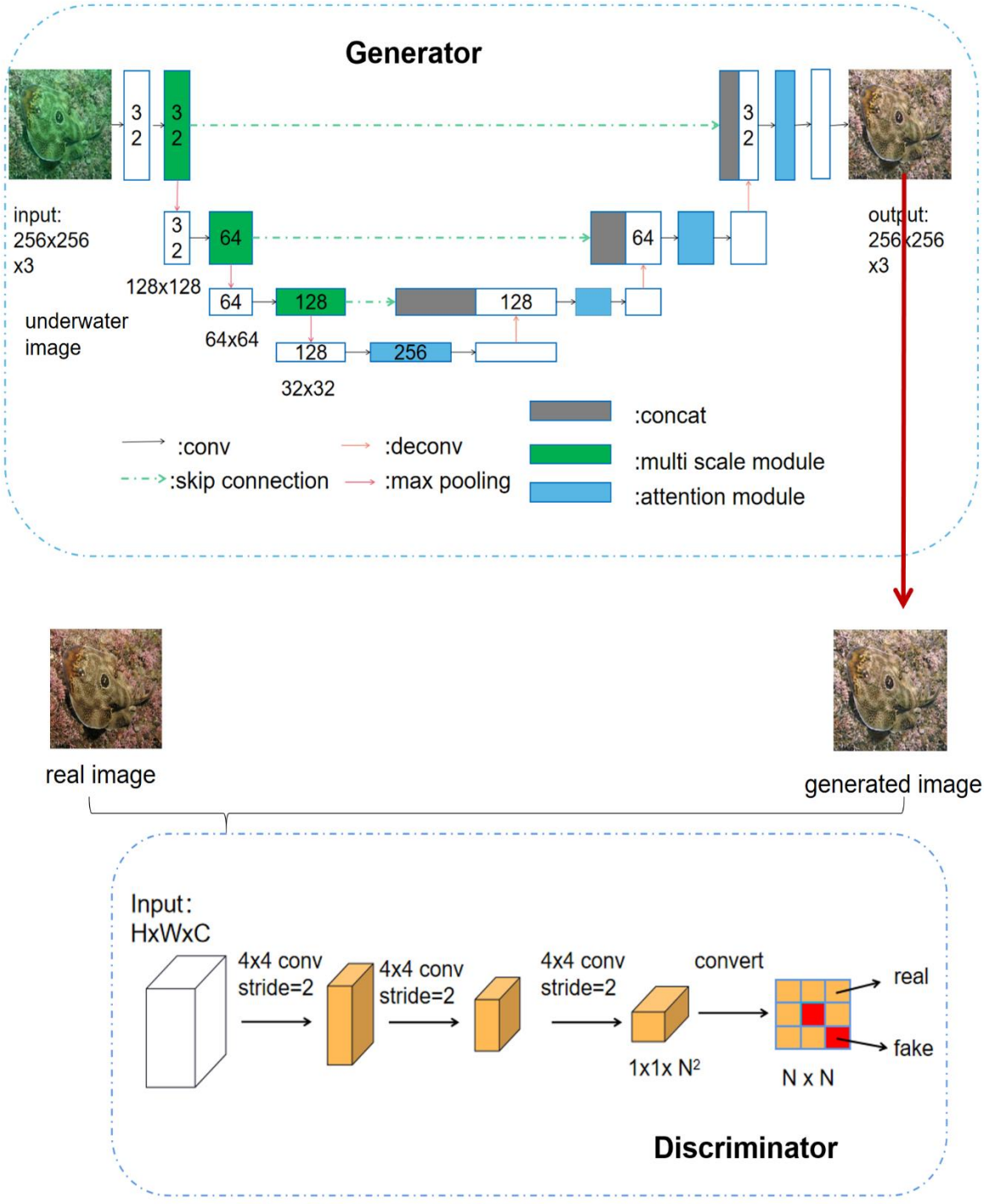


Figure 3.25: The network structure of our GAN

3.6 Loss Function

Table 3.1: Variables used in the part

Variable	Description
D	Discriminator
G	Generator
p_x	Data of datasets
p_z	Random vectors
$G(z)$	Generated data of G
$D(x)$	Result of D
$\mathbf{I}_c$	The original image
$\mathbf{I}_d$	Generated image by G
L_{adv}	Generating adversarial loss
L_{per}	Perceptual loss
L_{SSIM}	Structure similarity loss

For generating adversarial loss, the overall function is expressed as follows:

$$\min_{\mathrm{G}}\max_{\mathrm{D}} V(D,\mathbf{G}) = E_{\mathbf{x}\sim p_x(\mathbf{x})}[\log D(\mathbf{x})] + E_{z\sim p_z(z)}[\log(1 - D(\mathbf{G}(z)))] \quad (3.1)$$

Where the entire formula is made up of only two words. The true image is represented by $\mathbf{x}$, the noise input to the G network is represented by z, and the produced image by the G network is represented by $\mathbf{G}(z)$. $D(\mathbf{x})$ denotes the likelihood that the D network will determine if the true image is original (because x is original, so for D, the closer this value is to 1, the better). And $D(\mathbf{G}(z))$ is the chance that the D network will decide whether the image formed by G is original or not.

G's goal: As previously stated, $D(\mathbf{G}(z))$ is the likelihood that the D network will determine if the image formed by G is original, and G should hope that the image it makes is "as close to the actual as possible." To put it another way, G wants $D(\mathbf{G}(z))$ to be as big as possible, hence $V(D,G)$ will shrink at this point. As a result, we can observe that the formula's front mark is $\min_{G}$.

D's goal: The more D's ability, the larger $D(\mathbf{x})$ and the smaller $D(\mathbf{G}(z))$. $V(D,G)$

will grow in size at this point. As a result, D's formula is to maximize $\max_{D}$.

The input of the generator in this thesis is the underwater original image $\mathbf{I}_c$, and the image generated by the generator based on the generating adversarial loss is $\mathbf{I}_d$, so the new function L_{adv} is proposed as below:

$$\min_{\mathrm{G}}\max_{\mathrm{D}} V(G,D) = E_{x\sim \mathbf{I}_c}[\log D(\mathbf{x})] + E_{y\sim \mathbf{I}_d}[\log(1 - D(\mathbf{G}(y)))] \quad (3.2)$$

In addition to generating adversarial loss L_{adv}, in order to ensure that the content of the generated image is consistent with the original image, this chapter also introduces the L_1 loss:

$$L_1 = \frac{1}{\mathrm{CHW}}[\|\mathbf{I}_c - G(\mathbf{I}_d)\|_1] \qquad (3.4)$$

Where CHW means the size of feature map is W×H with C channels, $\mathbf{I}_c$ is a original image (ground truth), and $\mathbf{I}_d$ is an underwater distorted image.

However, using the L_1 function as the only optimization benchmark will result in blur artifacts on the image due to the average of the pixels in the pixel space. Therefore, this chapter additionally introduces Perceptual Loss, which is based on the difference between the feature map of the generated image extracted by the convolutional layer and the feature map of the target image. The use of perceptual loss can reduce the loss of high-frequency features caused by pixel averaging in L_1 loss, which is defined as follows:

$$L_{\text{per}} = \frac{1}{\mathrm{CHW}}\|F(\mathbf{I}_c) - F(\mathbf{I}_d)\|_2^2 \quad (3.5)$$

Where CHW means the size of feature map is W×H with C channels, and F is the high-level feature extracted by the last convolutional layer of the pretrained VGG19 network.

Generally, when calculating the difference between two images, the L_2 distance (Mean Square Error, MSE) is easily interfered by light and cannot measure the structural similarity of the images. In order to solve this defect, this article proposes the structure similarity method of SSIM to make the structure similarity between the image generated by the generator and the original image. SSIM is equivalent to normalizing

the data, calculating the illumination of the image block (the mean value of the image block), the contrast (the variance of the image block) and the normalized pixel vector, and finally multiplying the three. The experiment proves that it has better performance and improves the training effect. For two images $\mathbf{I}_c$, $\mathbf{I}_d$, the definition of SSIM is as follows:

$$L_{\text{SSIM}} = \text{SSIM}(\mathbf{I}_c, \mathbf{I}_d) = l(\mathbf{I}_c, \mathbf{I}_d) \cdot c(\mathbf{I}_c, \mathbf{I}_d) \cdot s(\mathbf{I}_c, \mathbf{I}_d) \quad (3.6)$$

Where $l(\mathbf{I}_c, \mathbf{I}_d)$ is the image illumination comparison part, $c(\mathbf{I}_c, \mathbf{I}_d)$ is the image contrast comparison part, $s(\mathbf{I}_c, \mathbf{I}_d)$ is the image structure comparison part. The specific formulas for these three parts as follow:

$$l(\mathbf{I}_c, \mathbf{I}_d) = \frac{2\mu_{\mathbf{I}_c}\mu_{\mathbf{I}_d} + C_1}{\mu_{\mathbf{I}_c}^2 \mu_{\mathbf{I}_d}^2 + C_2}, \quad C_1 = (K_1 L)^2 \quad (3.7)$$

$$c(\mathbf{I}_c, \mathbf{I}_d) = \frac{2\sigma_{\mathbf{I}_c}\sigma_{\mathbf{I}_d} + C_2}{\sigma_{\mathbf{I}_c}^2 + \sigma_{\mathbf{I}_d}^2 + C_2}, \quad C_2 = (K_2 L)^2 \quad (3.8)$$

$$s(\mathbf{I}_c, \mathbf{I}_d) = \frac{\sigma_{\mathbf{I}_c \mathbf{I}_d} + C_3}{\sigma_{\mathbf{I}_c}\sigma_{\mathbf{I}_d} + C_3}, \quad C_3 = C_2 / 2 \quad (3.9)$$

Where μ_{I_c}, μ_{I_d} are the average value of all pixels of the image block, and σ_{I_c}, σ_{I_d} are the variance of the image pixel value; $K << 1$, L is the gray dynamic range. SSIM has the following three properties: 1) Symmetry: $\text{SSIM}(\mathbf{I}_c, \mathbf{I}_d) = \text{SSIM}(\mathbf{I}_d, \mathbf{I}_c)$. 2) Boundedness: $\text{SSIM}(\mathbf{I}_c, \mathbf{I}_d) \leq 1$. 3) Unique extreme value:$\text{SSIM}(\mathbf{I}_c, \mathbf{I}_d) = 1$ if and only if $\mathbf{I}_c = \mathbf{I}_d$.

Combine (3.2), (3.4), (3.5) and (3.6) we get our final objective function:

$$L = L_{\text{adv}} + \alpha L_1 + (1 - \alpha) L_{\text{SSIM}} + \beta L_{\text{per}} \quad (3.10)$$

$$\alpha/(1-\alpha) = 3/7 \quad (3.11)$$

where α and β are weighting parameters that determine the trade-off between adversarial, MSE, and feature-based losses. The combination of L_1 and L_{SSIM} are proved the effectiveness in [56].

Chapter 4

Experimentation

4.1 Purpose of Experiment

By using the data set NYU mentioned in Chapter 3, which includes 10 underwater images of water quality and the corresponding original images, it is used to train our network model mentioned in Chapter 4 to learn the mapping relationship between underwater images and original images.

4.2 Method and Process

Our network structure is based on the GAN method, and the generator incorporates an attention module and a multi-scale module. The attention module aims to find the important parts of the image, while the multi-scale module aims to restore the details of the important parts. The generator and the discriminator are interactively trained. First, the generator is fixed to improve the discriminator's discriminative ability, and then the discriminator is fixed, and the generator is trained to improve the generation ability. Eventually the Nash equilibrium is reached, and the training loss tends to stabilize.

For the comparison experiments, we use six evaluation indicators of objective evaluation based on subjective evaluation. These include three evaluation indicators MSE, SSIM and PSNR that need to refer to ground truth, and three evaluation indicators UCIQE, UIQM, CIE that do not need to refer to ground truth. where the meanings of these indicators are as follows:

MSE: Mean squared error between estimated values and true values.

PSNR: Calculate the error of the corresponding pixels of the two images.

SSIM: Measure ground-truth and noisy images from brightness, contrast, and structure.

UCIQE: Evaluates the information richness of contrast, saturation, and texture.

UIQM: The evaluation is based on the linear combination of the three indicators of color, sharpness and contrast.

CIE2000: Calculates the perceptual color difference between two colors on sensation, that is, the distance between two-pixel points in the color space.

And we will present the specific formulas and instructions in Chapter 5.

4.3 The Training Condition

4.3.1 Experiment Environment

In this research experiment, the local Pycharm project was uploaded to the hard disk provided by VMware, and the Linux command was used to connect with JAIST Supercomputing to SSH to control the remote end to read the python file in the hard disk and run the project program.

The GPU: NVIDIA Tesla P100 of JAIST Supercomputer,

Memory: 128GB

Deep learning framework: version 3.6.0 of Pytorch

Language: version 3.6 of python

Optimizer: the Adam algorithm with a learning rate of 0.0001 to optimize network training.

Batch size: 16

The epoch: 200.

And every time the generator is updated, the discriminator is updated five times.

4.3.2 Data Preprocessing

For the data set NYU described in Chapter 3, the size of the image data is 640*460, which needs to be preprocessed for later training use. It includes three parts: Resize, Shuffle, Data Augmentation.

1. Resize: First, perform a unified resize process on the data set and change the size to 256*256.
2. Shuffle: To avoid the order of data input affecting network training, the data of the training model must be intermingled to increase randomness, improve the network's generalization performance, avoid the appearance of regular data, which causes the gradient of the weight update to be too extreme, and avoid over-fitting or under-fitting of the final model.
3. Data Augmentation: The data set was enlarged by data flipping and adding noise in order to improve the number of samples in the model and its generalization capabilities.

4.3.3 The Flow Chart of Training Process

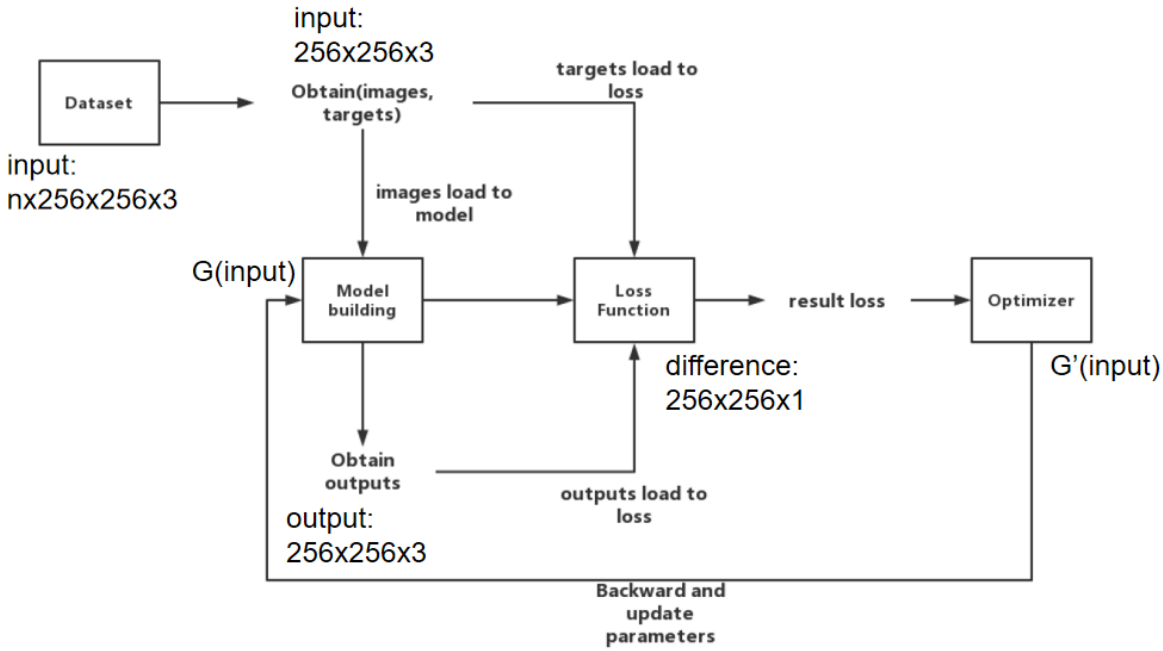


Figure 4.1: Description of the Training Process

The training process of a general network is roughly as follows: Firstly, obtain the degraded image and the target image through the preprocessed data set; Secondly, the degraded image is loaded into the model (have not trained completed) to obtain the output, and the loss function is used to calculate the loss value between the degraded image and the target image. Then the parameters of the model are updated through back propagation by calculating the gradient, so that the optimizer optimizes the loss value.

4.3.4 The training loss

Show the change of loss during training, where Loss_d is the total loss of the discriminator, and Loss_g is the total loss of the generator. The figure below shows the result after 10,000 steps (200 epochs), where the abscissa is step, and the ordinate is the value of the evaluation index, as shown below:

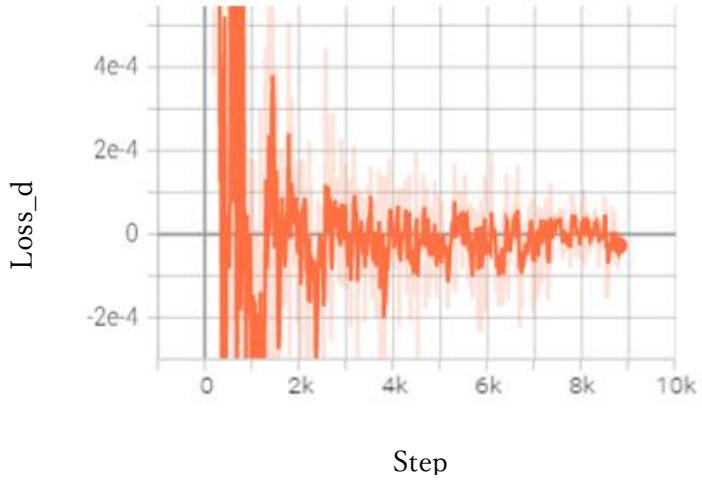


Figure 4.2: Discriminator loss calculation result graph

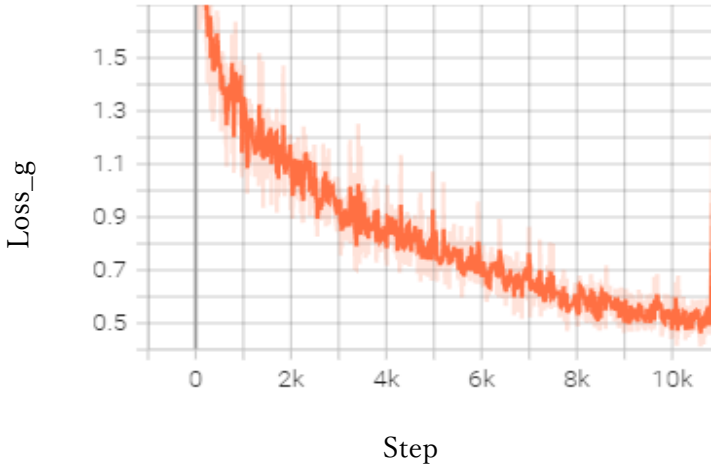


Figure 4.3: Generator loss calculation result graph

4.3.5 Training process

We performed a total of 200 epochs, in which the parameters of every 50 epochs were saved, and finally used the images in the test set to load different parameters to show the generated results of different epochs.

Table 4.1: The results of different epochs on the testing data set

Epoch	Test image 1	Test image 2	Test image 3	Test image 4
50				
100				
150				
200				

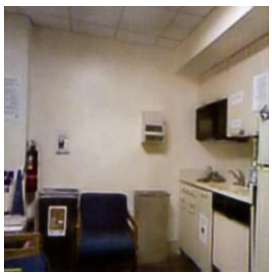
From the above results we can see that the results are more or less the same for different epochs, but if we look closely, we can see that the blue block in the whiteboard in Test

image 3 is gradually disappearing; especially in Test image 4, the green part gets a better color correction and the red block above the clock eventually disappears, and the wall returns to white.

4.4 The Results of the Training

After passing the complete training, the parameters of 200 epoch were retained and tested using ten water quality images of the same picture. the comparison results with UWGAN have some reference to prove the effectiveness of our method.

Table 4.2: The results by our method on the testing data set

Type	Underwater image	Our Result	UWGAN	Original image (Ground truth)
1				
3				
5				

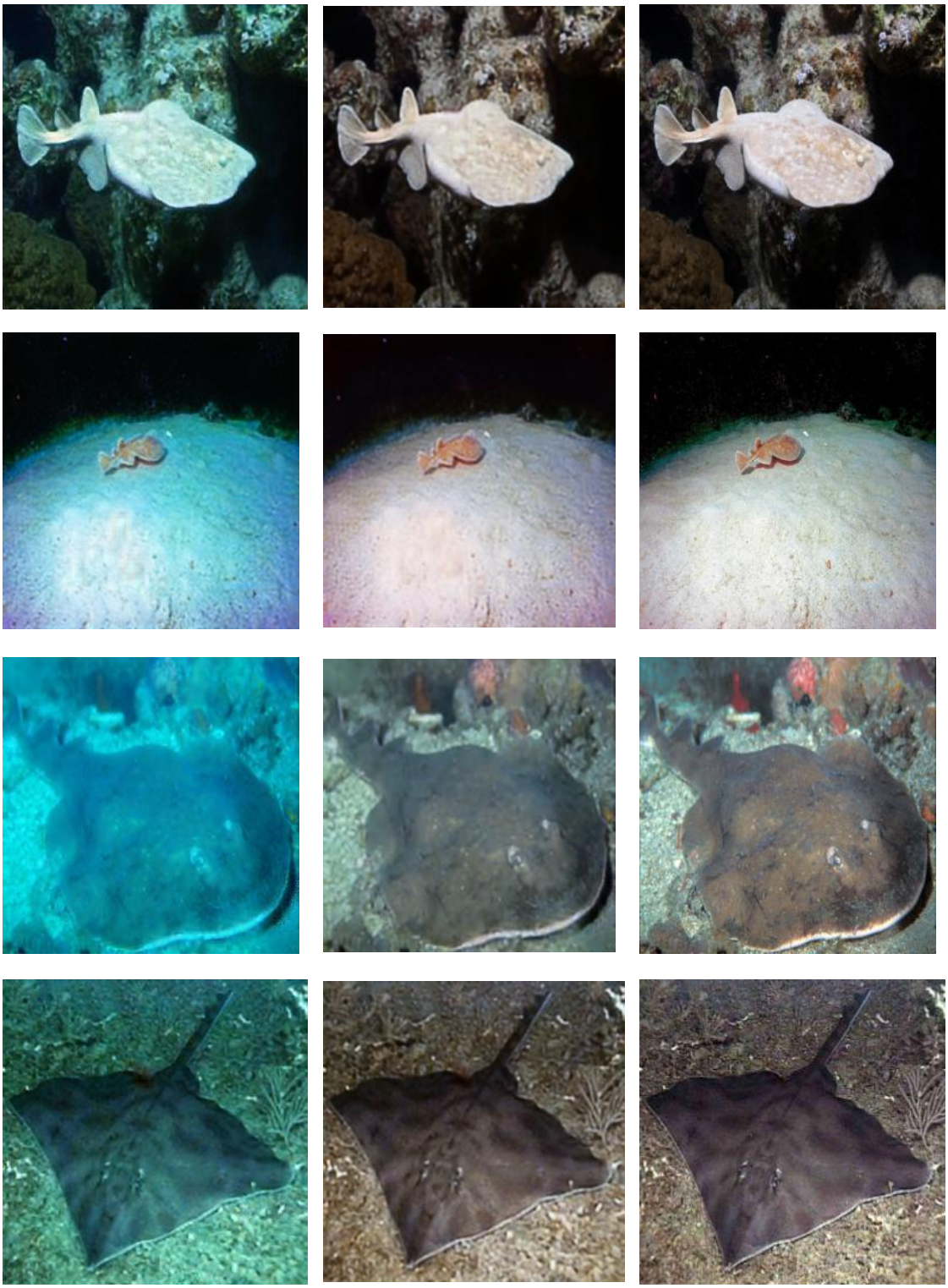
7				
9				
I				
IA				
IB				
II				

III				

From the results of the test set, for five kinds of marine water quality (1, 3, 5, 7, 9), our method can recover the color and object details of 1, 3, 5 water quality better, but for particularly severe cases 7 and 9 is not good recovery of image information; for five kinds of coastal water quality (I, IA, IB, II, III), our method can recover its correct color better, and for underwater images with exposure also recovered the exposure area as much as possible, but the result of recovery as IA is a little weak.

Table 4.3: The results by our method on the ImageNet data set

Underwater image	Our Result	CycleGAN
		
		



Chapter 5

Evaluation

5.1 Image Quality Evaluation indicators

Subjective and objective assessment methods are the two primary types of underwater picture quality evaluation methods used today. Subjective evaluation entails the tester observing the target picture and judging the underwater image's quality using subjective vision. The Mean Opinion Score (MOS) and the Differential Mean Opinion Score (DMOS) are indicators of subjective evaluation [57], but because subjective evaluation methods are affected by many factors and subjective evaluation is difficult to describe with a mathematical model, it is difficult to achieve high-quality evaluation.

The objective assessment approach involves using a set of algorithms and mathematical formulae to determine the image's visual quality. There are three types of image quality assessment: full-reference image quality assessment (FR-IQA), reduced-reference image quality assessment (RR-IQA), and no-reference image quality assessment (NR-IQA) [60]. RR-IQA and NR-IQA are mostly employed in the experimental section of this study. The following is a comprehensive overview of FR-IQA and NR-IQA.

5.2 Full-reference Image Quality Assessment (FR-IQA)

5.2.1 Peak Signal-to-Noise Ratio

Peak Signal-to-Noise Ratio (PSNR) [61] is to evaluate the quality of the image by calculating the error of the corresponding pixels of the two images. The greater the peak signal-to-noise ratio, the better the image quality. The calculation formula can be described in equation 3.2.

$$\mathrm{MSE} = \frac{1}{\mathrm{mn}} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} \|\mathbf{x}(i,j) - \mathbf{y}(i,j)\|^2 \qquad (3.1)$$

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{MAX_{\mathbf{x}}^2}{\mathrm{MSE}}\right) \qquad (3.2)$$

Where m and n represent the image size, MSE represents the mean square error of the true value image $\mathbf{x}$ and the noise image $\mathbf{y}$, $MAX_{\mathbf{x}}^2$ represents the maximum pixel value that can be obtained in the image $\mathbf{x}$.

5.2.2 Structural Similarity

To assess the original image $\mathbf{x}$ and the noisy image $\mathbf{y}$, structural similarity (SSIM) [62] is used to quantify the similarity of two images, primarily from the three features of brightness, contrast, and structure. The better the image quality, the greater the structural resemblance. SSIM= 1 when the two photos are precisely the same. Equation 3.3 is a description of the calculating formula.

$$\mathrm{SSIM} = \frac{(2\mu_{\mathbf{x}}\mu_{\mathbf{y}}+C_1)(2\sigma_{\mathbf{xy}}+C_2)}{(\mu_{\mathbf{x}}{}^2+\mu_{\mathbf{y}}{}^2+C_1)(\sigma_{\mathbf{x}}^2+\sigma_{\mathbf{y}}^2+C_2)} \quad (3.3)$$

Where $\mu_{\mathbf{x}}$, $\sigma_{\mathbf{x}}^2$ are the average and variance of the image $\mathbf{x}$ pixels, $\mu_{\mathbf{y}}$, $\sigma_{\mathbf{y}}^2$ are the average and variance of the image $\mathbf{y}$ pixels, $\sigma_{\mathbf{xy}}$ is the covariance of the image $\mathbf{x}$ and $\mathbf{y}$. $C_1=(k_1L)^2$, $C_2=(k_2L)^2$ are constants, where L is the range of pixel value, $k_1=0.01$, $k_2=0.03$.

5.3 No-reference Image Quality Assessment (NR-IQA)

5.3.1 UIQM

Karen et al. [63] suggested an Underwater Image Quality Measures (UIQM) system based on underwater image deterioration characteristics and a vision system. The Underwater Image Colorfulness Measure (UICM)[63], the Underwater Image Sharpness Measure (UISM)[63], and the Underwater Image Contrast Measure (UIConM)[63] are evaluated using a linear combination of the three components. UIQM's calculation formula is shown in Figure 3.5.

$$\mathrm{UIQM} = c_1 \times \mathrm{UICM} + c_2 \times \mathrm{UISM} + c_3 \times \mathrm{UIConM} \quad (3.5)$$

Where the weighting factors of the measurement components in the linear combination are c_1, c_2, and c_3. The weighting variables in this study are c_1 =0.0282, c_2 = 0.2953, and c_3 = 3.5753.

5.3.2 UCIQE

Yang Miao et al. introduced the Underwater Color Image Quality Evaluation (UCIQE) [64] as an image quality evaluation index. This indicator evaluates the richness of information such as contrast, saturation, and texture which need to be paid attention to for underwater image restoration. Its calculation formula can be described in 3.6.

$$\text{UCIQE} = c_1 \times \sigma_c + c_2 \times con_l + c_3 \times \mu_s \qquad (3.6)$$

Where σ_c is the chromaticity standard deviation, con_l is the brightness contrast, μ_s is the average saturation, and $c_1 = 0.4680$, $c_2 = 0.2745$, $c_3 = 0.2576$ are the values. The higher the UCIQE score, better the underwater image quality.

5.3.3 CIE76

$\mathbf{L_{ab}}$ is a color system of CIE, table color system, based on $\mathbf{L_{ab}}$ means based on top of this color system, basically used to determine the numerical information of a certain color. $\mathbf{L_{ab}}$ color space is based on the human eye's perception of color, can represent all the colors that the human eye can perceive. l means brightness, $\mathbf{a}^*$ means red-green difference, $\mathbf{b}^*$ means blue-yellow difference, the formular shows as below:

$$\Delta\mathbf{E}^*_{\mathbf{ab}} = \sqrt{(\Delta\mathbf{L}^*)^2 + (\Delta\mathbf{a}^*)^2 + (\Delta\mathbf{b}^*)^2} \qquad (3.7)$$

$$\Delta\mathbf{L}^* = \mathbf{L}^*_1 - \mathbf{L}^*_2 \qquad (3.8)$$

$$\Delta\mathbf{a}^* = \mathbf{a}^*_1 - \mathbf{a}^*_2 \qquad (3.9)$$

$$\Delta\mathbf{b}^* = \mathbf{b}^*_1 - \mathbf{b}^*_2 \qquad (3.10)$$

Where $\Delta\mathbf{E}^*_{\mathbf{ab}}$ is the total color difference, indicates the Euclidean distance of two colors in space, $\Delta\mathbf{L}^*$ is the difference between the brightness of two colors, $\mathbf{L}^*_1$ is the brightness of the original image, and $\mathbf{L}^*_2$ is the brightness of the test image; $\Delta\mathbf{a}^*$ is the difference between the red-green color of two colors, $\mathbf{a}^*_1$ is the red-green color of original image, and $\mathbf{a}^*_2$ is the red-green color of test image; $\Delta\mathbf{b}^*$ is the difference between the blue-yellow color of two colors, $\mathbf{b}^*_1$ is the blue-yellow color of the original image, and $\mathbf{b}^*_2$ is the blue-yellow color of the test image.

5.4 Comparison of Results of Different Methods

This part mainly evaluates the results from two aspects, one is subjective evaluation, and the other is objective evaluation. Firstly, the visual effects of the recovery results of different methods are observed intuitively through subjective evaluation, and then the differences of different methods are shown concretely through the value of objective evaluation

5.4.1 Subjective Evaluation Comparison

In this evaluation experiment, ensure the ground truth is original image, the SUID [48] data set was selected as the test data set. The original image of the ground is one half of SUID, while the other portion is a synthetic underwater image that contains greenish, blue, foggy, and low light. The specific information about SUID as below:

Table 5.1: SUID data set

	Ground Truth	Greenish	Bluish	Hazy	Low Light
Numbers	30	240	240	210	210

So, this experiment put the SUID underwater image and ground truth image on the top, and the rest of the methods on the bottom.

Four images were selected from the original underwater image to represent the four underwater scenes, including the bluish distortion scene, the greenish distortion scene, the hazy distortion scene and the low light distortion scene, and the results of this research are compared with the results of other methods:

(a) Underwater image (b) Original image(Ground Truth)

(c) UDCP[27] (d)RED[28] (e) Sea-thru[17] (f)UWCNN[65]

(g)UWGAN[66] (h)CycleGAN[47] (i)FUnIEGAN[50] (j)Ours

Figure 5.1：Comparison of bluish distortion scenes

(a) Underwater image (b) Original image(Ground Truth)

(c) UDCP[27] (d)RED[28] (e) Sea-thru[17] (f)UWCNN[65]

(g)UWGAN[66] (h)CycleGAN[47] (i)FUnIEGAN[50] (j)Ours

Figure 5.2：Comparison of greenish distortion scenes

(a) Underwater image (b) Original image(Ground Truth)

(c) UDCP[27] (d)RED[28] (e) Sea-thru[17] (f)UWCNN[65]

(g)UWGAN[66] (h)CycleGAN[47] (i)FUnIEGAN[50] (j)Ours

Figure 5.3：Comparison of hazy distortion scenes

(a) Underwater image (b) Original image(Ground Truth)

(c) UDCP[27] (d)RED[28] (e) Sea-thru[17] (f)UWCNN[65]

(g)UWGAN[66] (h)CycleGAN[47] (i)FUnIEGAN[50] (j)Ours

Figure 5.4：Comparison of low light distortion scenes

From the above results, For traditional methods, the UDCP method will lead to darkening and no correct color correction, but it performs well in the restoration of details; while the Sea-thru method has the highest contrast among all methods, but because the test data has no depth information, based on Sea-thru paper, the pixel value of the depth map generated by the provided monodepth method is too high, which is considered to be a very shallow area, resulting in no color processing. This is also a weakness that the degradation model requires camera parameters and a correct depth map. In contrast, the generalization ability of the GAN method is required to adapt to more data.

For GAN methods, most of the results generated by GAN methods are blurry, and some generated results have a certain color cast, which performs well on the training set but not on the validation data set. However, the result of our method is sharper and correctly restores certain colors.

In summary, it is proved that our method can adapt to more kinds of underwater images. In most of the data, the color of the underwater image and the details of the foreground objects are better restored, and obtained good results are better than most methods.

5.4.2 Objective Evaluation Comparison

Since there is no ground truth in the real underwater scene, we choose the SUID data set as our validation data set, which has ground truth to evaluate our method and other methods more convincing. And the quality of the restored underwater images generated by different methods is evaluated through the above five indexes of objective evaluation.

It can be seen from Table 5.1 that the total number of synthesized underwater data is 900 pieces.

By using the 900 images above, 900 results of different methods were obtained respectively, and the best 30 result images were selected in each method separately.

The MSE, SSIM and PSNR were calculated together with the 30 ground truth images, while UCIQE and UIQM were calculated with only 30 result images. The resulting table is as follows, with the largest value in bold red and the second largest in bold blue:

Table 5.2: Full-reference Image Quality Assessment

Methods	MSE ↓	SSIM ↑	PSNR ↑	CIE76 ↓
UDCP	17.8199	0.4301	9.6980	26.6554
RED	15.5624	0.4941	9.6232	26.8368
Sea-thru	8.6061	0.6662	14.0326	25.1599
UWCNN	2.3693	**0.8174**	**18.5349**	**11.9011**
UWGAN	5.1556	0.7302	17.7250	12.0037
CycleGAN	**1.6956**	0.7976	17.4086	12.8628
FUnIEGAN	3.1764	0.7281	16.5874	13.9663
Ours	**0.4216**	**0.8018**	**18.6679**	**10.1934**

Table 5.3: No-reference Image Quality Assessment

Methods	UCIQE↑	UIQM ↑
UDCP	0.3872	7.9249
RED	0.3838	5.4204
Sea-thru	**0.4413**	4.7925
UWCNN	0.3447	16.2557
UWGAN	0.3765	**20.9971**
CycleGAN	0.3223	16.6933
FUnIEGAN	0.4122	18.0230
Ours	**0.4171**	**22.5069**

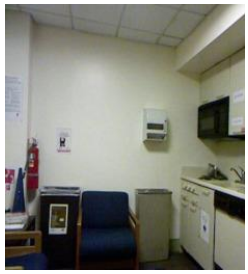
From the value of the objective evaluation indicators in the table, our method is better than most other methods on the same data set. It proves the effectiveness of the method of combining the multi-scale module and the attention module proposed in this research.

5.5 Ablation Experiment

Ablation experiments are utilized to demonstrate the efficiency of the attention module and multi-scale module offered by our technique.

The ablation experiment is a test to see if certain of the network structure's structures are effective. Ablation experiments are used to figure out what each component's purpose is. The results of deleting the attention module and the attention module are shown in the first column of the figure below, while the results of removing the multi-scale module and the multi-scale module are shown in the second column. The results of the experiments show that our proposed attention and multi-scale modules are successful.

Table 5.4: Ablation experiment results

Input image	No attention module	No multi scale module	Our result	Ground truth
				
1	2	3	4	5

From the results of the above ablation experiments,

1) the results generated by the removal of attention module (No.2) compensate the light for the wall, that is, the focus of the image is not found. However, the important part - the sofa is not well enhanced.
2) the result generated by removing the multi-scale module (No.3) is blurred as a whole, and no more detailed restoration of the important part is done.
3) And the result of our method (No.4) not only restores the color of the sofa better, but also improves the clarity and restores more details, which proves the effectiveness of the combination of attention module and multi-scale module in the thesis.

Chapter 6

Conclusion

In the field of underwater vision research, image enhancement plays an important role. A Generative Adversarial Network underwater picture improvement technique based on the integration of multi-scale and attention mechanisms is suggested to address the problems of blur, poor contrast, and color deviation in underwater photographs. A data set with ten different types of water quality is utilized to boost the data's generalization capacity.

An attention module and a multi-scale module are introduced to the generative network to highlight the important parts of the image and restore their features, hence improving the enhancement impact of underwater photographs. The generative network is utilized to create clear underwater photos, while the discriminative network is primarily used to aid the generative network in producing images with comparable visual perception to the reference image.

The results of this experiment can be roughly divided into three categories: subjective evaluation methods, objective evaluation methods, and ablation experiments.

1) For the results of subjective evaluation methods, our results are better than those of most methods in terms of correct color restoration in four scenes and recovery of more detailed information of objects.
2) For the results of objective evaluation methods, the values of our method on MSE, SSIM, PSNR, UIQM, UCIQE and CIE76 indexes are 0.4216, 0.8018, 18.6679, 0.4171, 22.5069 and 10.1934, respectively, which are higher than most methods. It proves that our method can improve the image sharpness, contrast, and chromatic aberration at the same time.
3) For the results of the ablation experiments, we removed the multiscale module and the attention module, but the generated results showed overall blurring and recovered objects in error, respectively, which proved the effectiveness of our method combining the multiscale module and the attention module.

In future work, we try to update the new network structure based on the current network as well as further optimize the loss function to generate better results.

Bibliography

[1] McGlamery B L. A computer model for underwater camera systems[C]//*Ocean Optics VI. International Society for Optics and Photonics*, 1980, 208: 221-231.

[2] Sharif M, Khan M A, Iqbal Z, et al. Detection and classification of citrus diseases in agriculture based on optimized weighted segmentation and feature selection[J]. Computers and electronics in agriculture, 2018, 150: 220-234.

[3] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[4] Fabbri C, Islam M J, Sattar J. Enhancing underwater imagery using generative adversarial networks[C]//2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2018: 7159-7165.

[5] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.

[6] Lu J, Li N, Zhang S, et al. Multi-scale adversarial network for underwater image restoration[J]. Optics & Laser Technology, 2019, 110: 105-113.

[7] Gao Y, Luo H, Zhu W, et al. Self-Attention Underwater Image Enhancement by Data Augmentation[C]//2020 3rd International Conference on Unmanned Systems (ICUS). IEEE, 2020: 991-995.

[8] Solonenko M G, Mobley C D. Inherent optical properties of Jerlov water types[J]. Applied optics, 2015, 54(17): 5392-5401.

[9] M. Yang, J. Hu, C. Li, G. Rohde, Y. Du and K. Hu, "An In-Depth Survey of Underwater Image Enhancement and Restoration," in *IEEE Access*, vol. 7, pp. 123638-123657, 2019, doi: 10.1109/ACCESS.2019.2932611.

[10] J. Y. Chiang and Y. Chen, "Underwater Image Enhancement by Wavelength Compensation and Dehazing," in *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1756-1769, April 2012, doi: 10.1109/TIP.2011.2179666.

[11] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, D. Tao, “An Underwater Image Enhancement Benchmark Dataset and Beyond,” *IEEE Trans. Image Process*., vol. 29, pp.4376-4389, 2019.

[12] N. Jerlov, ”Irradiance optical classification,” in *Optical Oceanography*, pp. 118120, 1968.

[13] wikipedia, *https://en.wikipedia.org/wiki/Visible_spectrum.*

[14] Akkaynak D, Treibitz T, Shlesinger T, et al. What is the space of attenuation coefficients in underwater computer vision?[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017: 4931-4940.

[15] C. P. Huynh and A. Robles-Kelly. Comparative colorimetric simulation and evaluation of digital cameras using spectroscopy data. In Digital Image Computing Techniques and Applications, *9th Biennial Conference of the Australian Pattern Recognition Society on, pages 309–316. IEEE*, 2007

[16] Akkaynak D, Treibitz T. A revised underwater image formation model[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018: 6723-6732.

[17] Derya Akkaynak, Tali Treibitz; Sea-Thru: A Method for Removing Water From Underwater Images. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1682-1691.

[18] D. Berman, T. Treibitz, and S. Avidan. Diving into hazelines: Color restoration of underwater images. *In Proc. British Machine Vision Conference (BMVC)*, 2017. 2, 3.

[19] Roser M, Dunbabin M, Geiger A. Simultaneous underwater visibility assessment, enhancement and improved stereo[C]. 2014 *IEEE International Conference on Robotics and Automation (ICRA). IEEE*, 2014: 3840-3847.

[20] Schechner Y Y, Averbuch Y. Regularized image recovery in scattering media[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, 29(9): 1655-1660.

[21] Acharya T, Ray A K, Gallagher A. Image Processing: Principles and Applications[J]. *Journal of Electronic Imaging*, 2006, 15(3): 9901-9911.

[22] Liu Y C, Chan W H, Chen Y Q. Automatic white balance for digital still camera[J]. *IEEE Transactions on Consumer Electronics*, 1995, 41(3): 460-466.

[23] Pisano E D, Zong S, Hemminger B M, et al. Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms[J]. *Journal of Digital imaging*, 1998, 11(4): 193-205.

[24] Ancuti C, Ancuti C O, Haber T, et al. Enhancing underwater images and videos by fusion[C]. 2012 *IEEE Conference on Computer Vision and Pattern Recognition. IEEE*, 2012: 81-88.

[25] Zhuang P, Ding X. Underwater image enhancement using an edge-preserving filtering Retinex algorithm[J]. *Multimedia Tools and Applications*, 2020, 79(1):1-21.

[26]He K, Sun J, Tang X. Single image haze removal using dark channel prior[J]. *IEEE transactions on pattern analysis and machine intelligence*, 2010, 33(12): 2341-2353.

[27]Drews P L J, Nascimento E R, Botelho S S C, et al. Underwater depth estimation and image restoration based on single images[J]. *IEEE computer graphics and applications*, 2016, 36(2): 24-35.
[28]Galdran A, Pardo D, Picón A, et al. Automatic red-channel underwater image restoration[J]. *Journal of Visual Communication and Image Representation*, 2015, 26: 132-145.
[29] Nan W, Zheng H, Bing Z. Underwater Image Restoration via Maximum Attenuation Identification[J]. *IEEE Access*, 2017, 5(99): 18941-18952.
[30] 梁淑芬, 刘银华, 李立琛. 基于 LBP 和深度学习的非限制条件下人脸识别算法[J]. *通信学报*, 2014, 35(6): 154-160.
[31]尹宝才, 王文通, 王立春. 深度学习研究综述[J]. *北京工业大学学报*, 2015, 41(1): 48-59.
[32]陈海鹏. 基于深度学习的视频中文字幕检测技术研究[D]. *北京: 北京邮电大学*, 2019.
[33]Sun X, Liu L, Li Q, et al. Deep pixel-to-pixel network for underwater image enhancement and restoration[J]. *IET Image Processing*, 2018, 13(3): 469-474.
[34] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//*Advances in neural information processing systems*. 2014: 2672-2680.
[35] Arjovsky M, Chintala S, Bottou L. Wasserstein gan[J]. *arXiv preprint arXiv*:1701.07875, 2017.
[36] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein gans[C]//*Advances in neural information processing systems*. 2017: 5767-5777.
[37] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. *arXiv preprint arXiv*:1511.06434, 2015.
[38] Mao X, Li Q, Xie H, et al. Least squares generative adversarial networks[C]//*Proceedings of the IEEE International Conference on Computer Vision*. 2017: 2794-2802.
[39] Jolicoeur-Martineau A. The relativistic discriminator: a key element missing from standard GAN[J]. *arXiv preprint arXiv*:1807.00734, 2018.
[40] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. *nature*, 2015, 521(7553): 436.
[41] Dong C, Loy C C, He K, et al. Image super-resolution using deep convolutional networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 38(2): 295-307.

[42] Ledig C, Theis L, Huszár F, et al. Photo-realistic single image super-resolution using a generative adversarial network[C]//*Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017: 4681-4690.

[43] Fu X, Huang J, Ding X, et al. Clearing the skies: A deep network architecture for single-image rain removal[J]. *IEEE Transactions on Image Processing*, 2017, 26(6): 2944-2956.

[44] Yang W, Tan R T, Feng J, et al. Joint Rain Detection and Removal from a Single Image with Contextualized Deep Networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2019.

[45] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//*International Conference on Medical image computing and computer-assisted intervention*. Springer, Cham, 2015: 234-241.

[46] Li J , Skinner K A , Eustice R M , et al. WaterGAN: Unsupervised Generative Network to Enable Real-time Color Correction of Monocular Underwater Images[J]. *IEEE Robotics and Automation Letters*, 2017:1-1.

[47] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//*Proceedings of the IEEE international conference on computer vision.* 2017: 2223-2232.

[48]Guojia Hou, Xin Zhao, Zhenkuan Pan, Huan Yang, Lu Tan, Jingming Li, June 29, 2020, "SUID: Synthetic Underwater Image Dataset", *IEEE Dataport,* doi: https://dx.doi.org/10.21227/agdr-y109.

[49]C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, D. Tao, “An Underwater Image Enhancement Benchmark Dataset and Beyond,” *IEEE Trans. Image Process*., vol. 29, pp.4376-4389, 2019.

[50]M. J. Islam, Y. Xia and J. Sattar, "Fast Underwater Image Enhancement for Improved Visual Perception," in *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3227-3234, April 2020, doi: 10.1109/LRA.2020.2974710.

[51] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor segmentation and support inference from rgbd images,” in *ECCV*, 2012.

[52] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2818-2826.

[53] Zhang, “Attention model in deep learning” aticle from the website *https://zhuanlan.zhihu.com/p/37601161*

[54] Sanghyun Woo , Jongchan Park , Joon-Young Lee, In So Kweon; *Proceedings of the European Conference on Computer Vision (ECCV),* 2018, pp. 3-19.
[55] PatchGAN from the Internet, *https://www.researchgate.net/figure/The-PatchGAN-structure-in-the-discriminator-architecture_fig5_339832261*
[56] Zhou Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," in *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, April 2004, doi: 10.1109/TIP.2003.819861.
[57] Kundu D, Evans B L. Full-reference visual quality assessment for synthetic images: A subjective study[C]//2015 IEEE International Conference on Image Processing (ICIP). IEEE, 2015: 2374-2378.
[58] Bosse S, Maniry D, Müller K R, et al. Deep neural networks for no-reference and full-reference image quality assessment[J]. IEEE Transactions on image processing, 2017, 27(1): 206-219.
[59] Wang S, Gu K, Zhang X, et al. Reduced-reference quality assessment of screen content images[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2016, 28(1): 1-14.
[60] Bosse S, Maniry D, Müller K R, et al. Deep neural networks for no-reference and full-reference image quality assessment[J]. IEEE Transactions on image processing, 2017, 27(1): 206-219.
[61] Matworks E. Compute peak signal-to-noise ratio (PSNR) between images[J].
[62] Brunet D, Vrscay E R, Wang Z. On the mathematical properties of the structural similarity index[J]. IEEE Transactions on Image Processing, 2011, 21(4): 1488-1499.
[63] Panetta K, Gao C, Agaian S. Human-visual-system-inspired underwater image quality measures[J]. IEEE Journal of Oceanic Engineering, 2015, 41(3): 541-551.
[64] Yang M, Sowmya A. An underwater color image quality evaluation metric[J]. IEEE Transactions on Image Processing, 2015, 24(12): 6062-6071.
[65] Li C, Anwar S, Porikli F. Underwater scene prior inspired deep underwater image and video enhancement[J]. Pattern Recognition, 2020, 98: 107038.
[66] Wang N, Zhou Y, Han F, et al. UWGAN: underwater GAN for real-world underwater color restoration and dehazing[J]. arXiv preprint arXiv:1912.10269, 2019.
[67] The website, *https://www.whoi.edu/oceanus/feature/a-smarter-undersea-robot/*
[68] The website, *https://towardsdatascience.com/sea-thru-removing-water- from-underwater-images-935288e13f7d*

[69] Yu X, Qu Y, Hong M. Underwater-GAN: Underwater image restoration via conditional generative adversarial network[C]//International Conference on Pattern Recognition. Springer, Cham, 2018: 66-75.