Underwater Image enhancement using residual networks

First Author
Institution1
Institution1 address
firstauthor@i1.org

Second Author
Institution2
First line of institution2 address
secondauthor@i2.org

Abstract

Due to physical phenomenon like refraction, absorption, and scattering of light by suspended particles in water, raw underwater images have low contrast, blurred details, and color distortion. Such degradation of images interferes with computer vision tasks like segmentation, object detection and classification. Thus, underwater image enhancement is carried out for tackling such phenomenons. In the realm of deep learning models majorly GANs and CNNs are used for the task. Through our study we work on drastically improving a residual network (UResnet) [12] which works at power with state-of-art GANs for image enhancement. Our proposed architecture works towards improving the quantitative image enhancement metric while reducing the model complexity and computation requirements. The same is carried out by experimenting with loss functions generally utilized for image enhancement tasks like super-resolution reconstruction and improving the residual network architecture using novel skip connections.

1. Introduction

Underwater images hold great importance for oceanographic tasks, ocean research and remotely operated underwater vehicles. Such analysis and exploration of the marine environments requires clear underwater images. Naturally obtained underwater images are degraded due to light absorption and scattering due to particles in the water. Light travelling in underwater scenario is composed of direct light, forward scattering light and backscattering light. Varying degrees of capture of these lights results in a dominating bluish or greenish tone to the underwater images.

These unclear images hinders the performance of underwater scene understanding and computer vision applications such as aquatic robot inspection and marine environmental surveillance. Thus, it is necessary to develop effective solutions to improve visibility, contrast, and color properties of underwater images for superior visual quality and appeal. This task of improving the underwater image encompasses tasks of image enhancement, image restoration and supplementary information specific methods. In our study we are focusing on the image enhancement task.

Underwater image enhancement task can promote reliability of vision tasks by improving the image contrast and reducing the degradation caused by scattering. Thus, to address these challenges we began by reproducing results from other deep learning tasks fro image enhancement based on GANs [8], CNNs [1] and Residual Networks [12]. Our study involved improving a base residual network UResNet [12] by improving its model architecture and the loss function utilized for training.

2. Literature Survey

2.1. Underwater Image Enhancement Method

A variety of methods have been proposed for image enhancement task. These can be broadly classified into three groups : Non-physical model, physical model-based method and deep learning methods.

Non-physical Model : Non-physical models deal with the improvement of visual effects by adjusting image pixel values rather than establishing a mathematical equation or physical model to simulate the image optical imaging characteristics. The literature has widely explained models such as Histogram equalization used for contrast enhancement, white balance used for color correction and classical Retinex image enhancement methods based on human vision brightness. The non-physical models are sufficient for in-air image processing but end up ignoring properties specific to underwater images. Thus, these methods can easily lead to color deviation wherein the ehanced images looks over or under saturated in disparate regions. They are also difficult to generalize for different scenarios and their robustness is poor.

Physical Model : Physical models on other hand models the degradation process of the image by estimating the parameters of the model. Because particle scattering in water is similar to scattering due to aerosols in atmosphere image dehazing algorithms majorly form this group. These models include multiple dehazing algorithms like dark channel algorithm [6], the non-local algorithm [3], the Fattal image dehazing algorithm [4] and other image enhancement algorithms based on atmosphere scattering model. Usually underwater images have more color deviation and are more blurred than in-air images. Thus, these dehazing algorithms many a times fall insufficient for image enhancement tasks.

Deep Learning Methods : Most deep learning methods for underwater image enhancement are weakly supervised learning methods solely focusing on color correction. These majorly revolve around generative adversarial networks, convolutional neural networks or residual networks. We have majorly reviewed two of the GANs and residual networks in detail whose components helped us improve our model.

1. **FUnIE-GAN** : FUnIE-GAN [8] is used to deblur the images which formulates the problem as an image-to-image translation problem by assuming that there exists a non-linear mapping between the distorted (input) and enhanced (output) images. Further, a conditional GAN-based model learns this mapping by adversarial training on a large-scale dataset.

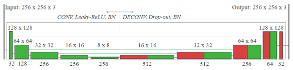

Figure 1. (a) Generator: five encoder-decoder pairs with mirrored skip-connections.[8]

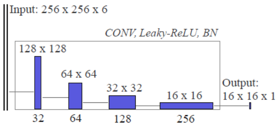

Figure 2. (b) Discriminator: a Markovian PatchGAN with four layers and a patch-size of 16 × 16.

Figure 3. Network architecture of FUnIE-GAN

Model Architecture [8]

(a) The generator network in Fig.1 is an encoder-decoder network (e1-e5,d1-d5) with connections between the mirrored layers, i.e., between (e1, d5), (e2, d4), (e3, d2), and (e4, d4). The outputs of each encoders are concatenated to the respective mirrored decoders.

(b) The input to the network is set to 256 X 256 X 3 and the encoder (e1-e5) learns 256 feature-maps of size 8 X 8. The decoder (d1-d5) utilizes these feature-maps and inputs from the skip connections to learn and generate a 256 X 256 X 3 enhanced image as output. Additionally, 2D convolutions with 4 X 4 filters are applied at each layer, which is then followed by a Leaky-ReLU non-linearity and Batch Normalization (BN).

(c) For the discriminator in Fig.2, a Markovian Patch-GAN architecture with a patch size of 16 X 16 is employed that only discriminates based on the patch-level information. This assumption is important to effectively capture high-frequency features such as local texture and style. Four convolutional layers are used to transform a 256 X 256 X 6 input (real and generated image) to a 16 X 16 X 1 output that represents the averaged validity responses of the discriminator. At each layer, 5 X 5 convolutional filters are used with a stride size of 2 followed by the nonlinearity and BN.

The major challenges of this network are that it incorrectly models sunlight and many a times amplifies the noise in the background. It also leads to either over-saturated or under-saturated images thus severely degrading the images. We have tried tackling this shortcoming of this network using a modified loss function and model improving the model architecture which would be explained further.

2. **UResnet** [12]: As majority of the methods based on GANs solely focus on color correction we studied a non-GAN based residual network UResnet which is a more comprehensive supervised learning method for underwater image enhancement.

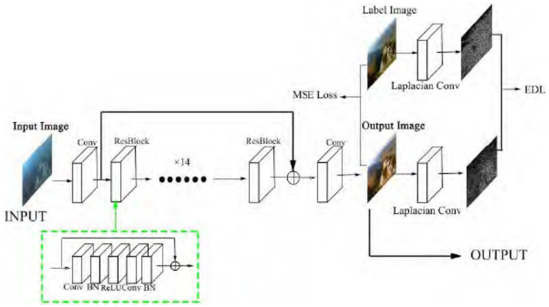

Figure 4. Network Architecture of UResnet with Edge Loss

Model Architecture [12]

(a) UResnet is a residual model [7] composed of

ResBlocks. It is composed of three sections : a head, body and tail.

(b) A long distance skip connection is included from head section outputs to the body section outputs. This connection adds feature information of input layer to output thus constraining the ResBlock modules to learn the difference between label images and input images.

(c) Each ResBlock is made of a sequence of Conv - Batch Norm - ReLU - Convolutional - Batch Norm layer.

(d) These blocks add the input of a particular layer to the output of the next layer. Thus establishing the skip connections helping the network to fully transfer information from the previous layer to the next layer.

(e) 16 such blocks are stacked together to ensure a deep network. As per the paper the network is majorly motivated from super-resolution reconstruction models EDSR [11] and SRResnet [10].

(f) The network uses a 3 × 3 convolution with stride of 1 pixel and zero - padding of 1 pixel to maintain shape of feature maps. This makes it easier for the network to obtain inputs with arbitary shapes.

(g) The paper also introduces a multi-term loss function called the Edge difference loss above the standard MSE loss. The EDL loss also forms the bases of one of the experiments in our proposed network.

Though very powerful the UResnet required a large amount of compute time and resources. In our study we have tried simplifying the same using a smaller network to reduce computation efforts while maintaining the performance. Also, due to the depth the network faced overfitting and saturation towards the later layer. Though relatively reduced due to the presence of residual blocks we tried further improving this network using a different multi-term loss function and a raw image based residual connection.

2.2. Underwater Image Quality Evaluation

We have briefly summarized the evaluation metric present in literature for evaluating image enhancement techniques. A more mathematical and detailed explanation of Full- reference Metric used for our study is provided in the approach section.

Full-reference Metric
For experiments with having ground truth image along with the unclear image mathematical metric like Means squared error (MSE), PSNR and SSIM [13] are used. These metrics work well for images which have been transformed using models to obtain the unclear images or wherein simulated images are used as unclear images.

Non-reference Metric Majority of the underwater images do not always have a clear or unclear image for calculation of the full-reference metric as it is challenging to obtain large amount of paired data. Thus, non-reference metric like image entropy, visible edges [5] and dynamic range independent image quality assessment [2] could be utilized.

3. Approach

Based on the literature survey we aimed at tackling the computational efficiency of the UResnet architecture along with trying to improve the vanishing gradient problem using an improved loss function and network architecture. A more detailed reasoning of the architecture used and the modifications carried out to the previous UResnet model are presented in this section.

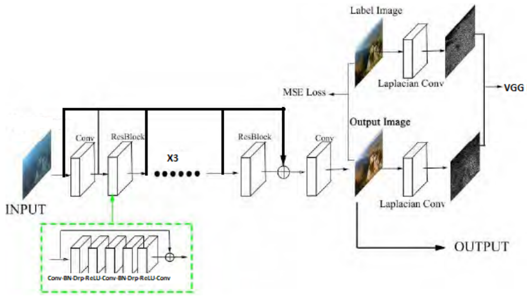

Figure 5. Modified UResnet Architecture with modifiec ResBlock and VGG Loss

3.1. Model Architecture

Following is a more granular explanation of our model demonstrated in Fig. 5

1. The architecture is a residual model composed of ResBlocks. It is composed of three sections : a head, body and tail.

2. **ResBlock** : Each ResBlock is made of a sequence of Conv - Batch Norm - Drop out - ReLU - Conv - Batch Norm - Drop Out - ReLU - Conv layer. The network uses a 3 × 3 convolution with stride of 1 pixel and zero - padding of 1 pixel to maintain shape of feature maps. This makes it easier for the network to obtain inputs with arbitary shapes. Unlike the other conv layers the

last conv layer in the ResBlock shrinks the number of channels from 64 to 61 to ease the concatenation of input image at the end of the ResBlock.

3. **Skip Connection** : Further, skip connections concatenating the input image at the end of each ResBlock is setup. These connections add feature information of input layer to the output and each residual block thus constraining the ResBlock modules to learn the difference between label images and input images.

4. Only 5 such modified ResBlocks are stacked together to ensure a deep enough network as well as ensuring requirement of lower compute for our model.

5. The multi-term loss function used in UResnet is replaced by a combination of standard MSE loss and perceptual loss. A more detailed account of perceptual loss has been covered in 3.1.2

3.1.1 Model Improvement

Post multiple experiments with the base architecture of UResnet following are the novel improvements incorporated into the model:

1. **Modified ResBlock and skip connection** : Instead of adding the first convoluted layer to the output and just setting up connections to let information flow from previous ResBlock to the next, we have setup skip connections to utilize the Raw input image. Also, instead of adding the Raw image at the end of each ResBlock layer it was concatenated to the input layer of each ResBlock. i.e. a 61 channel output from ResBlock was concatenated with the 3 channel Raw image to form the 64 channel output being added to the next ResBlock. In scenario of encountering vanishing gradient issue such a skip connection would impose larger weight on the channels associated with raw input image inplace of the channels outputed from the ResBlock. Thus, ensuring a learning from each block along with fully transfering information from the base image as well.

2. **Reduced layers :** We reduced the number of ResBlocks used by UResnet from 16 to 5. The proposed stacking of convolutional layers with the input data reduces the need for a very deep network and this property of stacking layers can be attributed to the superior performance of our proposed system even with a largerly smaller network.

3. **Dropout :** Although not intentionally very deep the network could still encounter overfitting due to the large number of parameters trained thus Drop out layers were introduced in the ResBlocks post the Batch Normalization layers.

3.1.2 Loss Metric Improvement

MSE and L1-loss function used by genreal image translational models provide per pixel difference loss functions. Thus, even though using these may improve the PSNR or SSMI values (i.e. quantitative evaluation metric) the generated images do not provide good visual effects. The MSE loss just averages differences at pixel level and fails to incorporate higher-level differences such as overall structure. To tackle this problem we experimented with two loss metric Edge difference loss proposed in the base paper of UResnet [12] and perceptual loss which is a loss metric very commonly used for style transfer or super-resolution tasks.

1. **Edge Difference Loss** : EDL loss was introduced to account for the information loss with regard to image edge information. The paper [12] introduced the same to penalize the models with EDL, so that the details of generated images are promoted to a higher level.

   To calculate the edge loss, a laplacian operator is used as an edge detector operator

$$lap = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -8 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

$$EDL = (I_{gt} \otimes lap - I_{en} \otimes lap)$$

   where,
   I_{gt} is the reference image
   I_{en} is the model enhanced image

   Total loss = MSE loss + $K_1 * SSIM$ loss + $K_2 * EDL$

2. **Perpetual Loss** : Taking inspiration from perceptual loss mentioned in [9], perceptual loss was added to the MSE loss. The perpetual loss was defined based on activation layer from VGG19 network trained on Imagenet dataset. This loss captures the difference between the feature representations of the enhanced image I_{en} and the reference image I_{gt}

   Perceptual loss, $L_j^{\phi} = \sum_{n=1}^{N} \left\| \phi(I_{en}^i) - \phi(I_{gt}^i) \right\|$

   Total loss = MSE loss + k* perpetual loss

   where,
   N = Number of each batch in the training procedure
   ϕ_j = j_{th} activation layer within the VGG19 network
   k = parameter obtained via hyperparameter tunning

Even though mentioned by the UResnet paper [12] utilization of EDL loss on our dataset didnot produce good results whereas utilization of perpetual loss improved the quantitative metric substantially.

3.2. Evaluation Metric

3.2.1 Quantitative Evaluation Metric

1. **Mean Squared Error (MSE)** : Mean square error is used for pixel to pixel comparison of the obtained clear image and the ground truth image. But the same may not be the right way to compare for an image enhancement task. The issue faced in this metric is tackled using multiple different loss metric mentioned in 3.1.2.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2$$

2. **Peak Signal to Noise ratio (PSNR)** : It is the peak signal-to-noise ratio, in decibels, between two images. It is used as a quality measurement between the original and the reconstructed image. The higher the PSNR, the better the quality of the reconstruction

$$PSNR = 10log_{10} \times \frac{L^2}{MSE}$$
$$L = \text{Dynamic range of image pixel intensities}$$

3. **Structural Similarity Index (SSIM)** : It is a method for measuring the similarity between the original and the reconstructed image. It compares the image patches based on three properties: luminance, contrast and structure

$$SSIM = l(x, y) \times c(x, y) \times s(x, y)$$
$$luminance, l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}$$
$$Contrasts, c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}$$
$$LocalStructures, s(x, y) = \frac{_{xy} + C_3}{\sigma_x + \sigma_y + C_3}$$

3.2.2 Qualitative Evaluation Metric

Post completing the decided experiments human analysis of images of the two base architectures was compared with the best performing architecture as a qualitative understanding of our process.

4. Experiment

4.1. Dataset - Enhancement of Underwater Visual Perception Dataset (EUVP)[8]

The dataset is a large collection of paired and unpaired images which contains poor and good quality images. This data was obtained during oceanic exploration and human-robot collaborative experiments in different locations under various visibility conditions. Additionally, some images are also obtained from $YouTube^{TM}$ videos. The paired dataset (i.e. with good and poor quality images) has been prepared using CycleGAN. Subset of images from Imagenet was used to train the CycleGAN based distortion model. Subsequently, a collection of good quality images are distorted using this model in order to generate the paired dataset.

4.1.1 Training Dataset

We have used 6128 EUVP Imagenet 6128 paired images for model training. Only a part of the whole 10K images were used to reduce the compute time required by model. The images are of various resolutions e.g. 800×600, 640×480, 256×256, and 224×224. The same are treated as per the procedure discussed 4.2 later for our network.

4.1.2 Testing Dataset

We used EUVP dark [8] which has 1000 paired images randomly selected from the EUVP dark dataset as our testing set as neither GANs or UResnet has been trained on this model whereas FUnIE-GAN has been trained on imagenet data.

4.2. Training

We performed multiple experiments with model architectures and loss functions. For each of the experiment, we resized the images to 256 X 256 and implemented models using Pytorch. CNN Models were trained with batch size = 1, learning rate = 2e-4, decay rate = 7e-1, step size = 400. Epochs and loss type varies with experiment.

1. **FUnIE-GAN**
Pre trained FUnIe-GAN model by [8] is used to enhance test images and calculate the performance metrics and used as one of the benchmark

2. **Vanilla UResnet**
UResnet model as mentioned in [7] is implemented to get benchmark scores for all metrics on the test set images

3. **Modified UResnet**
Improved the architecture as mentioned in section 3.1 with reduced number of stacked ResBlocks = 10

4. **Modified UResnet with perpetual loss**
Inspired by the perceptual loss mentioned in [**?**], we added perceptual loss to the improved UResnet model loss to capture the difference between the feature representations of the enhanced image I_c and the reference image I_g

5. **Modified UResnet with EDL loss**
   Inspired by the EDL loss mentioned in the base paper, we added the loss to the improved UResnet model loss.

6. **Modified UResnet perpetual loss and reduced layer**
   To improve further computationally, we further tried to reduce the number of stacked Res blocks to 5 along the addition of perpetual loss mentioned above. Reducing number of Res blocks still gave us similar lift to the image enhancement.

4.3. Results

4.4. Quantitative Results

As can be seen in table 1 the improved version of UResnet with the improvements suggested in section 3.1 the quantitative performance better than GANs as well as the vanilla UResnet being utilized as benchmark.

Model	MSE	PSNR	SSIM
FUnIE-GAN	205.80	26.18	0.91
Vanilla UResnet	283.03	24.69	0.88
Modified UResnet	226.55	25.56	0.89
Modified UResnet - EDL	11585	7.59	0.01
Modified UResnet - perpetual loss	195.72	25.95	0.90
Modified UResnet - perpetual loss & reduced layer	**193.93**	**25.98**	**0.90**

Table 1. Quantitative Evaluation Metric

4.5. Qualitative Results

As observed in figures, GAN performs a better color correction on blurred images than vanilla UResnet model (image 2 and image 3) but under saturates the image. While our modified UResnet with perpetual loss outperforms in both color correction and improving sharpness of the image as compared to GAN. (image 2 and image 4).

As we see in figure 6 image, GAN is not able to remove the sea hue completely and the object edges are not well defined (image 1). But vanilla UResnet helps to improve the color correction of the Res blocks but still have scope of improvement. Using perpetual loss using pretrained VGG19 model for object detection and classification helped to sharpen the image after color correction.

But as seen in figure 7, sharpness of image can still be improved using better pre trained object detection model. When the object is camouflaged with the background, we are not able to fully enhance the image.

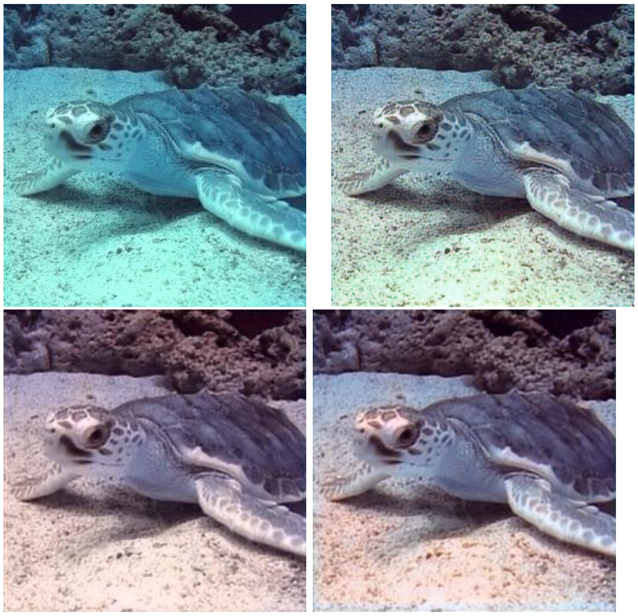

Figure 6. From left to right: Blurred image, Enhanced using FUnIE-GAN, Enhanced using Vanilla UResnet, Enhanced using Modified Uresnet with perpetual loss

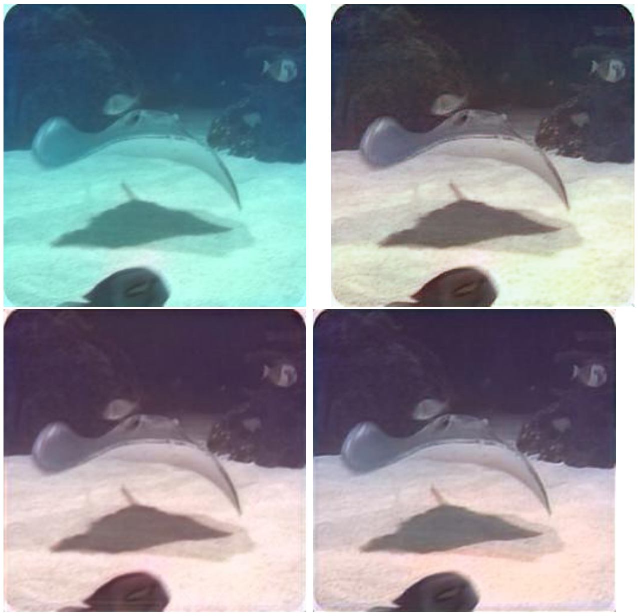

Figure 7. From left to right: Blurred image, Enhanced using FUnIE-GAN, Enhanced using Vanilla UResnet, Enhanced using Modified Uresnet with perpetual loss

5. Conclusion

5.1. Learnings

Experimenting with already available networks introduced us to the importance of task specific loss function. Loss being a crucial component of a network architecture did rightly play a significant role in our current performance improvement. We also learned the importance of improving the network without just increasing its length. Specifically, the saturation of results when the model was increased from 5 to 10 to 14 ResBlocks taught the importance of stronger component blocks than just stacking of layers.

5.2. Future Work

Following are a few idea which could be explored in the future :

1. We briefly dewelled on utilization of physical models like Histogram equalization and white balance during literature review. Such task specific transformations could be implemented in the network to help the network learner the precise shortcomings with an even smaller architecture.

2. We currently used VGG19 as one of the models for calculating our transformation loss. A more advanced network trained on larger and task specific datasets e.g. networks trained on only underwater images from Imagenet could be used as a better loss function.

3. We have not explored the possibility of utilizing transfer learning in a more concrete way. This could act as a much wider net of methods to explore.

References

[1] S. Anwar, C. Li, and F. Porikli. Deep underwater image enhancement. *arXiv preprint arXiv:1807.03528*, 2018.

[2] T. O. Aydin, R. Mantiuk, K. Myszkowski, and H.-P. Seidel. Dynamic range independent image quality assessment. *ACM Transactions on Graphics (TOG)*, 27(3):69, 2008.

[3] D. Berman, S. Avidan, et al. Non-local image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1674–1682, 2016.

[4] R. Fattal. Single image dehazing. *ACM transactions on graphics (TOG)*, 27(3):72, 2008.

[5] N. Hautiere, J.-P. Tarel, D. Aubert, and E. Dumont. Blind contrast enhancement assessment by gradient ratioing at visible edges. *Image Analysis & Stereology*, 27(2):87–95, 2008.

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

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] M. J. Islam, Y. Xia, and J. Sattar. Fast underwater image enhancement for improved visual perception. *arXiv preprint arXiv:1903.09766*, 2019.

[9] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[10] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[11] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

[12] P. Liu, G. Wang, H. Qi, C. Zhang, H. Zheng, and Z. Yu. Underwater image enhancement with a deep residual framework. *IEEE Access*, 7:94614–94629, 2019.

[13] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.