HIFI-Net: A Novel Network for Enhancement for Underwater Images

Jiajia Zhou, Junbin Zhuang, Yan Zheng*, Di Wu

Abstract—A novel network for enhancement to underwater images is proposed in this paper. It contains a Reinforcement Fusion Module for Haar wavelet images (RFM-Haar) based on Reinforcement Fusion Unit (RFU), which is used to fuse an original image and some important information within it. Fusion is achieved for better enhancement. As this network makes “Haar Images into Fusion Images”, it is called HIFI-Net. The experimental results show the proposed HIFI-Net performs best among many state-of-the-art methods on three datasets at three normal metrics and a new metric.

Index Terms—underwater images enhancement, CNN, image fusion, Haar wavelet.

I. METHOD

The Reinforcement Fusion Unit (RFU) is introduced in this section firstly. RFU is used to fuse information in two different kinds of images, in which one is an image containing base information, and the other is an image containing the additional reinforcement information. RFU generates the fusion image (or called fusion feature map), as shown in Fig.1. Then, Reinforcement Fusion Module for Haar (RFM-Haar), which fuses Haar images as reinforcement information into the original underwater image as base information, is introduced. As shown in Fig.2, RfM-Haar is based on RFU. Finally, a novel underwater image enhancement network, which uses RfM-Haar, is proposed as shown in Fig.3. As this network fuses Haar images into a fusion image, we call it HIFI-Net.

A. Reinforcement Fusion Unit

RFU can fuse a base image and a reinforcement image into a new image (which can also be seen as a 3D matrix), as shown in Fig. 1. RFU is universal for images containing different kinds of reinforcement information. Usually, the coordinates of the pixels must have the same physical meaning in a base image and its reinforcement image, in order to achieve ideal fusion.

The RFU has two inputs (Input1 and Input2) and one output (OutputU). Input1 is an image containing base information, called base image. Input2 is an image containing additional reinforcement information, called reinforcement image. OutputU is a fusion image. At first, RFU processes

This paper was supported by NSFC 62101156.

Manuscript received June 4, 2022. Yan Zheng is the corresponding author of this paper. Emails of authors are zhoujiajia@hrbeu.edu.cn;916800391@qq.com;yanzheng3@hrbeu.edu.cn;dwu.robotics@gmail.com. They are all in College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China.

\[ I_b = \text{Maxpool}(\phi_{5-c}(\phi_{1-c/2}(\text{Input1}))) \] (1)

where \( \phi_{k-c} \) is the abstract function of convolution processing to an image with \( c \) kernels of size \( k \) (size \( k \) means \( k \times k \)), and followed by the Relu activation function. Similarly, \( \text{Maxpool} \) denotes the max-pooling with a pooling kernel of 3 and a stride size of 1. The pooling parameters are the same in this paper.

On the other hand, RFU uses a Shared Multilayer Perceptron (MLP) to extract the features of Input1. The Shared MLP is composed of a \( 1 \times 1 \) convolution, a ReLU activation function and a \( 1 \times 1 \) convolution, as shown in Eq(2).

\[ \text{MLP}(\text{Input1}) = \phi''_{1-c}(\text{ReLU}(\phi''_{1-c/2}(\text{Input1}))) \] (2)

where \( MLP \) is the abstract function of the Shared MLP. A Shared MLP adjusts and aggregates the underlying information on a base image. \( \phi_{k-c} \) is convolution operation without activate.
Fig. 2. Reinforcement Fusion Module for Haar.

function. Then Sigmoid function filters the candidate features, as shown in Eq(3).

\[ M_b = \sigma(MLP(Input)) \]

where \( \sigma \) denotes the Sigmoid activation function. \( M_b \) can be viewed as a matrix of to base information.

Input2 is a reinforcement image of size \( M \times N \). RFU processes the image using \( 1 \times 1 \) and \( 3 \times 3 \) convolution kernels. Numbers of kernels in two sizes are all set to \( C \), as shown in Eq(4).

\[ M_a = \text{Maxpool} (\phi_{3-c} (\phi_{1-c} (Input))) \]

where max-pooling follows convolutions. The computation result \( M_a \) can be viewed as the matrix of coefficients to reinforce image.

The matrix \( M_a \) coming from Input2 and the matrix \( M_b \) coming from Input1 are summed. The fusion intensity control matrix \( M_f \) is shown in Eq(5).

\[ M_f = M_a \oplus M_b \]

where \( \oplus \) denotes each pair of corresponding elements in \( M_a \) and \( M_b \) are summed to form \( M_f \).

Next, \( M_f \) is used to adjust \( I_b \) and generate the fusion feature map, as show in Eq(6).

\[ OutputU = I_b \otimes M_f \]

where \( \otimes \) means that each element in \( I_b \) is multiplied by its corresponding element in \( M_f \) to form OutputU.

B. Reinforcement Fusion Module for Haar

RFM-Haar is designed based on RFU. It contains four RFUs and four residual convolution (RE) operations, as shown in Fig.2. The first-order two-dimensional Haar wavelet transform forms the original image of size \( M \times N \) into four sub-images of size \( M/2 \times N/2 \). These four sub-images after proper upsampling to size \( M \times N \) are fused with a base image to finally generate the fusion image with Haar reinforcement in RFM-Haar. The Haar wavelet transform result after proper upsampling and the original input image meet the spatial correspondence as a condition of RFU usage.

This module for fusing images in four steps is designed for Haar, as it has four sub-images. This module is equally applicable to other image fusion containing five inputs (one base and four reinforcement images). Even for fusing different numbers of images, RFM-Haar is only need to change the hyperparameters of the module appropriately. Therefore, though RFM-Haar is designed for Haar images fusion, the module can also be understood as RFM-RI (RI means reinforcement information). This may be a new way to combine CNN with traditional image processing.

RFM-Haar has five inputs (InputA, InputB, InputC, InputD and InputE), four RFUs (RFU1, RFU2, RFU3 and RFU4) and four residual blocks (RE1, RE2, RE3 and RE4 ).

InputA: An original underwater image (3 channels).

InputB: The upsampled zero-order image of Haar wavelet transformed InputA.

InputC: The upsampled first-order x-direction image (horizontal component) of Haar wavelet transformed InputA.

InputD: The upsampled first-order y-direction image (vertical component) of Haar wavelet transformed InputA.

InputE: The upsampled first-order xy-direction image (diagonal component) of Haar wavelet transformed InputA.

InputA is pre-decomposed into 4 sub-images using Haar wavelet. Then, these 4 sub-images are upsampled to size \( M \times N \) using the nearest neighbor interpolation method, as the size of InputA is \( M \times N \).

InputA is firstly used as Input1 of RFU1 to participate in the first fusion, as shown in Fig.2. InputB is used as Input2 of RFU1, and the OutputU of RFU1 is a fusion image of \( M \times N \times 16 \). On the other hand, InputA also enters the residual convolution with 16 kernels of size \( 3 \times 3 \). Here the residual convolution is to solve the network degradation problem.

InputA is used as the input of the first residual block (RE1). The output of RE1 is a feature map of \( M \times N \times 16 \). This feature
map is summed with the OutputU of RFU1 to generate Input1 of RFU2. The calculation process from InputA, InputB to the Input1 of RFU2 is shown in Eqs(7,8).

\[
RFU1.\text{Output} = F_{RFU}(\text{InputA}, \text{InputB}) \tag{7}
\]

\[
RFU2.\text{Input} = RFU1.\text{Output} \oplus \phi_{3-16}(\text{InputA}) \tag{8}
\]

where, \(F_{RFU}(\cdot, \cdot)\) is the abstract function of RFU. \(RFU_j.\text{Input1}, RFU_j.\text{Input2}, RFU_j.\text{OutputU}\) represent two inputs and one output of the \(j\)th RFU respectively, and \(j\) can be 1, 2, 3, and 4. \(\phi_{3-16}\) means residual convolution of \(3 \times 3\) kernels.

As Fig.2 shows, the inputs of RFU2 and RFU3 are the outputs of previous stages, InputC and InputD. The RFU4.\text{Input1} is a spliced feature map in the channel dimension from RFU3.\text{Input1} and RFU3.\text{OutputU} \(\oplus \phi_{3-16}(RFU3.\text{Input1})\). RFU2.\text{OutputU} and RFU2.\text{Input} \(\oplus \phi_{3-16}(RFU4.\text{Input1})\) is spliced to form the output of RFM-Haar. The specific operation is shown in Eq (9-14).

\[
RFU2.\text{Output} = F_{RFU}(RFU2.\text{Input}, \text{InputC}) \tag{9}
\]

\[
RFU3.\text{Input} = RFU2.\text{Output} \oplus \phi_{3-16}(RFU2.\text{Input}) \tag{10}
\]

\[
RFU3.\text{Output} = F_{RFU}(RFU3.\text{Input}, \text{InputD}) \tag{11}
\]

\[
RFU4.\text{Input} = RFU3.\text{Input} \odot (RFU3.\text{Output} \oplus \phi_{3-16}(RFU3.\text{Input})) \tag{12}
\]

\[
RFU4.\text{Output} = F_{RFU}(RFU4.\text{Input}, \text{InputE}) \tag{13}
\]

\[
\text{OutputM} = RFU2.\text{Input} \odot (RFU4.\text{Output} \oplus \phi_{3-16}(RFU4.\text{Input})) \tag{14}
\]

where \(\odot\) denotes channel-wise splicing and OutputM is the output of RFM-Haar.

C. HIFI-Net

We proposed an underwater image enhancement network that uses the RFM-Haar. This network makes “Haar Images into Fusion Image”, so we call it HIFI-Net. As Fig.3 shows, this network has five inputs, the same as RFM-Haar. These five inputs are fused into a new 32-channel feature image in RFM-Haar. This new image is subsequently processed by the CBAM.

CBAM is composed of a channel attention unit and a spatial attention unit.

\[
F_1 = F_0 \odot \sigma(\phi_{1-32}(\text{ReLU}(\phi_{1-2}(\text{Avgpool}(F_0)))) \tag{15}
\]

\[
\odot \phi_{1-32}(\text{ReLU}(\phi_{1-2}(\text{Maxpool}(F_0)))) \tag{15}
\]

where \(F_0\) is the output of RFM-Haar, \(F_1\) is the channel attention output; AvgPool is the average pooling operation, kernel size 3, stride size 1.

\[
F_1' = \sigma(\phi_{7-1}(\text{AvgPool}(F_1) \odot \text{MaxPool}(F_1))) \tag{16}
\]

\[
F_2 = F_1 \otimes F_1' \tag{17}
\]

where \(F_2\) is the spatial attention output. In Eq(17), 1-channel image \(F_1'\) is multiplied by \(c\)-channels image \(F_1\). In this situation, 1-channel image \(F_1'\) is copied to form a \(c\)-channels image \(I_m\), whose each channel is as the same as \(F_1'\), then \(F_1 \otimes F_1' = F_1 \otimes I_m\) is established.

\(F_2\) is convolved by \(\phi_{3-12}\). The 12-channel feature map \(F_3\) is output, as shown in Eq(18).

\[
F_3 = \phi_{3-12}(F_2) \tag{18}
\]

The output of the weight generator is a feature map of \(M \times N \times 12\). It is mainly composed of a series of convolutions of different scales. The details of the operation are shown in Eq(19).

\[
W_{\text{all}} = \phi_{5-12}(\phi_{3-32}(\phi_{7-64}(\phi_{5-32}(\phi_{3-32}(F_2)))))) \tag{19}
\]

where \(F_2\) denotes the output of the CBAM; \(W_{\text{all}}\) is the output of the weight generator.

Underwater images processed by CNN are prone to contain color casts and artifacts. Therefore, a gated fusion module is
used to solve this problem. The gated fusion module uses a weight generator to learn the confidence mapping and splits it into four $M \times N \times 3$ confidence matrices. $F_3$ is also split into four $M \times N \times 3$ feature maps. The importance of these four feature maps is determined using the confidence matrices. Finally, these four feature maps are pixel-wise summed. The operation is shown in Eq.(20).

$$I_{en} = (R_{c1} \otimes C_{w1}) \oplus (R_{c2} \otimes C_{w2}) \oplus (R_{c3} \otimes C_{w3}) \oplus (R_{c4} \otimes C_{w4})$$

where $I_{en}$ is the underwater enhanced image, i.e., the output of HIFI-Net. $R_{c1}, R_{c2}, R_{c3},$ and $R_{c4}$ are four feature maps. $C_{w1}, C_{w2}, C_{w3},$ and $C_{w4}$ are confidence matrices.

### D. Objective function

To train HIFI-Net, the loss function is divided into three parts, which are $L_{cha}$, $L_{ssim}$ and $L_{per}$. The first part need to make the whole network sample and learn from the global similarity space. The robust Charbonnier loss [9] function is $L_{cha}$. It mainly focuses on the overall feature similarity of the image content, which is described in Eq.(21).

$$L_{cha}(J_c, \hat{J}_c) = \sqrt{(J_c - \hat{J}_c)^2 + \varepsilon^2}$$

where $\hat{J}_c$ and $J_c$ are the ground truth and enhanced image respectively, and $\varepsilon$ is a very small constant.

$L_{ssim}$ is the similarity index (SSIM) loss function [10]. It improves the local structure and details. SSIM can be defined as:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1) (2\sigma_{xy} + C_2)}{\mu_x^2 + \mu_y^2 + C_1 (\sigma_x^2 + \sigma_y^2 + C_2)}$$

where $\mu$ and $\sigma$ denote the mean and standard deviation of an image. C1 and C2 are constants used to maintain the stability. The $L_{ssim}$ can be written as Eq.(23).

$$L_{ssim}(J_c, \hat{J}_c) = I - SSIM(J_c, \hat{J}_c)$$

The $L_{per}$ is the perceptual loss [11]. The perceptual loss function is expressed as:

$$L_{per}(J_c, \hat{J}_c) = |(VGG(J_c) - VGG(\hat{J}_c))|$$

where $VGG$ refers to the $VGG - 19$ network. The enhanced image and the ground truth image are passed to the last convolutional layer of the pretrained VGG network to get the feature representations. $|\cdot|$ is the $L_1$ distance. Integrating $L_{cha}, L_{ssim}$ and $L_{per}$, the final loss function is written as:

$$L_{Total} = \lambda_1 L_{cha} + \lambda_2 L_{ssim} + \lambda_3 L_{per}$$

where $\lambda_1$, $\lambda_2$ and $\lambda_3$ denote the coefficients that adjust the importance of each loss component. During the training period, the values are set to 1, 1, 1, 11 adjusted as a hyperparameter.

### II. EXPERIMENTAL RESULTS

HIFI-Net is compared with several SOTA (state-of-the-art) algorithms for experiments on three public available datasets. These three datasets are UFO-120 [12], EUVP [7], and UIEB [5], where UFO-120 has a training set of 1620 pairs and a test set of 120 pairs, EVUP has a training set of 4553 pairs and a test set of 613 pairs, and UIEB has a training set of 800 pairs and a test set of 90 pairs. These experimental results involve several commonly used evaluation metrics (MSE, PSNR, SSIM) and a new proposed ER3C (The error of the ratio of three channels). ER3C is an evaluation metric to evaluate restoration ability of RGB structure. Then, the subjective evaluation experiments use some images from three datasets to show the enhancement of algorithms on underwater images. Finally, several ablation experiments are used to demonstrate the effectiveness of each module in HIFI-Net.

### A. commonly-used metrics

MSE, PSNR, and SSIM are commonly employed for quantitative evaluations. MSE and PSNR are used to evaluate the restoration ability of algorithms in terms of image content. SSIM is used to evaluate the restoration ability of algorithms in terms of image structure and texture. The smaller the value of MSE, the stronger the algorithm. The larger the values of PSNR and SSIM, the stronger the algorithm.

As shown in Tab.I, the proposed HIFI-Net gets the lowest MSE meaning the best performance among all algorithms in three datasets. It is 26.7% lower than the suboptimal algorithm Water-net on EUVP, 35.9% lower than the suboptimal algorithm Water-net on UFO-120 and 52.7% lower than the suboptimal algorithm Cyclegan on UIEB.

As shown in Tab.II, the HIFI-Net algorithm has the highest PSNR meaning the best performance among all algorithms in three datasets. It is 5.6% higher than the suboptimal algorithm Water-net on EUVP, 7.3% higher than the suboptimal algorithm Water-net on UFO-120 and 16.8% higher than the suboptimal algorithm Cyclegan on UIEB.

| Metric | Datasets | Cycle-GAN | DCP | Mul-Fusion | Water-Net | HE | FunIE-GAN | UWNNet | Ours |
|--------|----------|-----------|-----|------------|-----------|----|-----------|--------|------|
| MSE ($10^3$) | EUVP | 0.6173 | 0.8591 | 2.2212 | 0.2663 | 1.9295 | 0.2966 | 0.3593 | **0.1952** |
| | UFO-120 | 0.4624 | 0.7217 | 2.4803 | 0.2358 | 1.9080 | 0.2639 | 0.2596 | **0.1512** |
| | UIEB | 0.7496 | 1.1419 | 1.8620 | 0.7975 | 1.1610 | 1.8135 | 0.8766 | **0.4683** |
TABLE II
FULL REFERENCE IMAGE QUALITY WAS EVALUATED WITH PSNR ON THE EUVP, UFO-120 AND UIEB TEST SETS.

| Mtriy | Datesets | Cycle-GAN | DCP | Mul-Fusion | Water-Net | HE | FuNie-GAN | UWNNet | Ours |
|-------|----------|------------|-----|------------|-----------|----|-----------|--------|------|
| EUVP  | 20.2253  | 18.7900    | 14.6648 | 23.8756    | 15.2762   | 23.4078 | 22.5752       | 25.2241 |
| UFO-120 | 21.4797  | 19.5469    | 14.1857 | 24.4047    | 15.3250   | 23.9164 | 23.9877       | 26.3342 |
| UIEB  | 19.3824  | 17.5542    | 15.4308 | 19.1130    | 17.4823   | 15.5455 | 18.7026       | 21.4253 |

TABLE III
FULL REFERENCE IMAGE QUALITY WAS EVALUATED WITH SSIM ON THE EUVP, UFO-120 AND UIEB TEST SETS.

| Mtriy | Datesets | Cycle-GAN | DCP | Mul-Fusion | Water-Net | HE | FuNie-GAN | UWNNet | Ours |
|-------|----------|------------|-----|------------|-----------|----|-----------|--------|------|
| EUVP  | 0.6853   | 0.7525     | 0.5594 | 0.8129     | 0.6194    | 0.7916 | 0.7899       | 0.8265 |
| UFO-120 | 0.8112   | 0.8209     | 0.6663 | 0.8652     | 0.7565    | 0.7796 | 0.7796       | 0.8819 |
| UIEB  | 0.6549   | 0.6934     | 0.6716 | 0.7971     | 0.6348    | 0.5979 | 0.6646       | 0.8012 |

TABLE IV
FULL REFERENCE IMAGE QUALITY WAS EVALUATED WITH ER3C ON THE EUVP, UFO-120 AND UIEB TEST SETS.

| Mtriy | Datesets | Cycle-GAN | DCP | Mul-Fusion | Water-Net | HE | FuNie-GAN | UWNNet | Ours |
|-------|----------|------------|-----|------------|-----------|----|-----------|--------|------|
| EUVP  | 5.2791   | 5.5019     | 6.1032 | 3.4509     | 8.4905    | 3.0251 | 6.6403       | 2.6190 |
| UFO-120 | 3.0739   | 5.1852     | 6.3273 | 3.0143     | 8.2079    | 2.8869 | 8.7498       | 2.2918 |
| UIEB  | 25.1984  | 26.8712    | 26.6906 | 20.5563    | 31.6845   | 80.3736 | 68.1454      | 17.9162 |

As shown in Tab.III, the HIFI-Net algorithm has the highest SSIM meaning the best performance among all algorithms in three datasets. It is 1.7% higher than the suboptimal algorithm Water-net on EUVP, 1.9% higher than the suboptimal algorithm Water-net on UFO-120 and 0.9% higher than the suboptimal algorithm Water-net on UIEB.

B. The ER3C metric.

The values of a pixel in an RGB image can be seen as the ratio, which means the RGB structure, and the average gray value. The scattering of the water causes the RGB structure be destroyed, therefore the color of an underwater object looks different from that in the air. This RGB degradation mathematically is a disruption to the ratio of three channel values. The underwater image enhancement algorithm should have RGB structure restoration ability. In this section, ER3C is proposed to evaluate the RGB structure restoration ability of the algorithms.

Assuming the image is $I$, where $I(x, y)$ is a pixel at the position $(x, y)$. The pixel values on three channel are $I(x, y, 1)$, $I(x, y, 2)$, $I(x, y, 3)$ respectively. $I$ is the ground truth image.

$$i_a(x, y) = \sum_{z=1}^{3} I(x, y, z) / 3$$

$$\hat{i}_a(x, y) = \sum_{z=1}^{3} \hat{I}(x, y, z) / 3$$

$$ER3C = \left(1 / MN\right) \sum_x \sum_y \sum_z \left| I(x, y, z) / i_a(x, y) - \hat{I}(x, y, z) / \hat{i}_a(x, y) \right|$$

where $i_a(x, y)$ is the average gray value of $I(x, y)$ pixels and $\hat{i}_a(x, y)$ is the average gray value of $\hat{I}(x, y)$ pixels, in the three channels.

As shown in Tab.IV, the HIFI-Net algorithm has the lowest ER3C meaning the best performance among all algorithms on all three datasets. The proposed HIFI-Net is 24.1% lower than the suboptimal algorithm WaterNet In EUVP, 24.0% lower than the suboptimal algorithm WaterNet in UFO-120 and 12.8% lower than the suboptimal algorithm WaterNet in the UIEB. The proposed HiFI-Net performs better than other algorithms in terms of RGB structure restoration ability.
**TABLE V**

**Numerical results of HIFI-Net structure ablation experiments on UIEB**

|       | Haar | CBAM | RFM-Haar | Convs | PSNR↑ | SSIM↑ |
|-------|------|------|----------|-------|-------|-------|
| Exp1  | ✓    | ✓    |          |       | 18.82±3.048 | 0.770±0.081 |
| Exp2  | ✓    |      | ✓        |       | 20.54±3.048 | 0.781±0.0744 |
| Exp3  | ✓    | ✓    |          | ✓     | 20.22±0.2.892 | 0.771±0.080 |
| Exp4  | ✓    | ✓    | ✓        |       | 21.42±2.638 | 0.801±0.062 |

Fig. 4. Columns from left to right show input underwater images and the results of Cycle-GAN, DCP, Mul-Fusion, Water-Net, HE, FUnIE-GAN, UWNNet, the proposed HIFI-Net. The last column shows GT (ground truth) images.

**C. Subjective evaluation**

As the Fig.4 shows, this section shows the enhancement of multiple algorithms on three datasets. The first two rows are on the EUVP dataset. The third and fourth rows are on the UFO-120 dataset. The fifth and sixth rows are on the UIEB dataset. It can be seen that HIFI-Net is better in color restoration ability, compared to other algorithms.

**D. Ablation experiments**

To further verify the effectiveness of the HIFI-Net algorithm, ablation experiments are conducted on the key modules. The verifying experiments mainly include: the effect of Haar wavelet transform; the effectiveness of the RFM-Haar; the effectiveness of CBAM on the algorithm. The Convs in Tab.V indicates the convolution block composed of six convolutions of different kernels. This convolution block is used to fuse the original image and Haar wavelet transform results without RFM.

In Exp3 in Tab.V, the underwater images are pre-processed using Haar wavelet transform. The pre-processing results are input into the network. Exp3 and Exp4 are conducted to verify the effectiveness of the RFM-Haar. In Exp3, Convs is
TABLE VI
NUMERICAL RESULTS OF RFM-HAAR STRUCTURE ABLATION EXPERIMENTS ON UIEB

|                | w/o Residual | w/o Maxpool | w/o MLP | w/o Base image | Ours   |
|----------------|--------------|-------------|---------|----------------|--------|
| PSNR↑          | 20.5612      | 18.9985     | 18.8140 | 19.3252        | 21.42  |
| SSIM↑          | 0.7809       | 0.7716      | 0.7731  | 0.7822         | 0.8012 |

used instead of RFM to ensure the fairness. It can be seen that simply combining Haar wavelet transform with CNN to enhance underwater images is difficult to obtain better results compared to with RFM-Haar, as the PSNR and SSIM values in Exp3 is worse than in Exp4. RFM-Haar results in an improvement in terms of enhancement ability and robustness.

In Exp2 in Tab.V, Haar wavelet and RFM are not used. Within Exp2 and Exp3, It can be seen that simply combining Haar can even lead to degradation of CNN, as the PSNR and SSIM values in Exp3 is worse than in Exp2.

Exp1 and Exp4 in Tab.V are conducted to verify the effectiveness of CBAM module on the network. In HIFI-Net, CBAM is used as a supplementary to the RFM-Haar. Exp4 shows improvement compared to Exp1. It can be seen that using RFM-Haar in combination with CBAM in underwater image enhancement is a reasonable choice.

Tab.VI shows the ablation experiments on the components within the RFU and the residual blocks used to connect the RFU. It can be seen that both metrics of the network decreased after removing residual or any component in RFU.

REFERENCES

[1] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “Cbam: Convolutional block attention module,” in Proceedings of the European conference on computer vision (ECCV), pp. 3–19, 2018.

[2] C. Li, J. Guo, and C. Guo, “Emerging from water: Underwater image color correction based on weakly supervised color transfer,” IEEE Signal processing letters, vol. 25, no. 3, pp. 323–327, 2018.

[3] K. He, J. Sun, and X. Tang, “Single image haze removal using dark channel prior,” IEEE transactions on pattern analysis and machine intelligence, vol. 33, no. 12, pp. 2341–2353, 2010.

[4] S. Mohan and P. Simon, “Underwater image enhancement based on histogram manipulation and multiscale fusion,” Procedia Computer Science, vol. 171, pp. 941–950, 2020.

[5] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao, “An underwater image enhancement benchmark dataset and beyond,” IEEE Transactions on Image Processing, vol. 29, pp. 4376–4389, 2019.

[6] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, “Adaptive histogram equalization and its variations,” Computer vision, graphics, and image processing, vol. 39, no. 3, pp. 355–368, 1987.

[7] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, “Fast and accurate image super-resolution with deep laplacian pyramid networks,” IEEE transactions on pattern analysis and machine intelligence, vol. 41, no. 11, pp. 2599–2613, 2018.

[8] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” IEEE transactions on image processing, vol. 13, no. 4, pp. 600–612, 2004.

[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, pp. 694–711, Springer, 2016.

[10] M. J. Islam, P. Luo, and J. Sattar, “Simultaneous enhancement and super-resolution of underwater imagery for improved visual perception,” arXiv preprint arXiv:2002.01155, 2020.