USWformer: Efficient Sparse Wavelet Transformer for Underwater Image Enhancement

Supplementary Material

Overview of Supplementary Material

The supplementary material includes:
Sec. 1. Detailed Explanation of Loss Functions
Sec. 2. More Qualitative Results on Various Underwater Image Enhancement Datasets

1. Detailed Explanation of Loss Functions

Training Losses

The total loss function, L_T used for the network training is defined as:

$$L_T = \lambda_1 L_1 + \lambda_2 L_2 + \lambda_3 L_3 + \lambda_4 L_4 \tag{1}$$

Where, $\lambda_{1,2,3,4} \in (2, 3, 1, 2.5)$ are set empirically. The components of the loss functions include Perceptual loss (L_1), Charbonnier loss (L_2), Multi-Scale Structural Similarity Index (MS-SSIM) loss (L_3), and Gradient loss (L_4).

Perceptual Loss (L_1):
Perceptual loss measures the perceptual similarity between generated and target images by utilizing feature representations from a pre-trained neural network. This approach has been shown to improve the quality of generated images across various image-generation tasks. Let O represent the target image and G_t represent the generated image. Using a pre-trained VGG19 [2] network (ϕ_i) we extract feature maps at different layers. The perceptual loss, L_1, is calculated as the difference between the feature maps of the target and generated images:

$$L_1 = \sum_{i=1}^{N=4} \|\phi_i(O) - \phi_i(G_t)\|_2^2 \tag{2}$$

Here, ϕ_i represents the feature extraction function at layer i of the CNN, and $(N = 4)$ is the total number of layers considered for perceptual loss calculation.

Charbonnier loss (L_2):

Training the network with MSE loss often results in blurry reconstructions because it maximizes the log-likelihood of a Gaussian distribution. To address this issue, we chose the Charbonnier loss, a differentiable version of the L_1 norm. The Charbonnier loss is computed between the restored images (O) and their corresponding ground-truth images (G_t), and it is defined as follows:

$$L_2 = \mathbb{E}_{O \sim Q(O), G_t \sim Q(G_t)} \sqrt{(O - G_t)^2 + \epsilon} \tag{3}$$

where, $Q(O)$ and $Q(G_t)$ are the distributions of the restored image (O) and the ground-truth image (G_t), respectively. Additionally, the value of ϵ is empirically set to 1×10^{-3}.

MS-SSIM loss (L_3):

The Structural Similarity (SSIM) loss primarily addresses a single input resolution. In contrast, the Multi-Scale SSIM (MS-SSIM) loss provides greater flexibility by taking into account different input resolutions.

$$L_3 = 1 - (MSSSIM(O, G_t)) \tag{4}$$

Gradient loss (L_4):

Generally, the Charbonnier loss prioritizes low-frequency components. However, when training the network to incorporate high-frequency details, the gradient loss becomes crucial. This second-order loss function enhances the sharpness of edges in the output [1]. Here, $\hat{G_O}$ and $\hat{G_{G_t}}$ represent the distributions of $Q(O)$ and $Q(G_t)$ respectively.

$$L_4 = \mathbb{E}_{\hat{G_O}\sim Q(O), \hat{G}_O \sim Q(G_t)} \left\| \hat{G}_{G_t} - \hat{G}_O \right\|_1 \tag{5}$$

2. More Qualitative Results on Various Underwater Image Enhancement Datasets

Please see Figure S 1, 2 for more qualitative results on various underwater datasets.

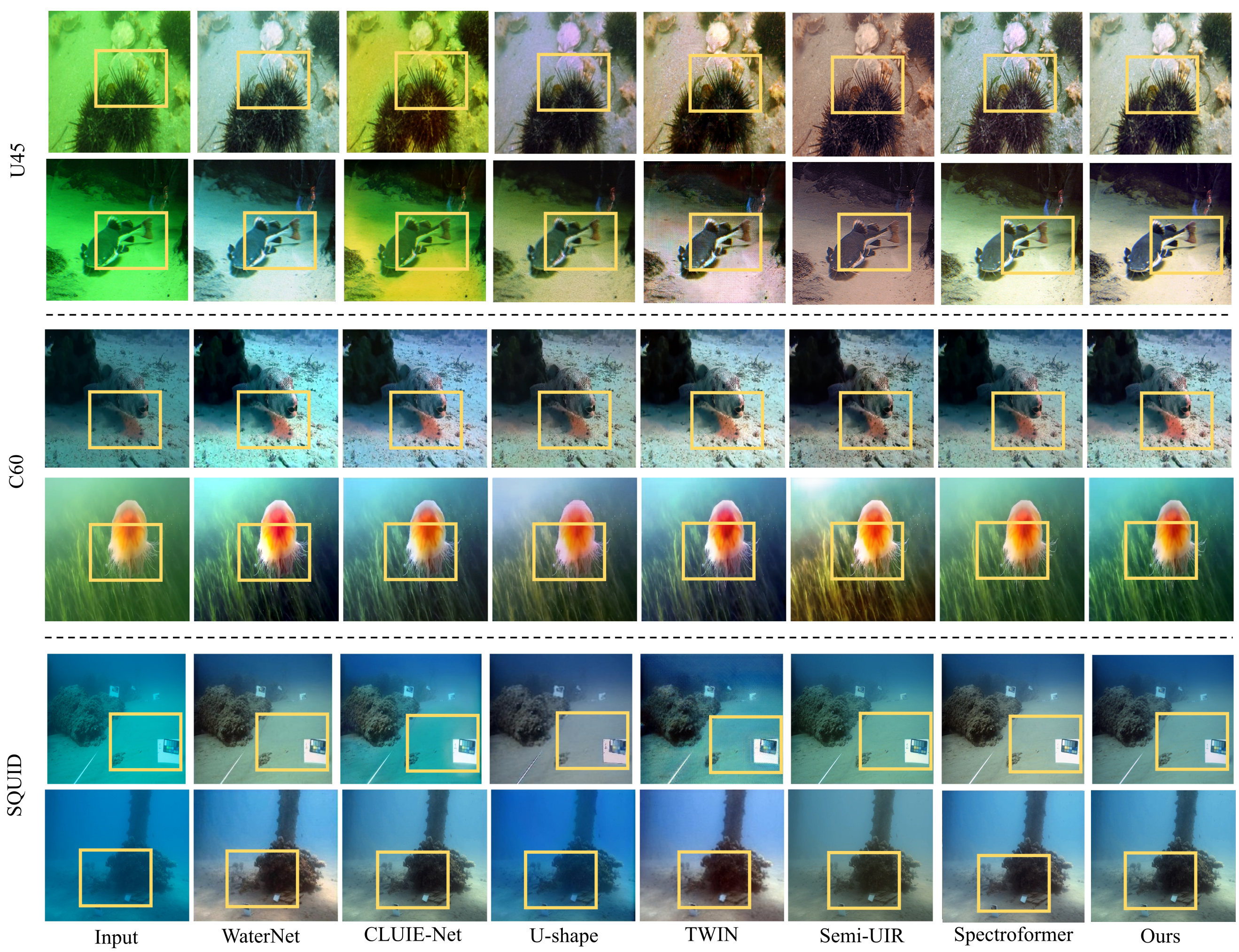

Figure S 1. Visual results on real-world datasets.

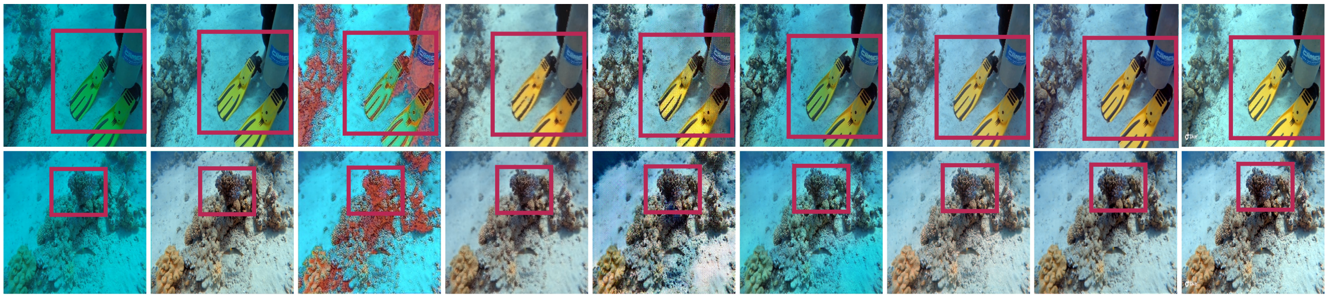

Figure S 2. Visual results on synthetic UIEB dataset.

References

[1] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015. 2

[2] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1