Enhancing underwater video from consecutive frames while preserving temporal consistency

Hu, K. ORCID: https://orcid.org/0000-0001-7181-9935, Meng, Y. ORCID: https://orcid.org/0000-0001-6901-8282, Liao, Z. ORCID: https://orcid.org/0009-0006-4686-3436, Tang, L. ORCID: https://orcid.org/0009-0003-2401-8520 and Ye, X. ORCID: https://orcid.org/0000-0002-1628-2060 (2025) Enhancing underwater video from consecutive frames while preserving temporal consistency. Journal of Marine Science and Engineering, 13 (1). 127. ISSN 2077-1312 doi: https://doi.org/10.3390/jmse13010127 Available at https://centaur.reading.ac.uk/120398/

Journal of *Marine Science and Engineering*

MDPI

Article

Enhancing Underwater Video from Consecutive Frames While Preserving Temporal Consistency

Kai Hu [1,2,*], Yuancheng Meng [1], Zichen Liao [1,3], Lei Tang [4] and Xiaoling Ye [1,2]

[1] School of Automation, Nanjing University of Information Science and Technology, Nanjing 210044, China; 202212490639@nuist.edu.cn (Y.M.); fj808642@student.reading.ac.uk (Z.L.); xyz.nim@163.com (X.Y.)
[2] Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET), Nanjing University of Information Science and Technology, Nanjing 210044, China
[3] University of Reading, Whiteknights, P.O. Box 217, Reading, Berkshire RG6 6AH, UK
[4] Information and Telecommunication Branch, State Grid Jiangsu Electric Power Company, Nanjing 211125, China; tanglei@js.sgcc.com.cn
* Correspondence: 001600@nuist.edu.cn

Abstract: Current methods for underwater image enhancement primarily focus on single-frame processing. While these approaches achieve impressive results for static images, they often fail to maintain temporal coherence across frames in underwater videos, which leads to temporal artifacts and frame flickering. Furthermore, existing enhancement methods struggle to accurately capture features in underwater scenes. This makes it difficult to handle challenges such as uneven lighting and edge blurring in complex underwater environments. To address these issues, this paper presents a dual-branch underwater video enhancement network. The network synthesizes short-range video sequences by learning and inferring optical flow from individual frames. It effectively enhances temporal consistency across video frames through predicted optical flow information, thereby mitigating temporal instability within frame sequences. In addition, to address the limitations of traditional U-Net models in handling complex multiscale feature fusion, this study proposes a novel underwater feature fusion module. By applying both max pooling and average pooling, this module separately extracts local and global features. It utilizes an attention mechanism to adaptively adjust the weights of different regions in the feature map, thereby effectively enhancing key regions within underwater video frames. Experimental results indicate that when compared with the existing underwater image enhancement baseline method and the consistency enhancement baseline method, the proposed model improves the consistency index by 30% and shows a marginal decrease of only 0.6% in enhancement quality index, demonstrating its superiority in underwater video enhancement tasks.

Keywords: underwater video enhancement; underwater image enhancement; optical flow prediction; temporal consistency

Academic Editor: Weicheng Cui

Received: 20 December 2024
Revised: 9 January 2025
Accepted: 10 January 2025
Published: 12 January 2025

Citation: Hu, K.; Meng, Y.; Liao, Z.; Tang, L.; Ye, X. Enhancing Underwater Video from Consecutive Frames While Preserving Temporal Consistency. *J. Mar. Sci. Eng.* **2025**, *13*, 127. https://doi.org/10.3390/jmse13010127

1. Introduction

Visual information is a critical resource that underwater robots can easily acquire, playing a key role in exploring and perceiving underwater environments. However, due to the numerous uncertainties inherent in aquatic environments and the absorption and scattering effects of water on light, the quality of raw underwater video footage often deteriorates significantly. These low-quality videos fail to meet human visual standards, impairing subsequent deep learning-based tasks such as video segmentation [1,2], object detection [3,4], multi-agent systems [5,6], image detection and classification [7,8], 3D image reconstruction [9], and medical image analysis [10].

With advances in underwater video capture and data communication technologies, real-time transmission of underwater videos is now feasible. Compared to underwater images, underwater videos offer greater potential for applications such as marine exploration owing to their spatiotemporal information and motion features. However, similar to underwater images, underwater videos are frequently affected by color distortion, edge blurring, low contrast, and uneven illumination due to optical constraints. Moreover, the influence of water currents on video capture equipment can weaken or even obscure texture features and details of moving objects. These issues severely hinder the ability of underwater video systems to accurately capture scene and object features.

Compared to underwater image enhancement, underwater video enhancement is inherently more complex, and is still in the early stages of research and development. When existing underwater image enhancement methods are directly applied to underwater video processing, they often result in temporal inconsistencies such as flickering and artifacts in the video. This is primarily because most underwater video enhancement techniques are straightforward extensions of single-frame image enhancement algorithms. In such approaches, each frame is enhanced independently before being combined into a complete video. Due to the absence of temporal modeling and processing between video frames, the enhanced frames lack coherence, fail to preserve temporal continuity effectively, and result in temporal artifacts and frame flickering.

Moreover, the complexity of underwater video scenes presents challenges for achieving temporal consistency, as current methods struggle to accurately extract features that address issues such as uneven illumination and dynamic changes in underwater environments. Underwater videos exhibit more intricate characteristics than terrestrial videos. Underwater scenes typically span multiple scales (e.g., detailed close-up objects versus distant backgrounds), yet many temporal consistency models process only single-scale motion information, which limits their ability to effectively fuse multi-scale features. Consequently, existing methods fail to capture both global information and detailed features in complex underwater videos, resulting in suboptimal enhancement outcomes.

This study aims to improve the temporal consistency of enhanced underwater videos while also enhancing the quality of video frames through training on underwater video frame images. Inspired by low-light video enhancement techniques [11], optical flow is introduced to simulate variations between adjacent frames in the absence of real video data. By predicting and applying optical flow, the model replicates motion trends across consecutive frames, ensuring temporal consistency in the enhancement process. However, a single U-Net model was found to be insufficient for complex underwater scenes, as it cannot effectively capture both global information and detailed features in underwater videos, leading to subpar enhancement performance. To address this limitation, we propose an underwater feature fusion module (WFM) and integrate it with a U-Net model to construct a novel underwater video enhancement network (WVEnet). WFM employs max pooling and average pooling to separately extract local and global features while using an attention mechanism to adaptively adjust the weights of different regions in the feature map, thereby enhancing key areas within video frames. Compared to traditional models based on a single U-Net, WVEnet achieves more effective feature enhancement at each layer of feature extraction, provides better management of issues such as uneven lighting and color distortion in underwater video frames, and reduces over-enhancement and artifacts that may arise in traditional U-Net models.

The effectiveness of this method was rigorously validated through extensive experiments. Experimental results on both synthetic and real-world data indicate that the proposed approach outperforms existing single-frame image enhancement methods and yields comparable results to video-based enhancement techniques. This demonstrates

that the proposed method effectively mitigates the issue of frame-to-frame flickering in underwater videos even without relying on real video data.

The primary contributions of this paper are as follows:

(1) To address the challenge of acquiring paired underwater video frames, clear 4K underwater video frames were extracted from publicly available YouTube videos. These frames were carefully selected to encompass a variety of scenes, water types, and natural lighting conditions. Additionally, a pretrained unsupervised CycleGAN network was employed to degrade unpaired underwater data, generating degraded underwater video frames with varying styles.

(2) An underwater feature fusion module was designed that integrates multiple convolutional layers, pooling operations, and feature fusion techniques. This module progressively refines input features, enhancing the model's capacity to understand and represent the data. Furthermore, this underwater feature fusion module was combined with a U-Net network to construct a novel network model. The network effectively retains important feature information during the downsampling process, excelling at handling complex features in underwater video frames such as light refraction, color distortion, and edge blurring. Through layer-by-layer enhancement, the model improves temporal consistency in single-frame inference, ensuring smooth transitions between frames and reducing flickering and temporal jitter.

(3) Motion between video frames was predicted using optical flow, then this information was applied to adjacent frames, ensuring smooth transitions and reducing flickering in video processing. The effectiveness of the method was validated through extensive comparative and ablation experiments, and the standard deviation of histograms for each video frame was used to visually demonstrate the results of the temporal consistency comparison.

2. Related Work

This section reviews and analyzes previous research on underwater visual enhancement.

2.1. Traditional Underwater Enhancement Methods

Common traditional underwater enhancement methods include Histogram Equalization (HE) and Retinex theory. Histogram Equalization (HE) [12] enhances the visual quality of an image by transforming its histogram from a narrow unimodal distribution to a more balanced one. In underwater imaging, Iqbal et al. [13] proposed an unsupervised color correction method (UCM) that combines color correction and selective histogram stretching to eliminate blue bias and enhance the brightness of the low-intensity red channel. However, unsupervised color correction methods often struggle to restore colors accurately without a thorough understanding of underwater light propagation characteristics, leading to oversaturation or color distortion in the corrected images. Ahmad et al. [14,15] introduced an adaptive histogram enhancement method that uses Rayleigh stretch contrast enhancement to improve image contrast, enhance details, and reduce over-saturated areas. However, while improving low-contrast regions, it is possible that noise may be amplified and artifacts may be introduced along edges, resulting in deterioration of image quality.

Retinex theory is based on color constancy, and aims to both eliminate the influence of lighting components on object colors and remove illumination unevenness to reveal the true color of the scene. Joshi et al. [16] applied Retinex theory to underwater images to enhance degraded images. Although Retinex enhancement improves brightness significantly, color distortion or shift often occurs. To address this issue, Mercado et al. [17] introduced Multiscale Retinex with Reverse Color Loss, incorporating color recovery and reverse color loss strategies. Li et al. [18] combined the MSRCR algorithm with a histogram

quantization-based color channel correction method. However, because MSR involves multiple parameters, their selection significantly influences the enhancement outcome, requiring tuning for different scenes and complicating the processing.

2.2. Deep Learning-Based Underwater Enhancement Methods

In deep learning-based underwater enhancement methods, Perez et al. [19] were the first to establish a paired dataset of degraded and clear underwater images. They utilized deep learning techniques to learn the mapping relationship for underwater image enhancement; however, this straightforward architecture proved inadequate for handling complex lighting variations and various image degradation phenomena in underwater environments.

Sun et al. [20] introduced a pixel-to-pixel deep learning model for underwater image enhancement. The model employs convolutional layers as encoders to filter noise and utilizes deconvolutional layers as decoders to recover lost details, optimizing the image pixel by pixel. While this method preserves low-level features through skip connections, artifacts or over-enhancement may occur during detail enhancement, particularly under high-noise levels or in complex scenes, which becomes more pronounced in such environments.

Li et al. [21] proposed the UWCNN method, which integrates physical models with deep learning by training a convolutional neural network on a synthetic dataset to enhance underwater images and videos. However, the physical model in UWCNN simplifies the complexities of light transmission in underwater environments and may fail to capture all underwater optical phenomena, leading to discrepancies in the enhanced image's color compared to the real scene. Furthermore, some fine image details may be erroneously identified as noise and removed, resulting in image distortion. In terms of video enhancement, UWCNN does not address temporal consistency between consecutive frames, which may lead to inconsistencies in video enhancement across frames, resulting in flickering or color jumps and affecting the overall visual experience.

Fabbri et al. [22] proposed a network called UGAN based on the use of Generative Adversarial Networks (GANs) to enhance the visual quality of underwater images. UGAN initially uses CycleGAN to generate distorted underwater images paired with relatively clear underwater images. In this way, it addresses the issue of insufficient real training data while incorporating gradient penalties to improve the stability of the generator and the quality of the generated images. Although UGAN's U-Net architecture preserves image details, it may result in over-smoothing of details or inadequate noise removal in certain cases, especially in images with rich details or high noise levels. When processing consecutive frames, it is possible for color inconsistencies, flickering, or other visual discontinuities to occur between frames.

Islam et al. [23] introduced a fast underwater enhancement model called FUnIE-GAN which utilizes the absolute error loss as the global loss and employs a pretrained VGG-19 network to extract high-level features for content loss. This algorithm performs well in color restoration and is computationally efficient. However, due to its inability to handle temporal sequences, frame consistency issues may arise when enhancing consecutive frames, in turn leading to flickering, color jumps, and other phenomena in the video and affecting the visual experience and stability.

Hu et al. [24] incorporated the Natural Image Quality Assessment (NIMA) metric into the GAN context to generate underwater images with higher contrast and enhance their visual appeal. However, this method still does not address the temporal instability between consecutive frames.

Tang et al. [25] proposed a model called AttU-GAN for underwater image enhancement. This approach integrates an attention gate mechanism into the U-Net architecture

to filter out irrelevant features and effectively capture important image attributes such as contours, textures, and styles. Although Attention U-Net performs excellently in filtering out irrelevant features, the enhanced images may still exhibit artifacts in complex and dynamic underwater environments. Moreover, this model does not take into account the temporal consistency between frames, leading to flickering issues in the enhanced video frames, which affects the overall visual experience. Li et al. [26] introduced UDA-Net, a network that combines multiscale grid convolutional neural networks and feature-level attention mechanisms, enabling it to adaptively allocate weights to different regions within each feature map to more effectively enhance severely degraded areas. However, UDA-Net still fails to fully leverage the temporal information between consecutive frames in videos.

In the domain of enhancing temporal consistency, Lai et al. [27] proposed a video enhancement method based on deep recurrent networks. This approach uses Convolutional Long Short-Term Memory (ConvLSTM) networks in combination with both short- and long-term temporal loss functions and perceptual loss, helping to maintain temporal stability in the video and its perceptual similarity to the processed frames. While the ConvLSTM module can capture temporal dependencies, its adaptability to the complex variations in underwater features is limited, potentially leading to unstable enhancement quality.

Based on the aforementioned research, we observed that when existing underwater image enhancement techniques are directly applied to video, each frame is typically enhanced individually and then stitched together to form a new video. However, the continuity of the enhanced video frames is not well preserved, which may lead to temporal artifacts and frame-to-frame flickering. Compared to underwater image enhancement techniques, underwater video enhancement is more complex, and the research in this field is still in its early stages; most existing underwater video enhancement methods are simple extensions of single-image enhancement algorithms, and do not address the issue of temporal consistency.

Additionally, the complexity of underwater videos makes it challenging for general temporal consistency enhancement models to effectively capture the key features of underwater video frames. Therefore, we train an image-based model using image data and implicitly embed optical flow information during the training process to ensure temporal consistency. Furthermore, we propose a novel feature fusion module that utilizes various convolutional layers, pooling operations, and feature fusion techniques to extract and process input features. The module also uses attention mechanisms to weigh the important regions of the input feature maps in order to effectively capture different types of degradation features in underwater images.

3. Method

Common underwater image enhancement networks primarily focus on improving the quality of individual images or frames while overlooking the consistency between consecutive frames; as a result, although the visual quality of individual frames may be enhanced, substantial discrepancies can still exist between consecutive frames, leading to jumps or flickering in the generated video. For example, the instability observed during the training of unsupervised GAN networks often results in the generated video frames lacking temporal consistency.

Horn and Schunck proposed that optical flow can effectively simulate motion changes between adjacent frames in video sequences [28]. An ideal temporal stability model should maintain consistency during transformations, meaning that the output processed by the model should exhibit the same transformation effect as the original regardless of the input's transformation. Models exhibiting this temporal stability property can process video frames continuously while preventing flickering issues. Based on this principle, we employ optical

flow sequences to simulate real-world motion in videos. Underwater video frame pairs are input into the network and optical flow is used to enforce consistency between the outputs of consecutive frames, effectively guiding the network to maintain temporal stability.

However, due to the complexity of underwater video scenes (e.g., light scattering, water body characteristics, and dynamic objects) and the intricate motion patterns between moving objects and the background, existing temporal stability models face significant challenges in extracting underwater features. Traditional optical flow or basic temporal consistency loss functions may be insufficient for handling these complex scenarios, as they often fail to capture the unique illumination inconsistencies and degradation features inherent in underwater environments.

Based on this, we designed a novel feature fusion module called WFM. This module leverages both max pooling and average pooling to extract distinct types of features, enhancing key local details through an attention mechanism. Simultaneously, WFM balances the outputs of max pooling and average pooling, enabling the network to focus on local detail information while retaining global features. This is crucial for underwater video enhancement, as underwater scenes require both the accurate restoration of local details (e.g., fish or seagrass) and the preservation of the natural perceptual consistency of the background water.

Furthermore, we integrated WFM with U-Net to construct a new dual-branch underwater video enhancement network. This network addresses U-Net's limitations in handling complex multiscale features and uses layer-by-layer enhancement to ensure that the temporal sequence derived from single-frame inference is more accurate and coherent. As a result, the smooth transition between video frames is further reinforced, reducing flickering and temporal jitter.

This section introduces the overall workflow and network architecture, followed by the detailed implementation specifics.

3.1. Dataset

Obtaining paired underwater video frames is a challenging task. To address this issue, clear underwater video frames were obtained from publicly available YouTube videos. The videos selected in this study were all 1920 $\times$ 1080 and 60 Hz, and were cropped to 512 $\times$ 512. During the cropping process, the cropping position of each video frame was kept consistent. Finally, 40 underwater videos were selected, totaling more than 4000 underwater video frames covering different scenes, water types, and lighting conditions.

A pretrained unsupervised CycleGAN network was employed to degrade the unpaired underwater data, generating underwater video frames with various styles. As shown in Figure 1, prior to training the network, reasonable optical flow was predicted based on high-quality ground truth. Although pretrained instance segmentation models from existing open-source toolkits such as Detectron2 [29] provide the necessary mask predictions, their segmentation performance is suboptimal when applied to underwater video frames, resulting in subpar visual outputs. In contrast, instance segmentation models such as WaterMask [30] are specifically designed for underwater images, and yield superior segmentation results; however, they do not directly generate the mask data format required for our task. Therefore, the WaterMask model was postprocessed to preserve the instance segmentation results while converting them into the necessary mask data format. After the estimated target masks were obtained, the unsupervised model CMP [31] was used to predict the optical flow. Figure 1 illustrates examples from our dataset along with segmentation results and optical flow predictions.

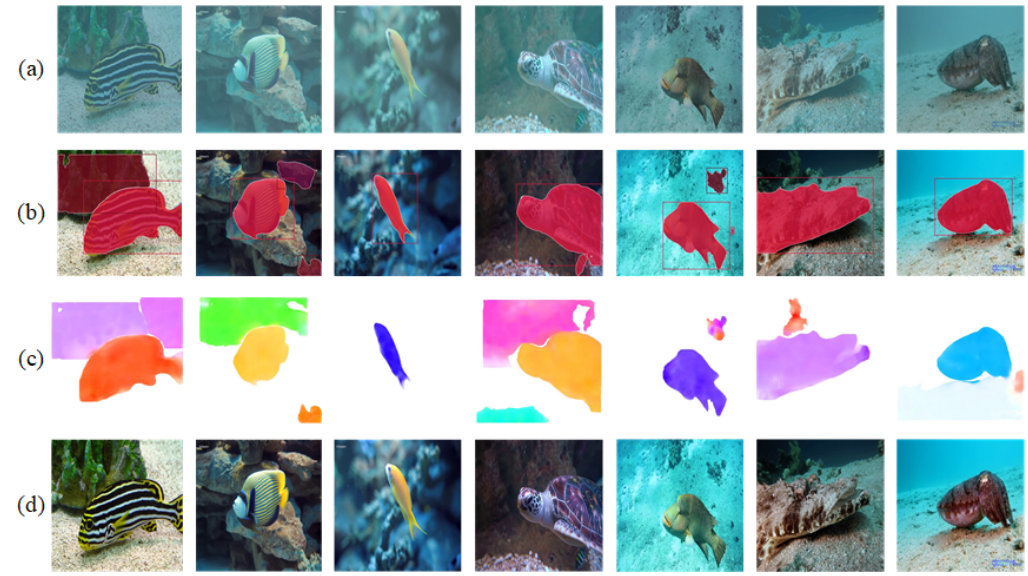

Figure 1. (**a**) Degraded simulated underwater video frame, (**b**) segmentation result, (**c**) optical flow prediction, and (**d**) ground truth.

3.2. Underwater Feature Fusion Module

As shown in Figure 2, the input to the WFM is the feature map f_{in}. First, f_{in} generates two feature maps f_{max} and f_{avg} through max pooling and average pooling operations, respectively, which capture different spatial feature information. Max pooling extracts prominent response values in spatially significant areas, while average pooling captures global average information. These two feature maps provide diverse representations for subsequent attention generation. Then, f_{max} and f_{avg} each enter two parallel convolution paths. The 1×1 convolution operation extracts compact global features, reducing the computational load and enhancing feature expression across channels, while the 3×3 convolution operation captures local spatial context information. To better integrate global and local information, these two parallel paths progressively combine features through cross-addition, making full use of the advantages of different convolution kernel sizes in feature extraction.

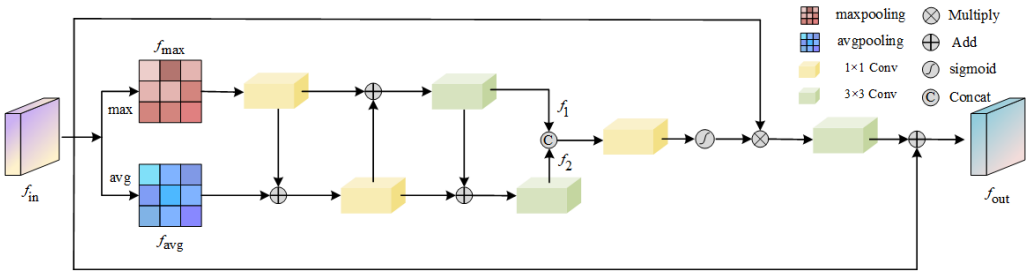

Figure 2. Underwater feature fusion module; this spatial attention mechanism is particularly important in underwater visual enhancement tasks, where it helps address issues such as uneven illumination, scattering effects, and color degradation in underwater scenes. Consequently, it significantly enhances the ability to focus on target information and improves the overall image quality.

The final fused features from the two parallel branches are f_1 and f_2, respectively; f_1 and f_2 are concatenated along the channel dimension to form a joint feature map, from which the spatial attention weight matrix is generated using the sigmoid activation function. The input feature f_{in} is then multiplied element-wise by the generated spatial attention weight matrix, thereby weighting the input feature and enhancing the feature representation of the salient areas. Next, a 3×3 convolution is applied to further enhance

the feature representation and the final output feature map f_{out} is obtained by element-wise addition with the original feature map.

The computation process for f_1 and f_2 is as follows:

$$f_1 = \text{conv}^3\{\text{conv}^1(f_{\max}) \oplus \text{conv}^1[\text{conv}^1(f_{\max}) \oplus f_{\text{avg}}]\} \quad (1)$$

$$f_2 = \text{conv}^3\{f_1 \oplus \text{conv}^1[\text{conv}^1(f_{\max}) \oplus f_{\text{avg}}]\} \quad (2)$$

where $\oplus$ denotes addition, $f_{\max}$ and f_{avg} are the features processed by global max pooling and global average pooling, respectively, and conv^1 and conv^3 represent the 1×1 convolution and 3×3 convolution operations, respectively.

Table 1 presents the detailed changes in spatial dimensions and the number of channels at each stage of the WFM.

Table 1. Dimensions and channel changes of the underwater feature fusion module.

Stage	Spatial Dimensions ($H \times W$)	Number of Channels (C)
Input Feature Map	$H \times W$	C
Average Pooling	$H \times W$	1
Max Pooling	$H \times W$	1
1×1 Convolution (Max Pooling)	$H \times W$	1
Addition (Max Pooling + Average Pooling)	$H \times W$	1
3×3 Convolution	$H \times W$	1
Addition	$H \times W$	1
Concatenation	$H \times W$	C
1×1 Convolution (Concatenated Features)	$H \times W$	2
Sigmoid Activation	$H \times W$	1
Weighted Operation (Input $\times$ Attention Weight)	$H \times W$	1
3×3 Convolution	$H \times W$	C
Residual Connection	$H \times W$	C

In this way, the model can dynamically adjust the weight distribution of the input feature map in the spatial dimension, significantly enhancing the feature expression of key areas while suppressing interference from irrelevant backgrounds. This mechanism is particularly crucial for underwater image enhancement tasks, as it helps to address issues such as uneven illumination, scattering effects, and color degradation in underwater scenes, both improving the model's ability to focus on target information and enhancing the visual quality of the images.

3.3. Underwater Video Enhancement Network

The underwater video enhancement network proposed in this paper adopts a two-stage structure. The first stage utilizes a U-Net architecture, as shown in Figure 3. This architecture is a fully convolutional structure without fully connected layers, offering fewer model parameters and faster processing speeds. Considering that U-Net may cause loss of details during the encoding–decoding process, the WFM underwater feature fusion module is introduced at each downsampling step.

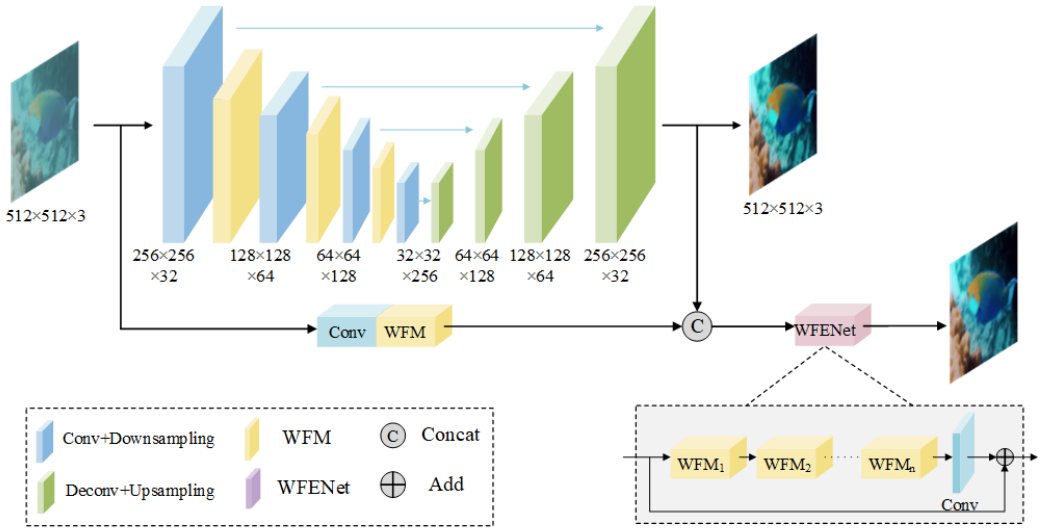

Figure 3. Underwater video enhancement network.

In the U-Net network, whenever the image generates features at different levels through the encoder, WFM fuses these features. In the encoder stage, the input underwater image is processed through convolution and pooling operations to extract features at multiple levels. The WFM performs weighted fusion on these feature maps to generate a more optimized feature representation. The weighting process can automatically adjust the importance of each feature point according to the specific characteristics of the underwater image. This process helps to suppress noise in the underwater environment while enhancing key detail areas.

The skip connections in U-Net enable low-level features to be fused with high-level features. The WFM further optimizes these fusion processes by processing multiscale feature information, allowing the network to fuse more information related to underwater environment characteristics in each skip connection. The network introduces various layers of underwater features such as color, brightness, texture, etc., to enhance both the local details and global consistency of the image, providing richer feature information for subsequent video frame reconstruction and enhancement. Additionally, by refining each layer, the temporal sequences derived from single frames become more accurate and coherent. The second stage introduces an enhancement network called WFENet, which takes the output of the first stage as input and performs further feature refinement on the video frame of the original resolution to generate images with more detailed content.

3.4. Implementation Details

An ideal temporally stable model should be able to maintain consistency of transformation, meaning that the output after model processing should exhibit the same transformation effect as the original output regardless of the input transformation. Only models that adhere to this principle will avoid flickering artifacts when processing videos frame by frame. Based on this condition, in this paper we attempt to use the motion information generated by optical flow to simulate the actual video sequence and enforce output consistency before and after distortion, thereby helping the network learn temporal stability.

Optical flow refers to the motion vector of each pixel in the image sequence. It describes the movement of objects between consecutive frames, and can simultaneously capture both global and local motion information. In underwater video enhancement, optical flow is used to capture motion information between frames. In this paper, we utilize this motion information to enhance temporal consistency and reduce temporal artifacts.

As shown in Figure 1, the Watermask segmentation model is first employed to separate the target from the background and obtain the target mask, after which it randomly samples ten guided motion vectors on each target area. The clear underwater video frames and motion vectors are input to the CMP optical flow prediction network. After providing the estimated object mask, optical flow predictions can be obtained through the CMP. Specifically, when predicting the ground truth optical flow through the pretrained CMP model, some guided motion vectors on the target are needed for initialization, and WaterMask can assist in obtaining these motion vectors:

$$f = CMP(y, V) \tag{3}$$

where CMP denotes the optical flow prediction model, y represents the ground truth video frame, and V represents the guiding motion vectors.

In this paper, we randomly extract ten guiding vectors for each object in the image to obtain the final prediction. Although randomly sampled guiding vectors cannot guarantee the quality of the optical flow prediction, and may even generate completely opposite motion vectors, the introduction of such interference helps to improve the robustness of the training.

The motion changes between video frames in dynamic scenes are represented by optical flow, and the images are warped accordingly to simulate adjacent frames. Given the original image and deformed image pair, the approach in this paper adopts a Siamese network for training, inputs them into the network one-by-one, and imposes consistency constraints between the output results, which helps the model to maintain stability in the temporal dimension. Based on the original optical flow prediction, various optical flow scenarios can be enhanced by changing the direction and position. According to the results, the deformed image can be obtained by the following method:

$$x_2 = W(x_1, f) \tag{4}$$

where f represents the predicted optical flow, x_1 denote the original image, and x_2 denotes the deformed image.

As shown in Figure 4, after preparing the necessary optical flow, image-based model is trained using a Siamese network. At the input layer of the network, the optical flow and image features are combined by weighted splicing. The optical flow map (motion vector of each pixel) is spliced with the deep features of the video frame to form a composite input containing temporal and spatial information. Different branches of the network process the spatial information of the image and the temporal information of the optical flow simultaneously, maintaining consistency in the temporal dimension.

In the upper branch, the degraded underwater video frame x_1 is input into the network $g(\cdot)$, producing an enhanced result $g(x_1)$ under the supervision of a clear underwater video frame (ground truth) y_1. To provide additional temporal information, the randomly generated optical flow f derived from the ground truth clear video frame is applied to warp the input image x_1. The warped image x_2 is then used as input to the lower branch, yielding the output $g(x_2)$, which is compared with the corresponding optical flow-warped ground truth y_1 for supervised training. The two branches share training weights. Finally, the same optical flow f is applied to warp the output $g(x_1)$ into $W(g(x_1), f)$, which is then compared with $g(x_2)$.

During the training process, optical flow information is used both as an input feature to guide video enhancement and as a training target for the supervisory signal optimization network. To maintain temporal consistency using optical flow information, this paper employs a loss function that includes the optical flow error. By employing optical flow as a

temporal consistency constraint, artifacts and blurring caused by temporal inconsistency can be reduced. All losses are computed using l_1 loss, and the total training loss for the network is defined as a combination of the enhancement loss $\mathcal{L}_e$ and the consistency loss $\mathcal{L}_c$:

$$\mathcal{L} = \mathcal{L}_e + \lambda \mathcal{L}_c \tag{5}$$

where λ is the weight that balances the constraints of the two loss components. We discuss the optimal value of λ in the subsequent ablation study. The enhancement loss $\mathcal{L}_e$ and consistency loss $\mathcal{L}_c$ can be expressed as follows:

$$\mathcal{L}_e = \sum_{i=1,2} \| g(x_i) - y_i \|_1 \tag{6}$$

$$\mathcal{L}_c = \| W(g(x_1), f) - g(x_2) \|_1 \tag{7}$$

where $g(\cdot)$ denotes the network convolution operation, x_i and y_i represent the i-th channel of the input and ground truth, respectively, and f is the optical flow generated for motion simulation.

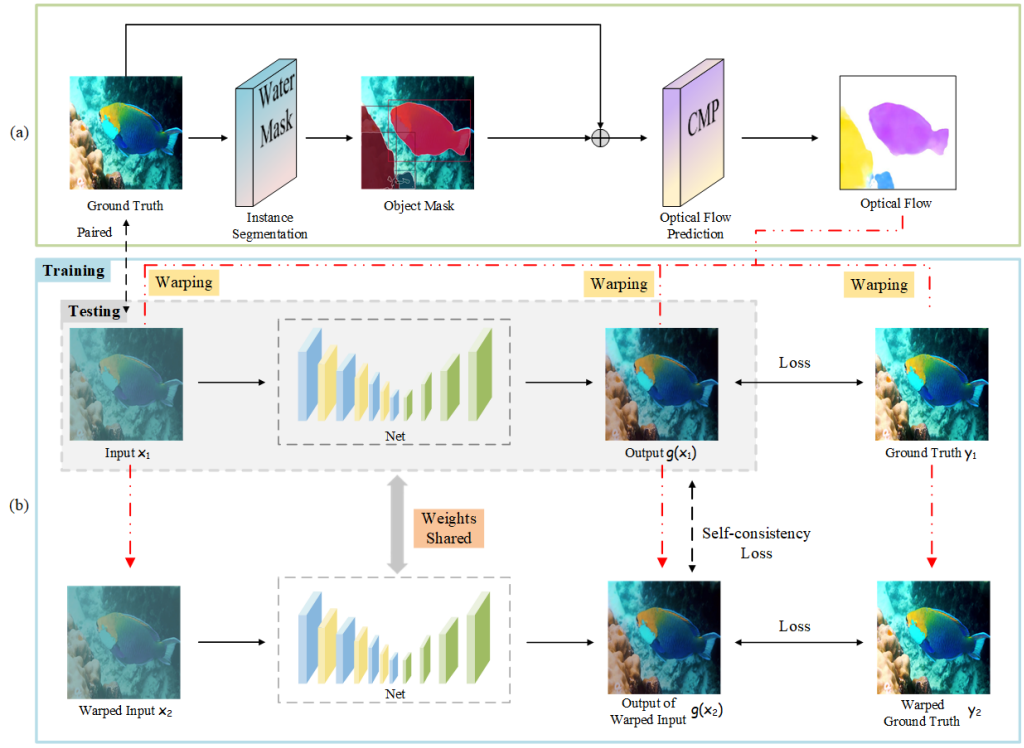

Figure 4. Overview of the full pipeline, consisting of two steps: (**a**) during optical flow prediction, Watermask is used to separate the object from the background and ten guided motion vectors are randomly sampled from each object area, after which the guided motion vectors and the clear underwater video frame are input to the CMP optical flow prediction model to obtain the predicted optical flow; (**b**) during training and testing, the network consists of two branches. The upper branch functions in both the training and testing phases, while the lower branch serves as an auxiliary branch used only during training to enforce temporal consistency. Images in the second branch are warped from those in the main branch using the same optical flow. During the testing phase, the network can directly take the input and predict the output without requiring optical flow prediction.

The visualization of our optical flow prediction results is shown in Figure 1c.

4. Experiments

The effectiveness of the proposed method in underwater video enhancement was validated through experiments focusing on both image quality and temporal consistency. This section first describes the equipment and environment used in the experiments, followed by a comparative study of several widely used underwater image enhancement algorithms. The visual effects, quality metrics and consistency metrics of the enhanced images are analyzed; additionally, the proposed method is compared with other algorithms that incorporate temporal consistency enhancement and the corresponding quality and consistency metrics are evaluated. Finally, ablation experiments are conducted to analyze the performance of the proposed algorithm and validate the effectiveness of each module.

4.1. Experimental Setup

The proposed algorithm was implemented on a Windows 10 operating system. The software environment included Python 3.9, and the network model was constructed and trained using the PyTorch 1.11 framework. The model was trained using the Adam optimizer with default parameters for 50 epochs on a single Nvidia RTX 3080 12 G GPU, with the learning rate set to 1×10^{-4} and a batch size of 1. Both the comparative experiments and the ablation study of the network model were evaluated using both subjective analysis and objective metrics to assess the effectiveness of the proposed method.

4.2. Comparative Experiments

Rather than being specifically designed for video enhancement, most existing underwater video enhancement methods are adaptations of single-image enhancement algorithms; therefore, in this paper we compare the proposed method with image-based enhancement methods and postprocessing enhancement methods utilizing temporal consistency. Six algorithms were selected from both categories: the traditional image enhancement method Multi-Scale Retinex (MSR) [32]; the deep learning-based image enhancement methods UGAN [22], UWCNN [21], FunieGAN [23], and SGUIENet [33]; and the postprocessing enhancement method utilizing temporal consistency proposed by Lai et al. [27] (referred to as BLIND in the evaluation metrics).

This paper uses the Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM), and UIQM no-reference-image quality assessment metric to evaluate the quality of the enhanced video frames. Additionally, the Learned Perceptual Image Patch Similarity (LPIPS) [34] and Warping Error (E_{warp}) [35] are used to assess the temporal consistency of the models.

The LPIPS is more reflective of human visual perception consistency than pixel-level metrics such as te MSE and PSNR. The smaller the LPIPS value, the smaller the pixel differences in the image and the less perceptible these differences are to the human visual system. The LPIPS metric is calculated by first computing the LPIPS between the second and third frames, then between the third and fourth frames, and so on until the last frame. The average LPIPS value is then obtained by summing all computed LPIPS values and dividing by the total number of comparisons, which eliminates randomness and helps to observe whether the flickering issue has been alleviated from a global perspective. The Warping Error is used to assess the similarity between the distorted image obtained via optical flow and the target image. A smaller warping error indicates smaller differences between adjacent frames.

To more clearly demonstrate the flickering issue in underwater video frames generated by different algorithms, the effectiveness of the proposed method in enhancing temporal consistency was validated using a multi-frame histogram overlay and calculating the standard deviation. Specifically,the RGB histograms of each frame are overlaid on the

same plot to visually display the variation in pixel value distributions between frames. Each channel (R, G, B) is represented by a different color line and the histograms of each frame are overlaid using semi-transparent lines to facilitate the observation of differences between frames. A statistical analysis is performed on each channel and the standard deviation is computed for each pixel value, which quantifies the fluctuations between frames. A higher standard deviation indicates greater fluctuation in the pixel value, leading to more noticeable flickering. The standard deviation curve quantifies the fluctuation of each pixel value between video frames. If the standard deviation of a pixel value is large, this indicates significant variation in that pixel value across different frames, meaning that it is more likely to exhibit flickering. The standard deviation curve is calculated using the following formula:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{8}$$

where x_i represents the pixel count of a specific pixel value in the histogram of each frame, μ is the mean of this pixel value across all frames, and N is the total number of frames.

As shown in Figure 5, all of the compared algorithms demonstrate significant enhancement effects on the input underwater images. As a traditional enhancement algorithm, MSR mitigates some color bias but suffers from overexposure, often due to the need for properly adjusted parameters to adapt to different underwater environments. UWCNN eliminates the blue–green color shift, but results in an overall darker image. The images enhanced by the SGUIENet algorithm exhibit noticeable flickering issues and still contain some blue–green color bias. Among the GAN-based algorithms, blurring and artifacts are commonly observed; UGAN generates severe artifacts and flickering, while FunieGAN effectively addresses color bias and brightness issues but still exhibits artifacts and significant flickering. GAN-enhanced images show blurred backgrounds with overlooked object edge details, resulting in insufficient clarity. BLIND performs well in mitigating flickering issues, but fails to effectively remove the blue–green color bias during underwater video frame enhancement, and the detailed information such as edges remains somewhat blurry. Due to space limitations, additional comparison results for other scenarios are provided in Appendix A.

From the histogram standard deviation in Figure 6, it can be observed that the MSR, UWCNN, SGUIENet, UGAN, and FunieGAN algorithms exhibit significant fluctuations, with the standard deviation curves reaching particularly high values in certain regions. This indicates that these algorithms result in substantial inter-frame variations in those areas of the enhanced video frames, leading to flickering issues.

Figure 7 shows the differences in the details of each model enhancement.

In contrast, the underwater feature fusion module designed in this paper enhances the model's understanding and expressive capability of the input data by progressively refining the salient features and background information. This effectively mitigates color bias, improves contrast, compensates for the shortcomings of GAN networks in capturing image details, and strengthens edge information, thereby contributing to a better visual experience. Additionally, by performing optical flow distortion on the images to simulate adjacent frames and balancing global feature weights through WFM, the temporal stability of underwater video frames is effectively enhanced.

As shown in Table 2, the deep learning-based algorithms outperform traditional image enhancement methods in terms of image quality metrics. However, regarding the continuity and temporal stability of video frame enhancement, the traditional methods are more satisfactory compared to GAN-based and other single-image-based deep learning algorithms. This is because the continuity of video frames results in minimal changes

in the aquatic environment, allowing traditional methods to achieve better continuity. Nevertheless, the limitations of traditional algorithms significantly affect continuity when processing video frames from various aquatic environments.

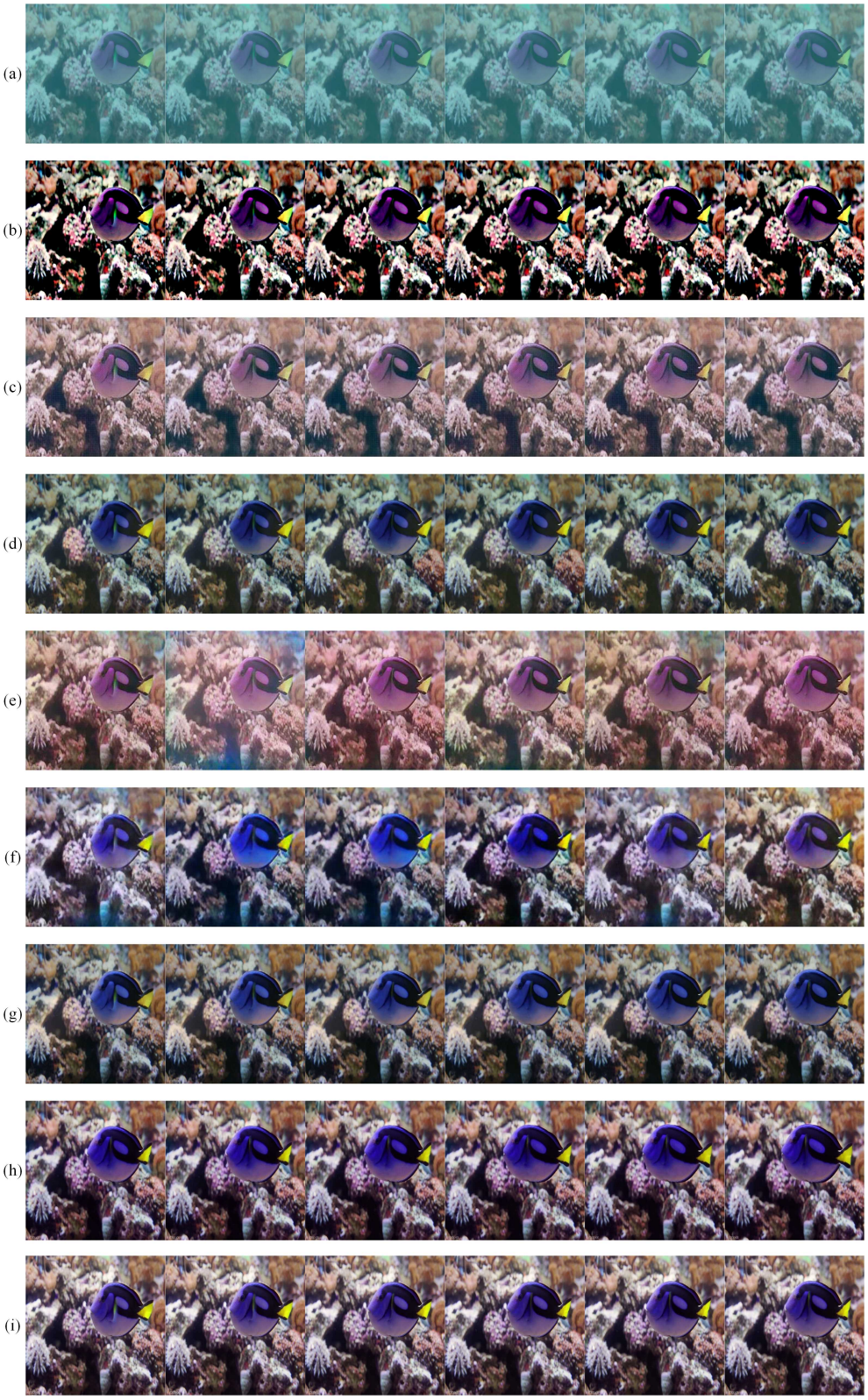

Figure 5. Comparison Experiment Scenario 1: (**a**) Input, (**b**) MSR, (**c**) UWCNN, (**d**) SGUIENet, (**e**) UGAN, (**f**) FunieGAN, (**g**) BLIND, (**h**) Ours, (**i**) GT. The six pictures in the figure are continuous video frames. They are shown together to demonstrate the enhancement results of the different models in terms of temporal consistency.

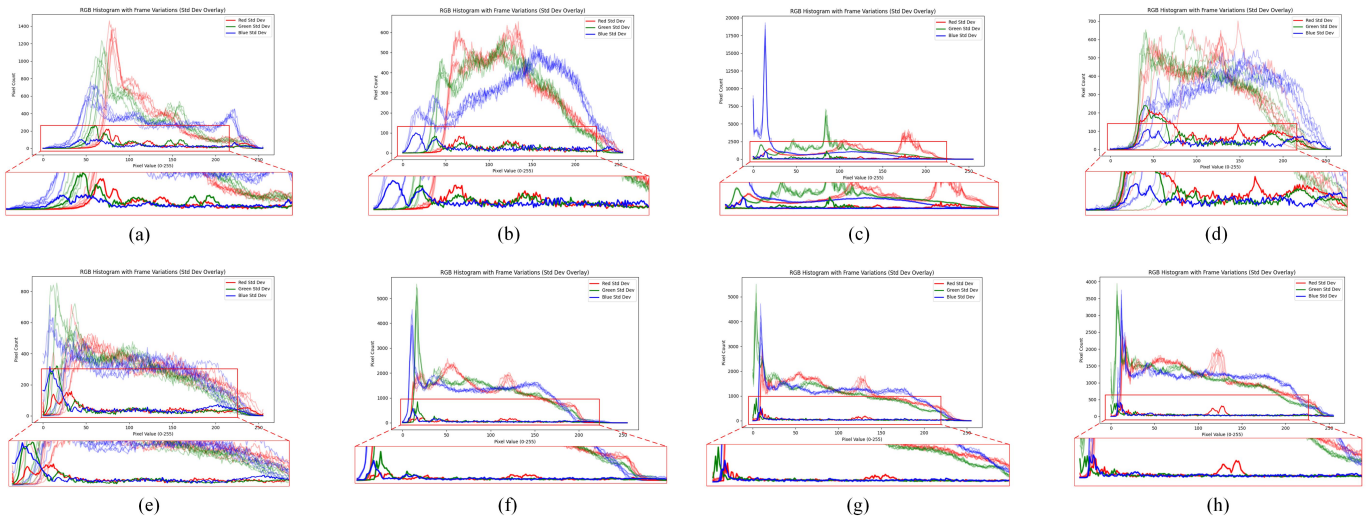

Figure 6. Histogram and standard deviation for Comparison Experiment Scenario 1: (**a**) MSR, (**b**) UWCNN, (**c**) SGUIENet, (**d**) UGAN, (**e**) FunieGAN, (**f**) BLIND, (**g**) Ours, (**h**) GT. The x-axis represents pixel values 0–255 and the y-axis represents the pixel count for each pixel value. The semi-transparent RGB histograms represent the pixel value distribution for the R, G, and B channels in each frame. The red, green, and blue thick lines represent the standard deviation curves for the R, G, and B channels, respectively.

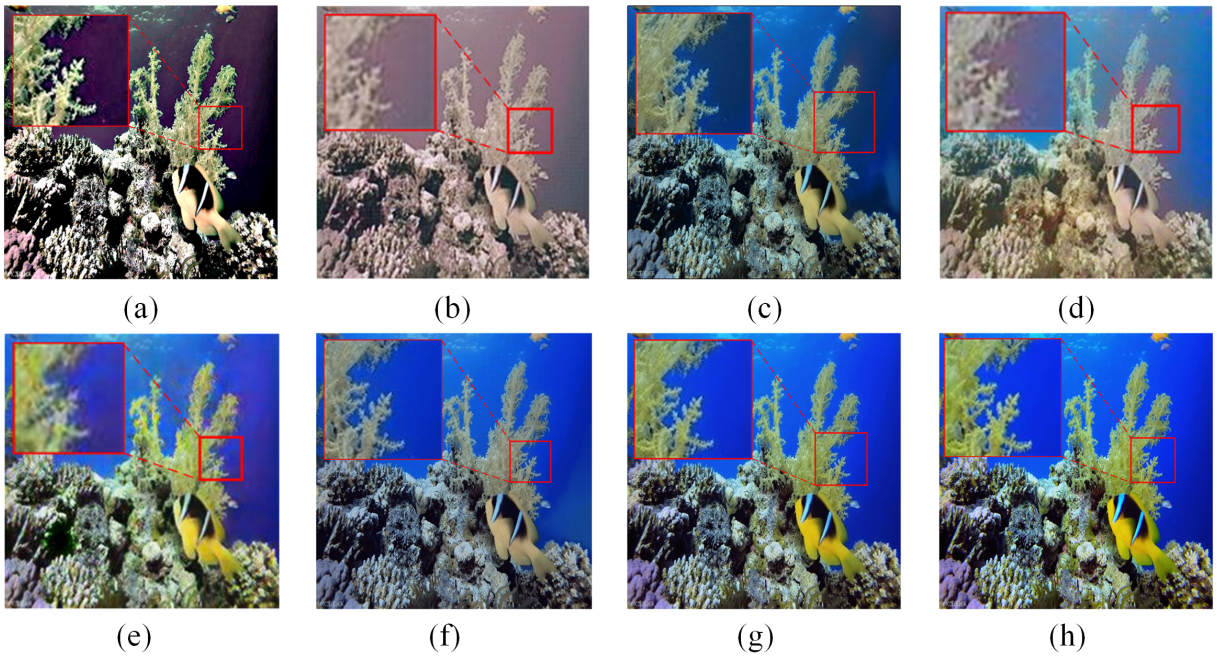

Figure 7. Comparison Experiment Scenario 1 details: (**a**) MSR, (**b**) UWCNN, (**c**) SGUIENet, (**d**) UGAN, (**e**) FunieGAN, (**f**) BLIND, (**g**) Ours, (**h**) GT. The figure shows the edge details and edge artifacts of video frames generated by different models.

Compared to FunieGAN, which performs the best in underwater single-image enhancement, our method only lags behind by 0.6% in terms of image quality metrics. However, in terms of perceptual consistency and temporal consistency, our method improves by 31% and 40%, respectively, surpassing both the traditional and deep learning-based enhancement algorithms.

Compared with single-frame image enhancement, the method in this paper significantly increases the workload. However, it is comparable to traditional image enhancement

methods in terms of computational time and efficiency, and even surpasses them in terms of speed.

Table 2. Quantitative metrics for enhancement quality and temporal stability in comparative experiments: PSNR, SSIM, and UIQM are indicators of the image quality generated by different models, LPIPS and E_{warp} are consistency indicators, and Time shows the operating efficiency of the different models. The experimental indicators of the model in this paper are in bold.

Model	PSNR ↑	SSIM ↑	UIQM ↑	LPIPS ↓	E_{warp} ↓	Time (s) ↓
MSR	17.669	0.834	2.383	0.139	0.0351	1.231
UWCNN	18.739	0.822	2.943	0.159	0.0453	0.831
SGUIENet	23.564	0.866	2.665	0.185	0.0625	0.751
UGAN	23.303	0.856	3.013	0.168	0.0622	0.341
FunieGAN	25.257	0.930	3.087	0.161	0.0621	0.092
BLIND	24.388	0.929	3.025	0.094	0.0262	1.132
Ours Model	**25.087**	**0.936**	**3.064**	**0.095**	**0.0265**	**0.391**

4.3. Ablation Experiments

4.3.1. Loss Function Weight

In underwater video enhancement, models not only need to enhance the visual quality of each frame, such as brightness, clarity, color, etc., but also need to ensure consistency between frames by avoiding flickering and artifacts. In order to find a balance between these two goals, in this paper we conducted a large number of experiments. It was found that a very small λ results in low temporal stability and leads to artifacts, whereas a very large λ increases computational costs, causing diminishing returns in enhanced image quality.

As shown in Figure 8 and Table 3, our extensive experiments revealed that the optimal parameter setting should be around $\lambda = 5$. Specifically, as the branch weight increases, the network's temporal stability improves compared to networks with smaller weights, while PSNR and SSIM also show improvements. However, after the weight exceeds a certain threshold, the benefits in image quality start to diminish, and further increases in temporal stability lead to a decrease in PSNR and SSIM.

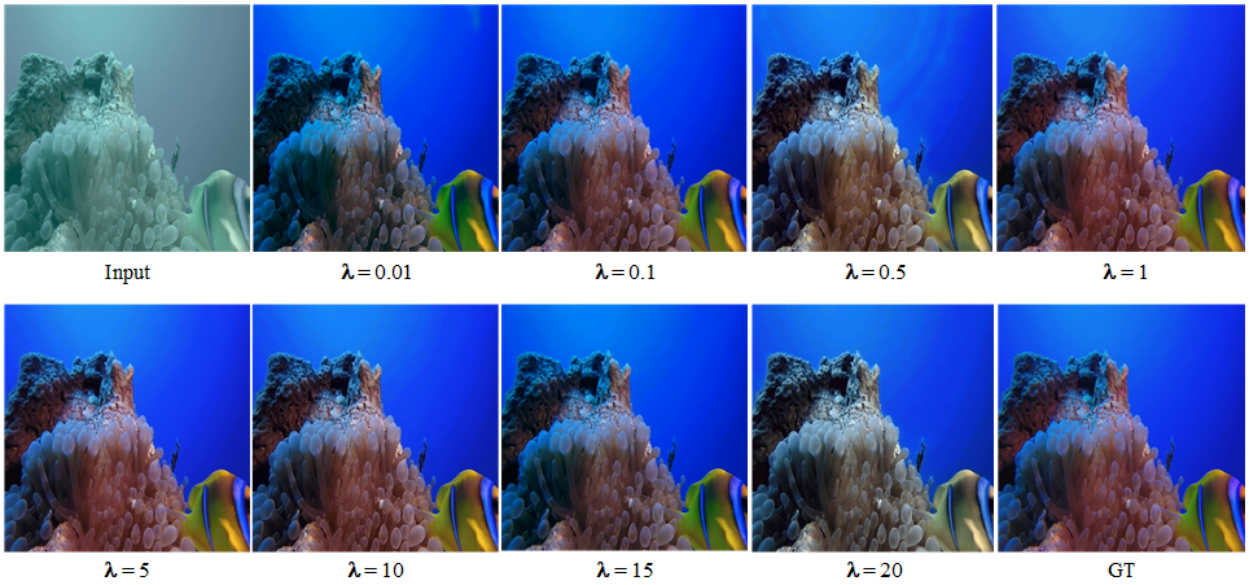

Figure 8. Demonstration of model performance variation with loss function weight λ: training a temporally stable image-based model is actually a compromise between visual quality and temporal stability. The optimal result lies in the balance of them.

Table 3. Average quantitative metrics of enhancement quality and temporal stability in comparative experiments: PSNR, SSIM, and UIQM are indicators of the image quality generated by different models, while LPIPS and E_{warp} are consistency indicators.

Weight	PSNR ↑	SSIM ↑	UIQM ↑	LPIPS ↓	E_{warp} ↓
$\lambda = 0.01$	23.411	0.917	2.659	0.123	0.0283
$\lambda = 0.1$	23.537	0.919	2.715	0.117	0.0279
$\lambda = 0.5$	23.793	0.922	2.757	0.113	0.0273
$\lambda = 1$	24.654	0.927	2.831	0.105	0.0269
$\lambda = 5$	25.087	0.936	3.013	0.095	0.0265
$\lambda = 10$	24.973	0.931	2.916	0.091	0.0262
$\lambda = 15$	24.613	0.920	2.798	0.088	0.0261
$\lambda = 20$	24.131	0.915	2.712	0.086	0.0258

4.3.2. Network Model

In order to better validate the effectiveness of the underwater feature fusion module and the underwater video enhancement network, ablation experiments were conducted to test three configurations: U-Net, U-Net + TripleWFM, and U-Net + Underwater Feature Enhancement Network (WFENet).

In this experiment, U-Net refers to the basic U-Net network used alone without any additional modules, as shown in Figure 3. U-Net + WFENet represents the upper part of Figure 3, which is the basic U-Net network model (without the underwater feature fusion module added after each downsampling) in parallel with the underwater feature enhancement network in the lower half. U-Net + TripleWFM uses the single-stage U-Net network from Figure 3 and incorporates the underwater feature fusion module after each downsampling. Due to space limitations, additional ablation experiment results for other scenarios are provided in Appendix A.

As shown in Figure 9 and Table 4, after training the dual-branch network with optical flow warping, each experimental result shows significant improvements in perceptual continuity and temporal stability. However, the standalone U-Net network still suffers from a blue–green color cast, insufficient contrast, and loss of edge information. U-Net + WFENet also shows artifacts and slight flickering. This reflects the difficulty encountered by U-Net in capturing global contextual information and all useful multiscale features within the image. Although the addition of Triple WFM improves U-Net's ability to capture global context and multiscale features, the enhanced video frames still occasionally exhibit small-scale artifacts and flickering. This instability is resolved after introducing WFENet, which prevents the loss of global features and maintains the stability of the global contextual information in the video frames.

Table 4. Ablation experiment results, showing quantitative metrics for enhancement quality and temporal stability. PSNR, SSIM and UIQM are indicators of the image quality generated by different models, LPIPS and E_{warp} are consistency indicators, and Time shows the operating efficiency of the different models. The experimental indicators of the model in this paper are in bold.

Model	PSNR ↑	SSIM ↑	UIQM ↑	LPIPS ↓	E_{warp} ↓	Time(s) ↓
U-Net	23.926	0.845	2.795	0.112	0.0283	0.316
U-Net + TripleWFM	24.402	0.923	2.823	0.098	0.0277	0.371
U-Net + WFENet	24.335	0.901	2.816	0.113	0.0285	0.387
Ours	**25.087**	**0.936**	**2.951**	**0.095**	**0.0265**	**0.405**

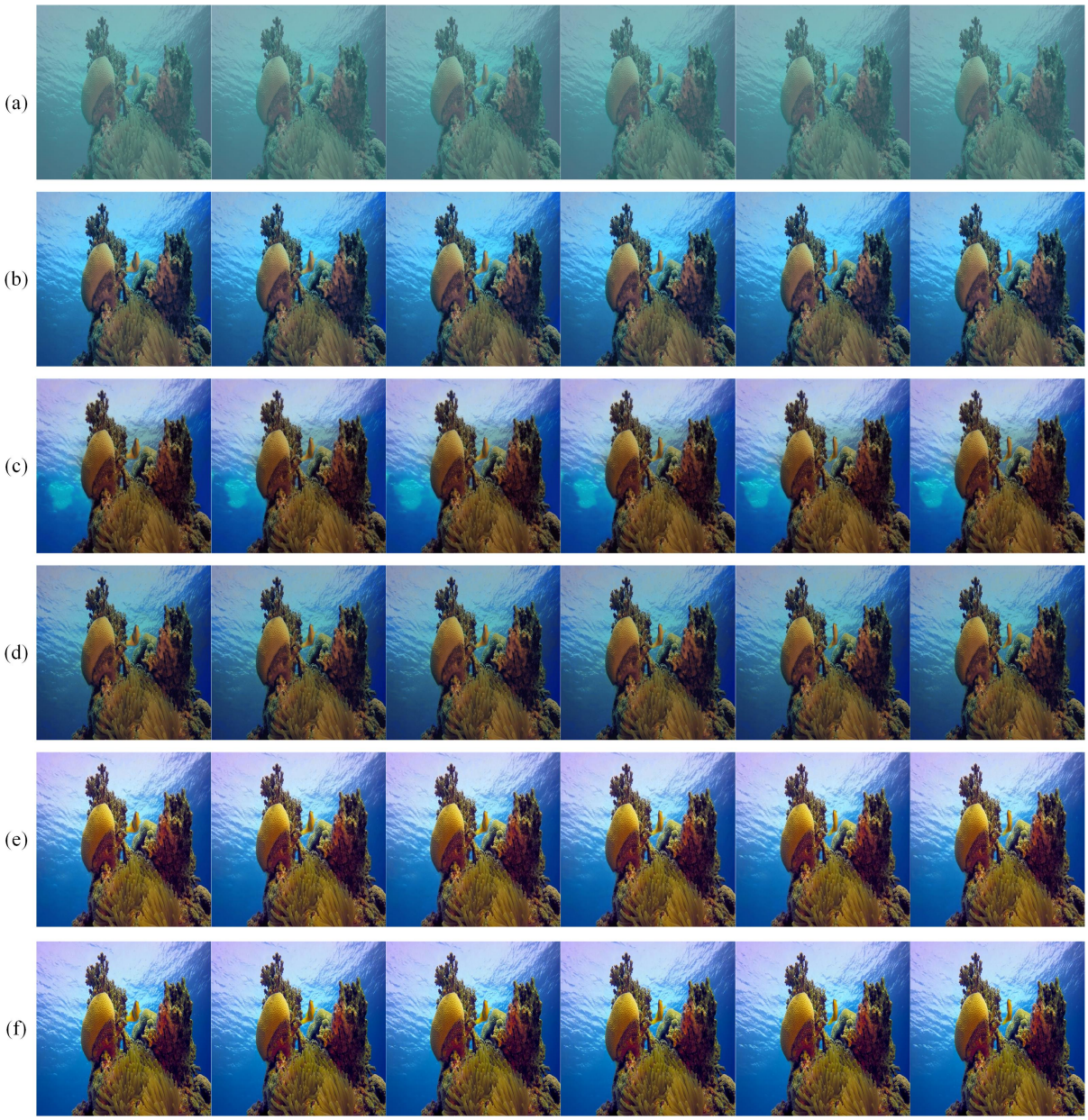

Figure 9. Ablation Experiment Scenario 1: (**a**) Input, (**b**) U-Net, (**c**) U-Net + TripleWFM, (**d**) U-Net + WFENet, (**e**) Ours, (**f**) GT. The six pictures in the figure are continuous video frames. They are presented together to show the enhancement results of the different models for temporal consistency.

Based on evaluation of the objective metrics, the proposed method demonstrates significant improvements in perceptual continuity and temporal stability compared to other image-based enhancement methods, which is consistent with the conclusions drawn from the subjective analysis. Additionally, the proposed underwater video enhancement network achieves superior performance in both enhancement quality metrics and temporal stability metrics. Through ablation experiments, the effectiveness of the underwater feature fusion module and the underwater video enhancement network structure in improving the temporal consistency of video frames has been validated, highlighting the necessity of addressing the challenges in capturing global context information and multiscale features within the U-Net network.

5. Conclusions

In this paper, we propose an underwater video enhancement method based on image optical flow distortion that alleviates flickering phenomena through temporal stability processing and leverages generated optical flow to guide the model in learning temporal consistency. Both quantitative and qualitative results demonstrate that the trained model performs well in enhancement quality and temporal stability. The proposed method effectively enhances individual video frames, improving both perceptual continuity and temporal stability. Furthermore we have designed a new underwater feature fusion module and underwater video frame network. The proposed module integrates features extracted through global average pooling and max pooling, utilizing various convolutional layers, pooling operations, and feature fusion techniques to process the input features. The gradual refinement of processing enhances the model's understanding and expressive ability of the input data, preventing the loss of content information in generated images and enabling them to contain richer detail features, which in turn helps to better reconstruct distorted information. Additionally, the increased weight of global features in the video frames enhances the temporal consistency of video frame sequences and aids in reducing flicker and jitter. Comparisons with several algorithms through both subjective visual assessments and objective evaluation metrics verify the superiority of the proposed method in terms of enhancement quality and temporal consistency.

6. Limitations

Current research primarily focuses on relatively ideal underwater environments, assuming clear water quality and the absence of strong currents or pollution. However, in practical applications, changes in water quality have a significant impact on the video enhancement process. The underwater environment may severely degrade image quality due to factors such as turbid water, suspended particles, and plankton, resulting in significant changes in brightness, contrast, and color. Existing methods may struggle to handle these non-ideal water conditions, particularly when illumination is extremely low or the water body is highly turbid, as the effectiveness of optical flow estimation and image enhancement can be considerably reduced.

Future research could develop more robust underwater video enhancement methods by integrating multimodal data such as laser scanning and depth images in order to better adapt to varying water conditions.

Additionally, current evaluation methods largely rely on self-generated synthetic datasets, which may introduce deviations from ground truth representations. This could potentially affecting the model's generalization ability, particularly in real-world underwater scenarios.

Although the proposed study introduces an effective underwater video enhancement method and addresses temporal consistency and motion compensation challenges in underwater videos to some extent using optical flow technology, there remain numerous challenges. Future research could further improve the robustness and practicality of underwater video enhancement by incorporating multiscale enhancement, non-rigid object motion modeling, noise suppression, and optical flow accuracy improvement. Furthermore, the use of more real-world datasets and advanced self-supervised learning techniques would provide stronger support and guidance for future developments in the field.

Author Contributions: Conceptualization, K.H.; methodology, K.H. and Y.M.; software, Y.M., L.T. and Z.L.; formal analysis, Y.M.; investigation, K.H., L.T. and X.Y.; writing—original draft preparation, Y.M.; writing—review, K.H.; editing, Y.M. and L.T.; visualization, Y.M. and L.T.; supervision, K.H.

and X.Y.; project administration, K.H. and X.Y. All authors have read and agreed to the published version of the manuscript.

Funding: The research in this article is supported by the National Natural Science Foundation of China (42275156).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and code used to support the findings of this study are available from the corresponding author upon request (001600@nuist.edu.cn).

Acknowledgments: The authors would like to express heartfelt thanks to the reviewers and editors who submitted valuable revisions to this article.

Conflicts of Interest: Author Lei Tang was employed by the company Information and Telecommunication Branch, State Grid Jiangsu Electric Power Company. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A. Additional Figures

The following figures provide additional information.

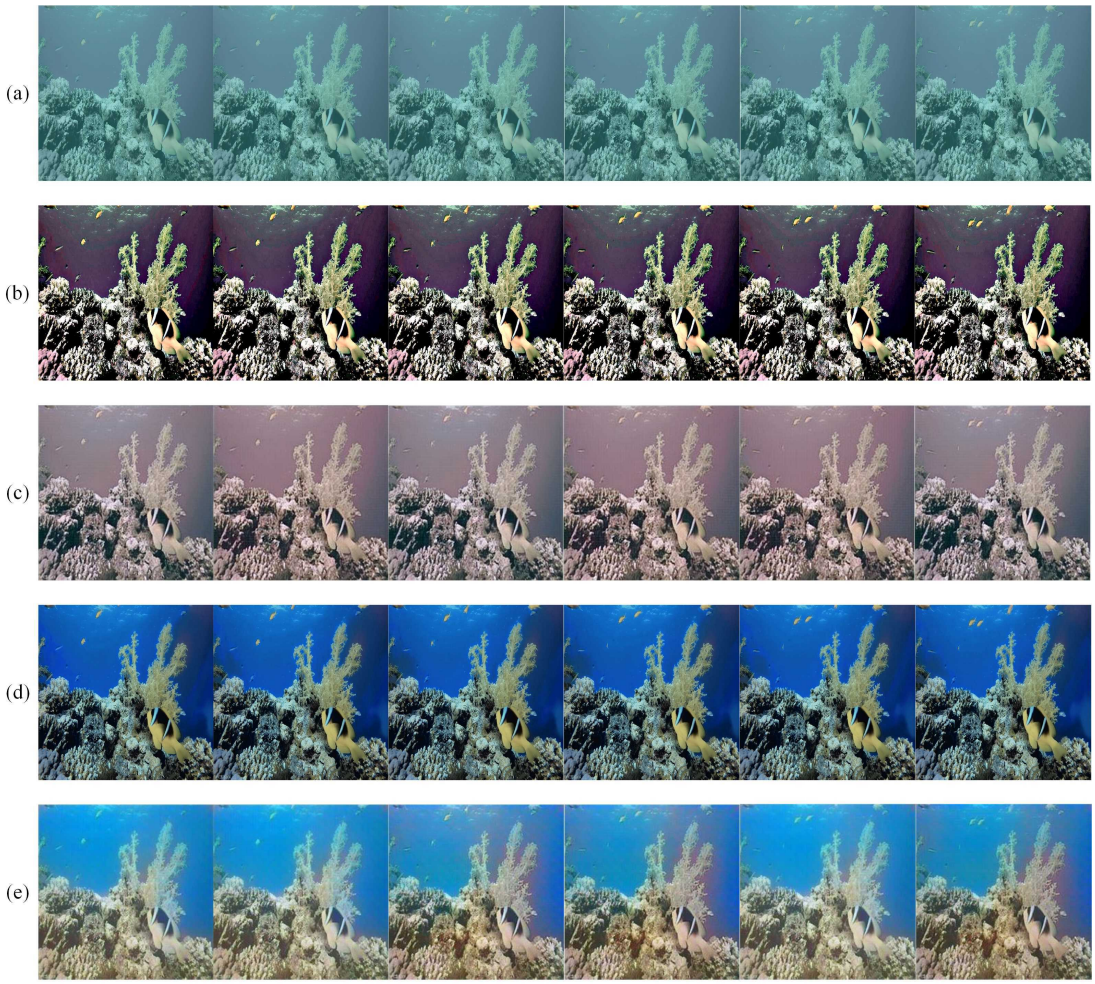

Figure A1. *Cont.*

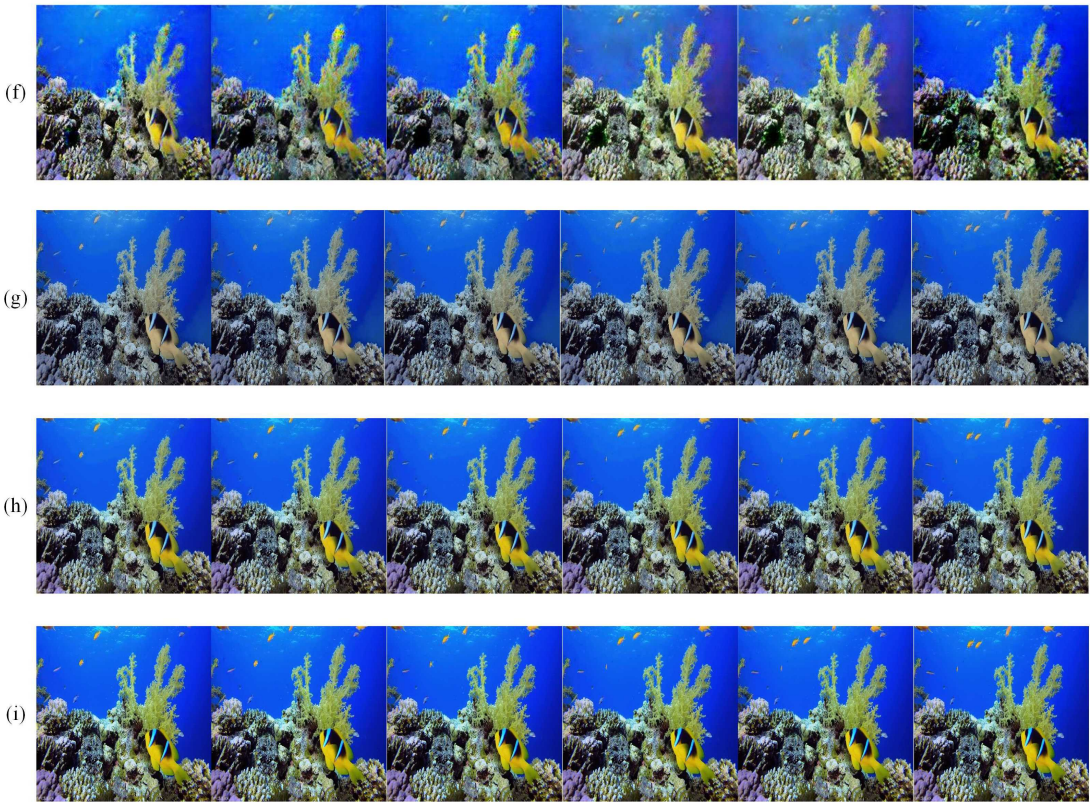

Figure A1. Comparison Experiment Scenario 2: (**a**) Input, (**b**) MSR, (**c**) UWCNN, (**d**) SGUIENet, (**e**) UGAN, (**f**) FunieGAN, (**g**) BLIND, (**h**) Ours, (**i**) GT.

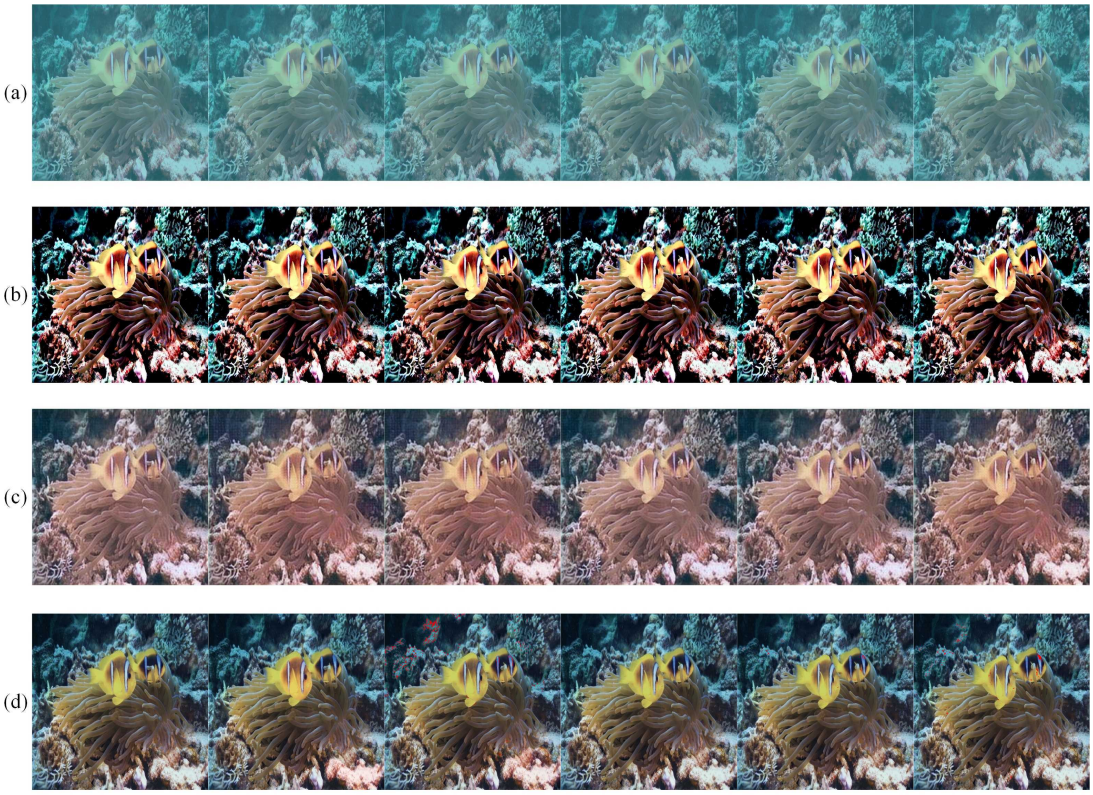

Figure A2. *Cont.*

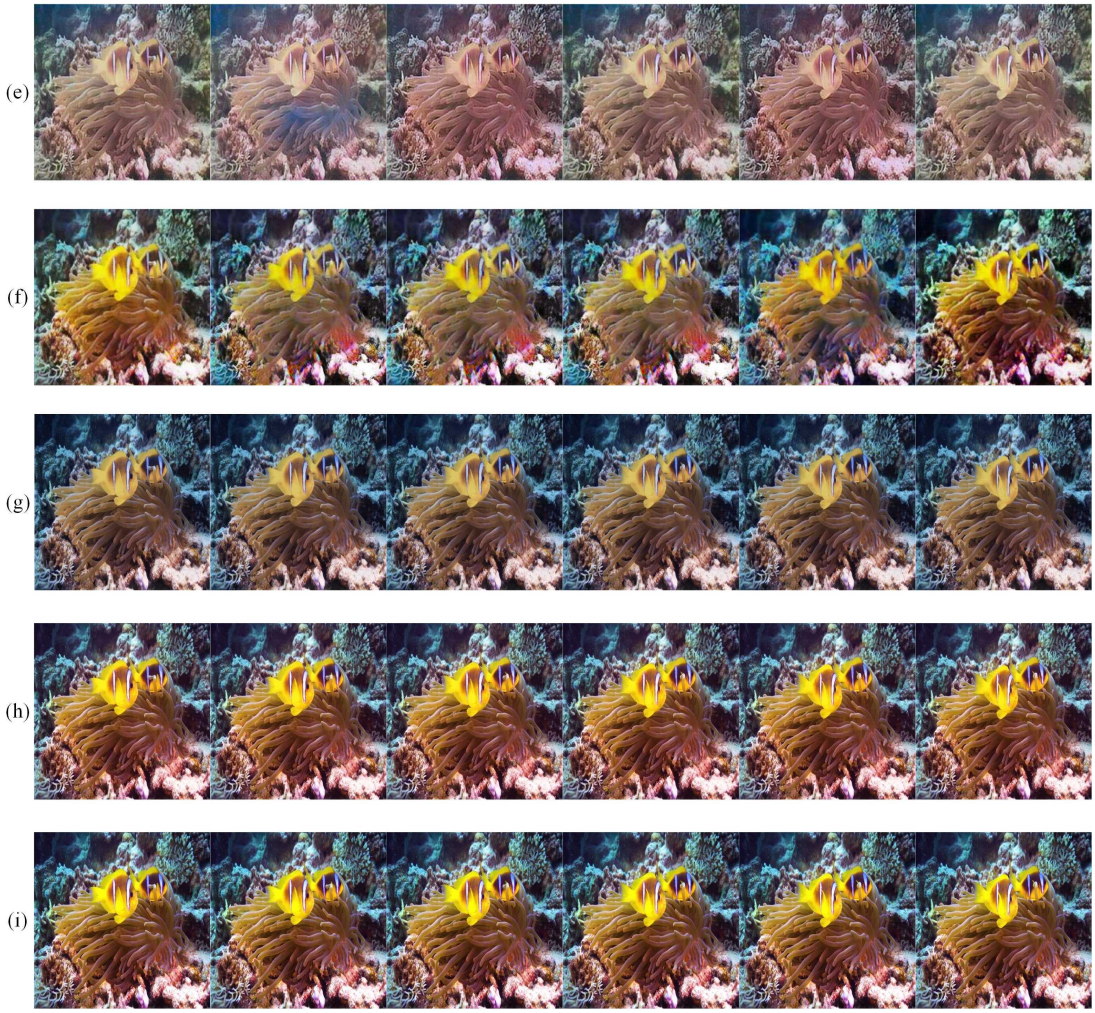

Figure A2. Comparison Experiment Scenario 3: (**a**) Input, (**b**) MSR, (**c**) UWCNN, (**d**) SGUIENet, (**e**) UGAN, (**f**) FunieGAN, (**g**) BLIND, (**h**) Ours, (**i**) GT.

Figure A3. *Cont.*

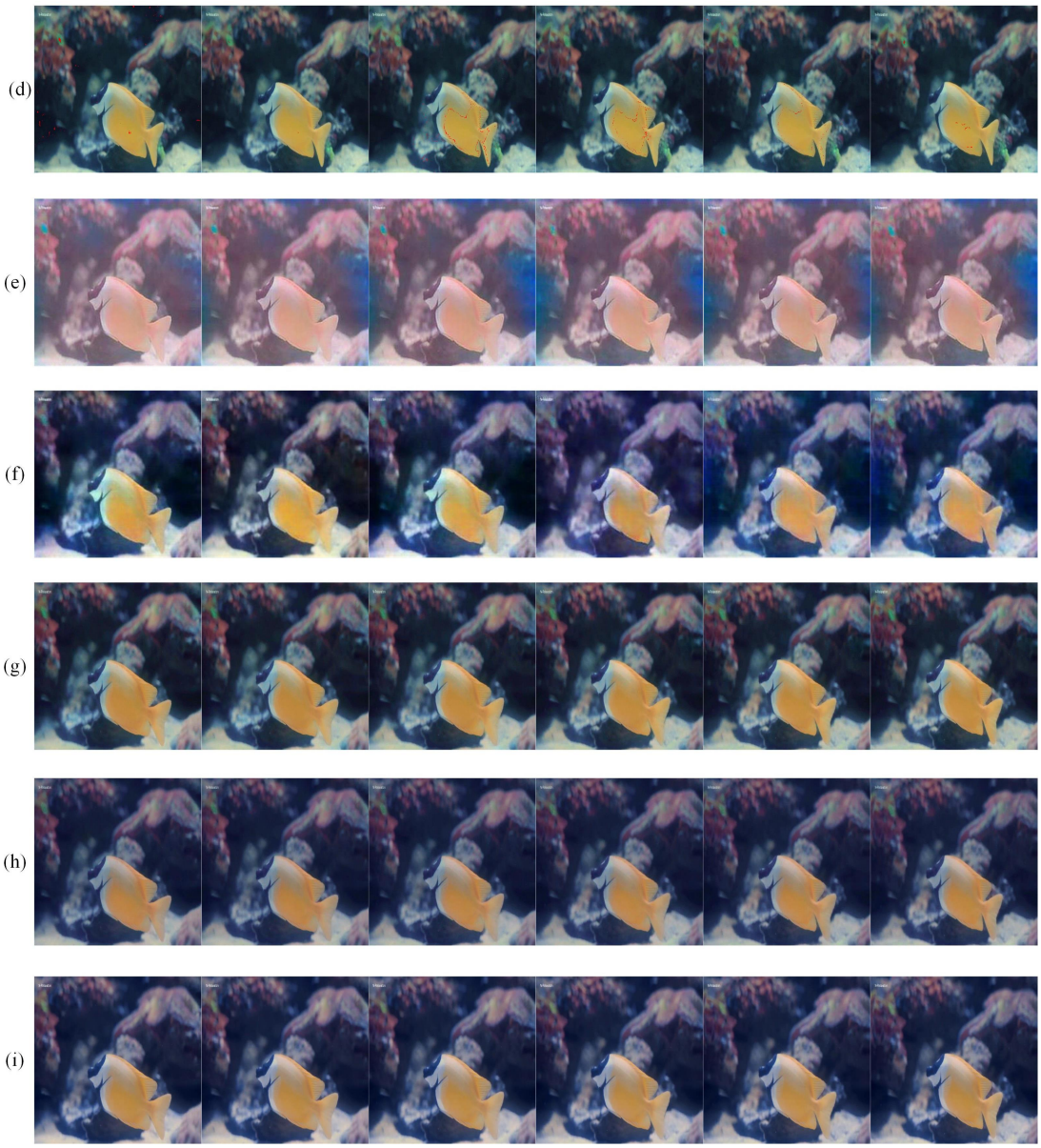

Figure A3. Comparison Experiment Scenario 4: (**a**) Input, (**b**) MSR, (**c**) UWCNN, (**d**) SGUIENet, (**e**) UGAN, (**f**) FunieGAN, (**g**) BLIND, (**h**) Ours, (**i**) GT.

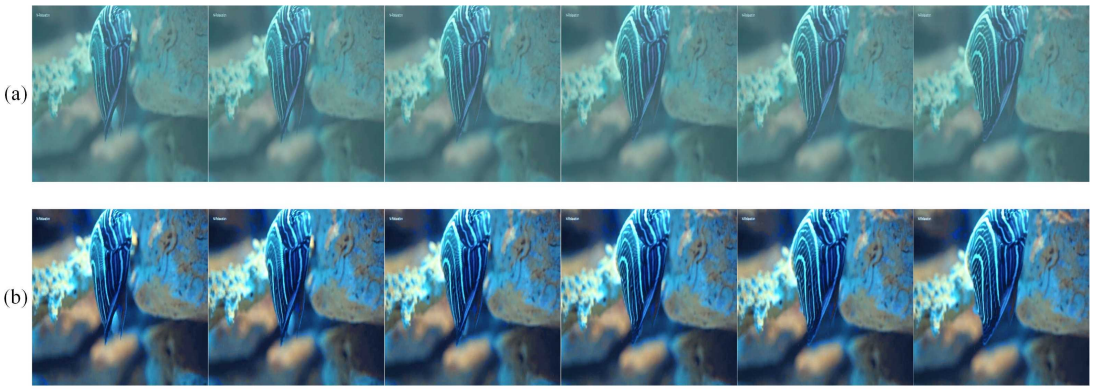

Figure A4. *Cont*.

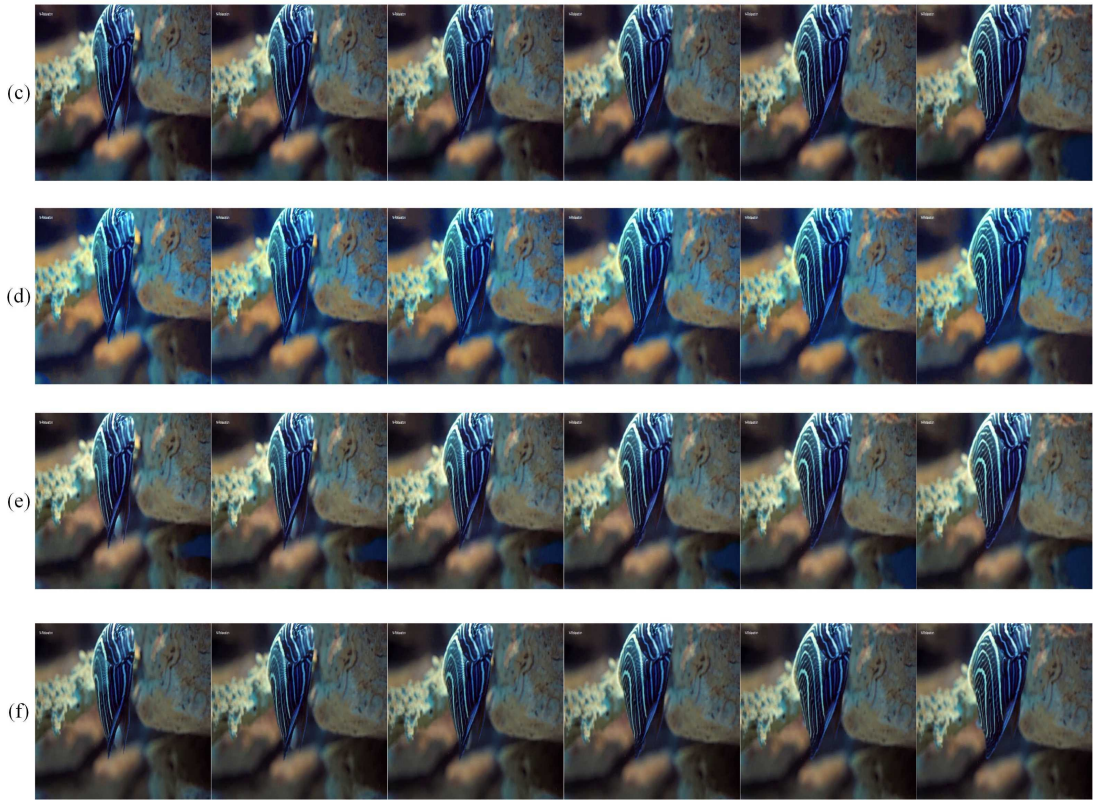

Figure A4. Ablation Experiment Scenario 2: (**a**) Input, (**b**) U-Net, (**c**) U-Net + TripleWFM, (**d**) U-Net + WFENet, (**e**) Ours, (**f**) GT.

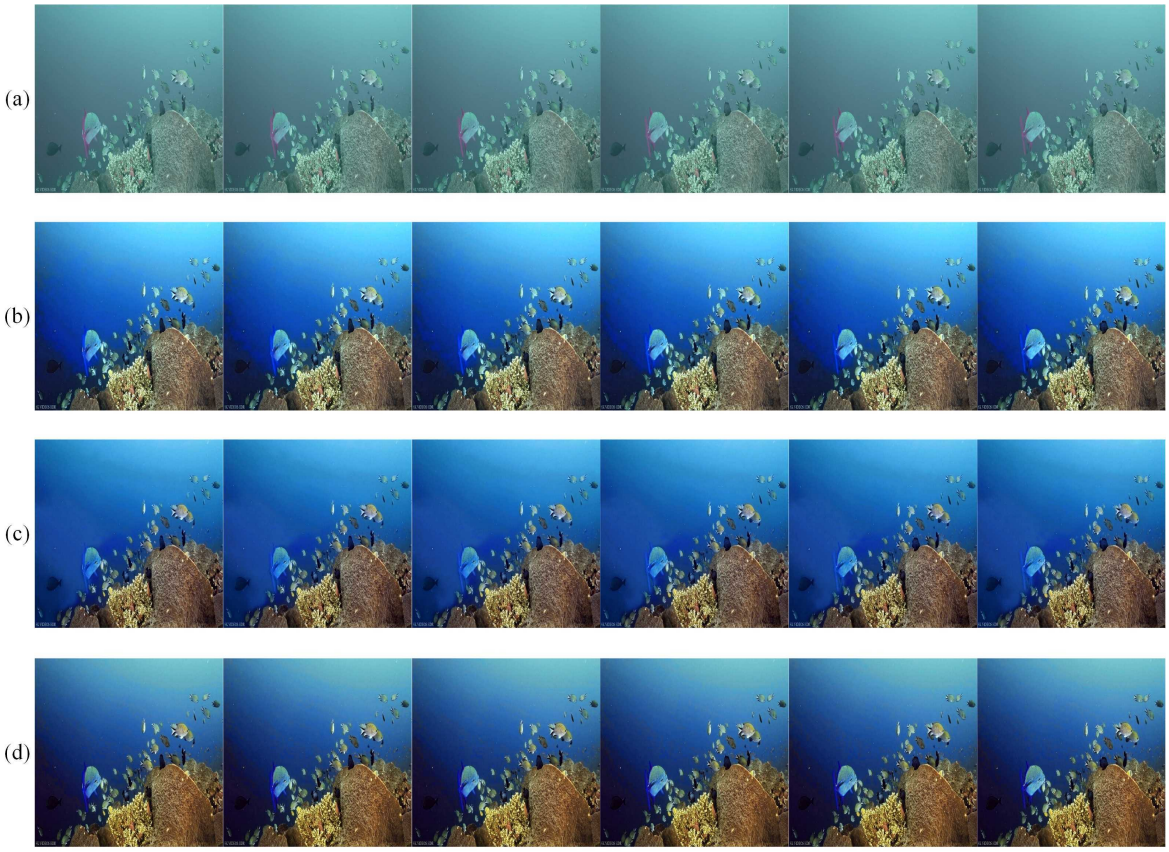

Figure A5. *Cont*.

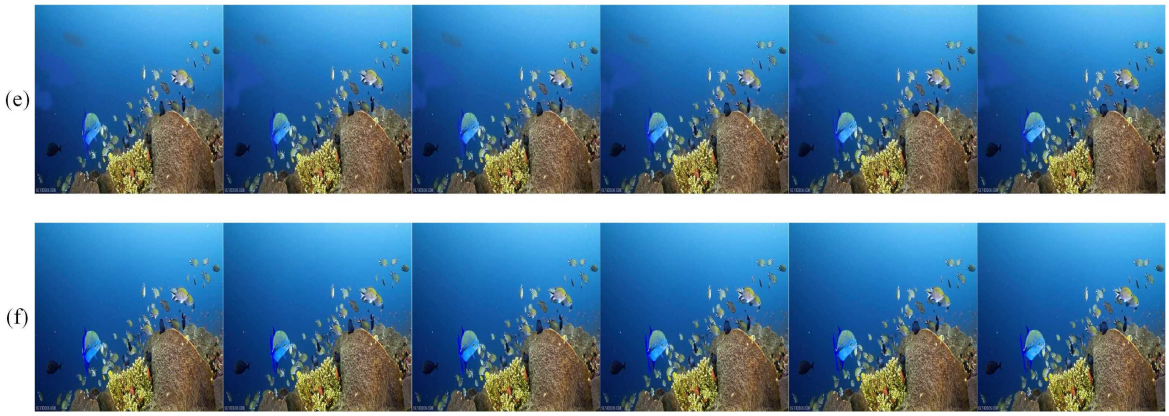

Figure A5. Ablation Experiment Scenario 3: (**a**) Input, (**b**) U-Net, (**c**) U-Net + TripleWFM, (**d**) U-Net + WFENet, (**e**) Ours, (**f**) GT.

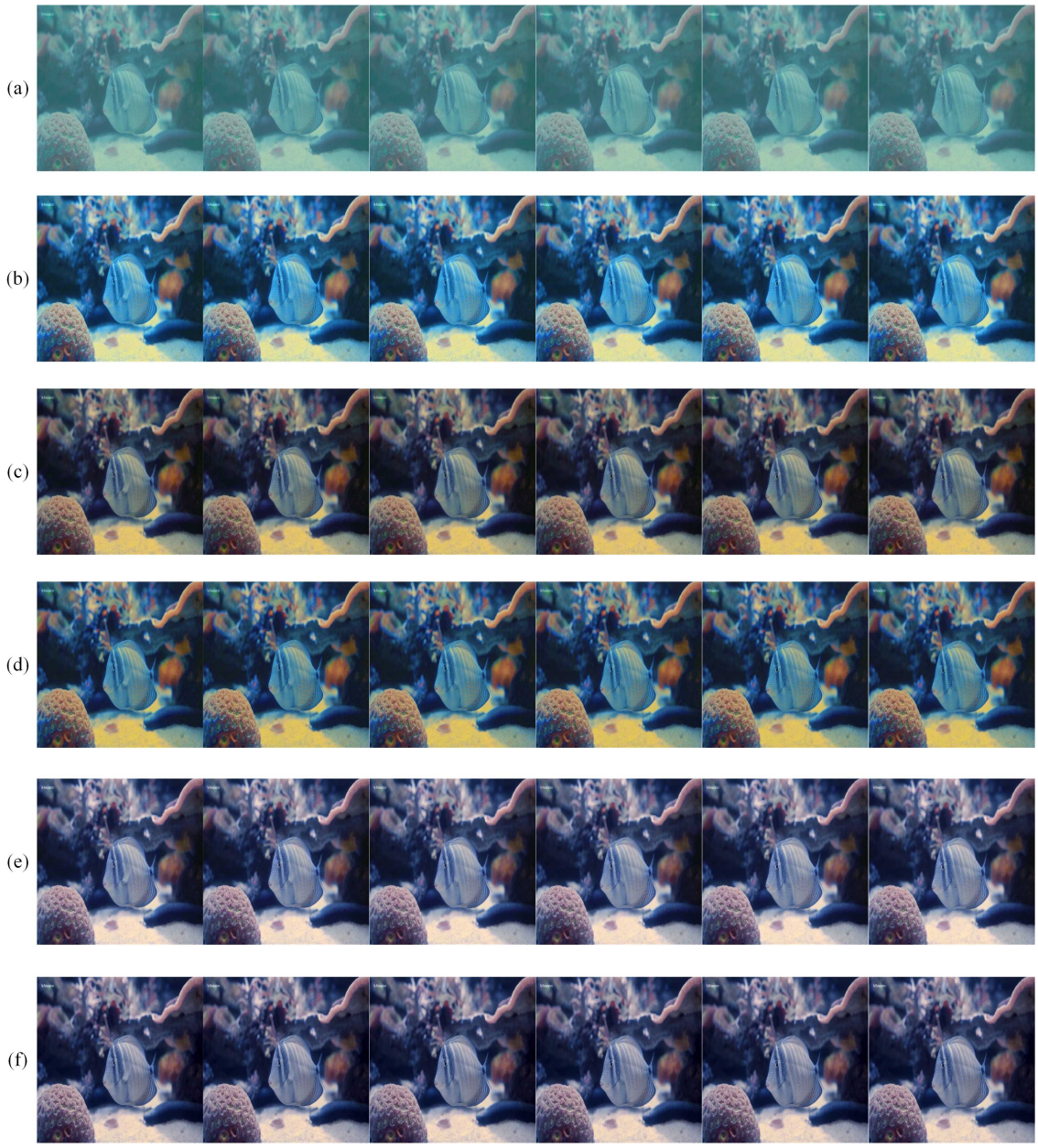

Figure A6. Ablation Experiment Scenario 4: (**a**) Input, (**b**) U-Net, (**c**) U-Net + TripleWFM, (**d**) U-Net + WFENet, (**e**) Ours, (**f**) GT.

References

1. Hong, Y.; Zhou, X.; Hua, R.; Lv, Q.; Dong, J. WaterSAM: Adapting SAM for Underwater Object Segmentation. *J. Mar. Sci. Eng.* **2024**, *12*, 1616. [CrossRef]
2. Xia, M.; Wang, T.; Zhang, Y.; Liu, J.; Xu, Y. Cloud/shadow segmentation based on global attention feature fusion residual network for remote sensing imagery. *Int. J. Remote Sens.* **2021**, *42*, 2022–2045. [CrossRef]
3. Zhao, X.; Wang, Z.; Deng, Z.; Qin, H. G-Net: An Efficient Convolutional Network for Underwater Object Detection. *J. Mar. Sci. Eng.* **2024**, *12*, 116. [CrossRef]
4. Huang, W.; Zhu, D.; Chen, M. A Fusion Underwater Salient Object Detection Based on Multi-Scale Saliency and Spatial Optimization. *J. Mar. Sci. Eng.* **2023**, *11*, 1757. [CrossRef]
5. Hu, K.; Li, M.; Song, Z.; Xu, K.; Xia, Q.; Sun, N.; Zhou, P.; Xia, M. A review of research on reinforcement learning algorithms for multi-agents. *Neurocomputing* **2024**, *599* , 128068. [CrossRef]
6. Hu, K.; Xu, K.; Xia, Q.; Li, M.; Song, Z.; Song, L.; Sun, N. An overview: Attention mechanisms in multi-agent reinforcement learning. *Neurocomputing* **2024**, *598*, 128015. [CrossRef]
7. Merugu, S.; Tiwari, A.; Sharma, S.K. Spatial–spectral image classification with edge preserving method. *J. Indian Soc. Remote Sens.* **2021**, *49*, 703–711. [CrossRef]
8. Haq, M.A.; Rahim Khan, M.A. DNNBoT: Deep neural network-based botnet detection and classification. *Comput. Mater. Contin.* **2022**, *71*, 1729–1750.
9. Hu, K.; Wang, T.; Shen, C.; Weng, C.; Zhou, F.; Xia, M.; Weng, L. Overview of underwater 3D reconstruction technology based on optical images. *J. Mar. Sci. Eng.* **2023**, *11*, 949. [CrossRef]
10. Bathula, A.; Gupta, S.K.; Merugu, S.; Saba, L.; Khanna, N.N.; Laird, J.R.; Sanagala, S.S.; Singh, R.; Garg, D.; Fouda, M.M.; et al. Blockchain, artificial intelligence, and healthcare: The tripod of future—a narrative review. *Artif. Intell. Rev.* **2024**, *57*, 238. [CrossRef]
11. Zhang, F.; Li, Y.; You, S.; Fu, Y. Learning temporal consistency for low light video enhancement from single images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4967–4976.
12. Hummel, R. Image enhancement by histogram transformation. *Comput. Graph. Image Process.* **1977** , *6*, 184–195. [CrossRef]
13. Iqbal, K.; Odetayo, M.; James, A.; Salam, R.A.; Talib, A.Z.H. Enhancing the low quality images using unsupervised colour correction method. In Proceedings of the 2010 IEEE International Conference on Systems, Man and Cybernetics, Istanbul, Turkey, 10–13 October 2010; pp. 1703–1709.
14. Ghani, A.S.A.; Isa, N.A.M. Enhancement of low quality underwater image through integrated global and local contrast correction. *Appl. Soft Comput.* **2015**, *37*, 332–344. [CrossRef]
15. Ghani, A.S.A.; Isa, N.A.M. Automatic system for improving underwater image contrast and color through recursive adaptive histogram modification. *Comput. Electron. Agric.* **2017**, *141*, 181–195. [CrossRef]
16. Joshi, K.; Kamathe, R. Quantification of retinex in enhancement of weather degraded images. In Proceedings of the 2008 International Conference on Audio, Language and Image Processing, Shanghai, China, 7–9 July 2008; pp. 1229–1233.
17. Mercado, M.A.; Ishii, K.; Ahn, J. Deep-sea image enhancement using multi-scale retinex with reverse color loss for autonomous underwater vehicles. In Proceedings of the OCEANS 2017-Anchorage, Anchorage, AK, USA, 18–21 September 2017; pp. 1–6.
18. Li, S.; Li, H.; Xin, G. Underwater image enhancement algorithm based on improved retinex method. *Comput. Sci. Appl.* **2018**, *8*, 9–15.
19. Perez, J.; Attanasio, A.C.; Nechyporenko, N.; Sanz, P.J. A deep learning approach for underwater image enhancement. In Proceedings of the Biomedical Applications Based on Natural and Artificial Computing: International Work-Conference on the Interplay Between Natural and Artificial Computation, IWINAC 2017, Corunna, Spain, 19–23 June 2017; Proceedings, Part II; Springer: Berlin/Heidelberg, Germany, 2017; pp. 183–192.
20. Sun, X.; Liu, L.; Li, Q.; Dong, J.; Lima, E.; Yin, R. Deep pixel-to-pixel network for underwater image enhancement and restoration. *IET Image Process.* **2019**, *13*, 469–474. [CrossRef]
21. Li, C.; Anwar, S.; Porikli, F. Underwater scene prior inspired deep underwater image and video enhancement. *Pattern Recognit.* **2020**, *98*, 107038. [CrossRef]
22. Fabbri, C.; Islam, M.J.; Sattar, J. Enhancing underwater imagery using generative adversarial networks. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, Australia, 21–25 May 2018; pp. 7159–7165.
23. Islam, M.J.; Xia, Y.; Sattar, J. Fast underwater image enhancement for improved visual perception. *IEEE Robot. Autom. Lett.* **2020**, *5*, 3227–3234. [CrossRef]
24. Hu, K.; Zhang, Y.; Weng, C.; Wang, P.; Deng, Z.; Liu, Y. An underwater image enhancement algorithm based on generative adversarial network and natural image quality evaluation index. *J. Mar. Sci. Eng.* **2021**, *9*, 691. [CrossRef]
25. Tang, P.; Li, L.; Xue, Y.; Lv, M.; Jia, Z.; Ma, H. Real-World Underwater Image Enhancement Based on Attention U-Net. *J. Mar. Sci. Eng.* **2023**, *11*, 662. [CrossRef]

26. Li, Y.; Chen, R. UDA-Net: Densely attention network for underwater image enhancement. *IET Image Process.* **2021**, *15*, 774–785. [CrossRef]
27. Lai, W.S.; Huang, J.B.; Wang, O.; Shechtman, E.; Yumer, E.; Yang, M.H. Learning blind video temporal consistency. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 170–185.
28. Horn, B. Determining optical flow. *Artif. Intell.* **1987**, *9*, 229–243.
29. Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.Y.; Girshick, R. Detectron2. 2019. Available online: https://github.com/facebookresearch/detectron2 (accessed on 12 October 2024).
30. Lian, S.; Li, H.; Cong, R.; Li, S.; Zhang, W.; Kwong, S. Watermask: Instance segmentation for underwater imagery. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 September 2023; pp. 1305–1315.
31. Zhan, X.; Pan, X.; Liu, Z.; Lin, D.; Loy, C.C. Self-supervised learning via conditional motion propagation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1881–1889.
32. Rahman, Z.u.; Jobson, D.J.; Woodell, G.A. Multi-scale retinex for color image enhancement. In Proceedings of the 3rd IEEE International Conference on Image Processing, Lausanne, Switzerland, 19–19 September 1996; Volume 3, pp. 1003–1006.
33. Qi, Q.; Li, K.; Zheng, H.; Gao, X.; Hou, G.; Sun, K. SGUIE-Net: Semantic attention guided underwater image enhancement with multi-scale perception. *IEEE Trans. Image Process.* **2022**, *31*, 6816–6830. [CrossRef]
34. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 586–595.
35. Lei, C.; Xing, Y.; Chen, Q. Blind video temporal consistency via deep video prior. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1083–1093.